PNAS RESEARCH ARTICLE | GENETICS

Causal interpretations of family GWAS in the presence of heterogeneous effects

Carl Veller[a,1], Molly Przeworski[b,c,1], and Graham Coop[d,1]

Affiliations are included on p. 9.

Edited by Marcus Feldman, Stanford University, Stanford, CA; received January 21, 2024; accepted July 26, 2024

Family-based genome-wide association studies (GWASs) are often claimed to provide an unbiased estimate of the average causal effects (or average treatment effects; ATEs) of alleles, on the basis of an analogy between the random transmission of alleles from parents to children and a randomized controlled trial. We show that this claim does not hold in general. Because Mendelian segregation only randomizes alleles among children of heterozygotes, the effects of alleles in the children of homozygotes are not observable. This feature will matter if an allele has different average effects in the children of homozygotes and heterozygotes, as can arise in the presence of gene-by-environment interactions, gene-by-gene interactions, or differences in linkage disequilibrium patterns. At a single locus, family-based GWAS can be thought of as providing an unbiased estimate of the average effect in the children of heterozygotes (i.e., a local average treatment effect; LATE). This interpretation does not extend to polygenic scores (PGSs), however, because different sets of SNPs are heterozygous in each family. Therefore, other than under specific conditions, the within-family regression slope of a PGS cannot be assumed to provide an unbiased estimate of the LATE for any subset or weighted average of families. In practice, the potential biases of a family-based GWAS are likely smaller than those that can arise from confounding in a standard, population-based GWAS, and so family studies remain important for the dissection of genetic contributions to phenotypic variation. Nonetheless, their causal interpretation is less straightforward than has been widely appreciated.

GWAS | G×E | epistasis | family GWAS | polygenic score

Significance

Family-based genome-wide association studies (GWASs) allow the effects of genetic variants carried by a child on the child's phenotype to be distinguished from other effects, notably those of their familial environment. It is often claimed that, because family studies leverage the randomness of genetic transmissions from parents to children, they provide unbiased estimates of average causal effects of genetic variants. We show that, because we only learn about variants' effects in children of heterozygous parents, the claim is not generally true for variants where children of heterozygotes and homozygotes experience different distributions of interacting environments. Thus, while family-based GWASs remain an important tool in teasing apart different sources of GWAS signals, their interpretation is less straightforward than has been assumed.

The standard genome-wide association study (GWAS) relies on a population sample to estimate the strength of associations between trait variation and loci across the genome. The approach does not only infer the direct genetic effects that are often of primary interest, however (i.e., the effects of alleles carried by a person on that person's trait value). Instead, estimates from population-based GWASs may also include indirect genetic effects of parents and other relatives, as well as absorb genetic and environmental confounding (1, 2). In the presence of these additional effects, population-based GWASs provide biased estimates of the direct genetic effects. The extent to which this bias is a concern depends on the particular application, but confounding is a clear impediment for studies aimed at identifying causal genetic mechanisms.

Because for many traits, there are a large number of GWAS associations or loci, each of which explains only a tiny proportion of variance, researchers often focus on aggregate properties of the GWAS loci, such as genetic correlations, or predict individual trait values by combining estimated effect sizes across loci into a "polygenic score" (PGS). The issues of confounding can be more pronounced for these aggregate measures, as systematic biases are compounded (3–7).

The possible contribution of factors other than direct genetic effects—in particular of environmental confounding—has motivated a turn toward family-based GWASs, which overcome the limitations of population-based GWASs by taking advantage of the randomness of Mendelian segregation from parent to child (2). By holding constant the differences in environments among families and randomizing alleles across the genetic backgrounds on other chromosomes, family-based GWASs provide estimates that are largely robust to the contribution of nondirect genetic effects (8–13). Given this property, many studies simply treat family GWASs as a tool to sidestep confounding in the estimation of genetic effects, with the implicit assumption that the estimand of a family GWAS is the same as that of an unconfounded population-based GWAS. Family studies present another nice feature, however: In their reliance on randomization, they resemble natural experiments or randomized controlled trials (RCTs) (14, 15), gold

Author contributions: C.V., M.P., and G.C. designed research; performed research; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: cveller@uchicago.edu, mp3284@columbia.edu, or gmcoop@ucdavis.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2401379121/-/DCSupplemental.

