NBER WORKING PAPER SERIES

SOCIAL-SCIENCE GENOMICS: PROGRESS, CHALLENGES, AND FUTURE DIRECTIONS

Daniel J. Benjamin
David Cesarini
Patrick Turley
Alexander Strudwick Young

Working Paper 32404
http://www.nber.org/papers/w32404

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2024

For helpful comments and suggestions, we thank Peter M. Visscher and the University of Queensland Statistical Genomics Lab Meeting. For excellent research assistance, we are grateful to Matthew Howell and Moeen Nehzati. For financial support, we thank the NIA/NIH (grants R24-AG065184, R01-AG042568, R00-AG062787, and R01-AG081518) and Open Philanthropy. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Social-Science Genomics: Progress, Challenges, and Future Directions
Daniel J. Benjamin, David Cesarini, Patrick Turley, and Alexander Strudwick Young
NBER Working Paper No. 32404
May 2024
JEL No. D87,I1,Q57

ABSTRACT

Rapid progress has been made in identifying links between human genetic variation and social and behavioral phenotypes. Applications in mainstream economics are beginning to emerge. This review aims to provide the background needed to bring the interested economist to the frontier of social-science genomics. Our review is structured around a theoretical framework that nests many of the key methods, concepts and tools found in the literature. We clarify key assumptions and appropriate interpretations. After reviewing several significant applications, we conclude by outlining future advances in genetics that will expand the scope of potential applications, and we discuss the ethical and communication challenges that arise in this area of research.

Daniel J. Benjamin
University of California Los Angeles
Anderson School of Management
and David Geffen School of Medicine
110 Westwood Plaza
Entrepreneurs Hall Suite C515
Los Angeles, CA 90095
and NBER
daniel.benjamin@anderson.ucla.edu

David Cesarini
Department of Economics
New York University
19 West 4th Street, Room 511
New York, NY 10012
and NBER
dac12@nyu.edu

Patrick Turley
Center for Economic and Social Research
University of Southern California
635 Downey Way
Los Angeles, CA 90089
pturley@usc.edu

Alexander Strudwick Young
University of California Los Angeles
Anderson School of Management
and David Geffen School of Medicine
110 Westwood Plaza
Los Angeles, CA 90095
alextisyoung@nber.org

This article reviews research at the intersection of genetics and social science, a field of research sometimes called social-science genomics.[1] In recent years, most areas of human genetics have experienced rapid progress, driven primarily by the extraordinarily steep declines in the costs of measuring genetic variation. The first human genome sequence was successfully completed a little over 20 years ago (Venter et al., 2001; International Human Genome Sequencing Consortium, 2001). Since then, the cost of sequencing a human genome has been falling at a rate faster than Moore's law, from many millions of dollars in the early 2000s to a few hundred dollars today (Wetterstrand, 2023). The cost of using so-called genotyping arrays, a less expensive technology for measuring most *common* genetic variation across individuals, has declined at a similar rate. These declines in cost have led to an explosion of available datasets with comprehensively genotyped individuals.

These increases in sample sizes have fueled brisk advances in knowledge about the links between particular genetic variants—regions of DNA that differ across individuals—and phenotypes, that is, measurable characteristics or outcomes (Visscher et al., 2017, 2012). These advances have primarily come from research designs called genome-wide association studies (GWASs), which produce estimates of the association between individual genetic variants and a phenotype of interest. Most GWASs have been conducted by medical geneticists whose primary focus is on disease, but the number of published GWASs of social and behavioral phenotypes has also been growing steadily. For applications in the social sciences, we contend that the main value of a GWAS is that the summary statistics it produces can be used to construct weights for DNA-based predictors of various phenotypes. Each such predictor is calculated as a weighted average of a person's genetic variants (typically millions, selected from across the genome), with the weight for each variant derived from its GWAS summary statistics.

In the literature, numerous labels—including polygenic scores (PGSs) or polygenic risk scores (PRSs)—have been used to describe these predictors. Following Becker et al. (2021), our preferred term for them is polygenic indexes (PGIs). Among the phenotypes that have the most highly predictive PGIs are (R^2's correspond to currently attainable level of predictive power of these PGIs in samples of European genetic ancestries): height ($R^2 \approx 0.45$; Yengo et al. 2022), body mass index ($R^2 \approx 0.15$; Zheng et al., 2022), educational attainment (PGI $R^2 \approx 0.15$; Okbay et al. 2022), age at first menses among women (PGI $R^2 \approx 0.12$; Becker

[1] Different names for this new area of research have been adopted in different social-science disciplines, reflecting their focus on particular applications and methodological approaches. In economics, genoeconomics (Benjamin et al., 2007); in political science, genopolitics (Fowler and Dawes, 2013); and in sociology, sociogenomics or social genomics (Mills and Tropf 2020; Conley 2016; however, sociogenomics and social genomics also describe a separate field of research on how social processes affect gene expression, as in Robinson, Grozinger and Whitfield, 2015). Psychology has a longstanding tradition of research on the role of genetics called behavior genetics (for a history, see Loehlin, 2009), and the psychologically oriented research we mention in this review falls under that rubric. Although this review reflects our economics perspective and the applications we discuss are economics-oriented, we use the more general term, social-science genomics, which may have originated in Rietveld et al. (2013), to emphasize the commonality of theory, data, and tools—which are what we focus on in this review.

et al., 2021) and self-rated health (PGI $R^2 \approx 0.05$; Becker et al., 2021).

The proliferation of genetic data is rapidly opening up new opportunities for social-science genomic research. Becker et al. (2021) provides an overview of some of the most commonly used datasets. By, far the most widely used today is the UK Biobank (Bycroft et al., 2018), which stands out primarily in terms of its sample size (roughly 500,000 individuals). Several commonly used datasets in economics also have genome-wide data on many of their participants, including the Health and Retirement Study (HRS), the English Longitudinal Study of Ageing (ELSA), the Panel Study of Income Dynamics (PSID), and the German Socio-economic Panel (GSOEP).

Most commonly, social-science genomics research analyzes PGIs constructed in a dataset such as one of those just mentioned (using PGI weights calculated from a GWAS conducted in an independent, non-overlapping, sample). Prior to the availability of measures of genetic variation, most research treated genetic influences as latent variables and sought to infer their effects by contrasting the phenotypic resemblance of twins, adoptees, and other kinships (Goldberger, 2005; Sacerdote, 2011; Cloninger, Rice and Reich, 1979). By contrast, PGIs are observed variables that can be incorporated directly into analyses.

In light of the advances over the past few years, we believe the time is ripe for a review paper. Two prior review papers in general-interest economics journals (Beauchamp et al., 2011*a*; Benjamin et al., 2012) were published before the first large-scale GWAS of a social-science phenotype (Rietveld et al., 2013). A third (Dias Pereira et al., 2022) provides an accessible, succinct and non-technical introduction to a subset of the topics we discuss here. More recent reviews have been published in sociology (Freese, 2018; Braudt, 2018; Martschenko, Trejo and Domingue, 2019; Conley, 2016) and psychology (Plomin et al., 2016). Relative to these, we seek to spell out connections to relevant underlying genetic theory more explicitly, and to provide a more self-contained and comprehensive treatment of technical details. Most of the applications we highlight are also from economics. Although our review is primarily oriented toward economists, it is written in the hope that the material will also be of utility to researchers from other disciplines. The theoretical focus of our paper makes it a natural companion to other texts whose main focus is on a number of important practical issues that arise in empirical analyses of large samples of comprehensively genotyped subjects (e.g., a recent textbook by Mills, Barban and Tropf, 2020).

This review highlights two recent trends in social-science genomics (and human genetics research more broadly). The first is an increased appreciation of the value that analysis of large samples with genotyped first-degree relatives can have in settings where the goal is to make causal inferences about genetic variation in general, and PGIs in particular. Geneticists have long understood that genes are transmitted from parents to offspring following laws that, in quasi-experimental parlance, can be leveraged to yield a powerful natural experiment enabling the identification of causal effects. Long before the modern GWAS era, Fisher (1952,

p. 7) remarked, "The different genotypes possible from the same mating have been beautifully randomised by the meiotic process. A more perfect control of conditions is scarcely possible, than that of different genotypes appearing in the same litter." This longstanding insight forms the basis of classical, pedigree-based analyses (e.g., Spielman, McGinnis and Ewens, 1993), but it has only recently been incorporated into the era of GWASs and PGIs. The main reason is that family-based samples are required to isolate the source of variation in genotype that is randomly assigned, and until recently, family samples large enough to produce estimates with meaningful precision were simply unavailable.

The second trend is conducting GWASs in samples from non-European-genetic-ancestry populations and constructing PGIs for those populations. The large majority of existing GWASs have been conducted in samples from European-genetic-ancestry populations (Mills and Rahal, 2019; Abdellaoui et al., 2023). For a variety of reasons that we discuss—some of them genetic, some non-genetic—findings from a population with one ancestry do not always generalize well to populations with different ancestries. An important example of this so-called "portability problem" arises with PGIs: decay in prediction accuracy is often substantial when a PGI constructed using a training sample of one genetic ancestry is evaluated in a validation sample composed of individuals from another genetic ancestry. Extending GWASs and PGIs to other populations is an important direction for ongoing research, both in order for the benefits of findings from genetic research to be distributed more equitably (Martin et al., 2019; Fatumo et al., 2022) and for social scientists to be able to use PGIs in a broader range of applications.

The main goal of this review is to bring interested researchers to the frontier of social-science genomics. To that end, we aim to provide a self-contained, streamlined, and integrated development from the underlying biology and definitions of genetic concepts, through GWAS, to PGIs and their applications in the social sciences. Another goal of this review is to offer a rigorous exposition using a unified framework that clarifies the assumptions underlying methods and the intuitions underlying key results. Throughout the review, we emphasize issues of causal inference and appropriate interpretation of genetic effects. In all of these respects, we believe our review is unique.

To keep the paper focused on getting to PGIs and their applications, we omit detailed discussions of a number of topics relevant to social-science genomics. For example, we do not discuss traditional twin, family, and adoption studies, except to clarify how these methods relate to the unified framework; these methods, and research linking genetic variation to social and behavioral phenotypes prior to GWAS, are covered in earlier reviews (e.g., Goldberger 1978; Otto, Christiansen and Feldman 1995; Beauchamp et al. 2011*a*; Benjamin et al. 2012; Cesarini and Visscher 2017). Similarly, we only briefly discuss research on gene expression and epigenetics. Similar caveats apply to research by economists on evolution and economic growth (for a review, see Ashraf and Galor, 2018) and work on comparative economic development that uses genetic variation as an empirical proxy

for diversity (e.g., Spolaore and Warcziag, 2009; Arbatli et al., 2020). Finally, we do not address the new economic questions that arise from the availability of PGIs, such as their impact on insurance markets (see, e.g., Karlsson Linnér and Koellinger, 2022), their use during in vivo fertilization to select embryos with lower disease risks or other characteristics (see, e.g., Turley et al. 2021*b*; Meyer et al. 2023*b*), or their potential role in personalized medicine.

This review is organized as follows. Section I begins with a rudimentary genetics primer. Section II lays out a general theoretical framework that is helpful for clarifying what it means for a genetic variant to *cause* phenotypic variation. We also use the framework to clarify some subtle interpretational issues that are prone to misunderstanding. In Section III, we turn to estimation of genetic effects. Having laid the groundwork in prior sections, Section IV defines, interprets, and analyzes PGIs. In Section V, we illustrate seven applications of genetic data in economics. In Section VI, we outline current trends in genetics research and what they imply about future applications in the social sciences. We conclude by highlighting some of the ethical, policy, and communication challenges that are intrinsic to research at the intersection of genetics and social science.

I Genetics Primer

This section provides genetics background relevant to what follows. Some readers may wish to skip the section and refer back to it as needed. Because we aim to provide close to the minimum amount of information needed to fully understand assumptions made elsewhere in the paper, we omit several nuances.[2] For readers interested in additional details, we recommend consulting a textbook in molecular genetics such as Strachen and Read (2018) and in population genetics such as Gillespie (2004).

A The Genome and SNPs

The *genome* usually refers to a person's genetic material at conception (prior to any mutations that occur throughout a person's life). Almost every cell in the body contains an exact copy of the entire genome.[3] The human genome has ∼21,000 *genes*. Genes contain sequences of DNA that code for amino acids, the building blocks of the body's proteins. Genes constitute only ∼2% of the genome. A much larger fraction of the genome affects when and how much genes are expressed (see Section I.G).

The genome is divided across 23 pairs of *chromosomes*, one sex chromosome pair and 22 non-sex chromosome pairs called autosomes. One chromosome in

[2]For example, we ignore mitochondrial DNA, which is technically part of the human genome but resides in mitochondria (outside the cell nucleus) and is inherited exclusively from the mother. We ignore it because mitochondrial DNA only contains 13 of the ∼21,000 human genes and is not generally included in the genotyping data we discuss.

[3]One important exception is germ cells, discussed in Section I.B. We also ignore mutations that cause small differences in DNA across cells.

each pair was inherited from the individual's mother (the maternal chromosome), and the other from the individual's father (the paternal chromosome).

Each chromosome consists of a pair of DNA strands that are bound together. Each strand is composed of a sequence of nucleotide molecules, referred to as bases. There are four bases: guanine (abbreviated G), cytosine (C), thiamine (T), and adenine (A). DNA bases always pair with their complementary base on the other strand: C with G, and A with T. Since the information is redundant, one strand is chosen by convention to be the reference strand, and the *base pair* is described by the base on the reference strand.

Each location in the genome can be described by its chromosome and base-pair position. At each position, an individual has one base pair (G, C, T, or A) from the (reference strand of the) maternal chromosome and one from the paternal chromosome. With rare exceptions, the biological function of the base pair does not depend on whether it was inherited from the mother for father. Thus, the *genotype* at a position—the composition of the genome at that position—can be described by a set of two bases (each on its reference strand), such as GC or TT, without reference to which was inherited from which parent.

By sequencing and assembling the genomes of several individuals, the Human Genome Project created reference human genomes, which serve as standard representations. At 99.8-99.9% of loci, depending on a person's genetic ancestry, there is no genetic variation from the reference human genome (Nurk et al., 2022). The parts of the genome that vary across individuals are called polymorphisms, or *genetic variants*. Definitions vary, but according to a typical definition, a *rare variant* is a genetic variant in which 99% or more of individuals have the same version of the genetic variant, and a *common variant* is one in which fewer than 99% of individuals have the same version.

People may vary in complex ways from a reference genome, such as due to sections being duplicated, deleted, or inverted. The simplest variation from the reference is a single nucleotide difference, called a *SNP* (*single-nucleotide polymorphism*). For example, the reference genome would have an A base, whereas some individuals would have a T base. One comprehensive analysis of global genetic variation (The 1000 Genomes Project Consortium, 2015) found that SNPs comprise roughly 95.5% of variants.

The bases that can occur at a SNP are called *alleles*, and one allele is inherited from each parent. At the vast majority of SNPs, there are only two alleles of non-negligible frequency in the population (either of which could be inherited from either parent). Whichever allele is less common in the population is called the *minor allele*. An individual's SNP genotype is often summarized by the minor allele count: 0, 1, or 2. SNP data for an individual typically comes as a vector of minor allele counts, with each element corresponding to a measured SNP at a particular locus. Following standard terminology, we will often refer to the minor allele count as the *genotype* and the vector of minor allele counts as the *genotype vector*.

B Genetic Inheritance

For reproduction, individuals produce *germ cells* (sperm in males, eggs in females). Unlike other cells, germ cells contain only one copy of each chromosome. Meiosis is a type of cell devision that produces germ cells, each containing a random half of the chromosome pair that a parent carries. An offspring is conceived when one germ cell from the father and one from the mother fuse. The resulting child then has a chromosome pair, with one chromosome coming from the father and one from the mother.

During meiosis, randomness is introduced in two distinct stages. First, within each chromosome pair, the chromosomes' arms cross a random number of times at random loci, and the chromosomes swap their DNA after the crossing points. This process is called crossing over, and the transfer of chunks of DNA is called *recombination*. Second, independently across the 23 chromosome pairs and after recombination, one among each pair is, with equal probability, transmitted to a given germ cell. This process is called *Mendelian segregation*.

These random processes have some important implications for our purposes. Fixing any given SNP, conditional on the parental genotypes, the offspring receives one of each parent's two alleles, with equal probability. For any two SNPs on different chromosome pairs, the transmission of alleles across the two SNPs are independent random processes. Finally, for any two SNPs on the same chromosome pair, the probability that alleles on the same parental chromosome are transmitted to the offspring is higher the closer the two loci are. This correlated inheritance of alleles on the same parental chromosome is called *linkage*.

C Linkage Disequilibrium (LD)

Linkage disequilibrium (LD) refers to correlation between the genotypes of genetic variants.[4] Under random mating, the only source of LD is linkage. In that case, no LD is expected between genetic variants on different chromosomes. Within each chromosome, the LD between two variants is decreasing with their physical distance. For nearby variants on a chromosome, the LD due to linkage can be very high, often reaching one or nearly one. Regions of the genome that are essentially perfectly correlated with each other in a given population are called *haplotype blocks*, and the different versions of a block effectively form a single genetic variant for that block.

[4]In other areas of genetics, LD refers more generally to the statistical association between genotypes, and measures other than correlation are sometimes used. Originally, LD was more specifically related to linkage than it is in modern usage. The concept of LD arose from considering what would happen after repeated recombinations. For example, beginning with a population where some individuals have an A allele at locus 1 and a T allele at locus 2 and other individuals have a C allele at locus 1 and a G allele at locus 2. Recombination between the loci will reduce the association between having an A allele and a T allele. In the limit, the genotype at locus 1 will become statistically independent of the genotype at locus 2 and remain so thereafter. This equilibrium state is called "linkage equilibrium." LD was meant to refer to deviations from that state (Sved and Hill, 2018). In a randomly mating population, loci on different chromosomes are expected to be in linkage equilibrium.

Non-random mating generates LD with different properties. Consider *assortative mating*: individuals who have some characteristic are more likely to mate with other individuals who have that characteristic. Most assortative mating processes that cause spousal resemblance on a phenotype will also induce a correlation between spousal genes associated with the phenotype. One example is height. Since genetic variants associated with being taller are scattered throughout the entire genome, assortative mating at the genetic level will lead to positive LD between height-associated alleles, including those located on different chromosomes. Another, related form (e.g. Bergstrom, 2013) of non-random mating is *population structure*: individuals within a subpopulation—e.g., geographic region, or with a shared ethnicity or language—tend to mate with each other. In that case, alleles that happen to be more common within the subpopulation will become correlated, regardless of their location in the genome.

D Complex Phenotypes

A *trait*, or *phenotype*, is any measurable characteristic, behavior, or outcome of an organism. A phenotype is called *monogenic* if most or all of the variation is controlled by a single gene. A phenotype is called *polygenic*, or *complex*, if it is affected by many genetic variants, not restricted to a single gene. Intermediate cases, where genetic variation is controlled by several genetic variants, also exist, as do hybrid cases. Late-onset Alzheimer's disease is an example of the latter: a single gene, *APOE*, has a relatively large effect, but most of the genetic influence is polygenic (Lambert et al., 2013).

Monogenic traits are featured in standard introductions to genetics. Classic examples dating back to Gregor Mendel's original experiments (Mendel, 1866) include whether a pea is green or yellow, or whether a pea is smooth or wrinkled. Monogenic diseases include phenylketonuria and Huntingdon's disease. Until roughly 2005 (when GWASs began to be conducted), progress in identifying specific genetic variants was restricted to monogenic traits, whose inheritance patterns can be traced through family pedigrees.

Most diseases and other phenotypes—including virtually all phenotypes of interest to social scientists—are complex phenotypes. Examples include height, educational attainment, and liability to diseases such as schizophrenia and Type 2 diabetes. Twin, family, and adoption studies have focused on complex phenotypes. Much recent progress in medical genetics has been in the domain of complex phenotypes, based on methods such as GWAS. This paper focuses on theory and methods relevant to complex phenotypes.

As a broad generalization, genetic influences on monogenic and complex phenotypes operate through different biological mechanisms. The variants that determine monogenic traits typically have a large impact on a particular gene/protein, such as causing the protein not to function at all. By contrast, most common genetic variants that affect complex phenotypes are in loci outside of genes and are believed to operate through regulating gene expression, typically with long

and complicated causal pathways from genetic variation to phenotype. As a hypothetical example, a genetic variant may affect brain development, which affects a child's propensity to be obedient, which affects how teachers react to the child, which affects their interest in school and, ultimately, their educational attainment. While it has been possible to completely characterize the causal (biological) pathways for some monogenic traits, studies of causal pathways for complex traits instead typically focus on describing a particular part of the causal chain (e.g., gene expression levels, or mediating psychological characteristics).

E Mutations, (Natural) Selection, and Genetic Drift

The genetic variation present in a population is the result of four processes: mutation, recombination, (natural) selection, and genetic drift. *Mutations* are changes in DNA that occur due to internal or external mutagenic agents (e.g., ultraviolet light) or errors that occur when DNA is copied. Mutations in germ cells can be transmitted to future generations. A mutation *not* present in an individual's parents at conception is called a *de novo* mutation. Empirical estimates suggest that most individuals carry 40-120 de novo mutations (e.g., Figure 1 in Jónsson et al., 2017), a tiny fraction of the 6 billion parental alleles. Although each individual carries only a small number of *de novo* mutations, most of these will not be present in any appreciable frequency in other individuals. This fact, combined with the recent increase in the human population, explains why most genetic variations are rare (Uricchio, 2020). Recombination—described above in Section I.B—generates novel combinations of alleles at different loci, thereby increasing the diversity of haplotypes (multi-locus genotypes).

Selection is the change in allele frequencies due to their association with *fitness*. Fitness is a concept in evolutionary biology that quantifies the reproductive success of an organism and its descendants. Fitness is typically defined in terms of the number of descendants an organism has in the next and/or subsequent generations, as well as the time between generations, with larger numbers of descendants and shorter generation times indicating higher fitness. Alleles that are positively associated with fitness will tend to increase in frequency over time, whereas alleles that are negatively associated with fitness will tend to decrease in frequency over time.

Selection is categorized into different types. For example, directional selection uniformly favors an increased or decreased level of the phenotype; stabilizing selection favors some particular (optimal) level of the phenotype; and diversifying selection favors extreme levels of the phenotype. Which type of selection occurs depends on how phenotype values map onto fitness.

If a phenotype is subject to either directional or stabilizing selection, common variants will have relatively weak effects on the phenotype (e.g., Sanjak et al., 2017; Simons et al., 2018). For example, consider a new mutation that has a large effect on the phenotype. Stabilizing selection will tend to keep that mutation at low frequencies, regardless of the mutation's direction of effect on the phenotype.

Directional selection will push the frequency toward 0% or 100%, depending on the mutation's direction of effect (for the possible role of directional selection in the evolution of economic preferences, see, e.g., Galor and Moav, 2002; Galor and Michalopoulos, 2012; Galor and Özak, 2016; Galor and Savitskiy, 2018).

Genetic drift is the random fluctuation over the generations in allele frequencies in a population of finite size, due to randomness (i.e., factors unrelated to genotype) in individuals' number of offspring and randomness in which alleles are transmitted to offspring. In contrast to natural selection, which systematically changes allele frequencies, the changes in allele frequencies due to genetic drift are random walks. For the vast majority of genetic variants, which are not subject to strong natural selection, genetic drift is a stronger influence on allele frequencies than natural selection. Thus, because of genetic drift, if a single population splits into two populations that do not mate with each other for many generations, allele frequency and LD differences between the populations are measures of how long ago the populations split (see Section III.D).

F Genomic Data: Sequencing and Genotyping

For research purposes, genetic data are typically obtained from a saliva or blood sample. The two main technologies for measuring DNA are genome sequencing and genotyping arrays.

Genome sequencing, or *sequencing*, refers to reading segments of DNA sampled from the genome sequence. Sequencing technologies differ from each other in terms of coverage (how much of the genome is read) and accuracy (related to how many times each segment is read on average, to distinguish true genetic variations from sequencing errors). For example, sequencing for clinical diagnostics has high coverage of the clinically relevant genetic variants and is usually highly accurate. By contrast, low-pass sequencing—which is much less expensive and adequate for many research purposes—has low coverage and lower accuracy. For most common genetic variants, the accuracy of low-pass sequencing can be greatly improved by using the imputation strategies described below (see Li et al., 2021).

Rather than from sequencing, most human genetic data today comes from *SNP arrays*, which measure a pre-specified set of SNPs. The array is chosen to have high coverage of the haplotype blocks and other common genetic variants in a particular population (or across several populations). Thus, the SNPs measured on an array are correlated with, or "tag," the vast majority of variation in the genome that is due to common variants (including common, non-SNP genetic variants). Typical arrays used today measure roughly 1 million SNPs.

Using a *reference panel*—a dataset containing high-coverage, high-accuracy whole-genome sequencing data on a sample of individuals—SNP array data can be used to impute genotypes not included on the array. The reference panel is used to infer the underlying haplotypes of the SNP genotypes, where the missing genotypes of unmeasured SNPs are "filled-in" (probabilistically) using the reference haplotypes (van Leeuwen et al., 2015). When the reference panel closely

matches the target sample in genetic ancestry, the imputation can be highly accurate (see, e.g., Marchini and Howie, 2010). For example, in the UK Biobank, for 98.5% of genetic variants with frequency above 0.1%, the imputation captures at least 80% of the variance.

Often, researchers wish to jointly analyze genetic data collected from different genotyping arrays. This creates a problem because the set of SNPs differs across arrays. The solution is to use a reference panel to impute the genotypes of a common set of SNPs, and then conduct analysis on the imputed data. The most common reference panels used today for this purpose are the 1000 Genomes (The 1000 Genomes Project Consortium, 2015), the HRC (Loh et al., 2016; McCarthy et al., 2016), and TOPMed (Taliun et al., 2021). Starting with roughly 1 million directly measured SNPs, typical raw imputed genotype data today contains tens of millions of SNPs. After applying standard *quality control* filters (e.g., Winkler et al., 2014)—such as dropping SNPs with insufficiently high imputation accuracy—the number of imputed SNPs remaining is usually around or in excess of 10 million. In subsequent sections, when we refer to *measured SNPs*, we really mean SNPs whose allele count is either observed directly or imputed with high accuracy. The same problem (different genetic variants are measured across individuals) and solution (genotype imputation) arises with sequencing. For example, low-pass sequencing can measure 50 million or more base pairs in a given individual, but these are essentially randomly sampled from across the genome, with little overlap from one individual to the next. After imputation to the set of SNPs in a reference panel, however, the same SNPs are available across individuals, and the data can be jointly analyzed with the data from SNP arrays that have been imputed to the same reference panel.

Like sequencing technologies, SNP array technologies have experienced sustained, rapid declines in cost over the past few decades. Almost all data used in genome-wide association studies (see Section III.E) have been from SNP arrays because sequencing has been much more expensive. However, today the cost of low-pass sequencing is the same as genotyping, both roughly $30 per participant, and produces data that are at least as good (Li et al., 2021). We anticipate that datasets will increasingly switch toward sequencing technologies in the coming years.

G Gene Expression and Epigenetics

Gene expression is when a protein is produced based on the amino acid sequence coded by a gene. Where, when, and how much a gene is expressed depends on many factors, including epigenetic modifications, discussed next. Gene expression can be measured using protein levels in a cell or using the molecules produced from DNA that are ultimately translated into protein, called messenger RNA.

Epigenetic modifications are molecules that attach to the genome, thereby promoting or suppressing gene expression (but not affecting the sequence of base pairs in the genome). The set of epigenetic modifications across the genome is

called the *epigenome*. Epigenetic modifications drive some crucial biological processes, such as the differentiation of a stem cell into a particular cell type (brain cell, muscle cell, skin cell, etc.). Aging (e.g., Chen et al., 2016) and certain environmental exposures, such as smoking (e.g., Gao et al., 2015), are associated with changes in the epigenome. There are several technologies for measuring different types of epigenetic modifications.

Gene expression and epigenetics are large, important areas of research. Research that relates gene expression and epigenetics with behavioral phenotypes faces two special challenges that do not arise in research that relates genotypes with behavioral phenotypes. First, gene expression and epigenetic modifications are different across cell types and across cells. For most research on behavior, the most relevant cells are in the brain, but the available sample sizes of brain cells are small and the cells are from individuals who have died, which limits the research questions that can be addressed. Second, it is more difficult to identify causal effects of gene expression and epigenetic modifications on social and behavioral phenotypes because there is no natural experiment analogous to Mendelian segregation. Moreover, gene expression and a phenotype may mutually affect each other and may both be affected by third variables. Nonetheless, using quasi-experimental approaches, there is active progress studying the causal effects of behaviors and the social environment on gene expression (e.g., Nelson-Coffey et al., 2017) and epigenetics (e.g., Schmitz and Duque, 2022).

II Theoretical Framework: Genetic Effects

In this section, we lay out a theoretical framework for understanding the relationship between genetic variants and behavior. The framework builds on classic treatments in Fisher (1918) and Falconer (1960). Relative to this earlier work, we have clarified the assumptions and interpretation of the framework, especially when and how it can be interpreted causally, by using modern conceptual apparatus such as the potential outcomes notation (e.g., Rubin, 1974).

A Setup

We denote individual i's genotype at genetic variant j by $x_{ij} \in \{0, 1, 2\}$ and the individual's vector of genotypes at all SNPs by $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$. When there is no risk of confusion, we refer to this vector as i's "genotype." Crucially for defining the theoretical concepts in this subsection, we assume this vector includes all genetic variants in the genome. In subsequent sections, we discuss the challenges that arise in practice from observing only a subset of genetic variants. (Implicitly, this setup also incorporates the simplifying assumption that each genotype is fully characterized by a reference allele count equal to 0, 1 or 2.) For simplicity, we also restrict our discussion to genetic influences on complex phenotypes (see Section I.D) that can be treated as continuous variables, such as height, educational attainment or neuroticism.

We denote individual i's phenotype value by y_i. Individual i's potential outcome for genotype $\mathbf{x}$ is denoted $y_i(\mathbf{x})$. The causal effect for individual i of changing from genotype $\mathbf{x}$ to genotype $\mathbf{x}'$ is $y_i(\mathbf{x}') - y_i(\mathbf{x})$. These causal effects can be understood as the outcomes of an experiment (hypothetical in humans) where the genotype is modified at conception. For individual i, only the potential outcome at the individual's actual genotype, $y_i = y_i(\mathbf{x}_i)$, is observed. Because causal effects for an individual cannot be identified, we focus on average causal effects in some population of individuals.

Throughout this section, we focus on population parameters, deferring estimation of parameters from a sample until Section III. Thus, when we refer to the distribution of genotypes $\mathbf{x}_i$ in the population, it corresponds to the true probability of observing each $\mathbf{x}_i$, not the frequency in a particular sample.

B The General Framework

Consider a large population of individuals, each of whom has a complete set of potential outcomes for every possible genotype, $\{y_i(\mathbf{x})\}$, and a genotype, $\mathbf{x}_i$. The average causal effect in the population of changing from genotype $\mathbf{x}$ to genotype $\mathbf{x}'$ is

$$\mathbb{E}\left[y_i\left(\mathbf{x}'\right) - y_i(\mathbf{x})\right],$$

where $y_i(\mathbf{x}') - y_i(\mathbf{x})$ is the causal effect on individual i. We define the *genetic factor* as the mean potential outcome in the population for genotype $\mathbf{x}$, denoted $G(\mathbf{x}) \equiv \mathbb{E}[y_i(\mathbf{x})]$. In words, $G(\mathbf{x})$ is what the mean phenotype in the population would be if we intervened at conception and changed everyone's genotype to $\mathbf{x}$. It follows from this definition that, for each $\mathbf{x}$, each i's potential outcome can be written as

$$y_i(\mathbf{x}) = G(\mathbf{x}) + \nu_i(\mathbf{x}), \tag{1}$$

where $\nu_i(\mathbf{x})$ is the deviation of individual i's potential outcome from the population mean. All factors causing the same genotype to lead to different outcomes for different individuals, including different exogenous environmental exposures, gene-environment interactions, and measurement error in y that is independent of genotype, are captured by $\nu_i(\mathbf{x})$. It follows from the properties of the conditional expectation that $\mathbb{E}[\nu_i(\mathbf{x})|\mathbf{x}] = 0$.

Informally, heritability refers to the fraction of variance in the phenotype that is due to genetic variation in the population. Several formal definitions have been proposed; we review the main ones here and in Sections II.C and II.I. The most inclusive notion is called *broad-sense heritability*, denoted H^2. It is defined as the fraction of variance in the phenotype values across individuals explained by the genetic factor, given the distribution of $\mathbf{x}_i$ in the population:

$$H^2 \equiv \frac{\text{Var}\left(G\left(\mathbf{x}_i\right)\right)}{\text{Var}(y_i)},$$

where the variances are taken over the actual distribution of $\mathbf{x}_i$ (and $y_i = y_i\left(\mathbf{x}_i\right)$) in the population. Heritability is sometimes treated as a property of the phenotype. In fact, it should *not* be thought of as a structural parameter but rather a descriptive statistic in a particular population. Heritabilities can vary across time, place, and people for a variety of reasons that include differences in $G\left(\mathbf{x}\right)$, $\nu_i\left(\mathbf{x}\right)$ (and hence, y_i), and/or differences in the distributions of $\mathbf{x}_i$ and $\nu_i\left(\mathbf{x}\right)$.

C The Additive Model

In the general framework, the genetic factor $G\left(\mathbf{x}\right)$ may be an arbitrary function. In practice, researchers almost always work instead with a linear model, which is referred to as the *additive model*. The additive model is, in a specific sense derived below, a best linear approximation to the genetic factor, given the distribution of genotypes in the population.

Without loss of generality, we assume in all that follows that y and $\mathbf{x}_i$ have been centered to have mean zero. The variance-covariance matrix of the actual genotypes in the population, $\mathbf{\Sigma} \equiv \text{Var}\left(\mathbf{x}_i\right) = \mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i'\right]$, is called the *linkage disequilibrium (LD) matrix*. Its diagonal elements correspond to the variances of each genotype. Each off-diagonal element corresponds to the LD between a pair of genotypes. For simplicity, we assume the LD matrix is of full rank.

The *additive genetic factor* is defined as the fitted population regression function from regressing $G\left(\mathbf{x}\right)$ onto $\mathbf{x}$:

$$g\left(\mathbf{x}\right) = \mathbf{x}\beta, \tag{2}$$

where $\beta \equiv \text{argmin}_{\mathbf{b}}\mathbb{E}\left(G\left(\mathbf{x}_i\right) - \mathbf{x}_i\mathbf{b}\right)^2$ and the expectation is taken with respect to the distribution of $\mathbf{x}_i$ in the population. If the number of non-zero elements of β is large and their magnitudes are small, then by the central limit theorem, the additive genetic factor is approximately normally distributed in the population. Unlike the function $G\left(\mathbf{x}\right)$, the regression function $g\left(\mathbf{x}\right)$ depends on the population distribution of $\mathbf{x}_i$. An important implication is that, in populations with different LD patterns, the vectors β will differ (even if their $G\left(\mathbf{x}\right)$'s are identical). The error from approximating the genetic factor by the additive genetic factor, $N\left(\mathbf{x}_i\right) \equiv G\left(\mathbf{x}_i\right) - g\left(\mathbf{x}_i\right)$, is called the *non-additive genetic factor*. The deviations from linearity captured by the non-additive genetic factor are categorized into two types: dominance and epistasis. *Dominance* refers to non-linearity in the effects of a genetic variant. *Epistasis* (or gene-gene interaction effects) refers to interactions between the genotypes of two or more genetic variants.[5] By the properties of linear

[5]To formalize dominance, for each variant $j = 1, 2, ..., J$, define i's 1×2 vector $\gamma_{ij} = \left(\mathbb{1}\left\{x_{ij} = k\right\}\right)_{k=1}^{2}$ where $\mathbb{1}\left(\cdot\right)$ is the indicator function. Stacking the J vectors horizontally yields i's "extended genotype vector" $\mathbf{\Gamma}_i = \left\{\gamma_{i1}, \gamma_{i2}, ..., \gamma_{iJ}\right\}$. Then, in analogy with the derivation of the additive genetic factor above,

regression, the vector of $N(\mathbf{x}_i)$ across individuals is orthogonal to span $(\{\mathbf{x}_i\}_i)$. In most of what follows, our focus is on the additive model:

$$y_i(\mathbf{x}) = \mathbf{x}\beta + \epsilon_i(\mathbf{x}), \tag{3}$$

where the error term is $\epsilon_i(\mathbf{x}) = N(\mathbf{x}) + \nu_i(\mathbf{x})$. Since $\nu_i(\mathbf{x})$ may not be orthogonal to span $(\{\mathbf{x}_i\}_i)$, $\epsilon_i(\mathbf{x})$ need not be either. The additive genetic factor $g(\mathbf{x}) = \mathbf{x}\beta$ can be interpreted as the best linear approximation (in the sense of projection) of the genetic factor. The elements of the vector β can be interpreted as the average causal effects of the genotypes, given the population distribution of $\mathbf{x}_i$. It is sometimes stated that the additive model assumes away dominance, epistasis, and gene-environment interactions. But as our derivation makes clear, gene-environment interactions are captured by $\nu_i(\mathbf{x})$, whereas dominance and epistasis are captured by $N(\mathbf{x})$. Hence, the parameter vector β is well defined without imposing any strong assumptions about non-additive genetic effects, or the nature and extent of gene-environment interactions.

The *narrow-sense heritability*, denoted h^2, is defined as the fraction of variance in the population explained by the additive genetic factor, given the distribution of $\mathbf{x}_i$ in the population:

$$h^2 \equiv \frac{\mathrm{Var}(g(\mathbf{x}_i))}{\mathrm{Var}(y_i)}.$$

Like broad-sense heritability, narrow-sense heritability depends on both the phenotype and the population. By construction, narrow-sense heritability is weakly smaller than broad-sense heritability, and strictly smaller in the presence of dominance or epistasis.

For polygenic phenotypes, there are theoretical reasons to expect that narrow-sense heritability h^2 will be close to, or only moderately smaller than, broad-sense heritability H^2, despite the many known examples of biologically meaningful dominance and epistatic deviations (for a review, see Hill, Goddard and Visscher, 2008; see also Mäki-Tanila and Hill, 2014). For example, one reason why dominance deviations are unlikely to explain much of the phenotypic variation is statistical. Consider dominance deviations in a single-variant model where the genetic variant's minor allele a has frequency p. In a large population with random mating, the frequency of individuals with genotype aa will be p^2. This frequency will be very low if p is small, as is true for most genetic variants under realistic allele frequency distributions (see Section I.E). For those genetic variants, most of the variance in the phenotype will be due to individuals with the other two genotypes, AA and Aa; therefore, the regression model (3) will mostly just try to fit the phenotype values for the genotypes AA and Aa, and in doing so will explain

let $h(\mathbf{\Gamma}) = \mathbf{\Gamma}\phi$, where $\phi \equiv \mathrm{argmin}_{\mathbf{b}} \mathbb{E}(G(\mathbf{x}_i) - \mathbf{\Gamma}_i\mathbf{b})^2$, and define i's "dominance genetic factor" as $d_i \equiv \mathbf{\Gamma}_i\phi - \mathbf{x}_i\beta$. The proportion of phenotypic variance explained by this dominance genetic factor is known as the dominance variance. Epistasis and epistatic variance can similarly be formalized.

most of the variance in the phenotype. Empirical evidence has borne out the prediction that dominance deviations explain only a small fraction of phenotypic variance for many phenotypes, in some cases a negligible fraction (Pazokitoroudi et al., 2021; Hivert et al., 2021; Okbay et al., 2022). Similar statistical arguments have been invoked to argue that epistatic deviations are also unlikely to explain much variance for polygenic phenotypes in humans. Empirical evidence to date also supports this prediction (Hivert et al., 2021), albeit with wider confidence intervals.

The additive model is attractive for applications because it provides a reasonably good approximation to the general framework, in the sense that h^2 is typically close to H^2, and it is far more tractable: there is a single parameter of interest for each genetic variant, β_j, referred to as the *additive effect of genetic variant* j, or when the meaning is unambiguous, the *genetic effect*.

D Heritability, the Environment, and "Genetic Endowments"

Here, we use the framework in the previous section to highlight five common misconceptions about genetics research.

First, the error term $\epsilon_i(\mathbf{x})$ is often referred to as the "environment." However, $\epsilon_i(\mathbf{x})$ can also capture variance due to factors uncorrelated with $\mathbf{x}$, including non-additive genetic effects and gene-environment interactions (for further discussion, see Benjamin et al., 2012). Relatedly, 100% minus the heritability is frequently misinterpreted as the fraction of variance explained by the environment. To correctly understand the role of the environment in Equations (1) and (3), following Jencks (1980) we distinguish between environmental variables that are causally influenced by the genotype vector, which we call *endogenous*, and those that are not, which we call *exogenous*. For example, if parental investment is partly a response to a child's genotype, then that component of parental investment is an endogenous environmental variable. The error term $\epsilon_i(\mathbf{x})$ includes non-environmental factors such as those mentioned above, but also all aspects of the environment that vary independently of $\mathbf{x}$. Place of birth and age are two examples of environmental variables that can usually be treated as exogenous. By contrast, endogenous environmental factors are *by definition* part of the causal effect of genotype, as we discuss in more detail below (the fourth misinterpretation).[6] Since heritability includes the variance explained by endogenous environmental factors, the fraction of variance explained by all environmental factors—the sum of endogenous and exogenous environmental factors—could be larger or smaller than 100% minus

[6]Rather than distinguishing between exogenous and endogenous environmental factors, following (Plomin, DeFries and Loehlin, 1977) the behavior genetics literature instead typically categorizes gene-environment correlation into three types: evocative (the genotype evokes a reaction from others that affects the phenotype), active (an individual's genotype causes them to select into particular environments), and passive (an individual's genotype is correlated with the environment in which they are raised, e.g., due to parental genetic effects). In our terminology, both active and evocative gene-environment correlation are causal effects of genotype that operate through endogenous environmental factors, whereas passive gene-environment correlation is correlation between genotype and exogenous environmental factors.

the heritability (Jencks, 1980). In many situations, it may be more accurate to label $\epsilon_i(\mathbf{x})$ as the error term.

Second, the concept of heritability is sometimes misunderstood to be a measure of "how genetic" a phenotype is, as if it were a fixed parameter. In fact, it is a descriptive statistic about the population under study, much like an R^2. A nice empirical illustration is Branigan, McCallum and Freese (2013), who find substantial variation in the heritability of educational attainment by nationality, sex, and birth cohort.

Third, heritability is often erroneously interpreted as an index of policy relevance, with higher heritability leaving less room for environmental interventions to have an effect. But as noted by Goldberger (1979, p. 344): "The policy-relevant effect of an explanatory variable is properly measured by its regression slope, not by its contribution to R^2..." His memorable example is that eyeglasses can make a big difference even though naked eyesight is highly heritable. The facts that nutrition has led to large improvements in average height and that variation in height at any point in time is largely explained by genetic factors are not contradictory (Visscher, Hill and Wray, 2008). Moreover, policy can itself change heritability. For example, if income were redistributed from those with a high genetic factor for income to those with a low genetic factor, then the heritability of post-transfer income would be reduced. Rimfeld et al. (2018) find that after Estonia gained independence from the Soviet Union, the heritability of educational attainment and occupational status increased, and they interpret this change as resulting from an increase in meritocracy.

Fourth, as highlighted by Jencks (1980), people often assume that genetic effects operate through purely biological mechanisms that are immutable. Many examples of genetic effects taught in school, such as single genes determining whether a pea is smooth or wrinkled, fit this conception. However, genetic effects may—and for behavioral and other complex phenotypes, likely do—operate through environmental mechanisms. Consider the role of genetic factors in educational attainment. One hundred years ago, there would have been a large negative effect of having two X chromosomes on educational achievement. The subsequent dismantling of many formal and informal barriers facing women in the education system has reduced male-female gaps in educational outcomes, sometimes reversing them entirely. The misconception that environmental mechanisms are distinct from genetic effects is often expressed, even by leading genetics researchers. For example, the abstract of a high-profile paper on the genetics of obesity states: "Although often attributed to unhealthy lifestyle choices or environmental factors, obesity is known to be heritable" (Khera et al., 2019). The statement implicitly assumes that genes do not impact BMI through pathways that are "environmental," but that assumption is implausible. For example, genes could impact BMI in part through their effects on diet, exercise and other modifiable lifestyle choices which are associated with genes.

Fifth, especially in economic applications, the genetic factor (or additive ge-

netic factor) for educational attainment or cognitive performance is sometimes referred to as an individual's "genetic endowment." We view this terminology as problematic because it ignores important conceptual distinctions. In economic models, a genetic endowment refers to an individual's human capital stock at an initial time period. The genetic factor captures all mechanisms by which an individual's genotype ends up affecting the phenotype, including not only the initial human capital stock, but also investments by the individual and by parents, teachers, and others that are caused by the individual's genotype (see Sanz-de Galdeano and Terskaya, Forthcoming Web Appendix E for a simple model). It can also capture preferences, physical appearance, other non-human-capital variables, and non-investment reactions by others that are influenced by genotype and affect the phenotype. For example, if admissions committees for schools discriminate against applicants with certain physical characteristics, genetic influences on those characteristics may be part of the genetic factor for educational attainment.

To avoid misunderstandings, we recommend researchers define key concepts clearly, avoid inaccurate terminology whenever possible, and take reasonable precautions to preempt common sources of misunderstanding about what can and cannot be concluded from a study's analyses.

E Gene Expression and Epigenetics in the Framework of Genetic Effects

The misconception that genetic effects are necessarily biological and immutable is, in some sense, the opposite of another source of conceptual confusion that sometimes arises: how to reconcile the idea of genetic effects with the fact that gene expression is dynamic and modifiable (see Section I.G). The definition of genetic effects in terms of potential outcomes does not depend on how often or in what situations a relevant gene is expressed biologically; it only requires that we specify a measurable outcome that we would compare in the hypothetical experiment of modifying a person's genotype at conception. Some genetic variants have effects because they code for a different protein when the relevant gene is expressed, others have effects because they influence gene expression.

Where variation in gene expression and epigenetics enters into the theoretical framework for genetic effects depends on its source. Anything that would be different if the person were conceived with a different genotype is part of the genetic factor in Equation (1); in that case, gene expression is a mechanism through which the genetic effect operates. Genotype could directly affect gene expression, affect epigenetic modifications that change gene expression, or lead to endogeneous environmental responses that in turn affect gene expression. On the other hand, if exogenous environmental factors affect gene expression (or cause epigenetic modifications that affect gene expression), these effects are part of the error term in Equation (1). Finally, if genetically influenced gene expression interacts with exogenous environmental factors, the genetic effects are averages over these interactions, and the deviations for an individual from the population average are in the error term.

F Parental, Sibling, and Other Interpersonal Genetic Effects

When we use the term "genetic effect," we mean what we have been discussing so far: the effect of an individual's genotype on that same individual's phenotype. However, we sometimes instead call it a *self genetic effect* to distinguish it from an *interpersonal genetic effect*: the effect of an individual's genotype on someone else's phenotype. For example, a parent's genotype may influence their child's educational attainment, for example by affecting the parent's nurturing behavior or income. In the literature, what we call self genetic effects are sometimes called "direct genetic" effects, and what we call interpersonal genetic effects are variously called "indirect genetic," "associative," or "genetic nurture" effects. For a review of the genetics literature, almost entirely focused on non-human examples, see chapter 22 in Walsh and Lynch (2018).

Interpersonal genetic effects are relevant to social science because they operate via the exogenous environment of the individual. Interpersonal genetic effects sidestep the reflection problem (Manski, 1993) that bedevils other approaches to studying interpersonal influences. Thus, understanding the magnitudes and mechanisms of interpersonal genetic effects offers the promise of becoming a broadly useful approach to learning about how individuals are affected by the behaviors and environments generated by people around them.

The two main types of interpersonal genetic effects that we discuss are *parental genetic effects* and *sibling genetic effects*, which refer to the effects of the genotype of an individual's parent or sibling, respectively, on the individual's phenotype. Other interpersonal genetic effects include grandparental genetic effects and friend genetic effects (e.g., Sotoudeh, Harris and Conley, 2019).

Sibling genetic effects can be defined analogously to self genetic effects: a sibling genetic effect refers to the effect on an individual's phenotype from the hypothetical experiment of changing their sibling's genotype at conception. However, two subtleties arise. First, the relevant population for which sibling genetic effects are well-defined is the population of individuals who have siblings, which is a subset of the population for which self genetic effects are well-defined. Second, more than one causal parameter of interest may exist. For example, in families with three or more children, one parameter corresponds to the experiment of changing a random sibling's genotype holding other siblings' genotypes constant, while another parameter corresponds to the experiment of changing all siblings' genotypes.

Parental genetic effects are more complicated to define. When a parent's genotype is changed at conception, this change will affect the offspring through two pathways. First, it could change the allele that is transmitted to the offspring, which would then have a self genetic effect on the offspring. Second, it will have a self genetic effect on the parent, and then the parent's phenotypes (e.g., their income and behaviors) affect the phenotype of the offspring. The parental genetic effect is defined as the component of the overall causal effect that operates through this second pathway (for formal treatments, see Shen and Feldman 2020 and Young, 2023). Note that, since the parent's altered genotype could be trans-

mitted to other offspring, the parental genetic effect will include part of the sibling genetic effects acting on the focal offspring (Young et al., 2022).

G Twin, Family, and Adoption Studies

Prior to the availability of genetic data, most empirical human genetics research focused on estimating heritabilities using twin, family, or adoption studies. Even today, the majority of research on the genetics of behavioral phenotypes includes such studies (Becker et al., 2021). This work, launched in economics by Taubman (1976), has been previously reviewed for an economics audience (e.g., Beauchamp et al., 2011*a*; Benjamin et al., 2012; Sacerdote, 2011). Here, we touch on this research only briefly, for two purposes: to relate it to our framework and to provide a benchmark for genomic-data-based analyses discussed later.

The basic idea of a twin, family, or adoption study is to infer heritability and other variance components from the phenotypic resemblances of pairs of relatives who differ in their environmental similarity and genetic relatedness. The expected phenotypic resemblance of each such kinship (e.g., mono- or dizygotic twins, half siblings, etc.) is derived given some causal theory of how genetic and non-genetic factors determine a phenotype. Given sufficiently strong assumptions, for example about how genes and environmental factors are transmitted intergenerationally, the degree of assortative mating, the degree of gene-environment correlation, the processes through which adoptees are assigned to their rearing families, or the extent to which twins face special environments, the theory can be used to predict how each kinship correlation depends on a few structural parameters. Each kinship thus provides a moment condition, and the set of moment conditions jointly identifies the parameters of the model.

Before the modern molecular genetic era, a range of models seemed capable of "explaining" the available kinship correlations, despite major differences in their underlying assumptions about, for example, dominance, assortative mating, and gene-environment correlation (Loehlin, 1978). Despite some spirited debates, efforts to distinguish between these models were largely unsuccessful. The modern GWAS era has enabled researchers to make substantial progress on the important question of which of these explanations is the least wrong. For outcomes such as educational attainment, we now have compelling evidence that gene-environment correlations (Young et al., 2018) and assortative mating are strong (Robinson et al., 2017; Lee et al., 2018), whereas dominance variance is not a major source of phenotypic variance (Okbay et al., 2022). This new evidence rules out models that could not be ruled out by the available kinship data. The more general lesson is that Goldberger (1978, p. 72) was right to caution that misspecified models will generally deliver biased estimates but "not necessarily bad fits"—and that dismissing concerns about model misspecification based on the result of some goodness-of-fit test is often a mistake.

H Estimating Heritability Using Genetic Data

Whereas twin, family, and adoption studies estimate heritability by examining the relationship between phenotypic resemblance and *expected* genetic relatedness, an alternative approach estimates heritability by examining the relationship between phenotypic resemblance and the *realized* relatedness between pairs of relatives—relying on genetic data to directly measure realized relatedness. Genetic relatedness can be measured by inferring the segments of DNA inherited from common ancestors, called identity-by-descent (IBD) segments.

For siblings, segments are deemed IBD if they are inherited from the same parental haplotype. An alternative approach uses the variation in genetic relatedness between siblings (Visscher et al., 2006), which is due to random segregations of genetic material during meiosis and so should be independent of almost all environmental effects (sibling genetic effects being one exception). However, this approach requires *at least* several tens of thousands of genotyped siblings to obtain reasonably precise estimates of heritability. Moreover, it cannot separately the identify additive and non-additive components of heritability (Young and Durbin, 2014). Relatedness Disequilibrium Regression (RDR) is a generalization of the sibling approach to all relative pairs, which can increase power and reduce the influence of sibling genetic effects and non-additive genetic effects. However, RDR can only be applied settings where parental genotypes are available and may miss some heritability due to very rare variants (Young et al., 2018).

I (Un)measured Variants, the Additive SNP Factor, and SNP Heritability

One challenge for researchers is that only a subset of genetic variants is measured. As mentioned in Section I.F, genetic data (after imputation and quality control) often include ∼10 million SNPs, which capture much of the common genetic variation in European-genetic-ancestry populations. However, most datasets do not contain information about most rare SNPs nor non-SNP genetic variants that could have causal effects on the phenotype. It is therefore useful to define concepts analogous to the additive genetic factor and narrow-sense heritability that account for the omission of genetic variants. We define the *additive SNP factor* as the fitted population regression function from regressing $G(\mathbf{x})$ onto the vector of measured SNP genotypes $\tilde{\mathbf{x}}$:

$$\tilde{g}(\mathbf{x}) = \tilde{\mathbf{x}}\tilde{\beta}, \tag{4}$$

where $\tilde{\beta} \equiv \text{argmin}_{\mathbf{b}} \mathbb{E}\left(G(\mathbf{x}) - \tilde{\mathbf{x}}\mathbf{b}\right)^2$ and the expectation is taken over the distribution of $\mathbf{x}_i$ in the population. The additive SNP model is

$$y_i = \tilde{\mathbf{x}}_i\tilde{\beta} + \tilde{\epsilon}_i, \tag{5}$$

where $\text{Cov}(\tilde{\mathbf{x}}_i, \tilde{\epsilon}_i) \neq 0$ is possible, for example, due to parental genetic effects or other sources of gene-environment correlation. We *cannot* write Equation

(5) in terms of the potential outcomes $y(\mathbf{x})$ because it does not have a causal interpretation. Instead, the vector $\tilde{\beta}$ should be interpreted as the best linear approximation (in the sense of projection) to the average causal effects of the genotype vector, given the genotypes of the measured SNPs and the distribution of $\mathbf{x}_i$ in the population. The measured SNPs proxy for, or "tag," the causal effects of unmeasured genetic variants that they are correlated with. Therefore, in populations with different LD matrices, the vectors $\tilde{\beta}$ will differ, even if the genetic-effect vectors β are the same.

The vector of coefficients $\tilde{\beta}$ is related to the vector of genetic effects β by the formula for omitted-variables bias: $\tilde{\beta} = \beta_{\tilde{\mathbf{x}}} + (\mathrm{Var}(\tilde{\mathbf{x}}_i))^{-1} \mathrm{Cov}(\tilde{\mathbf{x}}_i, \mathbf{x} \setminus \tilde{\mathbf{x}}_i) \beta_{\mathbf{x} \setminus \tilde{\mathbf{x}}}$, where $\mathbf{x} \setminus \tilde{\mathbf{x}}_i$ is the vector of unmeasured genetic variants, $\beta_{\tilde{\mathbf{x}}}$ is the subvector of β corresponding to measured SNPs, and $\beta_{\mathbf{x} \setminus \tilde{\mathbf{x}}}$ is the subvector of β corresponding to unmeasured genetic variants.

SNP heritability, denoted $\tilde{h}^2$, is defined as the fraction of variance in the population explained by the additive SNP factor, given the population distribution of $\mathbf{x}_i$:

$$\tilde{h}^2 \equiv \frac{\mathrm{Var}(\tilde{g}(\mathbf{x}_i))}{\mathrm{Var}(y_i)},$$

where the variances are taken with respect to the population distribution of $\mathbf{x}_i$. By construction, SNP heritability is weakly smaller than narrow-sense heritability, and strictly smaller if there are unmeasured genetic variants that affect the additive SNP factor that are not perfectly correlated with linear combinations of measured SNPs.

SNP heritability depends on the set of genetic variants included in the analysis (and hence on the genotyping technology, imputation method, and quality-control filters), and it is therefore less interpretable than narrow-sense heritability. However, the additive SNP factor and SNP heritability are central to understanding polygenic indexes (Section IV).

Several estimators for SNP heritability have been proposed. For some of them, the core idea is similar to the approaches described in Section II.H in that they compare phenotypic resemblence to genetic resemblance. However, this class of estimators uses samples of approximately unrelated individuals who share no (or very few) identifiable IBD segments. Therefore, genetic resemblance between a pair of individuals is measured as the correlation of their measured SNP genotypes. This type of similarity is referred to as identity-by-state (IBS). The relationship between phenotypic and IBS genetic resemblance can be estimated using least-squares approaches (generally referred to as Haseman-Elston regression, in reference to closely related work in Haseman and Elston (1972)) or maximum-likelihood approaches (e.g., Genomic-relatedness-matrix REstricted Maximum Likelihood, or GREML (Yang et al., 2010)).

Heritability estimates based on IBS are expected to be smaller than those based on IBD since shared IBD segments have identical (unmeasured) rare and non-SNP

variants, whereas IBS-based estimators only pick up the effects of such variants to the degree that they are correlated with observed SNPs. In order to capture the effects of rare and non-SNP variants, GREML has been extended to whole-genome data, where it is used to estimate a quantity closer to narrow-sense heritability (Wainschtein et al., 2022).

Regardless of the set of SNPs used, these methods are biased to the extent that genotypic resemblance is correlated with environmental similarity (Young et al., 2018) and biased by assortative mating (Young, 2022). These biases and strategies used to mitigate them are discussed in greater detail in Section III.A.

J Genetic Overlap and Genetic Correlation

Genetic overlap refers to the extent of shared genetic influences on two phenotypes.[7] Conceptually, genetic overlap addresses the question of to what degree genetic factors drive the similarity between a pair of phenotypes. When the two phenotypes are the *same* phenotype in two different populations, genetic overlap reflects hetereogeneity in the additive genetic effects, which could be due to gene-environment interaction or non-additive genetic effects.

Genetic correlation is a measure of genetic overlap. Somewhat confusingly, the term "genetic correlation" is used to refer to two different measures of genetic overlap that are rarely acknowledged as different from each other (our discussion here is based on Okbay et al., 2016, Supplementary Information section 3; see also Border et al., 2022*a*). Which of the two parameters is more appropriate depends on the specific setting and research question.

One measure is the correlation (across individuals in a population) of the additive genetic factor across the two phenotypes, A and B:

$$r_{\mathbf{x}\beta} = \frac{\mathbb{E}\left[\left(\mathbf{x}_i\beta_A\right)\left(\mathbf{x}_i\beta_B\right)\right]}{\sqrt{\mathbb{E}\left[\left(\mathbf{x}_i\beta_A\right)\left(\mathbf{x}_i\beta_A\right)\right]\mathbb{E}\left[\left(\mathbf{x}_i\beta_B\right)\left(\mathbf{x}_i\beta_B\right)\right]}} = \frac{\beta_A^{'}\mathbf{\Sigma}\beta_B}{\sqrt{\beta_A^{'}\mathbf{\Sigma}\beta_A\beta_B^{'}\mathbf{\Sigma}\beta_B}}, \tag{6}$$

where $\mathbf{\Sigma} \equiv \mathrm{Var}\left(\mathbf{x}_i\right) = \mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i'\right]$ is the LD matrix. This measure is the most relevant for assessing how well a polygenic index (discussed in Section IV below) constructed to predict phenotype A will predict phenotype B. Twin and family studies can be used to estimate $r_{\mathbf{x}\beta}$, but the methods require very strong assumptions (for discussion, see Beauchamp et al., 2011*b*). GREML and related methods for estimating SNP heritability can be adapted to estimate a variant of $r_{\mathbf{x}\beta}$: the correlation of the additive SNP factor across A and B (e.g. Deary et al., 2012). Intuitively, the idea is to estimate, across pairs of individuals, the relationship be-

[7]Genetic overlap is related to the concept of pleiotropy. Although the precise definition depends on context, a genetic variant is said to be *pleiotropic,* loosely speaking, when it affects more than one phenotype. Genetic overlap corresponds to the extent of pleiotropy genome-wide. Sometimes pleiotropy refers more specifically to when a genetic variant affects the phenotypes through distinct biological mechanisms, or to when the phenotypes are in some sense unrelated to each other.

tween the within-pair covariance of A and B and the pair's genotypic resemblance on measured SNPs.

The other measure is the correlation (across genetic variants) of the causal effects of the variants on A with the causal effects of the variants on B:

$$r_\beta = \frac{\beta_A' \beta_B}{\sqrt{\beta_A' \beta_A \beta_B' \beta_B}}. \tag{7}$$

This measure corresponds to a hypothetical experiment where a single, randomly chosen genotype is changed at conception and the effect of that change on two different phenotypes is compared. As such, this measure of genetic correlation is most relevant to the concept of pleiotropy because it measures whether the actual self genetic effects are correlated rather than the additive genetic factors; these may differ if, for example, the causal genetic variants for A and B differ but are in LD with each other. The genetic correlation in Equation (7) is what LD Score regression aims to estimate (Bulik-Sullivan et al., 2015*b*).

Genetic correlations can differ from phenotypic correlations in interesting ways. For example, in the context of risk preferences, Karlsson Linnér et al. (2019) find that the genetic correlations between a survey-based measure of general risk tolerance and a variety of risky behaviors remain higher than the corresponding phenotypic correlations even after adjustment of the phenotypic correlations for measurement error (their Supplementary Tables 8 and 9). Karlsson Linnér et al. interpret this finding as supporting the view that a general factor of risk tolerance partly accounts for cross-domain correlation in risky behavior (Einav et al., 2016; Frey et al., 2017) and implying this factor is genetically influenced. The relatively low phenotypic correlations, which have been interpreted as evidence against a general factor (Weber, Blais and Betz, 2002; Hanoch, Johnson and Wilke, 2006), appear to be driven by (non-measurement-error) domain-specific contributors to the additive SNP model's error term.

III Estimation of Genetic Effects

An estimate of the genetic-effect vector, β, is the key input for most downstream analyses, including constructing polygenic indexes. In this section, we discuss the theoretical issues and practical challenges that arise in estimation.

A The Challenge of Causal Inference

The additive model in Equation (3) cannot be estimated because only one potential outcome is observed for each individual. With a slight abuse of notation, we can rewrite the analog of Equation (3) in terms of observables:

$$y_i = \mathbf{x}_i \beta + \epsilon_i. \tag{8}$$

Whereas in Equation (3) we were holding fixed the genotype vector $\mathbf{x}$ for everyone in the population, in Equation (8), the value of $\mathbf{x}_i$ corresponds to individual i's actual genotype vector, and the residual $\epsilon_i = \epsilon(\mathbf{x}_i)$ is the deviation of individual i's phenotype value from the prediction of the additive model, $g(\mathbf{x}_i) = \mathbf{x}_i\beta$. In Equation (8), the ordinary least squares estimator of β will be biased in the presence of correlation between genotype and the residual: $\text{Cov}(\mathbf{x}_i, \epsilon_i) \neq 0$.[8]

In practice, the main source of bias is *gene-environment correlation*. There are two leading sources of gene-environment correlation:

- *Population Stratification*: Individuals with shared genetic ancestry have more similar genotypes, but these individuals also share similar environmental exposures. A classic, memorable example is a hypothetical study of genetic causes of the use of chopsticks in a population of individuals living in San Francisco (Lander and Schork, 1994; Hamer and Sirota, 2000). For many genetic variants, individuals with Asian genetic ancestries have different allele frequencies than individuals with European genetic ancestries (largely because of genetic drift that occurred after the populations became relatively separate). For cultural reasons, these individuals are also more likely to use chopsticks.

- *Parental Genetic Effects*: Individuals with phenotype-increasing alleles have parents with these alleles. Parents with a high phenotype value may create a rearing environment that increases (or perhaps, in some cases, decreases) the child's phenotype value. For example, individuals with alleles that increase educational attainment are likely to have more highly educated parents. Such parents may read more to their children, or they may earn higher incomes that enable them to live in better neighborhoods and send their children to better schools.

The shared genetic ancestry relevant to population stratification may be in the distant past, as with continental ancestries, but shared genetic ancestry going back only a handful of generations could be associated with environmental influences (e.g., Zaidi and Mathieson, 2020). Other interpersonal genetic effects (besides parental) could also generate gene-environment correlation. For example, sibling genetic effects can generate negative gene-environment correlation if parents differentially invest in their children to offset genetic differences.

Because random assignment of genotype is infeasible in studies of humans, researchers instead address the causal inference problem by including a vector of control variables $\mathbf{z}_i$. While additional complications discussed in Section III.E will necessitate altering the specification, to clarify the conceptual issues we begin by writing the regression that researchers would want to estimate if they could:

[8] In this subsection we ignore assortative mating, because we are considering regression equations that include the full vector of genotypes $\mathbf{x}_i$ for all genetic variants in the genome. As we discuss in Section III.E, assortative mating exacerbates the omitted-variables bias that arises in GWAS due to running a regression on a single SNP at a time. This omitted-variables bias is absent when the full vector of genotypes $\mathbf{x}_i$ is included in the regression.

$$y = \mathbf{x}\beta + \mathbf{z}\gamma + \epsilon, \tag{9}$$

where we drop the i subscripts to make the notation more compact. If the genotype vector $\mathbf{x}$ is as good as random conditional on $\mathbf{z}$, then this strategy fully addresses the challenge of causal inference. Otherwise, the inclusion of controls will not fully rule out confounding factors. In Sections III.B and III.C we describe how, given appropriate family data, it is possible to construct $\mathbf{z}$ such that $\mathbf{x}$ is as good as random conditional on $\mathbf{z}$. However, most research to date has relied on imperfect controls that we discuss in Section III.D.

B The Natural Experiment of Mendelian Segregation

At each locus in an offspring's genome, one of each parent's two alleles is transmitted in a random process called Mendelian segregation (see Section I.B). Mendelian segregation generates a natural experiment that can be exploited to draw causal inferences. To see how, denote the genotype vectors of an individual's parents by $\mathbf{x}_f$ (father) and $\mathbf{x}_m$ (mother). Then we have

$$\mathbb{E}\left[\mathbf{x} \mid \mathbf{x}_f, \mathbf{x}_m\right] = \underbrace{\left(\mathbf{x}_f + \mathbf{x}_m\right)/2}_{\mathbf{x}_p}.$$

Under Mendelian inheritance, the expected genotype at SNP j is simply the average of the parental genotypes, which we denote $\mathbf{x}_p$. Next, define the deviation from the parental midpoint as $\mathbf{x}_r \equiv \mathbf{x} - \mathbf{x}_p$. Then we can decompose the genotype vector into two terms, one representing the parental midpoint, $\mathbf{x}_p$, and one representing the deviation from this midpoint, $\mathbf{x}_r$:

$$\mathbf{x} = \mathbf{x}_p + \mathbf{x}_r.$$

where $\mathbf{x}_r$ is a random deviation from $\mathbf{x}_p$, $\mathbb{E}\left[\mathbf{x}_r|\mathbf{x}_p\right] = 0$ and $\mathbb{E}\left[\epsilon|\mathbf{x}_r\right] = 0$, where ϵ is the error in the additive model (3).[9] Therefore, if Regression (9) is run with $\mathbf{x}_p$ included among the controls, the residual variation in $\mathbf{x}$ will isolate the random component $\mathbf{x}_r$, and the β estimator will have a causal interpretation.