Published September 13, 2024.

standards of causal inference in the medical and social sciences. When viewed through that lens, family studies are instead seen as a way to estimate causal genetic effects.

In causal inference, the effect of interest is often that of a treatment, which can be defined in terms of a hypothetical manipulation, in which each person is moved from untreated to treated (16–19). In reality, it is not possible to observe both the treated and untreated outcomes for the same person. Instead, what an RCT provides is a comparison of treated and untreated individuals who, because of the randomization procedure, are assumed to be similar in all other regards (in expectation, or asymptotically). The mean difference in outcome between treated and untreated individuals provides an estimate of the causal effect, which has an interpretation in terms of the counterfactual thought experiment in which, other than the change in the treatment status, all else is held equal.

In many contexts, it cannot be assumed that the treatment effect is identical across individuals. In the presence of heterogeneous treatment effects, the question becomes whether the estimated effect has internal validity, i.e., whether it is an unbiased estimate of the mean effect for the entire sample—what in social sciences is known as an average treatment effect or average treatment effect (ATE)—or an unbiased estimate for a well-defined subsample; that is, a "local" ATE or LATE (20, 21). For instance, in an RCT, there may be "compliers" with the treatment and "noncompliers," and the estimate of the treatment effect is then a LATE for the compliers. A related question is the extent to which the estimate obtained from one sample generalizes to other samples from the population or to other populations, i.e., whether the estimate has external validity (22). Borrowing from this language, family-based GWASs are often presented as providing an ATE for causal (direct) genetic effects, either by envisaging the hypothetical swap of one or many causal alleles in the child or by invoking an analogy to RCTs, in which the treatment is the allele (or PGS value) inherited by the child and the outcome is the child's trait value (14, 15, 23–26).

Although it is common to present family-based GWASs as providing estimates of ATEs, there has been little or no discussion of the impact of heterogeneity in effect sizes on family-based GWAS estimates. Yet in genetics and quantitative genetics, the importance of heterogeneous genetic effects has long been recognized and, in contrast to many other contexts in which ATEs may be obtained, its possible sources are understood—notably, as arising from gene-by-environment interactions and epistasis. In plants and animals in which environments and/or genotypes can be controlled, such gene-by-environment (G×E) and gene-by-gene (G×G) effects are ubiquitous. Similarly, studies in a wide range of species have established norms of reaction, in which the phenotypic effect of a genotype depends on the context [Ch. 22 in ref. 27]. In human genetics, where similar manipulations are obviously infeasible, the tools to study these phenomena are much more indirect, relying entirely on statistical models. Using these approaches, the evidence for G×E and G×G is limited, but there are well-known examples for specific loci, and a number of lines of evidence for varying genetic effects across environmental settings (e.g., refs. 28–32). Even in the absence of G×E or G×G, apparent heterogeneous genetic effects at noncausal "tag" loci can arise from varying patterns of linkage disequilibrium (LD) across families. These observations suggest that genetic effects plausibly vary with the familial environment, raising the question of whether family-based GWAS estimates have internal and external validity when they do.

This question is all the more relevant because of the ways in which family data have been used to date. For most traits, sample sizes remain too small for family-based GWASs to be feasible. Instead, a PGS is built using effect-size estimates from a population-based GWAS. Then, in a second step, researchers use the data from families to test whether, and to what extent, the population-based PGS reflects direct genetic effects. Specifically, they test a null model that direct genetic effects do not contribute at all to the population PGS, by asking whether the PGS is predictive of trait differences within families—that is, whether a regression of the trait value on the PGS has a nonzero slope when controlling for parental PGSs. In addition, they often quantify the extent to which the direct genetic effects captured by the PGS contribute to variation in the trait by assuming that the slope of this regression provides an unbiased estimate of the direct causal effect of the PGS. The validity of these procedures relies on both internal validity of the estimates produced by the family-based analyses as well as their external validity for the sample in which the standard population-based GWAS was conducted.

Here, we interrogate these assumptions. To that end, we first define the causal effect on a trait of an allele at a single locus, in terms of counterfactual manipulations at the locus. Next, we examine under what conditions a family-based association study at a locus provides an unbiased estimate of the causal effect thus defined, either in the whole sample or a well-defined subset of the sample—i.e., we delimit when the estimates obtained can be considered ATEs or LATEs. We then consider aggregate effects of many loci, as combined in a PGS. Throughout, for comparison, we also calculate the analogous quantities for unconfounded population-based GWASs and PGS–trait associations.

Results

We assume that, for each individual in a population, the genetic contribution to some trait of interest is additive across causal loci, but that there is heterogeneity in the effects of alleles across individuals. An investigator is interested in the average causal effect of an allele or a PGS on the trait, in a sample of individuals taken from the population. They have the genotypes of the individuals in the sample as well as the genotypes of their parents. (Alternatively, for family-based studies, they might have the genotypes of full siblings.)

We consider two common uses of family-based GWAS: i) relying on the randomness of Mendelian segregation from parents to offspring in order to more cleanly estimate the average causal effect of an allele or PGS on the trait in the offspring, and ii) using the causal effect estimated in the sample to learn about the population from which the sample is drawn. The investigator may also wish to use the estimated effect to construct a genetic instrument, with the goal of identifying other causal relationships with the trait, as in Mendelian Randomization approaches (33).

To make things comparable, in discussing family GWASs, standard GWASs, or hypothetical manipulations, we assume that phenotypic outcomes are measured in the offspring generation.

The Effects of Alleles at a Single Locus. To generate intuition for the influence of G×E (and G×G) interactions on population and family-based GWAS estimates, we initially focus on a single biallelic locus, at which the two alleles have a direct causal effect on the trait of interest. We later consider the case where the alleles at the genotyped locus do not causally affect the trait, but instead tag alleles at a nearby causal locus. The two alleles are labeled A_1 and A_2, where A_2 is the "focal" allele. We assume that there is no genetic or environmental confounding, and no

indirect genetic effect via relatives or peers, making possible what we henceforth call an "unconfounded GWAS." To incorporate G×E interactions, we allow the effect sizes of the alleles at the locus to depend on the family environment. The phenotype of individual i in family f is

$$Y_i = Y^* + (\alpha + \alpha_f + \alpha_i)\, g_i + \epsilon_f + \epsilon_i, \qquad [1]$$

where g_i is the number of copies of the focal allele A_2 carried by individual i at the locus (0, 1, or 2), and ϵ_f and ϵ_i are family- and individual-specific environmental deviations in the trait's value, with mean zero. Y^* is an intercept value of the trait. α is the mean genetic effect of the focal allele. The family- and individual-specific deviations of the genetic effect due to G×E are α_f and α_i; we define their population means to be zero: $\mathbb{E}[\alpha_f] = \mathbb{E}[\alpha_i] = 0$.

Given this phenotypic model, we consider two questions. First, what is a sensible definition of the causal effect of allele A_2 on the phenotype? Second, how do the estimates obtained by a population-based or family-based association study compare to this definition of a causal effect?

Causal effects via manipulation. It is common to define causal effects in terms of counterfactual manipulations (17). To this end, we consider two thought experiments, one a randomization and one a more precise genetic manipulation.

We begin with a simple thought experiment in which we randomly reassign genotypes at the locus to individuals in the sample, independent of their environments or genetic backgrounds. This thought experiment resembles an RCT. Since, in expectation, there is zero covariance between g_i and α_f or α_i, the expected effect of A_2 is simply α. We note that some existing definitions of the causal effect of an allele in the presence of G×E in fact amount to this thought experiment (equation 15 in ref. 34; see *SI Appendix*, section 1.3).