A subtlety that has not been appreciated, to the best of our knowledge, is that controlling for the parental genotypes is equivalent to a two-stage least squares regression of y on $\mathbf{x}$ using $\mathbf{x}_r$ as an instrument for $\mathbf{x}$. Therefore, as Veller, Przeworski and Coop (2023) showed, the estimate will be a *local* average treatment effect (LATE, see Imbens and Angrist, 1996), i.e., a weighted average of treatment effects across individuals, with the individuals who contribute more identifying

[9] A subtlety here is that $\mathbf{x}_r$ is only mean independent of $\mathbf{x}_p$ and ϵ. That is because at any genetic variant j, only heterozygous parents contribute to variation in the offspring genotype. Therefore, the variance-covariance matrix of $\mathbf{x}_r$ and the variance of the non-additive genetic factor, which is part of ϵ, both depend on the parental genotype vector $\mathbf{x}_p$.

variation getting a larger weight.

For each genetic variant j, only individuals whose parents are heterozygous at j contribute identifying variation in genotype. In general, the LATE could differ from the average treatment effect in the population as a whole if the genetic effects differ across subpopulations with different frequencies of heterozygosity. Note that across genetic variants j, the estimated β_j's will be different weighted averages across individuals, depending on which individuals have heterozygous parents at j. See Appendix I for a more formal treatment.

In a randomly mating population without genetic drift, the amount of identifying variation in genotype $\mathbf{x}_r$ (which is sometimes denoted the "segregation variance"), $\text{Var}(\mathbf{x}_r)$, is half the total population variance in $\mathbf{x}$. This can be seen by first noting:

$$\text{Var}(\mathbf{x}_r) = \text{Var}(\mathbf{x}) - \text{Var}(\mathbf{x}_p).$$

Calculating the variance of the mean parental genotype therefore yields

$$\begin{aligned}\text{Var}(\mathbf{x}_p) &= \text{Var}\left(\frac{\mathbf{x}_f + \mathbf{x}_m}{2}\right) = \frac{1}{4}\left[\text{Var}(\mathbf{x}_f) + \text{Var}(\mathbf{x}_m) + 2\text{Cov}(\mathbf{x}_f, \mathbf{x}_m)\right] \\ &= \frac{1}{2}\text{Var}(\mathbf{x}),\end{aligned}$$

where $\text{Cov}(\mathbf{x}_f, \mathbf{x}_m) = 0$ and $\text{Var}(\mathbf{x}_f) = \text{Var}(\mathbf{x}_m) = \text{Var}(\mathbf{x})$ follow from our assumptions about random mating and zero genetic drift. If parents assortatively mate such that $\text{Cov}(\mathbf{x}_f, \mathbf{x}_m) > 0$, then $\text{Var}(\mathbf{x}_p) > \frac{1}{2}\text{Var}(\mathbf{x})$, and the amount of identifying variation is less than half the total population variance.

Rather than controlling for the mean parental genotype, including the father's and mother's genotypes separately as controls in Regression (9) would also identify the self genetic effect, because $\text{span}(\mathbf{x}_p) \subset \text{span}(\mathbf{x}_f, \mathbf{x}_m)$.[10] In some cases, estimating the coefficients separately may be of substantive interest, for example, in settings where the investigator has a hypotheses about their relative magnitudes. We highlight that neither the coefficients on the parental midpoint, $\mathbf{x}_p$, nor the coefficients on mother's and fathers' genotypes when estimated separately, should be interpreted causally (Shen and Feldman, 2020; Young, 2023).

Several caveats to the the natural experiment of Mendelian segregation should be noted. Depending on the age and recruitment mechanism of the study sample, differential survival by genotype may distort Mendelian segregation proportions and induce collider bias with other factors that affect survival. For example, gametes with genotypes that preclude conception will not become embryos, and embryos with lethal genotypes will not survive to birth, but such genotypes are rare (because natural selection acts against them). The same applies to genotypes that cause significant childhood mortality. However, some common variants af-

[10] In terms of precision, the choice of parental controls is unlikely to meaningfully matter. Theoretically, there are two opposing effects. On the one hand, controlling for $\mathbf{x}_f$ and $\mathbf{x}_m$ uses an additional degree of freedom. On the other, including both vectors can improve precision if one parental genotype vector absorbs more of the variation in ϵ. In practice, both effects are likely to be very small.

fect mortality later in life, and their segregation proportions could be distorted if study participants are recruited later in life. Moreover, segregation proportions will also be distorted for genotypes that affect recruitment into a study. The resulting difficulties are analogous to those that arise in the evaluation of a randomized controlled trial with endogenous attrition. These concerns do not apply if parents are recruited—whether randomly or not—and the analyses are done on the resulting offspring.

C Identification in Sibling Samples

Although samples with genotyped close relatives are becoming more common, such samples are still relatively scarce. The mean parental genotypes are only directly observed for genotyped parent-child trios, which are rare in the datasets currently available. Among genotyped family samples, sibling samples are the most common.

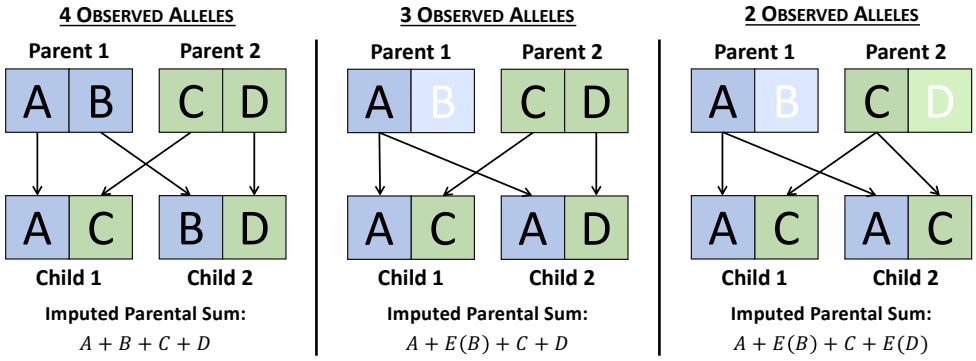
In genetic analyses with siblings, researchers have generally addressed the challenge of causal inference by using sibling fixed effects. An underappreciated point, highlighted by Young et al. (2022), is that the regression with sibling fixed effects is a biased estimator of the self genetic effect, β_j, if an individual's genotype also affects their sibling. The potential for bias arises because, instead of analyzing $\mathbf{x}_r$, a sibling comparison analyzes the deviation of each sibling's genotype vector from the mean sibling genotype. The identifying variation for the first sibling in a pair is thus given by:

$$\mathbf{x} - \left(\frac{\mathbf{x} + \mathbf{x}_{sib}}{2} \right) = \frac{1}{2} \left(\mathbf{x} - \mathbf{x}_{sib} \right),$$

where $\mathbf{x}_{sib}$ is the genotype vector of the focal individual's sibling. Since full siblings share a parental midpoint, variation in $\frac{1}{2}\left(\mathbf{x} - \mathbf{x}_{sib}\right) = \frac{1}{2}\left(\mathbf{x}_r - \mathbf{x}_{r,sib}\right)$ is random, with $\mathbb{E}\left[\left(\mathbf{x} - \mathbf{x}_{sib}\right)/2 | \mathbf{x}_p\right] = 0$ and $\mathbb{E}\left[\epsilon \left| \left(\mathbf{x} - \mathbf{x}_{sib}\right)/2 \right.\right] = 0$. However, when Regression (9) is run with the $\mathbf{z}$ being sibling fixed effects, the expectation of the within-sibling estimator is $\beta - \beta_{sib}$, where β_{sib} denotes the sibling genetic effect. Intuitively, the identifying variation in a sibling analysis is the focal individual's genotype vector relative to that of his/her sibling's; thus, the estimated coefficient is picking up both the effect of the individual having a higher genotype value and the effect of the sibling having a lower genotype value.

Young et al. (2022) proposed an alternative approach to analyzing sibling data, which enables separate identification of the self genetic effect: from the individual's and her sibling's genotype, impute the mean parental genotype (or, equivalently, the sum of the parental genotypes) using the laws of genetic inheritance (recombination and Mendelian segregation; see Section I.B), and then control for the mean parental genotype. Intuitively, this imputation enables identification because it adds information to the multiple regression, since the process of genetic inheritance is non-linear. The idea behind the imputation method is illustrated

Figure 1. : Mendelian Imputation of Parental Genotypes (Young et al., 2022).



Note: A, B, C, and D refer to arbitrary alleles and are colored white if not transmitted to either child.

in Figure 1. The first imputation case is when all four alleles are observed in the children. In this scenario, all four parental alleles can be recovered, and the mean parental genotype is known with certainty. In the remaining cases, only two or three of the parental alleles are observed in the children. In these situations, the expected mean parental genotype is imputed using sample allele frequencies in place of unobserved alleles. The imputation adds information when at least three of the four parental alleles are observed. Young et al. (2022) show that running Regression (9) where $\mathbf{z}$ includes the *imputed* mean parental genotype produces consistent and unbiased estimates of the self genetic effect (provided that the imputation itself is unbiased).

Even when there are no sibling genetic effects, controlling for the imputed parental genotypes is preferable to the sibling fixed-effects analysis because the former produces more precise estimates (Young et al., 2022).[11] The one subtlety is that sibling random effects must be included in Regression (9) to control for the unobserved differences across families that would be controlled for by sibling fixed effects. Young et al. (2022) also show how to obtain consistent and unbiased estimates of the self genetic effect via imputation of parental genotypes in samples with genotyped child-parent pairs, with or without any number of genotyped siblings.

[11]The fundamental reason is that the sibling fixed-effects analysis exploits less of the random identifying variation. In the imputed parent approach, the model is identified off of deviations from the mean of the observed parental genotypes. In a sibling fixed-effect approach, the model is identified off of deviations from the mean sibling genotype. These two approaches are identical when siblings inherit either perfectly overlapping or non-overlapping alleles from their parents (the outer cases in Figure 1). However, in the case when siblings inherit a common allele from one parent but distinct alleles from the other parent (the middle case in Figure 1), the common allele is double counted in the sibling mean, shading it toward the observed sibling genotypes and reducing the variance of the identifying variation. Formal details are in Young et al. (2022).

D *Imperfect Controls and Genetic Principal Components*

Since researchers have historically not had access to large genotyped family samples, the standard strategy for addressing concerns about one major source of gene-environment correlation, population stratification (see Section III.A), involved (i) restricting to samples of relatively homogeneous genetic ancestry, typically European-genetic-ancestry (see Section III.H), and (ii) relying on imperfect controls $\mathbf{z}$: typically the first several principal components of the genotype data—hereafter, *genetic PCs*—which we discuss below, and a small set of non-genetic controls, typically a polynomial in birth year (or age), sex, and their interactions. One major reason why the set of non-genetic controls is small is that genetic analyses have been meta-analyses that combined summary statistics from multiple datasets, many of them small, medical datasets with few covariates available. Birth year and sex are among the few covariates that are widely available and measured consistently across samples.

Often, the first several genetic PCs capture the orthogonal axes that explain the most genetic variance among the individuals in a particular sample (Menozzi, Piazza and Cavalli-Sforza, 1978). One early paper using PCs from SNP array data found that, when estimated in a sample of individuals throughout Europe living in the same place as their grandparents, to a remarkably high degree of accuracy the first PC captured a north-south axis of European geography and the second PC captured a west-east axis (Novembre et al., 2008). Other early papers found that, when estimated within a relatively genetically homogeneous sample such as individuals in Iceland, the genetic PCs similarly had natural interpretations corresponding to geographical axes (e.g. Price et al., 2009). Theoretical analyses showed that, in simple models where a population separates into distinct populations and then genetic drift causes the allele frequencies of the populations to diverge (see Section I.E), genetic PCs can identify which population an individual is from (Patterson, Price and Reich, 2006). These early findings, combined with the key advantage that genetic PCs can be calculated in any sample with SNP array data, quickly led to the adoption of the first few genetic PCs as standard control variables (Price et al., 2006). For a time, a common view among researchers was that controlling for genetic PCs was adequate to address population stratification.

In recent years, appreciation of the limitations of controlling for genetic PCs has been growing. We highlight three that we think are the most important. First, the first few PCs capture only gross features of genetic population structure (such as north-south and west-east location of genetic ancestry within Europe), but population-stratification bias could result from more subtle genetic population structure. For example, within Germany, Lutherans and Catholics may preferentially marry others of the same religion and have different cultural practices, but these two groups would not be differentiated by the first few genetic PCs. Researchers sometimes attempt to address this limitation by controlling for an even larger number of PCs (e.g., 100), but we are skeptical that this strategy will

be effective, in part due to the limitation we discuss next.

Second, large samples are needed to estimate more than a few PCs accurately (Patterson, Price and Reich, 2006; Bloemendal, 2019). Controlling for PCs that are estimated with substantial measurement error does not effectively control for those PCs. In practice, researchers often address this concern by estimating PC weights in some large, external dataset and then applying those weights in their own sample, but to effectively control for population structure, this approach requires that the sample in the external dataset and the analysis sample are sufficiently similar in genetic ancestry.

Third, genetic PCs computed in the standard way do not capture recent population structure (i.e., occurring within the last few generations; Zaidi and Mathieson, 2020; Abdellaoui et al., 2022*a*). Controlling for PCs thus does not correct for population stratification at the level of extended families, and controlling for genetic PCs does not address confounding of self-genetic effects from parental genetic effects. Therefore, even in a large sample in which many PCs can be accurately estimated and controlled for, potentially important sources of confounding remain.

The limitations of controlling for genetic PCs were made salient by several recent papers, which found substantial bias in analyses even after controlling for PCs (e.g., Sohail et al., 2019; Berg et al., 2019; Lee et al., 2018). The findings of these papers are one impetus for the growing interest in family-based genetic studies that control for parental genotypes, which, as discussed in Section III.B above, exploits a natural experiment that enables clean identification of self-genetic effects.

When imperfect controls (such as genetic PCs) are used, genetic studies do not have a clean causal interpretation and should instead be interpreted through a predictive framework. We define the optimal genetic predictor given a set of controls, $\mathbf{z}$, as the fitted population regression function from regressing y onto the vector of measured SNP genotypes, $\breve{g} \equiv \tilde{\mathbf{x}}\breve{\beta}$, where,

$$\left\{\breve{\beta}, \breve{\gamma}\right\} \equiv \arg\min_{\mathbf{b},\mathbf{a}} \left\{\mathbb{E}\left[(y - \tilde{\mathbf{x}}\mathbf{b} - \mathbf{z}\mathbf{a})^2\right]\right\}.$$

In what follows, we refer to $\breve{g}$ as the *optimal predictor* and $\breve{\beta}$ as the *optimal predictor weights*, without reference to the controls. Similarly, and parallel to the notion of SNP heritability from the causal model, we define $\breve{h^2} \equiv \mathrm{Var}\left(\breve{g}\right)$ as the *optimal predictive power* for y since it is the linear combination of observed SNPs with maximal predictive power for y given the set of controls. When the controls $\mathbf{z}$ are sufficient to control for potential biases (e.g., if $\mathbf{z}$ is the vector of mean parental genotypes), then the optimal predictor coincides with the additive SNP factor and the optimal predictive power coincides with the SNP heritability.

E Genome-Wide Association Studies (GWAS)

As discussed in Section III.A above, researchers would like to estimate the multivariate regression in Equation (9), restated here for convenience:

$$y = \mathbf{x}\beta + \mathbf{z}\gamma + \epsilon,$$

where $\mathbf{x}$ is a vector of *all* genetic variants in the genome and $\mathbf{z}$ is a vector of controls. Three complications arise when trying to make inferences about the parameters in this equation.

First, as mentioned in Sections I and I.F, only a subset of the genetic variants is measured.

Second, even among the SNPs that are measured, many are in high, sometimes perfect, LD with other genetic variants, especially those physically nearby on the same chromosome (see Section I.B). The inclusion of perfectly correlated SNPs would lead to failure of the full rank condition, making estimation of Equation (9) impossible.

Third, the number of individuals included in the regression is typically much smaller than the number of measured SNPs, also leading to failure of the full rank condition even if all SNPs are pairwise uncorrelated. Historically, large genetic studies were meta-analyses of GWASs, each with sample sizes of only a few thousand individuals. Even today, the largest datasets have 100,000 up to a few million individuals—still fewer than the number of measured SNPs. In principle, machine-learning methods can deal with the latter two challenges. While such approaches are actively explored (Vattikuti et al., 2013; Mieth et al., 2016), they have not (yet) been widely adopted.[12]

For these reasons, the standard approach—called *GWAS*—has been to estimate separate regressions by SNP,

$$y = \beta_j^{\text{GWAS}} x_j + \mathbf{z}\gamma_j + \eta_j, \tag{10}$$

for each of the measured SNPs j, where the vector of control variables $\mathbf{z}$ is the same as in Equation (9). These regressions generate coefficients β_j^{GWAS} that have a straightforward interpretation in terms of the parameters of interest β_j from Equation (9). If the residual x_{ij}'s after partialling out the controls $\mathbf{z}$ were mutually uncorrelated, then $\beta_j^{\text{GWAS}} = \beta_j$. More generally, however, β_j^{GWAS} differs from β_j because of LD between SNP j and all other $J-1$ genetic variants in the

[12]There are at least three major reasons for this. First, machine learning methods generally are computationally intensive and require individual-level data. Due to privacy and IRB restrictions for genetic data, to obtain large sample sizes researchers typically have to meta-analyze regression results across datasets (rather than pooling and analyzing individual-level data). Many machine-learning methods cannot be directly applied due to these restrictions. Second, many machine learning methods generate biased effect estimates, making them hard to interpret. Third, although machine learning may prove especially useful for prediction by flexibly accommodating non-linearities (e.g. Raben et al., 2023), for most phenotypes, a linear model is expected to capture most of the predicted variance (see Section II.C).

genome (including unmeasured variants):

$$\beta_j^{\text{GWAS}} = \sum_{k=1}^{J} \frac{r_{jk\perp\mathbf{z}}}{r_{jj\perp\mathbf{z}}} \beta_j, \tag{11}$$

where $r_{jk\perp\mathbf{z}}$ is the covariance between the residual genotype of SNP j and the residual genotype of genetic variant k, after partialling out the controls $\mathbf{z}$. To address this omitted-variables bias, as we discuss further below, virtually all analyses of GWAS results take into account the LD matrix of the population under study in some way.

If the vector of controls $\mathbf{z}$ isolates the random component of the genotype vector $\mathbf{x}_{r,i}$, then the relevant LD matrix is $\mathbf{\Sigma}_r = \mathbb{E}\left(\mathbf{x}_{r,i}\mathbf{x}_{r,i}'\right)$. In that case, covariances are zero for genetic variants located on different chromosomes, since chromosomes are inherited independently. Thus, Equation (11) can be written as $\beta_j^{\text{GWAS}} = \sum_{k \in chr(j)} \frac{r_{jk}}{r_{jj}} \beta_j$, where the r_{jk}'s are now the elements of $\mathbf{\Sigma}_r$. If the vector of controls $\mathbf{z}$ does not fully isolate the random component of the genotype vector—as is likely when imperfect controls such as genetic PCs are relied on—then the relevant LD matrix is $\mathbf{\Sigma}_{\perp\mathbf{z}} = \mathbb{E}\left(\mathbf{x}_{i\perp\mathbf{z}}\mathbf{x}_{i\perp\mathbf{z}}'\right)$, where $\mathbf{x}_{i\perp\mathbf{z}}$ is the residual genotype vector after partialling out the control vector $\mathbf{z}$. In this case, if the phenotype (or a correlated phenotype) is subject to assortative mating, the controls are unlikely to adequately address confounding from the LD that is due to assortative mating, including cross-chromosome LD (see Section I.C). Thus, for every SNP in the GWAS, Equation (11) needs to sum over the entire genome.

The primary output of a GWAS is the *GWAS summary statistics*: the vector of GWAS estimates for the K measured SNPs,

$$\hat{\beta}^{\text{GWAS}} = \left(\hat{\beta}_1^{\text{GWAS}}, \hat{\beta}_2^{\text{GWAS}}, ..., \hat{\beta}_K^{\text{GWAS}}\right)',$$

together with their standard errors or p-values.

In a GWAS, the conventional p-value threshold for statistical significance is 5×10^{-8}, which is called the *genome-wide significance threshold*. This threshold can be understood as the Bonferroni-corrected 0.05 threshold, given that there are roughly 1 million independent statistical tests for a GWAS, after accounting for the LD between the >1 million measured SNPs (Panagiotou and Ioannidis, 2012). Experience indicates that, to date, this threshold has kept the rate of false positives low in GWAS (e.g. Okbay et al., 2016). However, as genotyping technology improves and captures rarer SNPs (which necessarily have weaker LD with other SNPs), GWASs involve more than 1 million independent statistical tests in European-genetic-ancestry samples (Wu et al., 2017). Moreover, even with current genotyping technology, there are many more than 1 million independent statistical tests in samples where LD is on average weaker, such as African-genetic-ancestry samples. In these cases, lower p-value thresholds will be needed to keep

the rate of false positives as low as it has been.

The stringent significance threshold, combined with the very small fraction of variance explained by individual SNPs for polygenic phenotypes, means that GWAS sample sizes have had to be large to have adequate power. For example, the largest SNP associations with body mass index (BMI) have an R^2 of roughly 0.003 (Locke et al., 2015); those with educational attainment, roughly 0.0002 (Okbay et al., 2016). To attain 80% power to detect these effects at the genome-wide significance threshold requires sample sizes of ∼13,000 and ∼200,000, respectively.

When a SNP is genome-wide significant, many nearby SNPs will typically also be genome-wide significant, due to the high local LD. To avoid counting the same signal multiple times, there are several standard algorithms used to restrict attention to a more limited set of *lead SNPs*. For example, a simple, iterative approach is to identify the SNP with the lowest p-value as a lead SNP, dropping all genome-wide-significant SNPs whose absolute correlation with the lead SNP exceeds some threshold, then designating the SNP with the lowest p-value among the remaining SNPs as a lead SNP, and so on.

F GWAS Follow-Up Analyses

The exponential growth in genetic knowledge in recent years has been fueled, in large part, by follow-up analyses conducted on the results of increasingly large-scale GWASs. In the social sciences, the most important of these often use GWAS summary statistics to construct weights for polygenic indexes, the topic of Section IV. Here, we briefly describe other common follow-up analyses.

As noted in Section III.E, one challenge for the analysis of GWAS results is that individual-level data are often not available. Therefore, many of the follow-up analyses are designed to use the GWAS summary statistics instead. Another challenge is that the the LD matrix for the GWAS sample is often not made available. In its place, many of the follow-up analyses instead use the LD matrix estimated from a reference panel (see Section I.F).

In medical research, GWAS primarily aims to discover biological processes linked to diseases, thereby guiding the identification of potential targets for drug development. Many follow-up analyses therefore aim to identify the genes or regulatory elements of the genome that are causally responsible for the GWAS associations. The simplest such analysis looks for genes that are close to the set of lead SNPs. These genes are then looked up in databases listing the known functions of those genes, often validated in biology labs, providing some evidence that those functions may be important for the phenotype analyzed in the GWAS. However, more sophisticated methods have also been developed and applied. For example, some such "gene discovery" methods use the GWAS summary statistics and the LD matrix to conduct gene-level tests of association (e.g., de Leeuw et al., 2015), which aggregate across multiple SNPs. While identifying potentially relevant genes has value, most genetic effects on complex phenotypes are from genetic variants outside of genes that play some role in regulating gene expression (see,

e.g., Mostafavi et al., 2023, and the references therein). Another kind of analysis, called "fine-mapping," aims to localize the source of a GWAS association by jointly analyzing GWAS summary statistics and LD matrices (potentially from multiple genetic ancestry groups that have different LD patterns) (for a review, see Schaid, Chen and Larson, 2018). A large body of research in bioinformatics combines GWAS summary statistics and LD matrices with databases on gene expression, gene regulation, protein coding, and/or protein-protein interactions to refine our understanding of biological pathways from genotype to phenotype (e.g., Pers et al., 2015).

Another type of follow-up analysis uses the summary statistics to estimate SNP heritability (see Section II.I). The most commonly used method is LD Score Regression (Bulik-Sullivan et al., 2015*a*), which exploits the relationship in Equation (11) between β_j^{GWAS} and SNP j's LD with other SNPs. Specifically, under the assumptions of the model, the expected chi-squared test statistic (i.e., squared z-statistic) for any SNP j,

$$\mathbb{E}\left[\chi^2\right] = \mathbb{E}\left[\left(\frac{\beta_j^{\text{GWAS}}}{\sqrt{\text{Var}\left(\beta_j^{\text{GWAS}}\right)}}\right)^2\right],$$

is linear in the square of the SNP's "LD score," $\ell_j \equiv \sum_{k=1}^{K} r_{jk}^2 / (r_{jj} r_{kk})$, which is a measure of its total correlation with other SNPs. The coefficient on this relationship is a known linear function of the SNP heritability. Intuitively, for a complex phenotype diffusely influenced by genetic variants spread across the genome, the SNPs with higher LD scores—those that are in stronger LD with more SNPs—will be more strongly associated with the phenotype in a GWAS, with the slope of this relationship distinguishing between phenotypes that are more or less strongly genetically influenced overall. Because it can be applied using only GWAS summary statistics, LD Score Regression is the most widely used method of estimating SNP heritability (and genetic correlation, discussed next), but it relies on many strong assumptions, and plausible violations of them can cause substantial bias (see, e.g., Speed and Balding, 2019, Berg et al., 2019 and Border et al., 2022*b*). When imperfect controls are used in the GWAS, LDSC estimates the optimal predictive power, $\breve{h}^2$, rather than the SNP heritability.

Using two sets of GWAS summary statistics for different phenotypes, together with an LD matrix, another kind of analysis estimates the genetic overlap across phenotypes. The most commonly used method is bivariate LD Score Regression (Bulik-Sullivan et al., 2015*b*). Under some assumptions, for any SNP j, the expected value of the product of the t-statistics from the two GWASs is linear in the SNP's LD score, with the regression coefficient proportional to the genetic correlation r_β defined in Equation (7). Thus, a regression across all measured SNPs of the product of GWAS test statistics on the LD scores estimates the

genetic correlation r_β (or, when the summary statistics come from GWASs with imperfect controls, it estimates the correlation of optimal predictor weights).

Methods of using GWAS summary statistics to estimate heritability and genetic correlation can be extended to estimate how much of the heritability or genetic correlation is due to different types of SNPs. These extensions draw on continually updated databases that categorize SNPs according to whether or not, for example, they affect different tissues (musculoskeletal, brain, liver), they are involved in particular functional pathways, or the genomic sequence they are part of has been conserved by evolution. For example, extensions of LD Score Regression make use of sets of LD scores that represent how correlated a SNP is with other SNPs in a particular category (e.g., Finucane et al., 2015).

G Population, Sibling, and Family-Based GWAS

GWASs can be categorized according to the vector of control variables $\mathbf{z}$ included in regression (10):

- *Family-Based*: $\mathbf{z}$ includes the mean parental genotype at SNP j, $x_{par,j}$, or both the maternal and paternal genotypes at SNP j, $x_{m,j}$ and $x_{f,j}$.
- *Sibling*: $\mathbf{z}$ includes sibling fixed effects.
- *Population*: $\mathbf{z}$ includes other ancestry controls, such as genetic PCs.

All of these GWASs differ from the regression (9) on the entire genotype vector in that only one SNP at a time is included as a regressor. Although these GWASs differ from each other in terms of how well they control for various confounds, they all aim to estimate a particular weighted sum of causal effects. Specifically, the estimand of each GWAS is given by Equation (11) with the causal effects, repeated here for convenience:

$$\beta_j^{\text{GWAS}} = \sum_{k \in chr(j)} \frac{r_{jk}}{r_{jj}} \beta_j,$$

a weighted sum of causal effects, where the r_{jk}'s are now the elements of $\mathbf{\Sigma}_r = \mathbb{E}\left(\mathbf{x}_{r,i}\mathbf{x}_{r,i}'\right)$, which includes unmeasured variants. Arguments analogous to those in Sections III.B, III.C, and III.D tell us how well each type of GWAS identifies β_j^{GWAS}, as we summarize now.

Family-based GWAS identifies β_j^{GWAS} from the random component of the genotype at SNP j, the deviation of the genotype from its expectation: $x_{j,r} = x_j - \mathbb{E}\left[x_j | x_{j,p}\right] = x_j - x_{j,p}$. This identifying variation is mean-independent from the parental genotype and from the GWAS regression error: $\mathbb{E}\left[x_{j,r} | x_{j,p}\right] = 0$ and $\mathbb{E}\left[\eta_j | x_{j,r}\right] = 0$. Thus, family-based GWAS is an unbiased and consistent estimator for β_j^{GWAS}.

Sibling GWAS uses the difference between the individual's genotype at SNP j and his or her siblings' genotype. In the most typical case of sibling pairs, the

identifying variation is $x_j - \left(\frac{x_j + x_{j,sib}}{2}\right) = \frac{1}{2}(x_j - x_{j,sib})$. Although this variation is mean-independent from the parental genotype and from the GWAS regression error — $\mathbb{E}\left[\frac{1}{2}(x_j - x_{j,sib})\,|x_{j,p}\right] = 0$ and $\mathbb{E}\left[\eta_j|\frac{1}{2}(x_j - x_{j,sib})\right] = 0$ — it identifies $\sum_{k=1}^{K} r_{jk}(\beta_j - \beta_{j,sib})$ rather than $\sum_{k=1}^{K} r_{jk}\beta_j$. Thus, sibling GWAS is a biased estimator for β_j^{GWAS} whenever $\beta_{j,sib} \neq 0$. Rather than running sibling regression, imputing parental genotypes and then running family-based GWAS eliminates the bias, as long as family random effects are included to control for non-genetic family variation, and is also more efficient (Young et al., 2022).

Population GWAS is a biased estimator for β_j^{GWAS} to the extent that the controls $\mathbf{z}$ are an imperfect proxy for the parental genotypes. In this scenario, some omitted-variable bias persists since the genotype x_j conditional on $\mathbf{z}$ is not uncorrelated with the residual. If the controls do not perfectly capture genetic ancestry, then there will be some bias due to population stratification, parental genetic effects, assortative mating, and other sources of correlation between x_j and η_j.

Currently, nearly all published GWASs are population GWASs because datasets with genotyped family members have been too small to enable well-powered sibling or family-based GWASs. Holding fixed the number of individuals in the GWAS, sibling and family-based GWASs are usually much less powerful than population GWASs for three reasons. First, the population GWAS estimates tend to be biased away from zero. Second, the within-family genetic variation is one-half or less of the population genetic variation (see Section III.B), leading to larger standard errors. Third, if the study includes sibling pairs, some information is used to estimate the random or fixed effect for each sibling pair, reducing the degrees of freedom.[13] Roughly speaking, a GWAS using parental genotypes as controls requires roughly twice the number of individuals (not including parents) than a population GWAS to obtain comparably sized standard errors, and a sibling-based GWAS needs even more.

To date, only one large-scale GWAS based on primarily family data has been published Howe et al. (2022*b*). It is a meta-analysis of sibling GWASs for 25 phenotypes conducted in 19 datasets, with a total sample size ranging from roughly 13,000 to roughly 164,000 European-ancestry individuals, depending on the phenotype. For molecular phenotypes, such as low-density-lipoprotein cholesterol, the study's results were largely in line with previously reported findings in population GWASs. By contrast, for many of the social and behavioral phenotypes—educational attainment, age at first birth, number of children, cognitive ability, depressive symptoms, and smoking—the sibling-GWAS estimates of β_j^{GWAS} were smaller in magnitude on average than the population-GWAS estimates, consistent with what had previously been reported for some social and

[13] By contrast, if siblings have correlated residuals, or if the parental genes explain a high proportion of the residual variance, then standard errors will be reduced. In practice, however, these effects only generate modest efficiency gains and are never large enough to offset the loss in power from the three factors described above.

behavioral phenotypes. For example, Lee et al.'s (2018) population-based GWAS of educational attainment reported a follow-up analysis of 22,135 sibling pairs that found within-family associations were deflated by $\sim 40\%$. They found that assortative mating could explain at most one third of the observed deflation, with most remaining deflation likely explained by omitted-variable biases (from non-genetic omitted factors) in the original GWAS. For comparison, they conducted an analogous analysis of height, finding more modest deflation, all of which could plausibly be attributed to assortative mating (See also the follow-up analyses in Okbay et al., 2022).

Howe et al.'s (2022*b*) sibling GWAS also produced some notable differences in the GWAS follow-up analyses. For example, the genetic correlation between BMI and educational attainment, estimated from population-GWAS summary statistics, was -0.32, compared to -0.05 when the sibling summary statistics were used. A plausible interpretation of the discrepancy is that the $\mathbf{z}$'s used in the original population-based GWASs failed to eliminate all confounding from factors that have correlated effects on BMI and education. We anticipate that the coming years will see many more sibling and family-based GWASs, and other conclusions drawn from population GWASs will need to be updated accordingly.

H Eurocentric Bias

As noted above in Section III.D, in an attempt to mitigate bias from gene-environment correlation, GWASs have traditionally been restricted to samples of relatively homogeneous genetic ancestry. The largest such samples have been from countries in Europe, the UK, the US, Australia, and New Zealand (Mills and Rahal, 2019), partly because these countries are wealthy and had the resources to fund large-scale genotyping efforts. As of June 2023, based on data from the GWAS Diversity Monitor (Mills and Rahal, 2020), an online database of GWASs, the average percentage of European-genetic-ancestry subjects in published studies is $\sim$95%, compared to their ˜15% share of the global population. This disproportionate tilt of genetics research is called *Eurocentric bias*.

Eurocentric bias in GWASs is widely considered to be a major problem (e.g., Martin et al., 2019; Duncan et al., 2019). One concern is that genetics research will disproportionately benefit individuals of European genetic ancestries, for example, because the research has focused on diseases that are more prevalent among these individuals. Another concern is that the common practice of dropping data from other genetic ancestry groups is scientifically inefficient, especially since such data can be especially valuable for "fine-mapping" studies that aim to identify causal genetic variants (see Section III.F). A third concern is that, as we discuss in Section IV.C, the polygenic indexes constructed from existing GWAS results are less predictive among individuals with non-European genetic ancestries. In social-science applications, this "limited portability" of polygenic indexes often reduces their value.

Many efforts to mitigate Eurocentric bias are currently underway. Some of these

are national biobanking efforts in some non-European-genetic-ancestry counties, with the largest samples to date being the China Kadoorie Biobank (a study with ∼100,000 genotyped individuals currently) and Biobank Japan (200,000 genotyped individuals currently). In the U.S., initiatives such as the Million Veterans Project, All of Us, the Multi-Ethnic Study, and the Atlas Study are collecting relatively large minority samples. The Pan UKB Project has analyzed data from the UK Biobank for individuals with non-European genetic ancestries that would normally be discarded. Direct-to-consumer genetic testing companies, despite having a disproportionately European-genetic-ancestry customer base, nonetheless have many non-European-genetic-ancestry customers who have consented to participate in research. Some of these companies, such as 23andMe, are helping to mitigate Eurocentric bias by contributing to GWASs in diverse samples (Yengo et al., 2022). Funding agencies, including the U.S. National Institutes of Health, have prioritized collecting and analyzing genetic data from non-European-genetic-ancestry samples. Major journals for genetics research have prioritized publishing such work.

Unfortunately, the genetic-data-collection efforts in developing countries remain small. This hampers genetics research since populations in these countries, especially in Africa, harbor a huge fraction of the global genetic diversity (Mills and Rahal, 2019).

IV Polygenic Indexes

Most applications using genetic data in the social sciences use polygenic indexes (PGIs), and increasingly so in recent years. Although PGIs had been discussed earlier (Wray, Goddard and Visscher, 2007), the first paper in humans genetics to construct and analyze a PGI was a GWAS of schizophrenia published in 2009 (Purcell et al., 2009). Since 2009, PGIs have been increasingly used in research related to the genetics of behavioral phenotypes (Becker et al., 2021). There are two main reasons for the use of PGIs in social-science research, one statistical and one conceptual. Statistically, because PGIs generally explain much more variance than individual SNPs, analyses using a PGI will generally have much greater statistical power. Conceptually, as we discuss, the PGI is an empirical proxy for the additive SNP factor and thus captured the combined explanatory power of measured SNPs. In this section, we discuss PGIs, their predictive power, and their appropriate interpretation.

A PGI Definition and Interpretation

In general, we define a PGI as a standardized, weighted sum of the genotypes of a set of measured genetic variants:

$$g_{\mathbf{w}} \equiv \frac{\tilde{\mathbf{x}}\mathbf{w}}{\text{std}\left(\tilde{\mathbf{x}}\mathbf{w}\right)},$$

where $\tilde{\mathbf{x}}$ is the vector of measured genotypes, $\mathbf{w}$ is a vector of weights (the "PGI weights"), and $\text{std}(\cdot)$ takes the standard deviation of its argument across a population of individuals. Typically, the measured genotypes are SNPs, and the PGI weights are chosen with the goal of having the PGI approximate the standardized additive SNP factor for some phenotype (defined in Section II.I) as well as possible.[14]

The PGI would equal the standardized additive SNP factor if $\mathbf{w} = \tilde{\beta}$, where $\tilde{\beta}$ is the vector of population regression coefficients defined in Equation (5). However, the vector of weights cannot be set equal to $\tilde{\beta}$ because $\tilde{\beta}$ is unknown. In practice, researchers usually aim to set $\mathbf{w}$ equal to an estimate of $\tilde{\beta}$. Most of the commonly used estimators take as inputs a set of GWAS summary statistics ($\hat{\beta}^{\text{GWAS}}$ and their standard errors), an LD matrix estimated in some reference sample, and a prior distribution of effect sizes.[15] The estimators adjust the GWAS estimates to take into account correlation across SNPs, as captured by the LD matrix, and shrink them toward the prior. Specifically, the estimators set each SNP's PGI weight equal to the mean of its Bayesian posterior-effect distribution; the estimators differ from each other mainly in their assumptions about the prior distribution and, for computational tractability, in the assumptions and approximations they make about the LD matrix (e.g., Vilhjálmsson et al., 2015; Ge et al., 2019; Zhang et al., 2021; Lloyd-Jones et al., 2019). These differences affect finite-sample performance and computational speed but do not matter for the purposes of discussion here.

We denote the resulting weights by $\hat{\beta}$ and the corresponding PGI by

$$\hat{g} \equiv \frac{\tilde{\mathbf{x}}\hat{\beta}}{\text{std}\left(\tilde{\mathbf{x}}\hat{\beta}\right)}.$$

In practice, $\hat{\beta}$ is based on GWAS estimates using imperfect controls, meaning that it is an estimate of the optimal predictor weights, $\breve{\beta}$. (If the controls were sufficient for the GWAS to have a causal interpretation, then $\hat{\beta}$ would be an estimate of the additive SNP factor weights.) In what follows, to keep focus on the central issues, we assume the LD matrix is estimated in a sample drawn from the same population as the GWAS population, and that the LD-matrix-estimation sample grows at the same rate as the GWAS sample (as would be true, for example, if the LD matrix were estimated in the GWAS sample itself). In Section IV.C below,

[14] The discussion and analysis in this subsection apply also to PGI weights chosen such that the PGI approximates some other quantity. For example, the genetic PCs estimated from a sample (discussed in Section III.D) are PGIs that aim to approximate the population's (true) genetic PCs.

[15] A simpler approach, developed earlier and still widely used (especially in medical applications), is called "pruning and thresholding." In this approach, the PGI is constructed from a set of approximately mutually uncorrelated ("pruned") SNPs whose p-value is below some threshold. For highly polygenic phenotypes—including social and behavioral phenotypes—pruning-and-thresholding makes less sense than approaches that use all the measured SNPs because all the measured SNPs could add information to the PGI. In addition to Bayesian approaches and pruning-and-thresholding, which are the most commonly used, machine-learning approaches also exist (e.g., Widen et al., 2021; Zhao et al., 2021).

we discuss the implications of using PGI weights based on a population that is different from the population of the prediction sample.

As noisy estimates of the optimal predictor weights, the PGI weights can be expressed as $\hat{\beta} = \breve{\beta} + \mathbf{u}$ for some sample-error vector $\mathbf{u}$. The PGI can therefore be interpreted as a standardized, noisy measure of the optimal predictor $\breve{g}$:

$$\hat{g} = \frac{\tilde{\mathbf{x}}\hat{\beta}}{\text{std}\left(\tilde{\mathbf{x}}\hat{\beta}\right)} = \frac{\tilde{\mathbf{x}}\left(\breve{\beta} + \mathbf{u}\right)}{\text{std}\left(\tilde{\mathbf{x}}\hat{\beta}\right)} = \frac{\tilde{\mathbf{x}}\breve{\beta} + \tilde{\mathbf{x}}\mathbf{u}}{\text{std}\left(\tilde{\mathbf{x}}\hat{\beta}\right)} = \frac{\breve{g} + e}{\text{std}\left(\breve{g} + e\right)}, \tag{12}$$

where $e \equiv \tilde{\mathbf{x}}\mathbf{u}$ is noise that comes from the sampling error $\mathbf{u}$. If $\hat{\beta}$ were estimated by multivariate regression of the phenotype on $\tilde{\mathbf{x}}$ and $\mathbf{z}$, the noise e would be mean zero, uncorrelated with the optimal predictor $\breve{g}$, and independent of all variables in any independent prediction sample. Moreover, it follows from $\text{Cov}\left(\breve{g}, e\right) = 0$ that $\text{Var}\left(\breve{g} + e\right) = \text{Var}\left(\breve{g}\right) + \text{Var}\left(e\right)$. For the standard approaches to constructing a PGI discussed above, these properties do not hold, but they hold approximately if the GWAS sample size (the sample size underlying $\hat{\beta}^{\text{GWAS}}$ and the estimated LD matrix) is large. Becker et al. (2021, see Supplementary Materials 4) derives formulas for these approximations and calculates that the approximations are tight for the PGI derived from a recent GWAS of educational attainment (Lee et al., 2018). In that case, e can be treated as classical measurement error.

This result that the PGI is a standardized, noisy measure of the optimal predictor with classical measurement error is important—much of what follows below will rely on it—and perhaps surprising. One might have had the intuition that the measurement error would be non-classical because the PGI coefficients $\hat{\beta}$ are estimated less precisely for some SNPs (rarer SNPs, which have less genotypic variation) than others. If the SNPs' *genotypes* were measured with different amounts of error, the measurement error would indeed be non-classical, but different amounts of measurement error in the PGI weights do not cause the measurement error in the PGI to be non-classical.

B Predictive Power of a PGI

In this subsection, we derive an analytic formula for the predictive power of a PGI. In some applications, including clinical use of PGIs to assess disease risk (e.g., Khera et al., 2018), the predictive power of a PGI is central to its usefulness. In social-science research applications, the predictive power of a PGI is a central factor in the statistical power of the analysis.

We focus on a univariate regression of a phenotype y on the PGI $\hat{g}$ for that phenotype, and briefly discuss afterward how covariates complicate the analysis. Our measure of predictive power is the coefficient of determination (R^2) from a population regression of y on $\hat{g}$ in a prediction sample that is independent from the GWAS sample used to estimate the PGI weights. We begin by considering the case where the GWAS and prediction samples are randomly sampled from the

same population. Our derivation follows Daetwyler, Villanueva and Woolliams (2008) and generalizations in subsequent work (de Vlaming et al., 2017*a*; Okbay et al., 2022; Wang et al., 2023). In a univariate regression, R^2 is equal to the squared correlation coefficient:

$$R^2 = \frac{[\text{Cov}\,(y, \hat{g})]^2}{\text{Var}\,(y)\,\text{Var}\,(\hat{g})} = \breve{h}^2 \left(\frac{\breve{h}^2}{\breve{h}^2 + \text{Var}\,(e)\,/\text{Var}\,(y)} \right) = \breve{h}^2 \left(\frac{\breve{h}^2}{\breve{h}^2 + M/N} \right), \tag{13}$$

where M is a constant and N is the GWAS sample size underlying the PGI weights.[16] The second equality is derived in Appendix II, and the third equality follows because $\text{Var}\,(e)$ converges to zero with the GWAS sample size at rate $1/N$.

The two terms in Equation (13) have interpretations that will persist through the various cases we consider below. The first term, which is the optimal predictive power $\breve{h}^2$, is the R^2 from a hypothetical regression of the phenotype on the (unobserved) optimal predictor. Since it is the predictive power that would be achieved if the PGI weights were estimated from an infinite GWAS sample, we call it the *asymptotic-R^2 term*. The second term in Equation (13)—which we call the *estimation-precision term*—is between 0 and 1 and is related to the signal-noise ratio in the GWAS: the optimal predictive power $\breve{h}^2$ is a measure of the signal, and $\text{Var}\,(e)\,/\text{Var}\,(y)$ is a measure of the noise. The estimation-precision term is increasing in the GWAS sample size and asymptotes to 1.

The constant M depends on the LD matrix, which will vary depending on the population's genetic ancestry. Under our assumption that the LD matrix is full rank, M is equal to the number of SNPs in the PGI. Otherwise, M is smaller than that number. Using population genetic theory, some simplifying assumptions, and estimates of related quantities, M has been roughly estimated to be 60,000 to 70,000 in European-genetic-ancestry populations (Hayes, Visscher and Goddard, 2009; Rietveld et al., 2013; Wray et al., 2013). Alternatively, M can be estimated by fitting Equation (13) based on the N's of previous GWAS and the R^2's of the resulting PGIs (as in Okbay et al., 2022). After either calibrating $M \approx 70,000$ or estimating M from previous GWASs, Equation (13) can be used to forecast the predictive power of a PGI from a future GWAS with a larger sample size.

We now generalize Equation (13) to the case where the to-be-predicted outcome, denoted y_{pred}, may be different from the PGI phenotype, denoted y_{GWAS}, and the prediction population may be different from the GWAS population. In a first step, we continue to assume the two populations have a common LD matrix; we relax this assumption in Section IV.C below. We need to distinguish between the optimal predicitve power for y_{GWAS} in the GWAS population, denoted $\breve{h}^2_{GWAS}$, and the optimal predictive power for the prediction outcome in the prediction

[16] In its original formulation, Daetwyler, Villanueva and Woolliams (2008) assumes that the PGI weights are unbiased estimates of the additive SNP factor such that $\breve{h}^2 = \tilde{h}^2$.

population, $\breve{h}^2_{pred}$. In Appendix II, we show that the predictive power is now:

$$R^2 = \left(\breve{h}^2_{pred} r^2_{\mathbf{x}\beta}\right) \left(\frac{\breve{h}^2_{GWAS}}{\breve{h}^2_{GWAS} + M/N}\right), \tag{14}$$

where $r_{\mathbf{x}\beta}$ is the correlation between the optimal predictors across the two populations, a type of "genetic correlation" as defined in Section II.J.

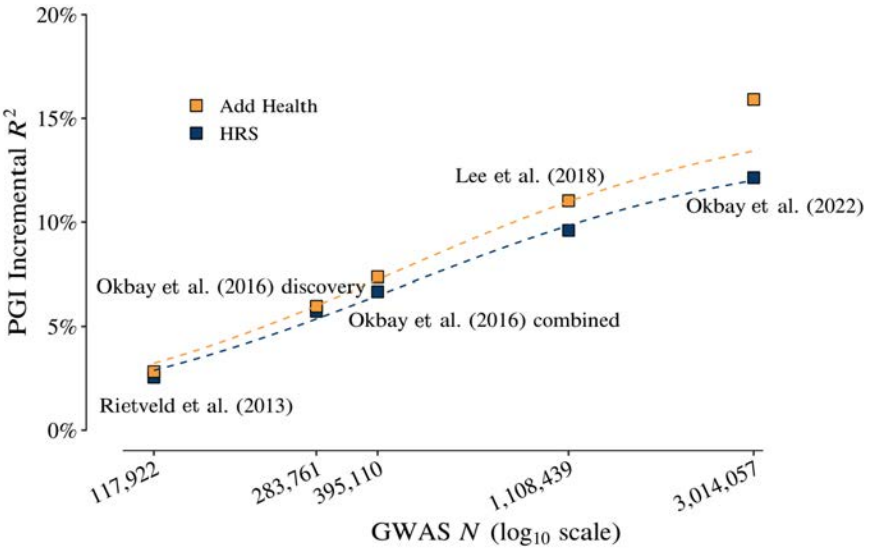
The estimation-precision term is the same as in Equation (13), with parameters that depend on the GWAS phenotype, population, and sample size. The asymptotic-R^2 term, however, is now the optimal predictive power for y_{pred} in the prediction population, $\breve{h}^2_{pred}$, multiplied by $r^2_{\mathbf{x}\beta}$. The optimal predictive power $\breve{h}^2_{pred}$ could be larger or smaller than $\breve{h}^2_{GWAS}$, depending on, among other things, what y_{pred} and y_{GWAS} are, how they are measured, and what the GWAS and prediction populations are. The attenuation factor $r^2_{\mathbf{x}\beta}$ is bounded above by 1—a bound achieved when, for example, the phenotype and populations in the prediction and GWAS samples are identical—and will be smaller than 1 to the extent that the optimal predictors for y_{pred} and y_{GWAS} differ.

Empirically, when researchers examine the predictive power of a PGI, they most commonly report the *incremental* R^2: the change in R^2 from adding the PGI to a regression of the phenotype on a baseline set of covariates. These baseline covariates are typically the same as those included in a GWAS: age, year of birth, and genetic principal components. To illustrate, Figure 2 shows how, for each of two prediction datasets, the incremental R^2 of the PGI for educational attainment has increased as GWAS discovery samples have increased from ∼100K individuals to ∼ 3M individuals. The two prediction datasets are the Health and Retirement Study (HRS), a U.S. nationally representative sample of older Americans, and the National Longitudinal Adolescent to Adult Health Study (Add Health), a U.S. nationally representative sample of younger Americans. In both datasets, the predictive power of the PGI is increasing in the GWAS sample size. At each sample size, the predictive power appears to be larger in Add Health.

Equation (13) is derived for a univariate regression, rather than for the incremental R^2 between two multivariate regressions, but it can nonetheless provide some useful insights. For example, the equation implies that the difference in predictive power across the datasets is due to a difference in $\breve{h}^2_{pred} r^2_{\mathbf{x}\beta}$; Okbay et al. (2022) indeed estimate a larger optimal predictive power $\breve{h}^2_{pred}$ in Add Health, albeit with large standard errors. The dashed line fits Equation (13) to the points in the figure separately for each dataset. The functional form implied by Equation (13) provides a good fit, except for the predictive power in Add Health from the most recent GWAS, which is larger than expected.

As another empirical illustration of Equation (13), Mostafavi et al. (2020) study PGIs for diastolic blood pressure, BMI, and educational attainment and document how their predictive power varies with the sex, age, and socioeconomic status of

Figure 2. Predictive Power of PGI for Educational Attainment as a Function of Sample Size



Note: The x-axis is the sample size of the GWAS on a log scale. The y-axis is the incremental R^2 of the EA PGI constructed from the GWAS summary statistics, in each of two prediction samples independent of the samples used in the original GWAS meta-analysis. Incremental R^2 is the increase in R^2 after adding the PGI to a regression of years of schooling on the following controls: a full set of dummy variables for year of birth, an indicator variable for sex, a full set of interactions between sex and year of birth and the first ten genetic PCs.

the prediction sample. Consistent with Equation (13), for each PGI separately, Mostafavi et al. (2023) find that the incremental R^2 is larger when the GWAS sample is demographically more similar to the prediction sample (implying higher $r^2_{\mathbf{x}\beta}$) and when the optimal predictive power is larger in the prediction sample.

Although an incremental R^2 relative to a baseline set of covariates can be a useful measure, it may be a misleading measure of the gain from including the PGI in social-science research for two reasons. First, such analyses often include a richer set of covariates that absorb more of the variation explained by the PGI. To illustrate this point, Lee et al. (2018, see their Supplementary Figure 12(b)) report how the incremental R^2 of the PGI for educational attainment declines as additional covariates are included in the regressions. With just the baseline covariates of age, sex, and genetic PCs, the incremental R^2 is roughly 11%, but it falls to roughly 5% when additionally controlling for marital status, income, mother's education, and father's education. Second, the incremental R^2 may not be the relevant measure of predictive power, depending on the purpose of the analysis. For example, we show in Section V.A that, for the purpose of increasing the precision of a treatment effect estimate, the efficiency gains from controlling for PGIs can be substantial, even when—indeed, *especially* when—a set of covariates explain much of the variation. In that context, the relevant measure of predictive power is the R^2 from a regression of the residual (of the outcome after controlling for the covariates) on the PGI. This R^2 is larger when the covariates explain more variation.

C Limited Portability of PGIs across Populations

For applications, one major limitation of PGIs is that, at present, their predictive power is typically much lower in populations of non-European ancestry individuals. In the literature, this issue is often called the problem of limited "portability."

To shed some light on some of its causes, we generalize Equation (14). A key difference when the GWAS and prediction samples consist of people with different genetic ancestries is that the LD matrix in the GWAS sample, $\mathbf{\Sigma}_{GWAS} \equiv \text{Var}\left(\tilde{\mathbf{x}}_{GWAS}\right)$, is not equal to the LD matrix in the prediction sample, $\mathbf{\Sigma}_{pred} \equiv \text{Var}\left(\tilde{\mathbf{x}}_{pred}\right)$. For expositional ease, define the LD-difference matrix: $\mathbf{\Delta}_{\mathbf{\Sigma}} \equiv \mathbf{\Sigma}_{pred} - \mathbf{\Sigma}_{GWAS}$, where the negative sign denotes standard elementwise subtraction. In Appendix II, we show

$$R^2 = \left(\breve{h}^2_{pred} r^2_g\right) \times \left\{ \frac{\breve{h}^2_{GWAS} + \frac{\breve{\beta}'_{GWAS} \Delta_{\Sigma} \breve{\beta}_{GWAS}}{\text{Var}(y_{GWAS})}}{\breve{h}^2_{GWAS} + \frac{\breve{\beta}'_{GWAS} \Delta_{\Sigma} \breve{\beta}_{GWAS}}{\text{Var}(y_{GWAS})} + \frac{\text{sum}\left(\Sigma_{pred} \circ \Sigma^{-1}_{GWAS}\right)}{N}} \right\}, \tag{15}$$

where $\breve{\beta}_{GWAS}$ is the vector of optimal predictor weights for y_{GWAS} in the GWAS sample (*not* the vector of GWAS summary statistics $\hat{\beta}^{GWAS}$), r^2_g is the correlation

between the optimal predictor in the prediction sample ($\tilde{\mathbf{x}}_{pred}\breve{\beta}_{pred}$) and a PGI in the prediction sample that uses the GWAS-sample optimal predictor weights ($\tilde{\mathbf{x}}_{pred}\breve{\beta}_{GWAS}$), $\circ$ denotes the elementwise multiplication matrix operator, and $\text{sum}(\cdot)$ is the "grandsum" matrix operator that returns the scalar sum of the matrix elements. As far as we are aware, Equation (15) has not been derived previously, though related formulas have been derived (Wientjes et al., 2016, 2015; Ding et al., 2023; Wang et al., 2020). For example, Ding et al. (2023) and Wang et al. (2020) derive approximations based on a model that treats the set of SNPs with non-zero effects as known, with random effect sizes drawn from a parametric distribution.

In addition to generalizing the squared genetic correlation parameter r_g^2, Equation (15) has two changes relative to Equation (14). The first is that a new term, $\breve{\beta}'_{GWAS}\Delta_\Sigma\breve{\beta}_{GWAS}$, appears in both numerator and denominator. We label it the *weighted LD-difference term*, because it is a weighted sum of the LD-difference matrix, Δ_Σ. Intuitively, it captures the difference between the samples in the frequencies of alleles that have larger coefficients in predicting the phenotype. Formally, a diagonal element of Δ_Σ contributes positively to this term if the SNP has greater genotypic variance—that is, minor allele frequencies closer to 50%—in the prediction sample than in the GWAS sample. The weight on a diagonal term is the SNP's squared optimal predictor weight in the GWAS sample. An off-diagonal element of Δ_Σ contributes positively to this term if the pair of SNPs has greater genotypic covariance—i.e., stronger LD—in the prediction sample than in the GWAS sample. The weight on an off-diagonal term is the product of the SNPs' optimal predictor weights. As discussed in Section I.E, the major reason different populations have different LD matrices is genetic drift. Under a model of genetic drift in which the prediction and GWAS populations diverged from a common ancestral population, the weighted LD-difference term will be zero in expectation. However, for any particular pair of GWAS and prediction samples, it could be positive or negative.[17] Depending on its sign and its magnitude relative to other terms, the weighted LD-difference term could either increase or decrease R^2.

Second, the estimation error (one piece of the estimation-precision term) becomes $\frac{1}{N}\text{sum}\left(\mathbf{\Sigma}_{pred}\circ\mathbf{\Sigma}_{GWAS}^{-1}\right)$. This term still vanishes at rate N, but for a given GWAS sample size N, it will generally be larger than when the GWAS and prediction populations coincide. Intuitively, the SNPs that have the largest genotypic variance in the GWAS sample—which are the SNPs that contribute most to prediction accuracy in the GWAS sample—are those whose optimal predictor weights are estimated most precisely, but those SNPs may not be the ones

[17]Three other forces also cause LD matrices to diverge between populations: natural selection, assortative mating, and mutation. Natural selection is discussed in Section I.E. Both directional selection and stabilizing selection increase the magnitude of correlation between alleles that have the strongest effect on the phenotype. As we discuss elsewhere, assortative mating increases genetic variance (see Section III.B) and the magnitude of correlation for alleles associated with the sorting phenotype (see Section I.C). Differences in LD matrices due to mutation are small over the time scale of modern human populations.

that have the largest genotypic variance in the prediction sample. Formally, if the GWAS and prediction populations have the same LD matrix, the estimation error equals $\frac{1}{N}\text{sum}\,(I) = \frac{M}{N}$ as in Equation (14) above. If the two populations diverged from a common ancestral population, under a model of genetic drift, we can think of $\mathbf{\Sigma}_{pred}$ and $\mathbf{\Sigma}_{GWAS}$ as random variables that are independent conditional on the LD matrix of the ancestral population $\mathbf{\Sigma}_{anc}$. In the special case where $\mathbf{\Sigma}_{pred}$ and $\mathbf{\Sigma}_{GWAS}$ are diagonal, it is straightforward to see that the estimation error is expected to be larger than $\frac{M}{N}$:

$$\begin{aligned}
\mathbb{E}_{\sigma^2_{pred},\sigma^2_{GWAS}}\left(\frac{1}{N}\sum_{j=1}^{M}\frac{\sigma^2_{pred,j}}{\sigma^2_{GWAS,j}}\right) &= \frac{M}{N}\mathbb{E}_{\sigma^2_{pred},\sigma^2_{GWAS}}\left(\frac{\sigma^2_{pred,j}}{\sigma^2_{GWAS,j}}\right) \\
&= \frac{M}{N}\mathbb{E}_{\mathbf{\Sigma}_{anc}}\left(\mathbb{E}_{\sigma^2_{pred},\sigma^2_{GWAS}}\left(\frac{\sigma^2_{pred,j}}{\sigma^2_{GWAS,j}}|\mathbf{\Sigma}_{anc}\right)\right) \\
&\geq \frac{M}{N}\mathbb{E}_{\mathbf{\Sigma}_{anc}}\left(\frac{\mathbb{E}_{\sigma^2_{pred}}\left(\sigma^2_{pred,j}|\mathbf{\Sigma}_{anc}\right)}{\mathbb{E}_{\sigma^2_{GWAS}}\left(\sigma^2_{GWAS,j}|\mathbf{\Sigma}_{anc}\right)}\right) = \frac{M}{N},
\end{aligned}$$

where the inequality follows from Jensen's inequality.

In practice, the GWASs underlying PGI weights are typically conducted in samples of individuals of European genetic ancestries (see Section III.H). On average across phenotypes, and for almost all phenotypes that have been studied, PGIs have less predictive power in samples of non-European-genetic-ancestry individuals. For example, Martin et al. (2019) find that, on average across 17 anthropometric and blood phenotypes, relative to the PGI R^2 in European-genetic-ancestry samples, the R^2 is $\sim 33\%$ smaller in native American and South Asian genetic-ancestry samples, $\sim 50\%$ smaller in East Asian genetic-ancestry samples, and $\sim 75\%$ smaller in African genetic-ancestry samples (Similar results are reported in Duncan et al., 2019 and Martin et al., 2017). The decline in the PGI R^2 from European-genetic-ancestry to African-genetic-ancestry populations is even more pronounced for educational attainment, roughly 85% (Lee et al., 2018; Okbay et al., 2022).

Consistent with the estimation-precision term in Equation (15), the average decline in predictive power tracks qualitatively with genetic distance from European genetic ancestry (and indeed, even among individuals with European genetic ancestry, average predictive power of a PGI is lower for individuals more distantly related to the GWAS sample; Ding et al., 2023). To estimate quantitatively the extent to which differences in LD matrices explain the decline in predictive power for various phenotypes, Wang et al. (2020) used an approximation to Equation (15), together with estimates of LD matrices from different populations and GWAS results for eight anthropometric and health-relevant phenotypes. They find that 70%-80% of the drop in PGI R^2 from European-genetic-ancestry

to African-genetic-ancestry populations can be accounted for by the LD-matrix differences. Following the same estimation strategy for educational attainment, Okbay et al. (2022) estimate a larger drop in PGI R^2 than can be accounted for by the LD-matrix differences.

Equation (15) implies that the remaining differences in PGI R^2 across populations are due to some combination of the weighted LD-difference term, $\breve{h}^2_{pred}$ differing from $\breve{h}^2_{GWAS}$, and $r_g^2 < 1$. The relative contributions of these factors is unknown. However, we expect $r_g^2 < 1$ because gene-environment interactions are probably common, (exogenous) environmental factors are often correlated with genetic ancestry, and non-additive genetic effects may play a role across genetic ancestries (particularly epistasis involving variants whose allele frequencies differ across genetic ancestries).

In the long term, constructing more predictive PGIs in non-European-genetic-ancestry populations will become possible as more genotyped samples from those populations become available. In those larger samples, GWASs can be conducted, and population-specific PGI weights can be obtained. In the shorter term, new statistical methods can partially substitute for larger GWAS samples (Turley et al., 2021*a*; Ruan et al., 2022; Miao et al., 2022). These methods leverage results from large-scale GWASs in European-genetic-ancestry populations to create synthetic GWAS results for other populations, using the populations' LD matrices to "translate" GWAS associations across populations.

The discussion above has focused on the problem of using PGIs trained in one population to predict phenotypic variation *within* a population with different genetic ancestry. Additional challenges arise when comparing the *level* of a PGI across individuals from different populations. Even when the two populations are genetically similar, such comparisons can be confounded by different mean levels of the phenotype (for non-genetic reasons), different true genetic effects across the populations (most notably due to gene-environment interactions), different patterns of gene-environment correlation, and different prediction-error variances. When the populations are from different genetic ancestries, the LD matrices differ—implying that the SNPs included in the PGIs will capture causal effects (including those of unmeasured genetic variants) to different degrees—and the non-genetic differences may be greater, exacerbating these challenges. Indeed, comparisons of PGI levels across populations with different genetic ancestry are unlikely to be valid in most cases (unless the ancestries are sufficiently similar). Martin et al. (2017) (their Figure 4A) provided a striking empirical example: they compared the distributions of height PGIs for several different populations with different genetic ancestries using data from the 1000 Genomes Project. They found that the African populations sampled are genetically predicted to be considerably shorter than all the European populations sampled, which contradicts empirical observations on measured height.

D Estimating and Interpreting the "Causal Effect of a PGI"

In this subsection, we discuss what is meant by the "causal effect of a PGI" and how it can be estimated. As far as we are aware, the material in this section is new.

We put "causal effect of a PGI" in quotes because we do not have in mind a hypothetical experiment in which we examine the effect of changing the PGI. Formulating such a hypothetical experiment is challenging, for two reasons.[18] First, the PGI weights typically do not represent causal effects of the SNPs included in the PGI. Even if the PGI weights were obtained from a family-based GWAS, the PGI weight on a SNP partly reflects the causal effects of unmeasured genetic variants that are correlated with measured SNPs. (For this same reason, in Section II.I, we could not define the additive SNP factor in the potential outcomes framework.) Second, the PGI is an index. Thus, even if the PGI weights were the causal effects of the included SNPs, which SNPs' genotypes were changed when considering a hypothetical experiment of changing the PGI by some amount could matter. Thus, two individuals whose PGIs changed by the same amount might have different hypothetical experiments that define the effect.[19] Instead, what we can hope for—and, in fact, what we can achieve because the PGI is an additive index—is to define the "causal effect of a PGI" as a weighted sum of causal effects of genetic variants (that are either themselves included in the PGI or in LD with SNPs that are included in the PGI due to linkage). An advantage of this definition (in contrast to some other possible definitions) is that it makes the "causal effect of a PGI" estimable. In this subsection, we derive and interpret this weighted sum of causal effects and show that it is estimated by a regression that controls for parental PGIs.

To begin, recall from Section III.B that an individual's genotype vector, $\mathbf{x}$, can be decomposed into the mean parental genotype vector, $\mathbf{x}_p \equiv \frac{\mathbf{x}_f+\mathbf{x}_m}{2}$, and a random deviation, $\mathbf{x}_r \equiv \mathbf{x} - \frac{\mathbf{x}_f+\mathbf{x}_m}{2}$. There, we considered the population regression of some outcome variable y on $\mathbf{x}$ and $\mathbf{x}_p$,

$$y = \mathbf{x}\beta + \mathbf{x}_p\mathbf{b}_p + \xi, \tag{16}$$

[18] Perhaps the most compelling hypothetical experiment would be to imagine that prospective parents create many embryos, one of which is chosen at random and results in a live birth. Each embryo has a different genotype vector randomly assigned conditional on the parents) and therefore a different PGI. The association between the PGIs and the potential outcomes has a causal interpretation—but it is an average treatment effect conditional on the parents, and it is not clear how it relates to a quantity that could be estimated. Moreover, it does not solve the two conceptual challenges to defining the "causal effect of a PGI" described in this paragraph.