In turn, to define a causal effect of A_2 via a counterfactual genetic manipulation, we choose a gamete that carries the A_1 allele randomly among the gametes produced by the parental generation, and we imagine flipping the allele in this gamete to A_2 (as if by CRISPR editing). If, among parents, p is the frequency of the focal allele A_2 and p_{11}, p_{12}, and p_{22} are the three genotype frequencies, then $1-p = p_{11} + p_{12}/2$ is the frequency of A_1 among parents, and so with probability $p_{11}/(1-p)$ the chosen A_1-bearing gamete derives from an A_1A_1 parent, while with probability $p_{12}/2(1-p)$ it derives from an A_1A_2 parent. The expected difference between the resulting offspring's phenotype if we were to flip the allele versus if we do not is therefore

$$\alpha^{\text{flip}}_{A_1\to A_2} = \alpha + \frac{p_{11}}{1-p}\mathbb{E}\left[\alpha_f \mid \text{parent } A_1A_1\right] + \frac{1}{2}\frac{p_{12}}{1-p}\mathbb{E}\left[\alpha_f \mid \text{parent } A_1A_2\right]. \qquad [2]$$

We can similarly define the causal effect of A_1 as the expected phenotypic difference caused by randomly selecting an A_2-bearing gamete and flipping the allele to A_1:

$$\alpha^{\text{flip}}_{A_2\to A_1} = -\alpha - \frac{p_{22}}{p}\mathbb{E}\left[\alpha_f \mid \text{parent } A_2A_2\right] - \frac{1}{2}\frac{p_{12}}{p}\mathbb{E}\left[\alpha_f \mid \text{parent } A_1A_2\right]. \qquad [3]$$

In general, the definitions Eqs. **2** and **3** need not be equal in magnitude, nor need they equal α. The reason is that each hypothetical allele flip occurs in the original environment of the genotype. For example, if we flip $A_1 \to A_2$ in a gamete, the resulting offspring gains an A_2 allele but still grows up in the environment in which the A_1 allele was originally found. If A_1 and A_2 alleles have different distributions of interacting environments, the two manipulations ($\alpha^{\text{flip}}_{A_1\to A_2}$ and $\alpha^{\text{flip}}_{A_2\to A_1}$) will differ in their effects and differ from α.

We can define a counterfactual manipulation that effectively randomizes alleles across environmental backgrounds and thus reconciles the randomization and allele-flipping definitions of the causal effect. To this end, we first note that the expected effect of the $A_1\to A_2$ flip defined above, $\alpha^{\text{flip}}_{A_1\to A_2}$, is the same, though opposite in sign, as that of flipping $A_2\to A_1$ in the environments experienced by A_1 alleles. Therefore, if we calculate the expected difference in an offspring's phenotype caused by choosing a gamete at random and flipping its allele, polarizing the difference by the allele that we flip, we obtain

$$\alpha^{\text{flip}} = (1-p)\alpha^{\text{flip}}_{A_1\to A_2} + p(-\alpha^{\text{flip}}_{A_2\to A_1}) = \alpha, \qquad [4]$$

(*SI Appendix*, section 1.1). Thus, we can define a sensible weighted average of the two allele-flip effects that returns α as the average causal effect at the locus. Note that α is an average effect for the population from which the sample is taken, and so is a property of the specific environments experienced by the population and their genetic backgrounds.