[19] Which genotypes were changed would not matter if two conditions are both satisfied: (i) the additive model is true, and (ii) the analysis focuses on the phenotype corresponding to the PGI (for example, an analysis of educational attainment using the PGI for educational attainment). While (i) may often be a reasonable approximation, most social-science applications of PGIs violate (ii). For example, consider a study of the effect of the PGI for educational attainment on income (as in Papageorge and Thom, 2020). To see how it may matter which genotypes were changed, suppose changing either of two SNPs would increase the PGI by one unit. If one of the SNPs affects income and the other does not, the two hypothetical experiments have different effects.

and discussed why the coefficient on the child's genotype vector, β, is the vector of causal genetic effects and why the coefficient on the parent's genotype vector, b_p, generally does not have a causal interpretation. Note that for what follows, whether y is the phenotype corresponding to the PGI or some other outcome makes no difference.

Using the same decomposition of the genotype vector, any PGI with weight vector $\mathbf{w}$ can be expressed as

$$g_{\mathbf{w}} \equiv \frac{\mathbf{x}\mathbf{w}}{\text{std}(\mathbf{x}\mathbf{w})} = \frac{\mathbf{x}_r\mathbf{w}}{\text{std}(\mathbf{x}\mathbf{w})} + \frac{\mathbf{x}_p\mathbf{w}}{\text{std}(\mathbf{x}\mathbf{w})}$$

where (unlike in Section IV.A above) we now express the PGI as a weighted sum of the genotypes of *all* genetic variants, with $w_j = 0$ for every unmeasured genetic variant j. Now consider a population regression of y on the individual's PGI $g_{\mathbf{w}}$ and the sum of parental PGIs, $p_{\mathbf{w}} \equiv \frac{\mathbf{x}_f\mathbf{w}}{\text{std}(\mathbf{x}\mathbf{w})} + \frac{\mathbf{x}_m\mathbf{w}}{\text{std}(\mathbf{x}\mathbf{w})}$:

$$y = \alpha_g g_{\mathbf{w}} + \alpha_p p_{\mathbf{w}} + u \tag{17}$$

(where we define $p_{\mathbf{w}}$ as a sum rather than an average because it makes the expressions for α_p and α_g symmetric). In Appendix III, we derive the relationship between the coefficients from the regressions in equations (16) and (17). Although our analysis there is more general and the resulting formulas correspondingly more complex, here we present the results under the assumption of a randomly mating population:

$$\alpha_g = \frac{\mathbf{w}'\mathbf{\Sigma}\beta}{\mathbf{w}'\mathbf{\Sigma}\mathbf{w}} \tag{18}$$

$$\alpha_p = \frac{\mathbf{w}'\mathbf{\Sigma}\mathbf{b}_p}{\mathbf{w}'\mathbf{\Sigma}\mathbf{w}}, \tag{19}$$

where $\mathbf{\Sigma} \equiv \text{Var}(\mathbf{x}) = \text{Var}(\mathbf{x}_f) = \text{Var}(\mathbf{x}_m)$ is the LD matrix, which is the same in the parents and children due to the random-mating assumption. Equations (18) and (19) state that the coefficients in Regression (17) are the coefficients from a generalized least squares regression of the respective coefficients from Regression (16)—either the genetic causal effects (the β_j's) or the parental coefficients (the b_{pj}'s)—onto their PGI weights (the w_j's).

Equation (18) has a key implication: the coefficient α_g from a regression of an outcome y on a child's PGI, controlling for the sum of parental PGIs, represents causal effects of genetic variants. Specifically, α_g is a weighted sum of the elements of β, which are causal effects. The relative weight on the causal effect of genetic variant j depends on its own PGI weight w_j and the PGI weights of the other (measured or unmeasured) SNPs with which it is correlated. Rare SNPs (which are only weakly correlated with measured SNPs) and unmeasured SNPs (whose causal effects are only captured via correlation with measured SNPs) will tend

to be weighted less heavily than common and measured SNPs. In Appendix III, we show that while the equation itself is more complicated, the key implication of Equation (18)—that α_g is a function only of the causal effects β and not of $\mathbf{b}_p$—holds more generally, including in situations with assortative mating.

In contrast to α_g, Equation (19) implies that α_p does *not* generally have a causal interpretation, since $\mathbf{b}_p$ does not. (As mentioned in Section III.B, having a causal interpretation for $\mathbf{b}_p$ would require controlling for the *grandparental* PGIs in regression Equation (16). In that case, the coefficient on the sum of parental PGIs would be a weighted sum of causal parental genetic effects (see Section II.F).) Furthermore, in the general case with assortative mating, we show that α_p is a function of both β and $\mathbf{b}_p$. Because it can partly (or even wholly) reflect β, interpreting α_p as the "non-genetic" or "environmental" effect is not accurate.

Just as β is identified if regression Equation (16) includes the father's and mother's genotypes separately rather than the mean parental genotype, α_g is identified if regression Equation (18) includes the father's and mother's PGIs separately rather than the sum of parental PGIs. Including the father's and mother's PGIs separately enables comparisons of the magnitudes of the father's and mother's coefficients. More generally, α_g remains identified when controls are added to regression Equation (18), as long as the controls are not themselves caused by genotypes, because the random component of the child's PGI is independent of such controls. For including the father's and mother's PGIs separately and more generally for including controls, there are opposing effects on precision: adding covariates adds degrees of freedom but can absorb more of the residual variation.

Currently, rather than controlling for parental PGIs, including siblings in the analysis and controlling for sibling fixed effects is more common. As in the case of estimating the effects of genotypes discussed in Section III.C, it is not widely appreciated that the regression with sibling fixed effects generates a biased estimator of the self genetic effect in the presence of sibling genetic effects and is inefficient relative controlling for the sum of parental PGIs. In the case of PGIs, the identifying variation with sibling fixed effects is the individual's PGI relative to the sibling mean, written here for the case of sibling pairs: $g - \left(\frac{g+g_{sib}}{2}\right) = \frac{1}{2}(g - g_{sib}) = \frac{1}{2}\frac{(\mathbf{x}-\mathbf{x}_{sib})\mathbf{w}}{\text{std}(\mathbf{xw})}$, where the subscript "sib" denotes an individual's sibling. Variation in $\frac{1}{2}(g - g_{sib})$ is random, but when y is regressed on $g_{\mathbf{w}}$ controlling for sibling fixed effects $\mathbf{z}$, the coefficient that is estimated is

$$\alpha_g = \frac{\mathbf{w}'\boldsymbol{\Sigma}(\beta - \beta_{sib})}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}},$$

where β_{sib} denotes the sibling genetic effect. Since the identifying variation is the individual's PGI relative to her sibling's, the coefficient is picking up both the effect of the individual having a higher PGI and the effect of the sibling having a lower PGI.

Analogous to the discussion in Section III.C, when sibling genotypes are ob-

served but parental genotypes are not, Young et al. (2022) shows that controlling for sibling fixed effects is dominated by imputing parental genotypes and controlling for the sum of parental PGIs (constructed from the imputed data), with random effects to control for family-specific means. This strategy generates a consistent and unbiased estimator of Equation (18) with greater precision than the sibling fixed-effects specification. With imperfect controls $\mathbf{z}$, such as principal components of the genetic data, the parameter being estimated is Equation (18), but with omitted-variables bias due to uncorrected-for gene-environment correlation, as well as bias due to assortative mating.

E Correcting for PGI Measurement Error in Applications

Typically, in a social-science application analyzing some PGI, it is not the PGI's effect that researchers are actually interested in. Instead, the PGI is used as an empirical proxy for the genetic influences on a phenotype. As we showed in Section IV.A, the PGI can be interpreted as a measure of the optimal predictor (and of the additive SNP factor when the GWAS has sufficient controls), but with measurement error that is (approximately) classical. The errors-in-variables bias that results from classical measurement error can distort the conclusions drawn from a statistical analysis in a number of ways (e.g., Gillen, Snowberg and Yariv, 2019). Moreover, different papers use PGIs based on different GWASs or constructed using different methods, so the amount of measurement error varies. All of these considerations bolster the case for facilitating comparability across studies by correcting for errors-in-variables bias in applications.

Two approaches have been developed to implement such a correction. First, DiPrete, Burik and Koellinger (2018) proposed an instrumental-variables approach. Two independent subsamples of the GWAS are used to construct two sets of PGI weights. These are then used to construct two PGIs in the prediction sample, and they are used to instrument for each other (as in Gillen, Snowberg and Yariv, 2019).

The other approach is a regression-disattenuation estimator, which uses external information on the amount of measurement error (Becker et al. (2021)). The core idea is as follows. Suppose a researcher is interested in the regression of some outcome y on some standardized, optimal predictor $\breve{g}/\text{std}\left(\breve{g}\right)$ (which could be the standardized, optimal predictor for a different phenotype than y). Although the optimal predictor is unobserved, the researcher has access to a PGI, $\hat{g}$, and runs the regression of y on $\hat{g}$. Recall from Section IV.A above that $\hat{g} = \frac{\breve{g}+e}{\text{std}(\breve{g}+e)}$, where e is (approximately) classical measurement error. A standard calculation shows that the regression coefficient of interest is

$$b \equiv \frac{\text{Cov}\left(\breve{g}/\text{std}\left(\breve{g}\right), y\right)}{\text{Var}\left(\breve{g}/\text{std}\left(\breve{g}\right)\right)} = \frac{\text{Cov}\left(\breve{g}, y\right)}{\text{std}\left(\breve{g}\right)},$$

but the estimated coefficient is

$$\begin{aligned}\frac{\text{Cov}\left(\hat{g},y\right)}{\text{Var}\left(\hat{g}\right)} &= \text{Cov}\left(\hat{g},y\right)\\ &= \frac{1}{\text{std}\left(\breve{g}+e\right)}\text{Cov}\left(\breve{g},y\right) = \frac{\text{std}\left(\breve{g}\right)}{\text{std}\left(\breve{g}+e\right)}b = \frac{1}{\rho}b,\end{aligned}$$

where $\rho \equiv \frac{\text{std}(\breve{g}+e)}{\text{std}(\breve{g})}$.[20] The expression is the well-known formula for attenuation bias for univariate regression, modified to reflect the fact that the PGI is standardized in our setting. An estimate of ρ is then used to calculate b.

The approach proposed by Becker et al. (2021) exploits the relationship

$$\begin{aligned}\rho^2 = \frac{\text{Var}\left(\breve{g}+e\right)}{\text{Var}\left(\breve{g}\right)} &= \frac{\text{Cov}\left(\breve{g},y\right)^2 / \left[\text{Var}\left(\breve{g}\right)\text{Var}\left(y\right)\right]}{\text{Cov}\left(\breve{g},y\right)^2 / \left[\text{Var}\left(\breve{g}+e\right)\text{Var}\left(y\right)\right]},\\ &= \frac{\text{Cov}\left(\breve{g},y\right)^2 / \left[\text{Var}\left(\breve{g}\right)\text{Var}\left(y\right)\right]}{\text{Cov}\left(\hat{g},y\right)^2 / \left[\text{Var}\left(\hat{g}\right)\text{Var}\left(y\right)\right]} = \frac{\breve{h}^2}{R^2},\end{aligned}$$

where $\rho^2 > 1$ because the PGI's actual predictive power, R^2, is smaller than the optimal predictive power, $\breve{h}^2$, when the PGI weights are estimated in a finite GWAS sample. The amount of measurement error determines how much the PGI's predictive power falls short of the optimal predictive power. R^2 can be estimated directly in the prediction sample at hand. The optimal predictive power $\breve{h}^2$ can also be estimated in the prediction sample if the sample size is large enough for a precise estimate; more commonly, an external estimate can be used.

Becker et al. extend this approach to obtain an analytic correction for the errors-in-variables bias in a multivariate regression of an outcome on a PGI, a set of non-genetic covariates, and possibly their interactions. They show the correction can be implemented using an estimated or assumed value of optimal predictive power in the prediction sample, together with quantities that are consistently estimated in the prediction sample. Becker et al. also derive analytic standard errors for the regression coefficients. Despite the fact that the standard errors ignore uncertainty in the optimal predictive power estimates, Sanz-de Galdeano and Terskaya (Forthcoming) (their Appendix F.5) show that standard errors will often be conservative (i.e., biased upward).

Relative to the regression-disattenuation estimator, the major advantage of the instrumental-variables estimator is that it does not require an estimate of the optimal predictive power. This advantage may be a particularly relevant for phenotypes with substantial assortative mating, which biases estimates of optimal predictive power. The major disadvantage of the instrumental-variables estimator is a loss of statistical power from having to split the GWAS sample.

[20] Note that our notation here, and in particular our definition of ρ, differs from that in Becker et al. (2021) because Becker et al. analyze the additive SNP factor in standardized units.

van Kippersluis et al. (2023) provide a detailed analysis and comparison of the two approaches.

Corrections for errors-in-variables bias require some additional assumptions when the regressors include multiple PGIs. Sanz-de Galdeano and Terskaya (Forthcoming) extend the Becker et al. approach to the important case of controlling for the parental PGIs and/or sibling PGI, to allow for estimating the causal effect of the optimal predictor. When controlling for the parental PGIs, an additional assumption is needed about the parent-child correlation of the optimal predictor. When controlling for the sibling PGI, an additional assumption is needed about the sibling correlation of the optimal predictor. Under random mating, these parameters are known and equal to $1/\sqrt{2}$ and $1/2$, respectively (Trejo and Domingue, 2018), but under assortative mating, the parameters depend on the degree of assortative mating.

We have focused here on analyses using PGIs to draw conclusions about what the results would have been if the optimal predictor were analyzed in place of the PGI. Rightly or wrongly, researchers typically interpret their results as if the PGI were an unbiased estimate of the additive SNP factor. Granting this premise (which is fully justified only if the GWAS controls were sufficient), a natural question is whether we could, by correcting for additional measurement error, draw conclusions about what the results would have been if the additive genetic factor (or even the genetic factor) had been analyzed. If so, this would generally be more interesting, since the relevant theory-driven hypotheses pertain to the effects of genetic variants overall, not merely those that can currently be measured. Moreover, it is feasible in many cases to correct for additional measurement error; the regression-disattenuation estimator above could be used with a twin-based heritability estimate in place of a SNP heritability estimate. The reason we have not focused on such an adjustment is that it is not fully justified: if we conceptualize the PGI as a standardized, noisy measure of the additive genetic factor, the measurement error is unlikely to be classical for two reasons. First, biases in population GWAS estimates are often correlated with the causal effect effects. Second, even if unbiased estimates were used, we do not have information about the genetic variants that are unmeasured and not highly correlated with measured SNPs. For example, if unmeasured rare variants—which have low correlations with measured SNPs—tend to affect the phenotype negatively and happen to be negatively correlated with the PGI, then the measurement error in the PGI would be correlated with the PGI. For the measurement-error correction to give the right answer, we would need to assume that the difference between the additive genetic factor and the additive SNP factor is uncorrelated with the PGI. While we believe it may be interesting to apply the correction also in such scenarios, it would be important to be transparent about the fact that an additional assumption is being imposed.

F PGIs As Social-Science Variables

Since PGIs are, at best, noisy measures of the additive SNP factor, and the additive SNP factor proxies for the additive genetic factor, all of the interpretational caveats from Section II.D apply to PGIs. In particular, the effects of a PGI will typically reflect a mix of all the mechanisms through which genetic variants (that are correlated with measured SNPs) operate. For many phenotypes, these mechanisms will include endogenous social and behavioral responses to phenotypes proximally affected by the PGI. Just as heritabilities are not measures of innateness, it is a mistake to assume that PGIs exclusively capture purely biological or innate characteristics. We caution against using labels such as "genetic endowment" to describe PGIs for the same reason that such labels inaccurately describe the genetic factor.

Some researchers, especially non-economists, have asserted that it is misleading to describing the effects of a PGI as "causal" because the mechanisms are largely or entirely unknown. Economists are well situated to provide a useful perspective, since economists often study the causal effects of environmental factors (and interventions) for which we have only a partial understanding of mechanisms. Of course, for many purposes, it is important to understand the mechanisms underlying a causal effect. That is why, after credibly identifying a causal effect, many economics papers go on to study potential mechanisms (often with evidence that is less air-tight than the identification of the causal effect).

As a variable that operates through many mechanisms, a PGI is like many other variables that social scientists study and incorporate into their theories. For example, an individual's biological sex has biological effects, such as body size and hormone levels, but it also affects an individual's behavior and outcomes through the reactions that other people have to the individual. While researchers need to bear these different possible mechanisms in mind when studying biological sex, biological sex is nonetheless an important and useful variable in social-science research. We believe PGIs can be important and useful in a similar way.

V Applications

In this section, we describe several applications of genetic data in economics. We have tried to cover a broad range, but our judgment of which applications to review is necessarily somewhat subjective and leaves out a number of valuable contributions, including, among others, Papageorge and Thom (2020), Barth, Papageorge and Thom (2020), and Arold, Hufe and Stoeckli (2022).

A Polygenic Indexes for Balance Tests and as Covariates

PGIs can be valuable even for research that is not directly related to genetics. For example, PGIs can be particularly useful variables for balance tests in RCTs and quasi-experiments, for three reasons. First, because the PGI is fixed at

conception, it cannot be affected by the treatment. Second, the cost of genotyping participants may be small relative to the cost of collecting some alternative measures used in balance tests (e.g., scores from long cognitive tests). Finally, once participants have been genotyped, it is possible to construct PGIs for multiple phenotypes and genetic principal components, which can all be used for (in some cases, uncorrelated) balance tests. The only example we are aware of in the literature is Barcellos, Carvalho and Turley (2018) (their Appendix B), who used PGIs for educational attainment and BMI, as well as 15 genetic principal components, as variables for a balance test for a regression discontinuity design (discussed below).

For the same reasons, PGIs may also be valuable control variables. For example, a PGI could be used to (partly) control for omitted variable bias in observational studies where the treatment of interest is correlated with genetic factors, e.g., studies of the association between parental behaviors and children's outcomes (e.g., Jami et al., 2021). Alternatively, even in randomized controlled trials (RCTs), which yield unbiased treatment effects without any control variables by virtue of randomizing the treatment, PGIs can be useful as controls that absorb residual variance, thereby making the treatment effect estimates more precise (Rietveld et al., 2013; Benjamin et al., 2012; Cesarini and Visscher, 2017).[21]

Rietveld et al. (2013) calculated the gains in effective sample size that could be obtained by controlling for PGIs in a simple RCT with two conditions. For example, if the set of baseline controls, absent the PGI, explain 20% of the variance in the outcome, they find that adding a PGI with an incremental R^2 of 15% would increase power equivalent to increasing the RCT sample size by 19%.

To date, only a handful of studies have used PGIs as control variables (e.g., Barcellos, Carvalho and Turley, 2018; Davies et al., 2018), perhaps due to lack of human capital for incorporating genetic-data collection into existing RCT research infrastructures. Some investigators may also be apprehensive about collecting sensitive data that is unrelated to the goals of the relevant RCTs. However, the cost-benefit profile of controlling for PGIs will only grow as genotyping becomes cheaper and PGIs become more predictive and available for more phenotypes.

B Heterogeneous Treatment Effects of Education on Health

Genetics provides natural dimensions of heterogeneity for studying heterogeneous treatment effects. One example is Barcellos, Carvalho and Turley (2018), who study heterogeneous treatment effects of education on health. For quasi-experimental variation in education, they use a regression-discontinuity design, exploiting the 1972 Raising Of School-Leaving Act, which raised the compulsory schooling age from 15 to 16. Their outcome variables are binarized measures of

[21] Controlling for PGIs can similarly increase the power of GWASs; see Bennett et al. (2021), Campos et al. (2023) and Jurgens et al. (2023). Relatedly, PGIs can be used for stratified sampling in an RCT, selecting extreme individuals to increase power for a given sample size (Fahed, Philippakis and Khera, 2022).

body size/obesity, lung health, and blood pressure. They use data from 253,377 individuals in the UK Biobank.

In a different dataset from the UK, Clark and Royer (2013) exploit the same law and find no evidence of effects of education on health. With their much larger dataset enabling more precise estimates, Barcellos et al. find that education reduces the risk of obesity (-6.0 percentage points, SE: 3.5, baseline risk: 21.5%) and lung disease (-8.9 percentage points, SE: 4.8, baseline risk: 28.7%) but increases the risk of hypertension (10.6 percentage points, SE: 5.1, baseline risk: 61.1%).

They then examine how the treatment effects vary with a PGI for BMI and find evidence for substantial heterogeneity in the effect of education on obesity. For example, an additional year of schooling reduced the risk of obesity by only 0.3 percentage points for those with a BMI PGI one standard deviation below the mean but reduced it by 11.7 percentage points for those one standard deviation above. This heterogeneity implies that the compulsory year of education reduced the health gap between high- and low-risk individuals, as measured by the BMI PGI.

C Assortative Mating

Assortative mating is of particular interest to economists due to its implications for societal inequalities. Because parents pass on their genes to their offspring in addition to their other forms of capital, genetics is potentially an important mechanism by which assortative mating affects inequalities. Even aside from this, genetic data provide promising new empirical tools to study assortative mating (kinship data can be similarly useful; see, e.g., Collado, Ortuno-Ortin and Stuhler, 2023).

Relative to using directly measured phenotypes, one key advantage of genetic data is that genotypes are fixed at conception and not influenced by the mate selection process. Thus, a positive correlation of a PGI for BMI between spouses must be due to factors that were in place prior to the match. By contrast, a positive phenotypic correlation of, say, spousal BMI could arise for a number of reasons, including assortative mating but also correlated spousal environments before or after matching.

One example of work that exploits this basic insight is Conley et al. (2016). Using data from the Health and Retirement study, they measure the phenotypic spousal correlation for education, height, BMI, and depression. They also measure the spousal correlation for PGIs for these four traits. They find moderate phenotypic correlations for all four phenotypes (as high as 0.53 for education and as low as 0.17 for height), but only the PGI correlations for education (0.13, CI: [.09,.17]) and height (0.30, CI: [0.27,0.34]) are statistically distinguishable from zero. They acknowledge, however, that the lack of correlations for the BMI and depression PGIs may be due to measurement error in the PGI, as described in section IV.E.

As another example, Abdellaoui et al. (2022*b*) consider a model where a person's socioeconomic status (SES) and their advantageous genes are assets in the marriage market, causing people to sort on both genes and SES. Such a model induces a correlation between genes and SES in subsequent generations because both are transmitted to offspring. To test this model, using the UK Biobank, they measure whether later-born children—who have lower SES on average than their older siblings—tend to marry people with lower average educational-attainment PGIs. They find weak evidence that later birth (and therefore lower expected SES) is associated with a person's spouse having a smaller educational-attainment PGI (-0.031, SE: 0.015 in their strongest specification), though this result is not robust across all specifications. They conclude that this model may partly explain long-run inequality.

A second advantage of genetic data is that it can be used to make inferences about assortment schemes in previous generations of individuals who were not genotyped. Such inferences are possible because if parents (or earlier ancestors) sort on a heritable phenotype, then alleles that are associated with increases in the phenotype will be correlated across the father and mother and thus correlated *within* their child's genome, even if the alleles are on different chromosomes. Yengo et al. (2018) made this observation and then exploited it to construct an estimator for the amount of assortative mating in some phenotype: they infer it from the correlation between a PGI that is constructed only from SNPs on even-numbered chromosomes and another constructed only from SNPs on odd-numbered chromosomes. Consistent with the results of Conley et al. (2016), which are based on observed spouse pairs, they find positive cross-chromosome correlation for height and educational attainment.

While we anticipate that approaches that exploit genetic data will prove useful for studying assortative mating even when researchers are primarily interested in sorting on phenotypes, it is important to keep in mind that the relationship between mates' genotypic correlation and mates' phenotypic correlation is nuanced. For example, a one-time, permanent increase in the amount of phenotypic assortment on height would generate a gradual increase in the correlation between mates' PGIs for height over several generations, asymptoting toward a higher level.

D Parental Investment and Sibling Differences

Sanz-de Galdeano and Terskaya (Forthcoming) use a PGI for EA to address the question of whether parental investments compensate for or reinforce childrens' ability differences. Theoretical analyses by Becker and Tomes (1976) and Behrman, Pollak and Taubman (1982) highlight two influences on parental investment decisions: parental concerns for efficiency versus equality, and the cost of investment. A body of prior work relied on measures of ability, such as cognitive test scores, that are themselves influenced by parental investments. Using a PGI instead has two key advantages: it is fixed at conception, prior to any

(even in-utero) parental investments, and it is randomly assigned, conditional on parental PGIs.

Sanz-de-Galdeano and Terskaya use data from 604 genotyped sibling pairs with European genetic ancestries in the National Longitudinal Adolescent to Adult Health Study (Add Health). Using the method developed by Young et al. (2022) (see Section III.B), they impute the mean parental genotype within each sibling pair and use it as a control variable in the analysis. Based on a survey conducted when the children were aged 12-20 that asked about how often the child engaged in various activities with the parents, Sanz-de-Galdeano and Terskaya (Forthcoming) construct an index of parental investment for each child. The paper's main regression specification is

$$I_0 = \beta_0 + \beta_1 \left(\text{PGI}_0 - \text{PGI}_y\right) + \beta_2 \text{PGI}_0 + \beta_3 \text{PGI}_{par} + \text{Controls} + u, \tag{20}$$

where the unit of analysis is the sibling pair, I_0 is the index of parental investment in the older sibling, PGI_0 and PGI_y are the older and younger sibling's respective PGIs, PGI_{par} is the (imputed) parental mean PGI, and u is an error term. Because PGI_0 and $(\text{PGI}_0 - \text{PGI}_y)$ are conditionally random given PGI_{par}, the coefficients β_1 and β_2 have causal interpretations. Moreover, in a structural model, Sanz-de-Galdeano and Terskaya (Forthcoming) show that β_1 captures a parental preference parameter for inequality aversion versus efficiency, and β_2 captures the cost of investment (the net effect of the child's PGI on parental investment is $\beta_1 + \beta_2$). As discussed in Section IV.E, estimating regression Equation (20) with the observed PGIs would generate coefficients that suffer from substantial errors-in-variables bias. Instead, Sanz-de-Galdeano and Terskaya develop and apply an extension of Becker et al.'s (2021) measurement-error correction, which simultaneously corrects for the measurement error in all three PGI terms.

In their full-sample analysis including all their control variables and after correcting for the measurement errors in the PGIs, Sanz-de-Galdeano and Terskaya estimate $\hat{\beta}_1 = -0.207$ (S.E. $= 0.102$), $\hat{\beta}_2 = 0.167$ (S.E. $= 0.140$), and $\hat{\beta}_3 = -0.029$ (S.E. $= 0.161$). The main result (albeit statistically weak) is the negative estimate of the parameter β_1, which suggests parents are inequality-averse over their childrens' human capital. The point estimate of β_2 is positive, which would imply parents that invest more when their children have higher ability, but the 95% confidence interval is large and includes zero. The estimate of β_3 is difficult to interpret because it picks up the effects of non-genetic variables that are correlated with PGI_{par}.

Sanz-de-Galdeano and Terskaya's estimates likely understate both the economic and statistical significance of childrens' genes for parental investment behavior. Statistically, as noted in Section IV.E, Sanz-de-Galdeano and Terskaya show that the measurement-error correction generates standard errors that are biased upward. Economically, relative to the theoretical concept of "initial ability" that the PGI proxies for, the PGIs contain additional measurement error that is not accounted for. For example, the theoretical concept corresponds to the "direct-

effect PGI" that could be produced from a family-based GWAS, rather than the PGI they use, which comes from a population GWAS (see Section III.G). The theoretical concept also would include the effects of all genetic variants, not just those captured by SNP array data (see Section II.I). Although these sources of measurement error are neither classical nor mean-zero, their primary effect on the main results is likely to be an attenuation bias.

E The Dynamics of Parental Investment and Child Cognitive Skills

Houmark, Ronda and Rosholm (2024) use a PGI for EA to model and estimate the joint evolution of cognitive skills and parental investments throughout early childhood. The paper builds on existing models of childhood skill formation (Cunha and Heckman, 2007, 2008). Like Sanz-de Galdeano and Terskaya (Forthcoming) discussed above, using a PGI as a measure of initial ability has the advantages that it is fixed prior to any parental investments, and it is randomly assigned, conditional on parental PGIs.

Houmark, Ronda and Rosholm (2024) use data from 4,510 genotyped children and their parents (with European genetic ancestries) in the Avon Longitudinal Study of Parents and Children, a birth cohort study based in Bristol, UK. Genetic data from both parents are available for 1,267 children. For other children, the authors use Young et al.'s (2022) method (see Section III.B) to impute the missing parent's genotype. Based on questionnaires sent regularly to the child's primary caregiver starting prior to birth, the authors construct measures of children's skills (e.g., ability to process new information and learn abstract concepts) and parental investments (the frequency with which the parent does certain activities with the child).

They model a child's initial cognitive skills, θ_{i0}, as a log-linear function of the child's, the mother's, and the father's additive SNP factors for educational attainment, G_i, G_i^m, and G_i^f, as well as a vector of controls, $\mathbf{X}_{i0}$, that include sex and 15 genetic PCs:

$$\ln(\theta_{i0}) = \alpha_1 G_i + \alpha_2 G_i^m + \alpha_1 G_i^f + \alpha_{x0}\mathbf{X}_{i0} + \varepsilon_{i0}. \tag{21}$$

In their preferred specification, the production function for cognitive skills in period t follows a translog specification:

$$\begin{aligned}\ln(\theta_{it}) =& \ln(A_t) + \delta_{1t}\ln(\theta_{it}) + \delta_{2t}\ln(I_{it}) \\ &+\delta_{3t}\ln(\theta_{it}) \times \ln(I_{it}) + \delta_{4t}G_i + \delta_{5t}G_i^m + \delta_{6t}G_i^f + \delta_{xt}\mathbf{X}_{it} + \varepsilon_{it},\end{aligned} \tag{22}$$

where $\ln(A_t)$ is total factor productivity and I_{it} is parental investment in period t. Parental investment behavior in period t is specified as a function of the child's cognitive skills in period t, the child's and parents' additive SNP factors, and

controls:

$$\ln(I_{it}) = \gamma_{1t}\ln(\theta_{it}) + \gamma_{2t}G_i + \gamma_{3t}G_i^m + \gamma_{4t}G_i^f + \gamma_{xt}\mathbf{X}_{i0} + \eta_{it}. \tag{23}$$

These structural equations are supplemented by a set of measurement equations that link latent skills, θ_{it}, and latent parental investments, I_{it}, to the observed measures of skills and investments in each period (as in Agostinelli and Wiswall, 2016; Cunha and Heckman, 2008). Identifying the latent factors requires standard but strong i.i.d. assumptions on the measurement errors of the observed measures. Houmark et al. similarly write down a measurement equation linking the (latent) additive SNP factor to the (measured) EA PGI, and in this way, they adjust for the measurement error in the EA PGI. Once the latent factors are estimated, Equations (21)-(23) are estimated by OLS, and the standard errors are obtained by bootstrapping.

The main results are about the effects of children's genotypes. These estimated effects have a causal interpretation due to the conditional random assignment of G_i, given G_i^m and G_i^f. The paper finds that genetic influences affect cognitive skills even for very young children, ages 0-2, and that the genetic influence on a child's cognitive skills is increasing with age. Previous work also reports increasing genetic influences with age (e.g., Bouchard, 2013; Belsky et al., 2016), but an alternative interpretation of the earlier findings—ruled out here—was that cognitive skills at younger ages are measured with more error. The paper also finds that children with higher additive SNP factors behave in ways that cause their parents to invest more in them. The parental investment responses magnify initial differences between children.

The other estimated effects should be interpreted more cautiously because they rely more heavily on the assumptions of the structural and measurement models, but they paint a rich picture of the dynamics of parental investment and childrens' cognitive skill accumulation. For example, the paper finds that parents with higher additive SNP factors invest more in their children (holding fixed the child's additive SNP factor) and that the returns to parental investments are substantially overestimated if genetic measures are omitted from the analysis.

F Mendelian Randomization

The first proposed use of genetic data in economics was as instrumental variables (Norton and Han, 2008; Ding et al., 2009; Fletcher and Lehrer, 2009; von Hinke Kessler Scholder et al., 2011). These early studies in economics used as instruments "candidate genes," whose reported associations with behavioral phenotypes subsequently came to be viewed with skepticism (see, e.g., Beauchamp et al., 2011*a*; Benjamin et al., 2012). In epidemiology, the strategy of using genetic variants as instruments is called Mendelian randomization (MR), which is the term we will use here.

The initial idea of MR appears to be due to Katan (1986), who proposed studying the causal impact of serum cholesterol on cancer risk by examining the association between the gene *APOE* (which affects serum cholesterol) and cancer. It was subsequently recognized that MR is a case of instrumental-variables regression (e.g., Thomas and Conti, 2004). There was already a substantial amount of MR research carried out in epidemiology by 2000 (Davey Smith and Ebrahim, 2003). MR has continued to grow in popularity in genetic epidemiology and medical genetics and now comprises a enormous literature. For a recent overview, see Sanderson et al. (2022).

There are many challenges to credibly using genetic variants as instruments, mainly because it is difficult to rule out violations of the exclusion restriction (e.g., Conley, 2009; Cawley, Han and Norton, 2011; McMartin and Conley, 2020). For one thing, genetic variants typically matter in multiple biological pathways, many of which are not fully understood, so the exclusion restriction will rarely be satisfied exactly. For another thing, genetic variants are typically in LD with many other variants (including distant variants, due to population structure and assortative mating) that could affect the outcome through other pathways. Gene-environment correlation is another reason the exclusion restriction may be violated. While controlling for parental genotypes would solve the problems due to gene-environment correlation, population structure, and assortative mating (Brumpton et al., 2020), it is rare in practice due to data limitations (exceptions are Fletcher and Lehrer, 2011; Howe et al., 2022*a*). In addition to violations of the exclusion restriction, weak instruments is another potential problem, particularly when the endogenous regressor is a complex phenotype. Because of these challenges, there have been few MR studies in economics.

While we view most published MR studies skeptically, there are exceptions. For example, one arguably persuasive MR study in epidemiology is Millwood et al. (2019). Prior observational studies of the effect of alcohol on health outcomes had found a non-monotonic relationship, with light drinking (especially of red wine) associated with better health than abstinence but heavy drinking associated with worse health. However, the association between abstinence and worse health could be due to reverse causation (unhealthy individuals drink less) or confounds. To obtain causal evidence, Millwood et al. use as instruments two SNPs that are well established to be associated with alcohol consumption due to their role in alcohol metabolism. The paper analyzes ∼160,000 Han Chinese individuals in the China Kadoorie Biobank. The more powerful of the two instruments (rs671 in the alcohol dehydrogenase gene *ALDH2*) commonly varies among Han Chinese, with one of its alleles causing people who drink alcohol to experience discomfort. The paper estimates a monotonic effect of alcohol consumption on cardiovascular health outcomes, with any amount of alcohol consumption worse than no alcohol consumption. To address concerns about violation of the exclusion restriction, Millwood et al. conduct a placebo test using the women in the sample, exploiting the fact that women in most regions of China largely abstain from alcohol. If

the effects of the genetic variants on health were due to correlation with other variants or gene-environment correlation, then we would expect an association between the genetic variants and health even in this subsample of non-drinkers. However, the paper finds no association between the genetic variants and health among women, only among men.

Economists tend to be very skeptical about instrumental variables except when violations of the exclusion restriction can be fully ruled out (see, e.g., Oster's (2022) critique of Millwood et al. and another, related study). In response, defenders of MR studies often argue that when randomized experiments are infeasible, as is common in epidemiology, MR studies are the best tool available and provide more persuasive evidence than observational studies.

For making the standard instrumental-variables assumptions most plausible, the best-case scenario is to use as an instrument a genetic variant with a relatively large effect size that operates through a known biological mechanism. Genetic variants involved in alcohol and nicotine metabolism are among the most promising for social-science applications and could be used to study effects of alcohol and cigarette consumption, respectively. However, even in these cases, the variants have other effects, besides affecting alcohol and nicotine metabolism. For example, the SNP rs1051730 tags the *CHRNA3* gene that is often used in MR studies of the effects of smoking (e.g., Skov-Ettrup et al., 2017). This gene codes for neuronal acetylcoline receptors, which nicotine binds to, but acetylcoline is a neurotransmitter that has a wide variety of functions. Placebo tests like in the above example will therefore generally be necessary for MR studies to be most persuasive.

In the epidemiology literature, the growing recent popularity of MR studies is due to the explosive growth of genetic associations identified by GWAS (Sekula et al., 2016). These studies look different from instrumental-variables studies in economics for two reasons. First, instead of the usual, single-sample instrumental-variables estimator, they use a two-sample instrumental-variables estimator: the effect of the instrument on the outcome is divided by the effect of the instrument on the endogenous regressor (Angrist and Krueger, 1992). Estimates of these effects are obtained from the summary statistics of two different published GWASs. For example, in an MR study of the effect of education on coronary heart disease (Tillmann et al., 2017), for each SNP used as an instrument, the coefficient from a GWAS of coronary heart disease was divided by the the coefficient from a GWAS of educational attainment. Second, these studies typically use a large number of instruments, for example, all of the genome-wide-significant SNPs in the GWAS of the outcome. Unlike in economics, where multiple instruments are used to test overidentifying restrictions or to identify different local average treatment effects, MR studies typically use multiple instruments in order to obtain causal estimates that are valid under assumptions that—while still strong—are weaker than the exclusion restriction. A variety of estimators have been developed for this purpose (e.g., Timpson et al., 2011; O'Connor and Price, 2018; Burgess et al., 2019;

Burgess et al., 2020; Bowden, Davey Smith and Burgess, 2015; Kang et al., 2016). Perhaps because of our jaded perspective as economists on instrumental variables, we are concerned by the proliferation of MR studies that draw causal conclusions whose validity hinges on assumptions that are rarely tested adequately.

VI Future Directions and Concluding Remarks

Over the last ten years, with the advent of GWAS for social and behavioral phenotypes, social-science genomics has come of age. PGIs are beginning to be used in social-science applications. In some cases, PGIs will be useful as control variables to increase statistical power (e.g., in randomized experiments) or to address confounds, for example, when studying the health-education gradient. In other cases, equipped with the PGI as a measure of genetic influences, economists and other social scientists will have greater leverage in addressing classic topics, such as the determinants and impacts of parental and school investments (e.g., Sanz-de Galdeano and Terskaya, Forthcoming; Houmark, Ronda and Rosholm, 2024), labor market returns to human capital (e.g., Papageorge and Thom, 2020), intergenerational transmission of skills (e.g., Barth, Papageorge and Thom, 2020), the determinants and consequences of migration (e.g., Abdellaoui et al., 2019), and assortative matching in marriage markets (e.g., Abdellaoui et al., 2022*b*). Progress on some of these topics is already underway, as illustrated in Section V.

While some of these applications will involve economists relatively straightforwardly importing PGIs from genetics into economics, in other applications, economists will need to build structural models to account for endogenous behavioral and social responses to genotypes. In such applications, economists will contribute to geneticists' understanding of the mechanisms through which PGIs matter. Sections V.D and V.E showcased some early examples.

To facilitate certain applications, we anticipate that social scientists will influence the genetics research that is conducted. This has already happened in the case of GWASs for social and behavioral phenotypes, which have been collaborations between social scientists and geneticists, driven (at least initially) by social scientists' interests in the phenotypes. Once genotyped samples become large enough for adequate power, we anticipate that social scientists will want to conduct GWASs in samples that contain randomized experiments or quasi-experiments (see also Schmitz et al., 2021). For example, in a sample where the curriculum is randomly assigned, economists may be interested in a GWAS of educational attainment in which the regressors include SNP-by-treatment interaction. A PGI can then be constructed from the coefficients on the interaction. This PGI would capture genetic influences on the effectiveness of the treatment. When based on a sufficiently well powered GWAS, such a PGI would be a better tool for targeting the curricular intervention than the current PGI for educational attainment.

We anticipate that four ongoing, related developments in statistical genetics will facilitate applications of genetic data in the social sciences. First is simply

more and larger GWAS samples. Larger samples will continue to enable better powered studies of all kinds, such as studies of gene-environment interactions, as well as more predictive PGIs for a larger set of phenotypes. In addition, at some point, larger samples will enable researchers to construct PGIs that capture some of the non-additive genetic variance, allowing the predictive power of a PGI to exceed the phenotype's SNP heritability. While we expect dominance variance to be negligible for most social and behavior phenotypes and epistatic variance to be small (see Section II.C), epistatic variance may well add a few percentage points of predictive power. Although the combinatorial explosion of potential gene-gene interactions is a challenge for efforts to credibly identify them, machine-learning methods should be able to capture some of their predictive power.

Second, the momentum in the field strongly favors family-based studies and methods. As larger genotyped family samples become available, researchers will be able to more precisely estimate (causal) self genetic effects, parental genetic effects, sibling genetic effects, and genetic effects of other family members. In regressions that control for parental PGIs, family-based PGIs will also overtake population PGIs in predictive power, enabling applications to be better powered and making the weights on genetic variants implicit in the "causal effect of the PGI" closer to the additive SNP factor weights.

Third, generating large samples with individual-level genetic data from all major underrepresented genetic ancestries is a high priority. This will likely have many benefits for medical genetics, including improving leverage for identifying causal genetic variants. For the social sciences, the main benefit will be PGIs that are more predictive for individuals with non-European genetic ancestries.

Fourth, genotyping will become denser, implying that the measured SNPs will capture a greater fraction of all of the genetic variation. Indeed, as the cost of sequencing continues to plummet, sequencing may soon overtake genotyping as researchers' preferred way of measuring genetic variation. As mentioned in Section I, sequencing can measure rare SNPs and non-SNP types of genetic variation much more accurately than genotyping and thus enables the discovery of very rare variants with large effects, e.g., on intellectual disabilities (e.g., Chen et al., 2023), that evade detection in association studies limited to common (mostly SNP) variants. Thus, due to sequencing, we will eventually have a much better catalog of the genes that, when disrupted, have a large impact on phenotypes relevant to the social sciences. We will need to rely less on imputation of unmeasured genetic variants and will therefore be able to better "fine map" causal variants, i.e., identify which of several mutually correlated genetic variants are causally responsible for their association with a phenotype. We expect that the main benefit of denser genotype measurement for social science will be the improved predictive power of PGIs. In the limit where a PGI includes all genetic variants, the PGI can be interpreted as a noisy measure of the additive genetic factor.

More so than in economics, progress in genetics has been propelled by technological advances in measurement that show no sign of slowing down. The

nearly 15 years since the last time a review of "genoeconomics" has been written (Beauchamp et al., 2011*a*; Benjamin et al., 2012) is an eternity in terms of the pace of genetics research. Although the coming 15 years may not produce transformational changes on par with those of the past decade and a half—a re-shaping of social-science genomics with GWAS and the resulting PGIs—there will surely be both progress and challenges that we cannot currently imagine.

We conclude by highlighting perhaps the most important ongoing challenge: conducting, interpreting, and communicating research at the intersection of genetics and social science responsibly. While these obligations apply to all researchers, researchers in social-science genomics bear additional responsibilities in light of how difficult it is to correctly interpret genetic associations—as highlighted by the extensive discussion of interpretation throughout this review—as well as the enduring legacy of eugenics (Rutherford, 2022). While far from sufficient, terminology can help to some degree. Researchers should be cognizant of the potential social harms of, and be especially careful about conducting and communicating, research that could be (mis)understood as comparing ethnic, racial, or other groups on socially valued phenotypes, such as cognitive performance or income. Given how easy it is to slip into genetic determinism, we believe it is helpful to continually remind readers of research papers that the effects of individual genetic variants are small (e.g., Chabris et al., 2015), can operate through environmental pathways (Jencks, 1980), and have no obvious bearing on the effectiveness of interventions (Goldberger, 1979). We believe it is useful to write a Frequently Asked Questions (FAQs) document along with a paper to explain to journalists and non-experts what the research does and does not find and to carefully address any ethical or policy questions raised by the research. Indeed, writing such FAQs has become standard practice in social-science genomics (Martschenko et al., 2021). We recommend a report published by the Hastings Center, a bioethics think tank (Meyer et al., 2023*a*), for helpful discussion of these and other best practices, as well as ethical issues related to social-science genomics more broadly.

REFERENCES

Abdellaoui, Abdel, Conor V Dolan, Karin JH Verweij, and Michel G Nivard. 2022*a*. "Gene–environment Correlations Across Geographic Regions Affect Genome-wide Association Studies." *Nature Genetics*, 54(9): 1345–1354.

Abdellaoui, Abdel, David Hugh-Jones, Loic Yengo, Kathryn E Kemper, Michel G Nivard, et al. 2019. "Genetic Correlates Of Social Stratification In Great Britain." *Nature Human Behaviour*, 3(12): 1332–1342.

Abdellaoui, Abdel, Loic Yengo, Karin JH Verweij, and Peter M Visscher. 2023. "15 Years of GWAS Discovery: Realizing the Promise." *American Journal of Human Genetics*, 110(2): 179–194.

Abdellaoui, Abdel, Oana Borcan, Pierre-Andre Chiappori, and David Hugh-Jones. 2022*b*. "Trading Social Status for Genetics in Marriage Markets: Evidence from UK Biobank." Human Capital and Economic Opportunity Working Group Working Paper No. 2022-018.

Agostinelli, Francesco, and Matthew Wiswall. 2016. "Estimating The Technology Of Children's Skill Formation." National Bureau of Economic Research Working Paper 22442.

Angrist, Joshua D., and Alan B. Krueger. 1992. "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples." *Journal of the American Statistical Association*, 87(418): 328–336.

Arbatli, Cemal Eren, Quamrul H. Ashraf, Oded Galor, and Marc Klemp. 2020. "Diversity and Conflict." *Econometrica*, 88(2): 727–797.

Arold, Benjamin W, Paul Hufe, and Marc Stoeckli. 2022. "Genetic Endowments, Educational Outcomes and the Mediating Influence of School Investments." *CESifo Working Paper No. 9841*.

Ashraf, Quamrul H, and Oded Galor. 2018. "The Macrogenoeconomics of Comparative Development." *Journal of Economic Literature*, 56(3): 1119–1155.

Barcellos, Silvia H., Leandro S. Carvalho, and Patrick Turley. 2018. "Education Can Reduce Health Differences Related To Genetic Risk Of Obesity." *Proceedings of the National Academy of Sciences of the United States of America*, 115(42): E9765–E9772.

Barth, Daniel, Nicholas W Papageorge, and Kevin Thom. 2020. "Genetic Endowments and Wealth Inequality." *Journal of Political Economy*, 128(4): 1474–1522.

Beauchamp, Jonathan, David Cesarini, Magnus Johannesson, Matthijs J H M van der Loos, Philipp D Koellinger, et al. 2011*a*. "Molecular Genetics And Economics." *Journal of Economic Perspectives*, 25(4): 57–82.

Beauchamp, Jonathan P., David Cesarini, Magnus Johannesson, Erik Lindqvist, and Coren Apicella. 2011*b*. "On The Sources Of The Height-Intelligence Correlation: New Insights From A Bivariate ACE Model With Assortative Mating." *Behavior Genetics*, 41: 242–252.

Becker, Gary S., and Nigel Tomes. 1976. "Child Endowments and the Quantity and Quality of Children." *Journal of Political Economy*, 84(S4): S143–S162.

Becker, Joel, Casper A. P. Burik, Grant Goldman, Nancy Wang, Hariharan Jayashankar, et al. 2021. "Resource Profile and User Guide of the Polygenic Index Repository." *Nature Human Behaviour*, 5(12): 1744–1758.

Behrman, Jere R, Robert A Pollak, and Paul Taubman. 1982. "Parental Preferences And Provision For Progeny." *Journal of Political Economy*, 90(1): 52–73.

Belsky, Daniel W., Terrie E. Moffitt, David L. Corcoran, Benjamin Domingue, HonaLee Harrington, et al. 2016. "The Genetics of Success." *Psychological Science*, 27(7): 957–972.

Benjamin, Daniel J., Christopher F. Chabris, Edward L. Glaeser, Vilmundur Gudnason, Tamara B. Harris, et al. 2007. "Genoeconomics." In *Biosocial Surveys*. , ed. Maxine Weinstein, James W Vaupel and Kenneth W Wachter, 304–335. National Academies Press.

Benjamin, Daniel J., David Cesarini, Christopher F. Chabris, Edward L. Glaeser, David I. Laibson, et al. 2012. "The Promises and Pitfalls of Genoeconomics." *Annual Review of Economics*, 4(1): 627–662.

Bennett, Declan, Donal O'Shea, John Ferguson, Derek Morris, and Cathal Seoighe. 2021. "Controlling For Background Genetic Effects Using Polygenic Scores Improves The Power Of Genome-Wide Association Studies." *Scientific Reports*, 11(19571).

Berg, Jeremy J, Arbel Harpak, Nasa Sinnott-Armstrong, Anja Moltke Joergensen, Hakhamanesh Mostafavi, et al. 2019. "Reduced Signal for Polygenic Adaptation of Height in UK Biobank." *eLife*, 8: e39725.

Bergstrom, Ted C. 2013. "Measures of Assortativity." *Biological Theory*, 8: 133–141.

Bloemendal, Alex. 2019. "A Primer on Random Matrix Theory." *YouTube Video*, URL: https://www.youtube.com/watch?v=B7ub92OLw1g. Accessed: November 14, 2023.

Border, Richard, Georgios Athanasiadis, Alfonso Buil, Andrew J Schork, Na Cai, et al. 2022*a*. "Cross-trait Assortative Mating Is Widespread And Inflates Genetic Correlation Estimates." *Science*, 378(6621): 754–761.

Border, Richard, Sean O'Rourke, Teresa de Candia, Michael E Goddard, Peter M Visscher, et al. 2022*b*. "Assortative Mating Biases Marker-based Heritability Estimators." *Nature Communications*, 13(1): 660.

Bouchard, Thomas J. 2013. "The Wilson Effect: the Increase in Heritability of IQ with Age." *Twin Research and Human Genetics*, 16(5): 923–930.

Bowden, J., G. Davey Smith, and S. Burgess. 2015. "Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression." *International Journal of Epidemiology*, 44(2): 512–525.

Branigan, Amanda R, Kenneth J McCallum, and Jeremy Freese. 2013. "Variation in the Heritability of Educational Attainment: An International Meta-Analysis." *Social Forces*, 9292(1): 109–140109–140.

Braudt, David B. 2018. "Sociogenomics in the 21st Century: An Introduction to the History and Potential of Genetically-Informed Social Science." *Sociology Compass*, 2(10): e12626.

Brumpton, Ben, Eleanor Sanderson, Karl Heilbron, Fernando Pires Hartwig, Sean Harrison, et al. 2020. "Avoiding Dynastic, Assortative Mating, And Population Stratification Biases In Mendelian Randomization Through Within-family Analyses." *Nature Communications*, 11(1): 3519.

Bulik-Sullivan, B., P.-R. Loh, H. Finucane, S. Ripke, J. Yang, et al. 2015*a*. "LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies." *Nature Genetics*, 47(3): 291–295.

Bulik-Sullivan, Brendan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, et al. 2015*b*. "An Atlas of Genetic Correlations across Human Diseases and Traits." *Nature Genetics*, 47(11): 1236–1241.

Burgess, Stephen, Christopher N. Foley, Elias Allara, James R. Staley, and Joanna M. M. Howson. 2020. "A Robust and Efficient Method for Mendelian Randomization with Hundreds of Genetic Variants." *Nature Communications*, 11(376).

Burgess, Stephen, George Davey Smith, Neil M Davies, Frank Dudbridge, Dipender Gill, et al. 2019. "Guidelines for Performing Mendelian Randomization Investigations: Update for Summer 2023." *Wellcome Open Research*, 4(186).

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, et al. 2018. "The UK Biobank resource with deep phenotyping and genomic data." *Nature*, 562(7726): 203–209.

Campos, Adrian I., Shinichi Namba, Shu-Chin Lin, Kisung Nam, Julia Sidorenko, et al. 2023. "Boosting The Power Of Genome-Wide Association Studies Within And Across Ancestries By Using Polygenic Scores." *Nature Genetics*, 55: 1769–1776.

Cawley, John, Euna Han, and Edward C Norton. 2011. "The Validity of Genes Related to Neurotransmitters as Instrumental Variables." *Health Economics*, 20(8): 884–888.

Cesarini, David, and Peter M. Visscher. 2017. "Genetics And Educational Attainment." *npj Science of Learning*, 2(1): 4.

Chabris, C. F., J. J. Lee, D. Cesarini, D. J. Benjamin, and D. I. Laibson. 2015. "The Fourth Law Of Behavior Genetics." *Current Directions in Psychological Science*, 24(4): 304–312.

Chen, Brian H, Riccardo E Marioni, Elena Colicino, Marjolein J Peters, Cavin K Ward-Caviness, et al. 2016. "DNA Methylation-based Measures Of Biological Age: Meta-analysis Predicting Time To Death." *Aging (Albany NY)*, 8(9): 1844–1859.

Chen, Chia-Yen, Ruoyu Tian, Tian Ge, Max Lam, Gabriela Sanchez-Andrade, et al. 2023. "The Impact Of Rare Protein Coding Genetic Variation On Adult Cognitive Function." *Nature Genetics*, 55: 927–938.

Clark, Damon, and Heather Royer. 2013. "The Effect Of Education On Adult Mortality And Health: Evidence From Britain." *American Economic Review*, 103(6): 2087–2120.

Cloninger, C. R., J. Rice, and T. Reich. 1979. "Multifactorial Inheritance With Cultural Transmission And Assortative Mating. II. A General Model Of Combined Polygenic And Cultural Inheritance." *American Journal of Human Genetics*, 31(2): 176–198.

Collado, M. Dolores, Ignacio Ortuno-Ortin, and Jan Stuhler. 2023. "Estimating Intergenerational and Assortative Processes in Extended Family Data." *Review of Economic Studies*, 90: 1195–1227.

Conley, D. 2009. "The Promise and Challenges of Incorporating Genetic Data into Longitudinal Social Science Surveys and Research." *Biodemography and Social Biology*, 55: 238–251.

Conley, Dalton. 2016. "Socio-Genomic Research Using Genome-Wide Molecular Data." *Annual Review of Sociology*, 42(1): 275–299.

Conley, Dalton, Thomas Laidley, Daniel W Belsky, Jason M Fletcher, Jason D Boardman, et al. 2016. "Assortative mating and differential fertility by phenotype and genotype across the 20th century." *Proceedings of the National Academy of Sciences*, 113(24): 6647–6652.

Cunha, Flavio, and James Heckman. 2007. "The Technology Of Skill Formation." *American Economic Review*, 97(2): 31–47.

Cunha, Flavio, and James J Heckman. 2008. "Formulating, Identifying And Estimating The Technology Of Cognitive And Noncognitive Skill Formation." *Journal of Human Resources*, 43(4): 738–782.

Daetwyler, Hans D., Beatriz Villanueva, and John A. Woolliams. 2008. "Accuracy Of Predicting The Genetic Risk Of Disease Using A Genome-wide Approach." *PLoS ONE*, 3(10): e3395.

Davey Smith, George, and Shah Ebrahim. 2003. "Mendelian Randomization: Can Genetic Epidemiology Contribute To Understanding Environmental Determinants Of Disease?" *International Journal of Epidemiology*, 32(1): 1–22.

Davies, Neil M., Matt Dickson, George Davey Smith, Gerard J. Van Den Berg, and Frank Windmeijer. 2018. "The Causal Effects of Education on Health Outcomes in the UK Biobank." *Nature Human Behaviour*, 2(2): 117–125.

Deary, Ian J, Jian Yang, Gail Davies, Sarah E Harris, Albert Tenesa, et al. 2012. "Genetic Contributions To Stability And Change In Intelligence From Childhood To Old Age." *Nature*, 482: 212–215.

de Leeuw, Christiaan A, Joris M Mooij, Tom Heskes, Danielle Posthuma, P M Visscher, et al. 2015. "MAGMA: Generalized Gene-Set Analysis of GWAS Data." *PLoS Computational Biology*, 11(4): e1004219.

de Vlaming, Ronald, Aysu Okbay, Cornelius A. Rietveld, Magnus Johannesson, Patrik K. E. Magnusson, et al. 2017*a*. "Meta-GWAS Accuracy and Power (MetaGAP) Calculator Shows that Hiding Heritability Is Partially Due to Imperfect Genetic Correlations across Studies." *PLOS Genetics*, 13(1): e1006495.

de Vlaming, Ronald, Aysu Okbay, Cornelius A. Rietveld, Magnus Johannesson, Patrik K. E. Magnusson, et al. 2017*b*. "Meta-GWAS Accuracy and Power (MetaGAP) Calculator Shows that Hiding Heritability Is Partially Due to Imperfect Genetic Correlations across Studies." *PLOS Genetics*, 13(1): e1006495.

Dias Pereira, Rita, Pietro Biroli, Titus Galama, Stephanie von Hinke, Hans van Kippersluis, et al. 2022. "Gene-Environment Interplay in the Social Sciences." In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press.

Ding, Weili, Steven F Lehrer, J Niels Rosenquist, and Janet Audrain-McGovern. 2009. "The Impact Of Poor Health On Academic Performance: New Evidence Using Genetic Markers." *Journal of Health Economics*, 28(3): 578–597.

Ding, Yi, Kangcheng Hou, Ziqi Xu, Aditya Pimplaskar, Ella Petter, et al. 2023. "Polygenic Scoring Accuracy Varies Across The Genetic Ancestry Continuum." *Nature*, 618: 774–781.

DiPrete, Thomas A, Casper A P Burik, and Philipp D Koellinger. 2018. "Genetic Instrumental Variable Regression: Explaining Socioeconomic And Health Outcomes In Nonexperimental Data." *Proceedings of the National Academy of Sciences*, 115(22): E4970–E4979.

Duncan, L., H. Shen, B. Gelaye, J. Meijsen, K. Ressler, et al. 2019. "Analysis Of Polygenic Risk Score Usage And Performance In Diverse Human Populations." *Nature Communications*, 10(1): 1–9.

Einav, By Liran, Amy Finkelstein, Iuliana Pascu, and Mark R Cullen. 2016. "How General Are Risk Preferences? Choices under Uncertainty in Different Domains." *American Economic Review,*, 102(6): 2606–2638.

Fahed, Akl C., Anthony A. Philippakis, and Amit V. Khera. 2022. "The Potential Of Polygenic Scores To Improve Cost And Efficiency Of Clinical Trials." *Nature Communications*, 13(1): 2922.

Falconer, Douglas C. 1960. *Introduction to Quantitative Genetics.* Edinburgh and London: Oliver and Boyd.

Fatumo, Segun, Tinashe Chikowore, Ananyo Choudhury, Muhammad Ayub, Alicia R Martin, et al. 2022. "A Roadmap To Increase Diversity In Genomic Studies." *Nature Medicine*, 28: 243–250.

Finucane, Hilary Kiyo, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, et al. 2015. "Partitioning heritability by functional category using GWAS summary statistics." *Nature Genetics*, 47: 1228–1235.

Fisher, Ronald. 1952. "Statistical Methods in Genetics." *Heredity*, 6: 1–12.

Fisher, Ronald A. 1918. "The Correlation between Relatives on the Supposition of Mendelian Inheritance." *Transactions of the Royal Society of Edinburgh*, 52(02): 399–433.

Fletcher, Jason M, and Steven F Lehrer. 2009. "The Effects of Adolescent Health on Educational Outcomes: Causal Evidence Using Genetic Lotteries between Siblings." *Forum for Health Economics & Policy*, 12(2).

Fletcher, Jason M, and Steven F Lehrer. 2011. "Genetic lotteries within families." *Journal of Health economics*, 30(4): 647–659.

Fowler, James H., and Christopher T. Dawes. 2013. "In Defense of Genopolitics." *American Political Science Review*, 107(2): 326–374.

Freese, Jeremy. 2018. "The Arrival of Social Science Genomics." *Contemporary Sociology: A Journal of Reviews*, 47(5): 524–536.

Frey, Renato, Andreas Pedroni, Rui Mata, Jörg Rieskamp, and Ralph Hertwig. 2017. "Risk preference shares the psychometric structure of major psychological traits." *Science Advances*, 3(10): e1701381.

Galor, Oded, and Omer Moav. 2002. "Natural selection and the origin of economic growth." *The Quarterly Journal of Economics*, 117(4): 1133–1191.

Galor, Oded, and Ömer Özak. 2016. "The agricultural origins of time preference." *American Economic Review*, 106(10): 3064–3103.

Galor, Oded, and Stelios Michalopoulos. 2012. "Evolution and the growth process: Natural selection of entrepreneurial traits." *Journal of Economic Theory*, 147(2): 759–780.

Galor, Oded, and Viacheslav Savitskiy. 2018. "Climatic Roots Of Loss Aversion." National Bureau of Economic Research Working Paper No. 25273.

Gao, Xu, Min Jia, Yan Zhang, Lutz Philipp Breitling, and Hermann Brenner. 2015. "DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies." *Clinical Epigenetics*, 7: 113.

Ge, Tian, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller. 2019. "Polygenic Prediction via Bayesian Regression and Continuous Shrinkage Priors." *Nature Communications*, 10(1): 1776.

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy*, 127(4): 1826–1863.

Gillespie, John H. 2004. *Population Genetics: A Concise Guide.* . Second ed., The Johns Hopkins University Press.

Goldberger, Arthur S. 1978. "Models and Methods in the IQ Debate: Part I and II. Revised." Social Systems Research Institute Working Paper No. 7801.

Goldberger, Arthur S. 2005. "Structural Equation Models in Human Behavior Genetics." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg.* , ed. Donald W K Andrews and James H Stock. Cambridge University Press.

Goldberger, Arthur S AS. 1979. "Heritability." *Economica*, 46(184): 327–347.

Hamer, D, and L Sirota. 2000. "Beware the Chopsticks Gene." *Molecular Psychiatry*, 5(1): 11–13.

Hanoch, Yaniv, Joseph G. Johnson, and Andreas Wilke. 2006. "Domain Specificity in Experimental Measures and Participant Recruitment: An Application to Risk-Taking Behavior." *Psychological Science*, 17(4): 300–304.

Haseman, Joseph K, and Robert C Elston. 1972. "The Investigation Of Linkage Between A Quantitative Trait And A Marker Locus." *Behavior Genetics*, 2(1): 3–19.

Hayes, Ben John, P M Visscher, and M E Goddard. 2009. "Increased accuracy of artificial selection by using the realized relationship matrix." *Genetics Research*, 91(01): 47–60.

Hill, William G, Michael E Goddard, and Peter M Visscher. 2008. "Data and theory point to mainly additive genetic variance for complex traits." *PLoS Genetics*, 4(2): e1000008.

Hivert, Valentin, Julia Sidorenko, Florian Rohart, Michael E Goddard, Jian Yang, et al. 2021. "Estimation Of Non-additive Genetic Variance In Human Complex Traits From A Large Sample Of Unrelated Individuals." *American Journal of Human Genetics*, 108(5): 786–798.

Houmark, Mikkel Aagaard, Victor Ronda, and Michael Rosholm. 2024. "The Nurture of Nature and the Nature of Nurture: How Genes and Investments Interact in the Formation of Skills." *American Economic Review*, 114(2): 385–425.

Howe, Laurence J, Ben Brumpton, Humaira Rasheed, Bjørn Olav Åsvold, George Davey Smith, et al. 2022*a*. "Taller height and risk of coronary heart disease and cancer: A within-sibship Mendelian randomization study." *eLife*, 11: e72984.

Howe, Laurence J, Michel G Nivard, Tim T Morris, Ailin F Hansen, Humaira Rasheed, et al. 2022*b*. "Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects." *Nature Genetics*, 54(5): 581–592.

Imbens, Guido W, and Joshua D Angrist. 1996. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

International Human Genome Sequencing Consortium. 2001. "Initial Sequencing And Analysis Of The Human Genome." *Nature*, 409(6822): 860–921.

Jami, Eshim S., Anke R. Hammerschlag, Meike Bartels, and Christel M. Middeldorp. 2021. "Parental Characteristics And Offspring Mental Health And Related Outcomes: A Systematic Review Of Genetically Informative Literature." *Translational Psychiatry*, 11(197).

Jencks, Christopher. 1980. "Heredity, environment, and public policy reconsidered." *American Sociological Review*, 45(5): 723–736.

Jónsson, Hákon, Patrick Sulem, Birte Kehr, Snaedis Kristmundsdottir, Florian Zink, et al. 2017. "Parental Influence On Human Germline De Novo Mutations In 1,548 Trios From Iceland." *Nature*, 549: 519–522.

Jurgens, Sean J, James P Pirruccello, Seung Hoan Choi, Valerie N Morrill, Mark Chaffin, et al. 2023. "Adjusting For Common Variant Polygenic Scores Improves Yield In Rare Variant Association Analyses." *Nature Genetics*, 55(4): 544–548.

Kang, Hyunseung, Anru Zhang, T Tony Cai, and Dylan S Small. 2016. "Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization." *Journal of the American Statistical Association*, 111(513): 132–144.

Karlsson Linnér, Richard, and Philipp D. Koellinger. 2022. "Genetic Rsk Scores in Life Insurance Underwriting." *Journal of Health Economics*, 81: 102556.

Karlsson Linnér, Richard, Pietro Biroli, Edward Kong, S Fleur W Meddens, Robbee Wedow, et al. 2019. "Genome-wide Association Analyses Of Risk Tolerance And Risky Behaviors In Over 1 Million Individuals Identify Hundreds Of Loci And Shared Genetic Influences." *Nature Genetics*, 51(2): 245–257.

Katan, Martjin B. 1986. "Effects Of Cholesterol-Lowering Diets On The Risk For Cancer And Other Non-Cardiovascular Diseases." In *Atherosclerosis VII: Proceedings of the Seventh International Atherosclerosis Symposium.* , ed. Noel H. Fidge and Paul J. Nestel. Elsevier.

Khera, Amit V, Mark Chaffin, Kaitlin H Wade, Sohail Zahid, Joseph Brancale, et al. 2019. "Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood." *Cell*, 177(3): 587–596.e9.

Khera, Amit V., Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, et al. 2018. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations." *Nature Genetics*, 50(9): 1219–1224.

Lambert, Jean-Charles, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, et al. 2013. "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease." *Nature Genetics*, 45(12): 1452–1458.

Lander, Eric S, and Nicholas J Schork. 1994. "Genetic Dissection Of Complex Traits." *Science*, 265: 2037–48.

Lee, James J., Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, et al. 2018. "Gene Discovery And Polygenic Prediction From

A 1.1-million-person Gwas Of Educational Attainment." *Nature Genetics*, 50(8): 1112–1121.

Li, Jeremiah H, Chase A Mazur, Tomaz Berisa, and Joseph K Pickrell. 2021. "Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays." *Genome Research*, 31(4): 529–537.

Lloyd-Jones, Luke R., Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, et al. 2019. "Improved polygenic prediction by Bayesian multiple regression on summary statistics." *Nature Communications*, 10(1): 5086.

Locke, Adam E, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, et al. 2015. "Genetic studies of body mass index yield new insights for obesity biology." *Nature*, 518(7538): 197–206.

Loehlin, John C. 1978. "Heredity-environment analyses of Jencks's IQ correlations." *Behavior Genetics*, 8(5): 415–436.

Loehlin, John C. 2009. "History of Behavior Genetics." In *Handbook of Behavior Genetics*. , ed. Yong-Kyu Kim. Springer New York.