Effect-size estimates from association study designs. Next we consider whether various GWAS designs provide an unbiased estimate of the average causal effect. The genotype of each offspring can be written as the average of the maternal and paternal genotypes, $(g_m + g_p)/2$, plus a zero-mean term ς that accounts for the randomness of segregation in transmissions from the parents:

$$g = \frac{1}{2}(g_m + g_p) + \varsigma. \qquad [5]$$

If we perform a family-based association study by regressing the trait values of offspring on their genotypes at the locus, controlling for their parental genotypes, we obtain, in expectation, an effect-size estimate for allele A_2 of

$$\hat{\alpha}^{\text{fam}} = \alpha + \mathbb{E}\left[\alpha_f \mid \text{parent } A_1A_2\right], \qquad [6]$$

where the second term—the deviation of the family-based estimate $\hat{\alpha}^{\text{fam}}$ from α—is the average family deviation conditional on a parent being heterozygous at the focal locus (*SI Appendix*, section 1.2). This result is intuitive: Parent–offspring studies rely on contrasting the associations of transmitted and untransmitted alleles with the phenotype of the offspring (35). When parents are homozygous, no such comparison can be made, since their transmitted and untransmitted alleles are the same. The family-based estimate therefore makes use only of heterozygous parents. In other words, we can define the family-based estimate at a single locus as a LATE in the subpopulation of children who are the offspring of heterozygous parents; in the language of causal inference theory, these are the "compliers." If heterozygous parents are nonrandomly distributed across environments, however, the estimate obtained from a family-based GWAS may not be an unbiased estimate for the whole sample, which includes children of homozygous parents, and may not have external validity for the sample used in the standard population-based GWAS.

We note that the estimate in Eq. **6** is the same, in expectation, as would be obtained in a sibling study where the phenotypic

differences between full siblings are regressed on their genotypic differences at the locus, since both designs amount to regressing offspring phenotypes on their segregation deviations ς (again, assuming no indirect effects of siblings; *SI Appendix*, section 1.2; also see refs. 36 and 37).

If we instead perform a population-based association study by regressing the trait values of the offspring on their genotypes, without controlling for the parental genotypes, then, in an unconfounded GWAS, we obtain an expected effect-size estimate that can be written in the form

$$\hat{\alpha}^{\text{pop}} = \frac{\text{Cov}(Y,g)}{\text{Var}(g)} = \frac{\text{Cov}(Y,\frac{1}{2}(g_m+g_p))}{\text{Var}(g)} + \frac{\text{Cov}(Y,\varsigma)}{\text{Var}(g)} = \frac{\text{Cov}(Y,\frac{1}{2}(g_m+g_p))}{\text{Var}(g)} + \frac{1}{2}\cdot\frac{1-F}{1+F}\hat{\alpha}^{\text{fam}}, \quad [7]$$

where F is the inbreeding coefficient at the locus, which, along with the frequency p of A_2, we have assumed to be the same among offspring and their parents (*SI Appendix*, section 1.2). The first term on the right-hand side of Eq. 7 is the contribution of the genotypic variance between families to the population slope, while the second term is the contribution from the genotypic variance within families (due to random segregation). The first term (and, more generally, Eq. 7 written out explicitly) is a complicated weighted sum of the effects of A_2 in the environments experienced by the offspring of the three possible parental genotypes (*SI Appendix*, section 1.2). In general, $\hat{\alpha}^{\text{pop}}$ differs from the causal effect of the alleles at the locus, as defined in our hypothetical manipulation above (Eq. **4**), because the association study examines the effects of the alleles at the locus in the particular environments in which they are found. The degree of the difference between $\hat{\alpha}^{\text{pop}}$ and the causal effect will depend on the strength of the correlation between the genotypes at the locus and the environments that modify the effects of the alleles at the locus—see *SI Appendix*, Eq. **A.12**.

Summary. In the presence of heterogeneous allelic effects at a locus, the effect-size estimates produced by family- and (unconfounded) population-based association studies need not be the same on average, nor will they equal in expectation the causal effects of the alleles defined via our hypothetical experimental manipulations. The reason is that the quantities obtained in these different study designs and hypothetical manipulations are averages of allelic effects over different distributions of environments.