Loh, Po-Ru, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, et al. 2016. "Reference-based phasing using the Haplotype Reference Consortium panel." *Nature Genetics*, 48(11): 1443–1448.

Lovell, Michael C. 1963. "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis." *Journal of the American Statistical Association*, 58(304): 993–1010.

Mäki-Tanila, Asko, and William G. Hill. 2014. "Influence of Gene Interaction on Complex Trait Variation with Multilocus Models." *Genetics*, 1: 355–367.

Manski, Charles F. 1993. "Identification of endogenous social effects: The reflection problem." *The Review of Economic Studies*, 60(3): 531–542.

Marchini, Jonathan, and Bryan Howie. 2010. "Genotype Imputation For Genome-Wide Association Studies." *Nature Reviews Genetics*, 11: 499–511.

Martin, Alicia R., Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, et al. 2017. "Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations." *American Journal of Human Genetics*, 100(4): 635–649.

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, et al. 2019. "Clinical Use Of Current Polygenic Risk Scores May Exacerbate Health Disparities." *Nature Genetics*, 51(4): 584–591.

Martschenko, Daphne Oluwaseun, Benjamin W Domingue, Lucas J Matthews, and Sam Trejo. 2021. "FoGS provides a public FAQ repository for social and behavioral genomic discoveries." *Nature Genetics*, 53(9): 1272–1274.

Martschenko, Daphne, Sam Trejo, and Benjamin W Domingue. 2019. "Genetics and Education: Recent Developments in the Context of an Ugly History and an Uncertain Future." *AERA Open*, 5(1).

McCarthy, Shane, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, et al. 2016. "A Reference Panel Of 64,976 Haplotypes For Genotype Imputation." *Nature Genetics*, 48(10): 1279–1283.

McMartin, Andrew, and Dalton Conley. 2020. "Commentary: Mendelian Randomization and Education–Challenges Remain." *International Journal of Epidemiology*, 49(4): 1193–1206.

Mendel, Gregor. 1866. "Versuche über Pflanzen-Hybriden." *Verhandlungen des naturforschenden Vereines in Brünn*, 4: 3–47.

Menozzi, Paolo, Alberto Piazza, and Luigi Cavalli-Sforza. 1978. "Synthetic Maps of Human Gene Frequencies in Europeans: These maps indicate that early farmers of the Near East spread to all of Europe in the Neolithic." *Science*, 201(4358): 786–792.

Meyer, Michelle N, Paul S Appelbaum, Daniel J Benjamin, Shawneequa L Callier, Nathaniel Comfort, et al. 2023*a*. "Wrestling with Social and Behavioral Genomics: Risks, Potential Benefits, and Ethical Responsibility." *Hastings Center Report*, 53: S2–S49.

Meyer, Michelle N., Tammy Tan, Daniel J. Benjamin, David Laibson, and Patrick Turley. 2023*b*. "Public Views On Polygenic Screening Of Embryos." *Science*, 379(6632): 541–543.

Miao, Jiacheng, Hanmin Guo, Gefei Song, Zijie Zhao, Lin Hou, et al. 2022. "Quantifying Portable Genetic Effects And Improving Cross-ancestry Genetic Prediction With GWAS Summary Statistics." *Nature Communications*, 14(832).

Mieth, Bettina, Marius Kloft, Juan Antonio Rodríguez, Sören Sonnenburg, Robin Vobruba, et al. 2016. "Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies." *Scientific Reports*, 6: 36671.

Mills, Melinda C, and Charles Rahal. 2019. "A Scientometric Review Of Genome-wide Association Studies." *Communications Biology*, 2(1): 9.

Mills, Melinda C., and Charles Rahal. 2020. "The GWAS Diversity Monitor Tracks Diversity By Disease In Real Time." *Nature Genetics*, 52(3): 242–243.

Mills, Melinda C., and Felix C. Tropf. 2020. "Sociology, Genetics, and the Coming of Age of Sociogenomics." *Annual Review of Sociology*, 46(1): 553–581.

Mills, Melinda C., Nicola Barban, and Felix C. Tropf. 2020. *An Introduction to Statistical Genetic Data Analysis.* The MIT Press.

Millwood, Iona Y, Robin G Walters, Xue W Mei, Yu Guo, Ling Yang, et al. 2019. "Conventional and Genetic Evidence on Alcohol and Vascular Disease Aetiology: A Prospective Study of 500,000 Men and Women in China." *The Lancet*, 393: 1831–1842.

Mostafavi, Hakhamanesh, Arbel Harpak, Ipsita Agarwal, Dalton Conley, Jonathan K Pritchard, et al. 2020. "Variable prediction accuracy of polygenic scores within an ancestry group." *elife*, 9: e48376.

Mostafavi, Hakhamanesh, Jeffrey P. Spence, Sahin Naqvi, and Jonathan K Pritchard. 2023. "Systematic Differences In Discovery Of Genetic Effects On Gene Expression And Complex Traits." *Nature Genetics*, 55: 1866–1875.

Nelson-Coffey, S Katherine, Megan M Fritz, Sonja Lyubomirsky, and Steve W Cole. 2017. "Kindness in the blood: A randomized controlled trial of the gene regulatory impact of prosocial behavior." *Psychoneuroendocrinology*, 81: 8–13.

Norton, Edward, and Euna Han. 2008. "Genetic Information, Obesity and Labor Market Outcomes." *Health Economics*, 17: 1089–1104.

Novembre, John, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, et al. 2008. "Genes Mirror Geography Within Europe." *Nature*, 456(7218): 98–101.

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, et al. 2022. "The Complete Sequence Of A Human Genome." *Science*, 376(6558): 44–53.

O'Connor, Luke J., and Alkes L. Price. 2018. "Distinguishing Genetic Correlation from Causation across 52 Diseases and Complex Traits." *Nature Genetics*, 50: 1728–1734.

Okbay, Aysu, Jonathan P. Beauchamp, Mark A Fontana, James J Lee, Tune H Pers, et al. 2016. "Genome-wide association study identifies 74 loci associated with educational attainment." *Nature*, 533(7604): 539–542.

Okbay, Aysu, Yeda Wu, Nancy Wang, Hariharan Jayashankar, Michael Bennett, et al. 2022. "Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals." *Nature Genetics 2022 54:4*, 54(4): 437–449.

Oster, Emily. 2022. "New Study on Alcohol Consumption and Heart Disease." *Blog Post*, Accessed: November 14, 2023.

Otto, Sarah P., Freddy B. Christiansen, and Marcus W. Feldman. 1995. "Genetic and Cultural Inheritance of Continuous Traits." Morrison Institute for Population and Resource Studies Working Paper No. 64.

Panagiotou, Orestis a, and John P a Ioannidis. 2012. "What Should The Genome-wide Significance Threshold Be? Empirical Replication Of Borderline Genetic Associations." *International Journal of Epidemiology*, 41(1): 273–286.

Papageorge, Nicholas W, and Kevin Thom. 2020. "Genes, Education, and Labor Market Outcomes: Evidence from the Health and Retirement Study." *Journal of the European Economic Association*, 18(3): 1351–1399.

Patterson, Nick, Alkes L Price, and David Reich. 2006. "Population Structure and Eigenanalysis." *PLoS Genetics*, 2(12): e190.

Pazokitoroudi, Ali, Alec M. Chiu, Kathryn S. Burch, Bogdan Pasaniuc, and Sriram Sankararaman. 2021. "Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data." *American Journal of Human Genetics*, 108(5): 799–808.

Pers, Tune H, Juha M Karjalainen, Yingleong Chan, H.-J. Harm-Jan Westra, Andrew R Wood, et al. 2015. "Biological interpretation of genome-wide association studies using predicted gene functions." *Nature Communications*, 6(1): 5890.

Plomin, R, JC DeFries, VS Knopik, and JM Neiderhiser. 2016. "Top 10 Replicated Findings From Behavioral Genetics." *Perspectives in Psychological Science*, 11(1): 3–23.

Plomin, Robert, John C DeFries, and John C Loehlin. 1977. "Genotype-environment Interaction And Correlation In The Analysis Of Human Behavior." *Psychological Bulletin*, 84(2): 309–322.

Price, Alkes L, Agnar Helgason, Snaebjorn Palsson, Hreinn Stefansson, David St Clair, et al. 2009. "The impact of divergence time on the nature of population structure: an example from Iceland." *PLoS Genetics*, 5(6): e1000505.

Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, et al. 2006. "Principal Components Analysis Corrects For Stratification In Genome-Wide Association Studies." *Nature Genetics*, 38(8): 904–909.

Purcell, Shaun M, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, et al. 2009. "Common Polygenic Variation Contributes To Risk Of Schizophrenia And Bipolar Disorder." *Nature*, 460(7256): 748–752.

Raben, Timothy G., Louis Lello, Erik Widen, and Stephen H Hsu. 2023. "Biobank-scale methods and projections for sparse polygenic prediction from machine learning." *Scientific Reports*, 13: 11662.

Rietveld, Cornelius A, Sarah E Medland, Jaime Derringer, Jian Yang, and Tõnu Esko et al. 2013. "GWAS of 126,559 individuals identifies genetic variants associated with educational attainment." *Science*, 340(6139): 1467–1471.

Rimfeld, Kaili, Eva Krapohl, Maciej Trzaskowski, Jonathan RI Coleman, Saskia Selzam, et al. 2018. "Genetic Influence On Social Outcomes During And After The Soviet Era In Estonia." *Nature Human Behaviour*, 2(4): 269–275.

Robinson, Gene E., Christina M. Grozinger, and Charles W. Whitfield. 2015. "Sociogenomics social life in molecular terms." *Nature Reviews Genetics*, 6: 257–270.

Robinson, Matthew R., Aaron Kleinman, Mariaelisa Graff, Anna A. E. Vinkhuyzen, David Couper, et al. 2017. "Genetic evidence of assortative mating in humans." *Nature Human Behaviour*.

Ruan, Yunfeng, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, et al. 2022. "Improving Polygenic Prediction In Ancestrally Diverse Populations." *Nature Genetics*, 54(5): 573–580.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, 66(5): 688.

Rutherford, Adam. 2022. *Control: The Dark History and Troubling Present of Eugenics.* WW Norton & Company.

Sacerdote, Bruce. 2011. "Nature and Nurture Effects On Children's Outcomes: What Have We Learned From Studies of Twins And Adoptees?" In *Handbook of Social Economics.* , ed. Jess Benhabib, Alberto. Bisin and Matthew O. Jackson, Chapter 1, 1–29. Elsevier/North-Holland.

Sanderson, Eleanor, M Maria Glymour, Michael V Holmes, Hyunseung Kang, Jean Morrison, et al. 2022. "Mendelian Randomization." *Nature Reviews Methods Primers*, 2(1): 6.

Sanjak, Jaleal S., Julia Sidorenko, Kevin R. Thornton, Matthew R. Robinson, and Peter M. Visscher. 2017. "Evidence of directional and stabilizing selection in contemporary humans." *Proceedings of the National Academy of Sciences*, 115(1): 151–156.

Sanz-de Galdeano, Anna, and Anastasia Terskaya. Forthcoming. "Sibling Differences in Genetic Propensity for Education: How do Parents React?" *The Review of Economics and Statistics, forthcoming.*

Schaid, Daniel J, Wenan Chen, and Nicholas B Larson. 2018. "From genome-wide associations to candidate causal variants by statistical fine-mapping." *Nature Reviews Genetics*, 19(8): 491–504.

Schmitz, Lauren L, and Valentina Duque. 2022. "In utero exposure to the Great Depression is reflected in late-life epigenetic aging signatures." *Proceedings of the National Academy of Sciences*, 119(46): e2208530119.

Schmitz, Lauren L., Julia Goodwin, Jiacheng Miao, Qiongshi Lu, and Dalton Conley. 2021. "The Impact Of Late-Career Job Loss And Genetic Risk On Body Mass Index: Evidence From Variance Polygenic Scores." *Scientific Reports*, 11(7647).

Sekula, Peggy, M Fabiola Del Greco, Cristian Pattaro, and Anna Köttgen. 2016. "Mendelian randomization as an approach to assess causality using observational data." *Journal of the American Society of Nephrology: JASN*, 27(11): 3253.

Shen, Hao, and Marcus W Feldman. 2020. "Genetic nurturing, missing heritability, and causal analysis in genetic statistics." *Proceedings of the National Academy of Sciences*, 117(41): 25646–25654.

Simons, Yuval B., Kevin Bullaughey, Richard R. Hudson, Guy Sella, and Yuval B. Simons. 2018. "A Population Genetic Interpretation Of Gwas Findings For Human Quantitative Traits." *Plos Biology*, 16(3): e2002985.

Skov-Ettrup, Lise S., Børge G. Nordestgaard, Christina B. Petersen, and Janne S. Tolstrup. 2017. "Does High Tobacco Consumption Cause Psychological Distress? A Mendelian Randomization Study." *Nicotine & Tobacco Research*, 19(1): 32–38.

Sohail, Mashaal, Robert M Maier, Andrea Ganna, Alex Bloemendal, Alicia R Martin, et al. 2019. "Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies." *Elife*, 8: e39702.

Sotoudeh, Ramina, Kathleen Mullan Harris, and Dalton Conley. 2019. "Effects of the peer metagenomic environment on smoking behavior." *Proceedings of the National Academy of Sciences*, 116(33): 16302–16307.

Speed, Doug, and David J. Balding. 2019. "SumHer better estimates the SNP heritability of complex traits from summary statistics." *Nature Genetics*, 51: 277–284.

Spielman, Richard S., Ralph E. McGinnis, and Warren J. Ewens. 1993. "Transmission Test For Linkage Disequilibrium: The Insulin Gene Region And Insulin-Dependent Diabetes Mellitus (IDDM)." *American Journal of Human Genetics*, 52(3): 506–516.

Spolaore, Enrico, and Romain Warcziag. 2009. "The Diffusion of Development." *Quarterly Journal of Economics*, 124(2): 469–529.

Strachen, Tom, and Andrew P Read. 2018. *Human Molecular Genetics.* . 5th ed., Garland Science.

Sved, John A, and William G Hill. 2018. "One hundred years of linkage disequilibrium." *Genetics*, 209(3): 629–636.

Taliun, Daniel, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, et al. 2021. "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program." *Nature*, 590(7845): 290–299.

Taubman, Paul. 1976. "The Determinants Of Earnings: Genetics, Family, And Other Environments: A Study Of White Male Twins." *American Economic Review*, 66(5): 858–870.

The 1000 Genomes Project Consortium. 2015. "A Global Reference for Human Genetic Variation." *Nature*, 526(7571): 68–74.

Thomas, Duncan C, and David V Conti. 2004. "Commentary: the concept of Mendelian Randomization." *International Journal of Epidemiology*, 33(1): 21–25.

Tillmann, Taavi, Julien Vaucher, Aysu Okbay, Hynek Pikhart, Anne Peasey, et al. 2017. "Education and coronary heart disease: Mendelian randomisation study." *British Medical Journal*, 358: j3542.

Timpson, Nicholas J, Børge G Nordestgaard, Roger M Harbord, Jeppe Zacho, Tim M Frayling, et al. 2011. "C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization." *International Journal of Obesity*, 35(2): 300–308.

Trejo, Sam, and Benjamin W Domingue. 2018. "Genetic nature or genetic nurture? Introducing social genetic parameters to quantify bias in polygenic score analyses." *Biodemography and Social Biology*, 64(3-4): 187–215.

Turley, Patrick, Alicia R. Martin, Grant Goldman, Hui Li, Masahiro Kanai, et al. 2021*a*. "Multi-Ancestry Meta-Analysis Yields Novel Genetic Discoveries And Ancestry-Specific Associations." *bioRχiv*.

Turley, Patrick, Michelle N. Meyer, Nancy Wang, David Cesarini, Evelynn Hammonds, et al. 2021*b*. "Problems with Using Polygenic Scores to Select Embryos." *New England Journal of Medicine*, 385(1): 78–86.

Uricchio, Lawrence H. 2020. "Evolutionary Perspectives On Polygenic Selection, Missing Heritability, And GWAS." *Human Genetics*, 139: 5–21.

van Kippersluis, Hans, Pietro Biroli, Rita Dias Pereira, Titus J. Galama, Stephanie von Hinke, et al. 2023. "Overcoming attenuation bias in regressions using polygenic indices." *Nature Communications*, 14(4473).

van Leeuwen, Elisabeth M., Alexandros Kanterakis, Patrick Deelen, Mathijs V. Kattenberg, The Genome of the Netherlands Consortium, et al. 2015. "Population-specific genotype imputations using minimac or IMPUTE2." *Nature Protocols*, 10: 1285–1296.

Vattikuti, Shashaank, James J. Lee, Stephen D. H. Hsu, and Carson C. Chow. 2013. "Application of compressed sensing to genome wide association studies and genomic selection." *GigaScience*, 3(10).

Veller, Carl, Molly Przeworski, and Graham Coop. 2023. "Causal Interpretations Of Family GWAS In The Presence Of Heterogeneous Effects." *bioRxiv*.

Venter, CJ, MD Adams, EW Myers, PW Li, and RJ Mural. 2001. "The sequence of the human genome." *Science*, 291: 1304–1351.

Vilhjálmsson, Bjarni J., Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindström, et al. 2015. "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores." *The American Journal of Human Genetics*, 97(4): 576–592.

Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. "Five years of GWAS Discovery." *The American Journal of Human Genetics*, 90(1): 7–24.

Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, et al. 2017. "10 Years of GWAS Discovery: Biology, Function, and Translation." *The American Journal of Human Genetics*, 101(1): 5–22.

Visscher, Peter M, Sarah E Medland, Manuel A R Ferreira, Katherine I Morley, Gu Zhu, et al. 2006. "Assumption-free Estimation Of Heritability From Genome-wide Identity-by-descent Sharing Between Full Siblings." *PLoS Genetics*, 2(3): e41.

Visscher, Peter, William Hill, and Naomi Wray. 2008. "Heritability in the Genomics Era - Concepts and Misconceptions." *Nature Reviews Genetics*, 9: 255–266.

von Hinke Kessler Scholder, Stephanie, George Davey Smith, Debbie A Lawlor, Carol Propper, and Frank Windmeijer. 2011. "Mendelian randomization: the use of genes in instrumental variable analyses." *Health Economics*, 20(8): 893–896.

Wainschtein, Pierrick, Deepti Jain, Zhili Zheng, L Adrienne Cupples, Aladdin H Shadyab, et al. 2022. "Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data." *Nature Genetics*, 54(3): 263–273.

Walsh, Bruce, and Michael Lynch. 2018. "Associative Effects: Competition, Social Interactions, Group and Kin Selection." In *Evolution and Selection of Quantitative Traits*. Chapter 22. Oxford:Oxford University Press.

Wang, Xiaotong, Alicia Walker, Joana A Revez, Guiyan Ni, Mark J Adams, et al. 2023. "Polygenic risk prediction: why and when out-of-sample prediction R^2 can exceed SNP-based heritability." *The American Journal of Human Genetics*, 110(7): 1207–1215.

Wang, Ying, Jing Guo, Guiyan Ni, Jian Yang, Peter M. Visscher, et al. 2020. "Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations." *Nature Communications*, 11(1): 1–9.

Weber, Elke U, Ann-rené E Blais, and Nancy E Betz. 2002. "A Domain-Specific Risk-Attitude Scale: Measuring Risk Perceptions and Risk Behaviors." *Journal of Behavioral Decision Making*, 15(4): 263–290.

Wetterstrand, KA. 2023. "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)." *Available at: www.genome.gov/sequencingcostsdata. Accessed April 4, 2023.*

Widen, Erik, Timothy G. Raben, Louis Lello, and Stephen D. H. Hsu. 2021. "Machine Learning Prediction of Biomarkers from SNPs and of Disease Risk from Biomarkers in the UK Biobank." *Genes*, 12(7).

Wientjes, Yvonne C. J., Piter Bijma, Roel F. Veerkamp, and Mario P. L. Calus. 2016. "An Equation to Predict the Accuracy of Genomic Values by Combining Data from Multiple Traits, Populations, or Environments." *Genetics*, 202(2): 799–823.

Wientjes, Yvonne CJ, Roel F Veerkamp, Piter Bijma, Henk Bovenhuis, Chris Schrooten, and Mario PL Calus. 2015. "Empirical And Deterministic Accuracies Of Across-Population Genomic Prediction." *Genetics Selection Evolution*, 47(5): 1–14.

Winkler, Thomas W, Felix R Day, Damien C Croteau-Chonka, Andrew R Wood, Adam E Locke, et al. 2014. "Quality Control And Conduct

Of Genome-Wide Association Meta-Analyses." *Nature Protocols*, 9(5): 1192–1212.

Wray, Naomi R, Jian Yang, Ben J Hayes, Alkes L Price, Michael E Goddard, et al. 2013. "Pitfalls of predicting complex traits from SNPs." *Nature Reviews Genetics*, 14(7): 507–15.

Wray, Naomi R, Michael E Goddard, and Peter M Visscher. 2007. "Prediction of individual genetic risk to disease from genome-wide association studies." *Genome Research*, 17(10): 1520–1528.

Wu, Yang, Zhili Zheng, Peter M Visscher, and Jian Yang. 2017. "Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data." *Genome Biology*, 18(1): 86.

Yang, Jian, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, et al. 2010. "Common SNPs explain a large proportion of the heritability for human height." *Nature Genetics*, 42(7): 565–569.

Yengo, Loic, Matthew R Robinson, Matthew C Keller, Kathryn E Kemper, Yuanhao Yang, et al. 2018. "Imprint of Assortative Mating on the Human Genome." *Nature Human Behaviour*, 2(12): 948–954.

Yengo, Loïc, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, et al. 2022. "A Saturated Map Of Common Genetic Variants Associated With Human Height." *Nature*, 610: 704–712.

Young, Alexander I. 2022. "Discovering missing heritability in whole-genome sequencing data." *Nature Genetics*, 54: 224–226.

Young, Alexander I, and Richard Durbin. 2014. "Estimation of Epistatic Variance Components and Heritability in Founder Populations and Crosses." *Genetics*, 198(4): 1405–1416.

Young, Alexander I., Michael L. Frigge, Daniel F. Gudbjartsson, Gudmar Thorleifsson, Gyda Bjornsdottir, et al. 2018. "Relatedness Disequilibrium Regression Estimates Heritability Without Environmental Bias." *Nature Genetics*, 50(9): 1304–1310.

Young, Alexander I., Seyed Moeen Nehzati, Stefania Benonisdottir, Aysu Okbay, Hariharan Jayashankar, et al. 2022. "Mendelian Imputation Of Parental Genotypes Improves Estimates Of Direct Genetic Effect." *Nature Genetics*, 54(6): 897–905.

Young, Alexander Strudwick. 2023. "Estimation Of Indirect Genetic Effects And Heritability Under Assortative Mating." bioRχiv.

Zaidi, Arslan A, and Iain Mathieson. 2020. "Demographic history mediates the effect of stratification on polygenic scores." *eLife*, 9: e61548.

Zhang, Qianqian, Florian Privé, Bjarni Vilhjálmsson, and Doug Speed. 2021. "Improved Genetic Prediction of Complex Traits from Individual-Level Data or Summary Statistics." *Nature Communications*, 12(1): 1–9.

Zhao, Zijie, Jie Song, Tuo Wang, and Qiongshi Lu. 2021. "Polygenic Risk Scores: Effect Estimation and Model Optimization." *Quantitative Biology*, 9(2): 133–140.

Zheng, Zhili, Shouye Liu, Julia Sidorenko, Loic Yengo, Patrick Turley, et al. 2022. "Leveraging Functional Genomic Annotations And Genome Coverage To Improve Polygenic Prediction Of Complex Traits Within And Between Ancestries." *biorXiv*.

"Technical Appendix"

January 2024

I Local Average Treatment Effect (LATE)

Here, we show that estimates from specifications that include controls for parental genotypes have a LATE interpretation. For expositional ease, we consider the single-locus case, but the results below generalize to the multi-locus case. Consider a model where the effect of a SNP on a person varies by person, such that

$$y_i = x_i\beta_i + z_i\gamma_i + \epsilon_i,$$

where y_i is the phenotype, x_i is the genotype, β_i is the causal effect of that SNP for person i, z_i is the average parental genotype, γ_i is the coefficient on parental genotypes for person i, and ϵ_i is the residual. Due to Mendelian segregation, we can decompose a person's genotype into

$$x_i = z_i + x_{r,i}$$

where $x_{r,i}$ is the random component of person i's genotype.

We are interested in the coefficient on the child's genotype from an OLS regression of y_i onto x_i and z_i in some population. By the Frisch-Waugh-Lovell theorem (Lovell, 1963), this can be obtained by regressing x_i onto z_i, regressing y_i onto z_i, and then regressing the first set of residuals on the second set. We first calculate the coefficient from a regression of x_i onto z_i, giving

$$\begin{aligned}
\frac{\text{Cov}\left(x_i, z_i\right)}{\text{Var}\left(z_i\right)} &= \frac{\text{Cov}\left(z_i + x_{r,i}, z_i\right)}{\text{Var}\left(z_i\right)} \\
&= \frac{\text{Var}\left(z_i\right) + \text{Cov}\left(x_{r,i}, z_i\right)}{\text{Var}\left(z_i\right)} \\
&= \frac{\text{Var}\left(z_i\right)}{\text{Var}\left(z_i\right)} \\
&= 1.
\end{aligned}$$

The third step follows because $x_{r,i}$ and z_i are uncorrelated. Therefore, the residual is:

$$\begin{aligned}
\dot{x}_i &= x_i - z_i \\
&= z_i + x_{r,i} - z_i \\
&= x_{r,i}.
\end{aligned}$$

Next we evaluate the regression of y_i onto z_i:

$$\begin{aligned}
\frac{\text{Cov}\left(y_i, z_i\right)}{\text{Var}\left(z_i\right)} &= \frac{\text{Cov}\left(x_i\beta_i + z_i\gamma_i + \epsilon_i, z_i\right)}{\text{Var}\left(z_i\right)} \\
&= \frac{\text{Cov}\left[x_{r,i}\beta_i + z_i\left(\beta_i + \gamma_i\right) + \epsilon_i, z_i\right]}{\text{Var}\left(z_i\right)} \\
&= \frac{\text{Cov}\left(x_{r,i}\beta_i, z_i\right) + \text{Cov}\left[z_i\left(\beta_i + \gamma_i\right), z_i\right] + \text{Cov}\left(\epsilon_i, z_i\right)}{\text{Var}\left(z_i\right)} \\
&= \frac{\text{Cov}\left[z_i\left(\beta_i + \gamma_i\right), z_i\right]}{\text{Var}\left(z_i\right)} \\
&= \frac{\int\left(\beta_i + \gamma_i\right)\left[z_i - \mathbb{E}\left(z_i\right)\right]^2 dF_{z_i}}{\int\left[z_i - \mathbb{E}\left(z_i\right)\right]^2 dF_{z_i}} \\
&= \mathbb{E}_{z^2}\left(\beta_i + \gamma_i\right),
\end{aligned}$$

where $\mathbb{E}_{z^2}\left(\beta_i + \gamma_i\right)$ is the weighted average of $\beta_i + \gamma_i$, with weights equal to $\left[z_i - \mathbb{E}\left(z_i\right)\right]^2$. Therefore, the residual is:

$$\begin{aligned}
\dot{y}_i &= y_i - z_i\mathbb{E}_{z^2}\left(\beta_i + \gamma_i\right) \\
&= x_i\beta_i + z_i\gamma_i + \epsilon_i - z_i\mathbb{E}_{z^2}\left(\beta_i + \gamma_i\right) \\
&= x_{r,i}\beta_i + z_i\left[\beta_i + \gamma_i - \mathbb{E}_{z^2}\left(\beta_i + \gamma_i\right)\right] + \epsilon_i.
\end{aligned}$$

Finally, regressing $\dot{y}_i$ onto $\dot{x}_i$ gives:

$$\begin{aligned}
\hat{\beta} &= \frac{\text{Cov}\left(\dot{y}_i, \dot{x}_i\right)}{\text{Var}\left(\dot{x}_i\right)} \\
&= \frac{\text{Cov}\left(x_{r,i}\beta_i + z_i\left[\beta_i + \gamma_i - \mathbb{E}_{z^2}\left(\beta_i + \gamma_i\right)\right] + \epsilon_i, x_{r,i}\right)}{\text{Var}\left(x_{r,i}\right)} \\
&= \frac{\text{Cov}\left(x_{r,i}\beta_i, x_{r,i}\right)}{\text{Var}\left(x_{r,i}\right)} \\
&= \frac{\int\beta_i\left[x_{r,i} - \mathbb{E}\left(x_{r,i}\right)\right]^2 dF_{x_{r,i}}}{\int\left[x_{r,i} - \mathbb{E}\left(x_{r,i}\right)\right]^2 dF_{x_{r,i}}} \\
&= \frac{\int\beta_i x_{r,i}^2 dF_{x_{r,i}}}{\int x_{r,i}^2 dF_{x_{r,i}}}
\end{aligned}$$

So we see that the coefficient on the child's genotype in a regression of the phenotype onto both the child's and parental genotypes yields a weighted average of the causal effect of the genotypes, weighted by the squared random component of a child's genotype. We can further improve our intuition for this expression by splitting the sample into three sets, H^0, H^1, and H^2, corresponding to the individuals who have zero, one, or two heterozygous parents. Notice that if an

individual has no heterozygous parents, then they will always have a genotype equal to the mean parental genotype. So $x_{r,i}^2 = 0$ for all i. For individuals with one heterozygous parent, $x_{r,i} \in \left\{-\frac{1}{2}, \frac{1}{2}\right\}$ and therefore $x_{r,i}^2 = \frac{1}{4}$ for all i. For individuals with two heterozygous parents, $x_{r,i} \in \{-1, 0, 1\}$ with probabilities of $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$ for each element, respectively. This means that $x_{r,i}^2 \in \{0, 1\}$ with a probability of $\frac{1}{2}$ for each state and hence $\mathbb{E}\left(x_{r,i}^2 | H^2\right) = \frac{1}{2}$. Thus,

$$
\begin{aligned}
\hat{\beta} &= \frac{\int \beta_i x_{r,i}^2 dF_{x_{r,i}}}{\int x_{r,i}^2 dF_{x_{r,i}}} \\
&= \frac{\mathbb{E}_{x_{r,i}}\left(\beta_i x_{r,i}^2\right)}{\mathbb{E}_{x_{r,i}}\left(x_{r,i}^2\right)} \\
&= \frac{\mathbb{E}_{x_{r,i}}\left(\beta_i x_{r,i}^2 | H^0\right)\pi_0 + \mathbb{E}_{x_{r,i}}\left(\beta_i x_{r,i}^2 | H^1\right)\pi_1 + \mathbb{E}_{x_{r,i}}\left(\beta_i x_{r,i}^2 | H^2\right)\pi_2}{\mathbb{E}_{x_{r,i}}\left(x_{r,i}^2\right)} \\
&= \frac{\mathbb{E}_{x_{r,i}}\left(\beta_i | H^1\right)\mathbb{E}_{x_{r,i}}\left(x_{r,i}^2 | H^1\right)\pi_1 + \mathbb{E}_{x_{r,i}}\left(\beta_i | H^2\right)\mathbb{E}_{x_{r,i}}\left(x_{r,i}^2 | H^2\right)\pi_2}{\mathbb{E}_{x_{r,i}}\left(x_{r,i}^2\right)} \\
&= \frac{\frac{1}{4}\pi_1 \mathbb{E}_{x_{r,i}}\left(\beta_i | H^1\right) + \frac{1}{2}\pi_2 \mathbb{E}_{x_{r,i}}\left(\beta_i | H^2\right)}{\mathbb{E}_{x_{r,i}}\left(x_{r,i}^2\right)} \\
&= \frac{\frac{1}{4}\pi_1 \mathbb{E}_{x_{r,i}}\left(\beta_i | H^1\right) + \frac{1}{2}\pi_2 \mathbb{E}_{x_{r,i}}\left(\beta_i | H^2\right)}{\frac{1}{4}\pi_1 + \frac{1}{2}\pi_2} \\
&= \frac{\pi_1}{\pi_1 + 2\pi_2}\mathbb{E}_{x_{r,i}}\left(\beta_i | H^1\right) + \frac{2\pi_2}{\pi_1 + 2\pi_2}\mathbb{E}_{x_{r,i}}\left(\beta_i | H^2\right)
\end{aligned}
$$

where π_0, π_1, and π_2 are the fraction of individuals in the population with zero, one, and two heterozygous parents, respectively.

This expression makes clear a few key points. First, individuals with homozygous parents receive no weight in this regression. So to the degree that individuals with homozygous parents have systematically different genetic effect sizes, family-based estimates will not generalize to those individuals. Second, individuals with two heterozygous parents receive double the weight as those with one heterozygous parent. Third, in genetic studies with diverse-ancestry samples, particular ancestry groups will get more weight for some genetic variants than others. That is because certain genotypes will be more common in certain groups. As a result, even if the samples are relatively balanced between the different populations represented, populations with genotype frequencies close to one-half will tend to have relatively more individuals with one or two heterozygous parents, so the estimated average effect for that genetic variant will give more weight to such

populations.

II Derivations of Formulae for PGI Predictive Power

Here, we derive analytic formulae for the predictive power of a PGI. To begin, we provide here the derivation of the main text's Equation (13), which follows Daetwyler, Villanueva and Woolliams (2008):

$$
\begin{aligned}
R^2 &= \frac{\left[\mathrm{Cov}\left(y,\hat{g}\right)\right]^2}{\mathrm{Var}\left(y\right)\mathrm{Var}\left(\hat{g}\right)} = \frac{\left[\mathrm{Cov}\left(y,\frac{\breve{g}+e}{\rho}\right)\right]^2}{\mathrm{Var}\left(y\right)} \\
&= \frac{\left[\mathrm{Cov}\left(y,\breve{g}\right)\right]^2}{\mathrm{Var}\left(y\right)\left[\mathrm{Var}\left(\breve{g}\right)+\mathrm{Var}\left(e\right)\right]} = \left(\frac{\left[\mathrm{Cov}\left(y,\breve{g}\right)\right]^2}{\mathrm{Var}\left(y\right)\mathrm{Var}\left(\breve{g}\right)}\right)\left(\frac{\mathrm{Var}\left(\breve{g}\right)}{\mathrm{Var}\left(\breve{g}\right)+\mathrm{Var}\left(e\right)}\right) \\
&= \left(\frac{\left[\mathrm{Cov}\left(y,\breve{g}\right)\right]^2}{\mathrm{Var}\left(y\right)\mathrm{Var}\left(\breve{g}\right)}\right)\left(\frac{\mathrm{Var}\left(\breve{g}\right)/\mathrm{Var}\left(y\right)}{\mathrm{Var}\left(\breve{g}\right)/\mathrm{Var}\left(y\right)+\mathrm{Var}\left(e\right)/\mathrm{Var}\left(y\right)}\right) \\
&= \left(\frac{\left[\mathrm{Cov}\left(y,\breve{g}\right)\right]^2}{\mathrm{Var}\left(y\right)\mathrm{Var}\left(\breve{g}\right)}\right)\left(\frac{\breve{h}^2}{\breve{h}^2+\mathrm{Var}\left(e\right)/\mathrm{Var}\left(y\right)}\right) = \breve{h}^2\left(\frac{\breve{h}^2}{\breve{h}^2+M/N}\right), \qquad (24)
\end{aligned}
$$

where M is a constant and N is the GWAS sample size underlying the PGI weights. The second equality follows from $\mathrm{Var}\left(\hat{g}\right)=1$ and main text Equation (12). The third equality follows from the the approximation $\mathrm{Cov}\left(y,e\right)=0$ discussed in main text Section IV.A. The second-to-last equality follows from the definition of the optimal predictor, which implies $\mathrm{Cov}\left(y,\breve{g}\right)=\mathrm{Cov}\left(\breve{g}+\varepsilon,\breve{g}\right)=\mathrm{Var}\left(\breve{g}\right)$, and the definition of optimal predictive power: $\breve{h}^2\equiv\mathrm{Var}\left(\breve{g}\right)/\mathrm{Var}\left(y\right)$. As mentioned in the main text, the last equality follows because $\mathrm{Var}\left(e\right)$ converges to zero with the GWAS sample size at rate $1/N$.

In the remainder of this appendix, we generalize the results in Daetwyler, Villanueva and Woolliams (2008) by relaxing two key assumptions. First, following de Vlaming et al. (2017*b*), we allow for imperfect genetic correlation *and* different optimal predictive powers between the GWAS and prediction samples. Second, we relax the assumption that the GWAS and prediction samples have a common LD matrix. Wang et al. (2020) and Ding et al. (2023) also relaxed both assumptions but did so in a random-effects framework. Hence, their derivations are valid given their parametric assumptions on the joint distribution of effect sizes across the two samples. Like us, Wientjes et al. (2015,2016) relaxed both assumptions without making parametric assumptions but do not formally define and interpret all the parameters.[22]

[22]For example, Wientjes et al. (2015,2016) introduce a term that they call the "genetic correlation between populations" but that object is never clearly defined, and does not correspond to any of objects in our basic framework.

We begin by establishing some notation. First, let

$$y_{pred} = \tilde{\mathbf{x}}_{pred}\breve{\beta}_{pred} + \tilde{\varepsilon}_{pred}.$$

Here, y_{pred} is the phenotype in the prediction sample, $\tilde{\mathbf{x}}_{pred}$ is the vector of observed SNP genotypes, $\breve{\beta}_{pred}$ is the vector of optimal predictor weights in the prediction sample, and $\tilde{\varepsilon}_{pred}$ is a residual that is uncorrelated with the genotypes. Next, let

$$\hat{g} = \frac{\tilde{\mathbf{x}}_{pred}\hat{\beta}_{GWAS}}{\text{std}\left(\tilde{\mathbf{x}}_{pred}\hat{\beta}_{GWAS}\right)} = \frac{\tilde{\mathbf{x}}_{pred}\left(\breve{\beta}_{GWAS} + \mathbf{u}_{GWAS}\right)}{\text{std}\left(\tilde{\mathbf{x}}_{pred}\hat{\beta}_{GWAS}\right)}$$

denote a PGI constructed in the prediction sample using estimates of PGI weights from the GWAS sample, $\hat{\beta}_{GWAS}$, and let $\mathbf{u}_{GWAS}$ denote the estimation error from such a projection in a finite sample. Finally, let $\mathbf{\Sigma}_{pred} \equiv \text{Var}\left(\tilde{\mathbf{x}}_{pred}\right)$ and $\mathbf{\Sigma}_{GWAS} \equiv \text{Var}\left(\tilde{\mathbf{x}}_{GWAS}\right)$ denote the LD matrices in the prediction and GWAS populations, respectively. Using this notation, the R^2 from a regression of the phenotype on the PGI in the prediction sample is:

$$\begin{aligned}
R^2 &= \frac{\left[\text{Cov}\left(y_{pred}, \hat{g}\right)\right]^2}{\text{Var}\left(y_{pred}\right)\text{Var}\left(\hat{g}\right)} \\
&= \frac{\left[\text{Cov}\left(\tilde{\mathbf{x}}_{pred}\breve{\beta}_{pred} + \tilde{\varepsilon}_{pred}, \frac{\tilde{\mathbf{x}}_{pred}\left(\breve{\beta}_{GWAS}+\mathbf{u}_{GWAS}\right)}{\text{std}\left(\tilde{\mathbf{x}}_{pred}\hat{\beta}_{GWAS}\right)}\right)\right]^2}{\text{Var}\left(y_{pred}\right)\text{Var}\left(\frac{\tilde{\mathbf{x}}_{pred}\left(\breve{\beta}_{GWAS}+\mathbf{u}_{GWAS}\right)}{\text{std}\left(\tilde{\mathbf{x}}_{pred}\hat{\beta}_{GWAS}\right)}\right)} \\
&= \frac{\left[\text{Cov}\left(\tilde{\mathbf{x}}_{pred}\breve{\beta}_{pred}, \tilde{\mathbf{x}}_{pred}\breve{\beta}_{GWAS}\right)\right]^2}{\text{Var}\left(y_{pred}\right)\text{Var}\left(\tilde{\mathbf{x}}_{pred}\hat{\beta}_{GWAS}\right)} \\
&= \underbrace{\frac{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\breve{\beta}_{pred}\right)}{\text{Var}\left(y_{pred}\right)}}_{=\breve{h}^2_{pred}} \underbrace{\frac{\left[\text{Cov}\left(\tilde{\mathbf{x}}_{pred}\breve{\beta}_{pred}, \tilde{\mathbf{x}}_{pred}\breve{\beta}_{GWAS}\right)\right]^2}{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\breve{\beta}_{pred}\right)\text{Var}\left(\tilde{\mathbf{x}}_{pred}\breve{\beta}_{GWAS}\right)}}_{=r^2_g} \frac{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\breve{\beta}_{GWAS}\right)}{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\hat{\beta}_{GWAS}\right)},
\end{aligned}$$

where $\breve{h}^2_{pred}$ is the optimal predictive power in the prediction sample. Hence, we have that:

$$R^2 = \breve{h}^2_{pred} r^2_g \frac{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\breve{\beta}_{GWAS}\right)}{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\hat{\beta}_{GWAS}\right)}. \tag{25}$$

For some intuition on how to interpret the parameter r_g^2, consider first the special case when $\mathbf{\Sigma}_{pred} = \mathbf{\Sigma}_{GWAS}$, $\breve{\beta}_{GWAS} = \tilde{\beta}_{GWAS}$, and $\breve{\beta}_{pred} = \tilde{\beta}_{pred}$. Then r_g^2 is the squared correlation between two additive SNP factors, one based on the GWAS weights ($\tilde{\beta}_{GWAS}$) and one based on the prediction sample weights ($\tilde{\beta}_{pred}$), so r_g is an instance of the genetic correlation parameter $r_{\boldsymbol{x}\beta}$ defined in Equation (6). In the more general case when $\mathbf{\Sigma}_{pred} \neq \mathbf{\Sigma}_{GWAS}$, $\breve{\beta}_{GWAS} \neq \tilde{\beta}_{GWAS}$, and $\breve{\beta}_{pred} \neq \tilde{\beta}_{pred}$, r_g^2 is the correlation between the optimal predictor in the prediction sample and a PGI in the prediction sample that uses the GWAS-sample optimal predictor weights.

We proceed with our derivation by rewriting Equation (25) as:

$$
\begin{aligned}
R^2 &= \breve{h}^2_{pred} r_g^2 \frac{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\breve{\beta}_{GWAS}\right)}{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\breve{\beta}_{GWAS}\right) + \text{Var}\left(\tilde{\mathbf{x}}_{pred}\mathbf{u}_{GWAS}\right)} \\
&= \breve{h}^2_{pred} r_g^2 \left(\frac{\frac{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\breve{\beta}_{GWAS}\right)}{\text{Var}(y_{GWAS})}}{\frac{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\breve{\beta}_{GWAS}\right)}{\text{Var}(y_{GWAS})} + \frac{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\mathbf{u}_{GWAS}\right)}{\text{Var}(y_{GWAS})}} \right). \qquad (26)
\end{aligned}
$$

Next, we consider the common term in the numerator and denominator of the fraction:

$$
\begin{aligned}
\frac{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\breve{\beta}_{GWAS}\right)}{\text{Var}\left(y_{GWAS}\right)} &= \frac{\breve{\beta}'_{GWAS}\mathbf{\Sigma}_{pred}\breve{\beta}_{GWAS}}{\text{Var}\left(y_{GWAS}\right)} \\
&= \frac{\breve{\beta}'_{GWAS}\left(\mathbf{\Sigma}_{pred} - \mathbf{\Sigma}_{GWAS} + \mathbf{\Sigma}_{GWAS}\right)\breve{\beta}_{GWAS}}{\text{Var}\left(y_{GWAS}\right)} \\
&= \breve{h}^2_{GWAS} + \frac{\breve{\beta}'_{GWAS}\left(\mathbf{\Sigma}_{pred} - \mathbf{\Sigma}_{GWAS}\right)\breve{\beta}_{GWAS}}{\text{Var}\left(y_{GWAS}\right)} \\
&= \breve{h}^2_{GWAS} + \frac{\breve{\beta}'_{GWAS}\mathbf{\Delta}_{\mathbf{\Sigma}}\breve{\beta}_{GWAS}}{\text{Var}\left(y_{GWAS}\right)} \qquad (27)
\end{aligned}
$$

where, in the last step, we substituted in the LD-difference matrix parameter, defined as $\mathbf{\Delta}_{\mathbf{\Sigma}} \equiv \mathbf{\Sigma}_{pred} - \mathbf{\Sigma}_{GWAS}$.

Next, note that by the properties of least-squares projection we have:

$$
\text{Var}\left(\mathbf{u}_{GWAS}\right) \approx \frac{\text{Var}\left(y_{GWAS}\right)}{N}\mathbf{\Sigma}^{-1}_{GWAS}.
$$

This approximation requires that the GWAS association of each SNP be small such that $\text{Var}\left(y_{GWAS}\right)$ is approximately equal to the variance of the residual of

each univariate GWAS regression and that the sample size is large enough that the GWAS estimates have converged to their asymptotic distribution. Therefore, we anticipate that this approximation will be extremely good for virtually all PGIs for complex phenotypes constructed from a large-sample GWAS. Furthermore, $\tilde{\mathbf{x}}_{pred}$ and $\mathbf{u}_{GWAS}$ are mean-zero and independent. The second term in the denominator can therefore be expressed as follows:

$$\begin{aligned}\frac{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\mathbf{u}_{GWAS}\right)}{\text{Var}\left(y_{GWAS}\right)} &= \frac{\text{sum}\left[\text{Var}\left(\tilde{\mathbf{x}}_{pred}\right)\circ\text{Var}\left(\mathbf{u}_{GWAS}\right)\right]}{\text{Var}\left(y_{GWAS}\right)} \\ &\approx \frac{\text{sum}\left[\text{Var}\left(\tilde{\mathbf{x}}_{pred}\right)\circ\frac{\text{Var}(y_{GWAS})}{N}\boldsymbol{\Sigma}_{GWAS}^{-1}\right]}{\text{Var}\left(y_{GWAS}\right)} \\ &= \frac{1}{N}\text{sum}\left(\boldsymbol{\Sigma}_{pred}\circ\boldsymbol{\Sigma}_{GWAS}^{-1}\right), \qquad (28)\end{aligned}$$

where $\circ$ denotes the element-wise multiplication operator and $\text{sum}\left(\cdot\right)$ denotes the grand sum (i.e., the sum over all the elements of the matrix). Substituting (27) and (28) into (26) and rearranging yields the analytic formula for the generalized Daetwyler projection given in the main text:

$$R^2 = \breve{h}^2_{pred}r_g^2\left(\frac{\breve{h}^2_{GWAS}+\frac{\breve{\beta}'_{GWAS}(\Delta_\Sigma)\breve{\beta}_{GWAS}}{\text{Var}(y_{GWAS})}}{\breve{h}^2_{GWAS}+\frac{\breve{\beta}'_{GWAS}(\Delta_\Sigma)\breve{\beta}_{GWAS}}{\text{Var}(y_{GWAS})}+\frac{\text{sum}\left(\boldsymbol{\Sigma}_{pred}\circ\boldsymbol{\Sigma}_{GWAS}^{-1}\right)}{N}}\right). \qquad (29)$$

A Remarks on Generalized Formula

For some intuition on the properties of the generalized formula, it is instructive to consider the special case where the LD matrices in the populations are both diagonal. Then

$$\frac{\text{Var}\left(\tilde{\mathbf{x}}_{pred}\mathbf{u}_{GWAS}\right)}{\text{Var}\left(y_{GWAS}\right)} \approx \frac{1}{N}\sum_{j=1}^{M}\frac{\sigma^2_{pred,j}}{\sigma^2_{GWAS,j}}.$$

In what follows, we will treat $\sigma^2_{pred,j}$ and $\sigma^2_{GWAS,j}$ as identically distributed random variables.[23] We consider two cases, one where $\sigma^2_{pred,j}$ and $\sigma^2_{GWAS,j}$ are equal

[23] We believe this assumption is reasonable for PGIs constructed using the Bayesian methods we focus on in this paper which use all measured SNPs, as long as the main driver of allele frequency differences between the prediction and GWAS populations is genetic drift. However, this assumption is likely to be violated for PGIs that are constructed using a "pruning-and-thresholding" approach, in which only a set of approximately uncorrelated SNPs that meet some statistical-significance threshold in the GWAS are included in the PGI. Under genetic drift, even though SNP effect sizes are equal across the prediction and GWAS populations, SNP allele frequencies will randomly differ, and hence $\sigma^2_{GWAS,j}$ and $\sigma^2_{pred,j}$ will randomly differ. Because inclusion in the PGI is conditioned on statistical significance, SNPs with a high $\sigma^2_{GWAS,j}$ are more likely to be included in the PGI since those SNPs will have a smaller standard error. By regression to the mean, $\sigma^2_{GWAS,j} \geq \sigma^2_{pred,j}$ for these SNPs on average, so $\sigma^2_{GWAS,j}$ and $\sigma^2_{pred,j}$ would not be identically distributed.

and one where they are independent. The first would arise if the GWAS and prediction populations are the same. The second is an extreme case that may arise if the two populations had been separated for an arbitrarily long time and there are no selective forces that cause allele frequencies to be similar.

Under first scenario $\left(\sigma^2_{GWAS,j} = \sigma^2_{pred,j}\right)$, this expression equals:

$$\frac{1}{N}\sum_{j=1}^{M}\frac{\sigma^2_{pred,j}}{\sigma^2_{GWAS,j}} = \frac{1}{N}\sum_{j=1}^{M}\frac{\sigma^2_{pred,j}}{\sigma^2_{pred,j}} = \frac{1}{N}\sum_{j=1}^{M}1 = \frac{M}{N}.$$

This expression is consistent with the analytical results reported in Daetwyler, Villanueva and Woolliams (2008) and de Vlaming et al. (2017*a*). To see this, note that if $\sigma^2_{pred,j} = \sigma^2_{GWAS,j}$, then Δ_Σ is a null matrix and $r_g = r_{\boldsymbol{x}\beta}$.Therefore,

$$R^2 = \breve{h}^2_{pred} r^2_{\boldsymbol{x}\boldsymbol{\beta}} \left(\frac{\breve{h}^2_{GWAS}}{\breve{h}^2_{GWAS} + \frac{M}{N}}\right),$$

which is exactly the formula derived by de Vlaming et al. (2017*a*).

Under the second scenario, the expected value of the $\text{Var}\left(\tilde{\mathbf{x}}_{pred}\mathbf{u}_{GWAS}\right)/\text{Var}\left(y_{GWAS}\right)$ term is:

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{N}\sum_{j=1}^{M}\frac{\sigma^2_{pred,j}}{\sigma^2_{GWAS,j}}\right) &= \frac{1}{N}\sum_{j=1}^{M}\mathbb{E}\left(\frac{\sigma^2_{pred,j}}{\sigma^2_{GWAS,j}}\right) \\
&= \frac{1}{N}\sum_{j=1}^{M}\mathbb{E}\left(\sigma^2_{pred,j}\right)\mathbb{E}\left(\frac{1}{\sigma^2_{GWAS,j}}\right) \\
&\geq \frac{1}{N}\sum_{j=1}^{M}\frac{\mathbb{E}\left(\sigma^2_{pred,j}\right)}{\mathbb{E}\left(\sigma^2_{GWAS,j}\right)} \\
&= \frac{1}{N}\sum_{j=1}^{M}1 \\
&= \frac{M}{N}.
\end{aligned}$$

where the inequality follows from Jensen's inequality since the function $f(x) = 1/x$ is convex. When the GWAS and prediction samples have different LD structures, we would expect the $\text{Var}\left(\tilde{\mathbf{x}}_{pred}\mathbf{u}_{GWAS}\right)/\text{Var}\left(y_{GWAS}\right)$ term to be larger, reducing the predictive power of the PGI.

III Causal Interpretation of PGI

Here, we show that in a regression of some phenotype on the child and parental polygenic indexes, the coefficient associated on the child polygenic index will be a weighted sum of causal effects of the child genotypes. Let $\mathbf{x}$ denote a vector of genotypes for some person, $\mathbf{x}_m$ denote the vector of genotypes for the person's mother, and $\mathbf{x}_f$ denote the vector of genotypes of the person's father. We use $\mathbf{x}_p$ to denote the combined genotypes of the parents such that

$$\mathbf{x}_p = \mathbf{x}_m + \mathbf{x}_f.$$

To begin, we evaluate the variance-covariance (VCV) matrices for the genotype vectors. We model the generations in discrete time but otherwise allow for a flexible model of assortative mating; for example, we do not impose that the VCV matrices are in equilbrium between the parent's and child's generation. To do this, we split each of the genotype vectors (with elements in $\{0, 1, 2\}$) into the sum of vectors, each with elements in $\{0, 1\}$, corresponding to the alleles inherited from each of their parents. We use $\mathbf{x}^{(m)}$, $\mathbf{x}_m^{(m)}$, and $\mathbf{x}_f^{(m)}$ to denote the maternally inherited genotypes for the individual, their mother, and their father, respectively. Similarly, we use $\mathbf{x}^{(f)}$, $\mathbf{x}_m^{(f)}$, and $\mathbf{x}_f^{(f)}$ to denote the respective paternally inherited genotypes. By definition,

$$\begin{aligned}
\mathbf{x} &= \mathbf{x}^{(m)} + \mathbf{x}^{(f)}, \\
\mathbf{x}_m &= \mathbf{x}_m^{(m)} + \mathbf{x}_m^{(f)}, \\
\mathbf{x}_f &= \mathbf{x}_f^{(m)} + \mathbf{x}_f^{(f)}.
\end{aligned}$$

Let $\mathbf{\Sigma} \equiv \text{Var}\left(\mathbf{x}_m^{(m)}\right) = \text{Var}\left(\mathbf{x}_m^{(f)}\right) = \text{Var}\left(\mathbf{x}_f^{(m)}\right) = \text{Var}\left(\mathbf{x}_f^{(f)}\right)$ denote the VCV matrix of the maternally or paternally inherited alleles of the parents. Due to assortative mating, the genotypes of the mother and father may be correlated. We use $\mathbf{A} \equiv \text{Cov}\left(\mathbf{x}_m^{(m)}, \mathbf{x}_f^{(m)}\right) = \text{Cov}\left(\mathbf{x}_m^{(m)}, \mathbf{x}_f^{(f)}\right) = \text{Cov}\left(\mathbf{x}_m^{(f)}, \mathbf{x}_f^{(m)}\right) = \text{Cov}\left(\mathbf{x}_m^{(f)}, \mathbf{x}_f^{(f)}\right)$ to denote this covariance. Because there may have been assortative mating in the generation of the grandparents (or in previous generations), the paternally and maternally inherited genotypes within each genotype may also be correlated. We use $\mathbf{B} \equiv \text{Cov}\left(\mathbf{x}_m^{(m)}, \mathbf{x}_m^{(f)}\right) = \text{Cov}\left(\mathbf{x}_f^{(m)}, \mathbf{x}_f^{(f)}\right)$ to denote this covariance.

We now use this notation to express the VCV matrix for each parent's geno-

types. We calculate

$$\begin{aligned}\mathrm{Var}\left(\mathbf{x}_m\right) &= \mathrm{Var}\left(\mathbf{x}_m^{(m)}\right) + \mathrm{Var}\left(\mathbf{x}_m^{(f)}\right) + 2\mathrm{Cov}\left(\mathbf{x}_m^{(m)}, \mathbf{x}_m^{(f)}\right) \\ &= 2\boldsymbol{\Sigma} + 2\mathbf{B} \\ &= 2\left(\boldsymbol{\Sigma} + \mathbf{B}\right).\end{aligned}$$

Similarly,

$$\mathrm{Var}\left(\mathbf{x}_f\right) = 2\left(\boldsymbol{\Sigma} + \mathbf{B}\right).$$

Next, we calculate

$$\begin{aligned}\mathrm{Cov}\left(\mathbf{x}_m, \mathbf{x}_f\right) &= \mathrm{Cov}\left(\mathbf{x}_m^{(m)}, \mathbf{x}_f^{(m)}\right) + \mathrm{Cov}\left(\mathbf{x}_m^{(m)}, \mathbf{x}_f^{(f)}\right) \\ &\quad + \mathrm{Cov}\left(\mathbf{x}_m^{(f)}, \mathbf{x}_f^{(m)}\right) + \mathrm{Cov}\left(\mathbf{x}_m^{(f)}, \mathbf{x}_f^{(f)}\right) \\ &= 4\mathbf{A}.\end{aligned}$$

Using these results,

$$\begin{aligned}\mathrm{Var}\left(\mathbf{x}_p\right) &= \mathrm{Var}\left(\mathbf{x}_m\right) + \mathrm{Var}\left(\mathbf{x}_f\right) + 2\mathrm{Cov}\left(\mathbf{x}_m, \mathbf{x}_f\right) \\ &= 2\left(\boldsymbol{\Sigma} + \mathbf{B}\right) + 2\left(\boldsymbol{\Sigma} + \mathbf{B}\right) + 8\mathbf{A} \\ &= 4\left[\left(\boldsymbol{\Sigma} + \mathbf{B}\right) + 2\mathbf{A}\right].\end{aligned}$$

Next, we calculate the VCV matrix for the child's genotypes. To do this, we first consider the VCV matrix for their maternally or paternally inherited alleles separately. Considering a pair of alleles inherited from a particular parent, if they both had been inherited from that parent's mother or both from that parent's father, those genotypes would have a covariance given by some element of the $\boldsymbol{\Sigma}$ matrix. If they had been inherited from different parents, those genotypes would have a covariance given by some element of the $\mathbf{B}$ matrix. We let $\mathbf{P}$ denote a matrix, each of whose entries isthe probability that a pair of genotypes from $\mathbf{x}^{(m)}$ are drawn from the same grandparent; this same matrix also encodes the probability that a pair of genotypes from $\mathbf{x}^{(f)}$ are drawn from the same grandparent. By the laws of Mendelian inheritance, $\mathbf{P}$ will have values of one along the diagonal, will have values of $1/2$ for any pair of genotypes corresponding to different chromosomes, and will have values between these two values for genotypes on the same chromosome. Since the means of $\mathbf{x}^{(m)}$ and $\mathbf{x}^{(f)}$ are the same no matter which grantparent they are inherited from,

$$\mathrm{Var}\left(\mathbf{x}^{(m)}\right) = \mathrm{Var}\left(\mathbf{x}^{(f)}\right) = \mathbf{P} \circ \boldsymbol{\Sigma} + (1 - \mathbf{P}) \circ \mathbf{B},$$

where $\circ$ is element-wise multiplication. Therefore,

$$\begin{aligned}\text{Var}(\mathbf{x}) &= \text{Var}\left(\mathbf{x}^{(m)}\right) + \text{Var}\left(\mathbf{x}^{(f)}\right) + 2\text{Cov}\left(\mathbf{x}^{(m)}, \mathbf{x}^{(f)}\right) \\ &= 2\left[\mathbf{P} \circ \mathbf{\Sigma} + (1 - \mathbf{P}) \circ \mathbf{B} + \mathbf{A}\right].\end{aligned}$$

Finally, we calculate the covariance between the child's and parental genotype vectors:

$$\begin{aligned}\text{Cov}(\mathbf{x}, \mathbf{x}_p) &= \text{Cov}\left(\mathbf{x}^{(m)} + \mathbf{x}^{(f)}, \mathbf{x}_m + \mathbf{x}_f\right) \\ &= \text{Cov}\left(\mathbf{x}^{(m)}, \mathbf{x}_m\right) + \text{Cov}\left(\mathbf{x}^{(f)}, \mathbf{x}_f\right) \\ &+ \text{Cov}\left(\mathbf{x}^{(m)}, \mathbf{x}_f\right) + \text{Cov}\left(\mathbf{x}^{(f)}, \mathbf{x}_m\right) \\ &= 2\left[(\mathbf{\Sigma} + \mathbf{B}) + 2\mathbf{A}\right].\end{aligned}$$

Projecting y onto $\mathbf{x}$ and $\mathbf{x}_p$, we obtain the following regression equation:

$$y = \mathbf{x}\beta + \mathbf{x}_p\mathbf{b}_p + e,$$

where e is the residual. Because of the random assignment of genotypes conditional on the genotypes of the parents, the entries of β are (local average) causal effects of each genotype on y (see Appendix I). The vector $\mathbf{b}_p$ must then pick up, in addition to parental genetic effects, any gene-environment correlations (including population stratification).

Suppose we construct a polygenic index with weight vector $\mathbf{w}$. The polygenic index of the individual is $\mathbf{xw}$, and the parental polygenic index is $\mathbf{x}_p\mathbf{w}$. We will show that when we regress y on $\mathbf{xw}$ and $\mathbf{x}_p\mathbf{w}$, the coefficient associated with $\mathbf{xw}$ will only be a weighted sum of the elements of the causal effect vector β and not a function of $\mathbf{b}_p$.

Let $\alpha = [\alpha_g; \alpha_p]$ denote the population coefficients from regressing y onto $\mathbf{xw}$

and $\mathbf{x}_p\mathbf{w}$. We calculate:

$$
\begin{aligned}
\alpha &= \begin{bmatrix} \operatorname{Var}(\mathbf{xw}) & \operatorname{Cov}(\mathbf{xw}, \mathbf{x}_p\mathbf{w}) \\ & \operatorname{Var}(\mathbf{x}_p\mathbf{w}) \end{bmatrix}^{-1} \begin{bmatrix} \operatorname{Cov}(\mathbf{xw}, y) \\ \operatorname{Cov}(\mathbf{x}_p\mathbf{w}, y) \end{bmatrix} \\
&= \begin{bmatrix} 2\mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (1-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\mathbf{w} & 2\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \\ & 4\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \end{bmatrix}^{-1} \\
&\times \begin{bmatrix} \operatorname{Cov}(\mathbf{xw}, \mathbf{x}\beta + \mathbf{x}_p\mathbf{b}_p + e) \\ \operatorname{Cov}(\mathbf{x}_p\mathbf{w}, \mathbf{x}\beta + \mathbf{x}_p\mathbf{b}_p + e) \end{bmatrix} \\
&= \begin{bmatrix} 2\mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (1-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\mathbf{w} & 2\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \\ & 4\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \end{bmatrix}^{-1} \\
&\times \begin{bmatrix} 2\mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (1-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\beta + 2\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \\ 2\mathbf{w}'\left[\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right]\beta + 4\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (1-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\mathbf{w} & \mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \\ & 2\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \end{bmatrix}^{-1} \\
&\times \begin{bmatrix} \mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (1-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\beta + \mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \\ \mathbf{w}'\left[\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right]\beta + 2\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \end{bmatrix} \\
&= \frac{1}{D}\begin{bmatrix} 2\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} & -\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \\ & \mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (1-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\mathbf{w} \end{bmatrix} \\
&\times \begin{bmatrix} \mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (1-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\beta + \mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \\ \mathbf{w}'\left[\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right]\beta + 2\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \end{bmatrix}
\end{aligned}
$$

where, after simplifying,

$$D = \mathbf{w}'\left[(2\mathbf{P}-1)\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\mathbf{w}\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}$$

is the determinant of the inverted matrix in the first line of the above derivation..

Thus, for the child-PGI coefficient, we obtain

$$
\begin{aligned}
\alpha_g &= \frac{1}{D}\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}\mathbf{w}'\left[(2\mathbf{P}-1)\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\beta \\
&= \frac{\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}\mathbf{w}'\left[(2\mathbf{P}-1)\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\beta}{\mathbf{w}'\left[(2\mathbf{P}-1)\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\mathbf{w}\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}} \\
&= \left(\mathbf{w}'\left[(2\mathbf{P}-1)\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\mathbf{w}\right)^{-1}\mathbf{w}'\left[(2\mathbf{P}-1)\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\beta
\end{aligned}
$$

which is a weighted sum of the true causal effects β. More specifically, it is a generalized least squares (GLS) regression coefficient of the true effect sizes onto the PGI weights, with GLS weight matrix $(2\mathbf{P}-1)\circ(\boldsymbol{\Sigma}-\mathbf{B})$.

For the parental PGI coefficient, we obtain

$$\begin{aligned}\alpha_p &= \frac{1}{D}\left\{\mathbf{w}'\left[(2\mathbf{P}-1)\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\mathbf{w}\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p\right. \\ &\quad \left.+\mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma}+(1-\mathbf{P})\circ\mathbf{B}+\mathbf{A}\right]\left(\mathbf{w}\beta'-\beta\mathbf{w}'\right)\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}\right\} \\ &= \mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \\ &\quad +\frac{\mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma}+(1-\mathbf{P})\circ\mathbf{B}+\mathbf{A}\right]\left(\mathbf{w}\beta'-\beta\mathbf{w}'\right)\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}}{\mathbf{w}'\left[(2\mathbf{P}-1)\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\mathbf{w}\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}},\end{aligned}$$

which is a function of both β and $\mathbf{b}_p$.

We next consider two special cases.

First, suppose that we use the true genetic effect sizes on y as the PGI weights (and all genetic variants with causal effects on y are included in the PGI), such that $\mathbf{w}=\beta$. In this case,

$$\alpha_g = \left(\beta'\left[(2\mathbf{P}-1)\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\beta\right)^{-1}\beta'\left[(2\mathbf{P}-1)\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\beta = 1.$$

Also $(\mathbf{w}\beta'-\beta\mathbf{w}')=\mathbf{0}$, so

$$\alpha_p = \left[\beta'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\beta\right]^{-1}\beta'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p,$$

which means that the coefficient on the parental PGI is simply a GLS regression (with weight matrix $\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}$) of the parental coefficients onto the causal genetic effects.

Second, suppose that there is no assortative mating, such that $\mathbf{A}=\mathbf{B}=\mathbf{0}$. Then:

$$\alpha_g = \left(\mathbf{w}'\left[(2\mathbf{P}-1)\circ\boldsymbol{\Sigma}\right]\mathbf{w}\right)^{-1}\mathbf{w}'\left[(2\mathbf{P}-1)\circ\boldsymbol{\Sigma}\right]\beta.$$

Recall that $P_{ij}=1/2$, implying that $2P_{ij}-1=0$ for each pair of SNPs, i and j, that are on different chromosomes. Within a chromosome, if there is random mating, then Σ_{ij} decays much more quickly than P_{ij} with distance between the SNPs. This is because the matrix $\mathbf{P}$ is approximately fixed across generations since it is related to the probability that an odd number of recombinations events will have occurred between a pair of SNPs in the genome. In contrast, the elements of $\boldsymbol{\Sigma}$ will decay by a factor of $\mathbf{P}$ in each generation of random mating. (This is because there is a P_{ij} chance that the ij correlation within a parent will be broken by recombination events in each generation.) Thus, we expect the approximation

$$(2\mathbf{P}-1)\circ\boldsymbol{\Sigma}\approx\boldsymbol{\Sigma}$$

to be very accurate. This gives us the expressions in the main text,

$$\begin{aligned}\alpha_g &= \left(\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}\right)^{-1}\mathbf{w}'\boldsymbol{\Sigma}\beta \\ \alpha_p &= \left(\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}\right)^{-1}\mathbf{w}'\boldsymbol{\Sigma}\mathbf{b}_p,\end{aligned}$$

where each coefficient has a generalized-least-squares interpretation.