In Fig. 1, we illustrate with a simple example how G×E can lead population- and family-based association studies to produce different estimates when the study sample is drawn from two populations inhabiting different environments. The differences are more pronounced for SNPs with larger allele-frequency differences between the populations. Current family-GWASs focus on samples from relatively genetically homogenous sets of individuals, so the bias at a single locus will likely be small. However, as family studies come to be used to mitigate confounding in less genetically homogenous samples, greater caution in the interpretation of effect sizes will be warranted.

LD differences. In practice, a locus with a signal of association in a GWAS will often not affect the trait itself. Instead, it will be in linkage disequilibrium with—and thus "tag"—one or more nearby loci that causally affect the trait. If patterns of LD between the marker and causal loci differ across individuals, then, even in the absence of G×E at the causal loci, the effects estimated for the marker locus can differ between study designs in ways that resemble those of G×E at the marker locus.

Consider a model with two loci, a genotyped marker locus that does not causally affect the trait of interest and an ungenotyped causal locus that does. The alleles at the marker locus are m and M, and the alleles at the causal locus are a and A. Here, we assume that there is no G×E at the causal locus: The effect of A is to increase the trait's value by α in expectation, independent of the environmental setting.

First, we consider a definition of the counterfactual effect of allele M at the marker locus analogous to the counterfactual definition of the causal effect of the allele A at the causal locus, laid out above. To this end, we imagine swapping haplotypes of a given physical length around the genotyped marker: Specifically, randomly choosing a gamete that carries the m allele and flipping its haplotype to a random haplotype in the sample containing M. If the length of the flipped haplotype is too short to contain the causal locus, the effect of this flip is simply $\alpha^{\text{flip}}_{m\to M} = 0$, since the alleles at the marker locus do not affect the trait. However, if the flipped haplotype is long enough to include the causal locus, then the swap might change the allele at that locus too—either $a\to A$ or $A\to a$—and therefore affect the phenotype. The manipulation effect will then depend on the proportion of marker allele y-containing haplotypes that also contain causal allele x, $p_{x|y}$; the average effect of the flip is

$$\alpha^{\text{flip}} = (p_{A|M} - p_{A|m})\,\alpha, \quad [8]$$

as shown in *SI Appendix*, section 1.5. Since we are assuming that there is no G×E at the causal locus, the average phenotypic effect of the reverse $M\to m$ haplotype switch would be the same (though opposite in sign) as Eq. **8**, which is why we have omitted the $m\to M$ subscript in Eq. **8**. While Eq. **8** is defined in terms of a single causal locus, it naturally generalizes to the case where the marker locus tags multiple causal loci within the flipped region.

In the absence of confounding, a population-based association study at the marker locus returns an effect-size estimate that is equal to the allele-flipping effect in expectation:

$$\hat{\alpha}^{\text{pop}} = \alpha^{\text{flip}} = (p_{A|M} - p_{A|m})\,\alpha. \quad [9]$$

(Note that the quantities in Eqs. **8** and **9** can also be written in terms of coefficients of linkage disequilibrium—see *SI Appendix*, section 1.5.)

In contrast, a family-based association study at the marker locus returns an expected effect-size estimate of

$$\hat{\alpha}^{\text{fam}} = (1-2r)\left(p'_{A|M,Mm} - p'_{A|m,Mm}\right)\alpha \approx (p_{A|M,Mm} - p_{A|m,Mm})\,\alpha, \quad [10]$$

where r is the recombination fraction between the marker and causal loci, $p'_{A|M,Mm}$ and $p_{A|M,Mm}$ are the proportions of M containing haplotypes that also contain A in heterozygous Mm parents and offspring, respectively, and $p'_{A|m,Mm}$ and $p_{A|m,Mm}$ are analogously defined for allele m (*SI Appendix*, section 1.5). The approximation in Eq. **10** holds under the assumptions that the marker and causal loci are tightly linked ($r \approx 0$) and that genotype frequencies are similar in the parental and offspring generations.

In summary, when the marker and causal loci are tightly linked, the family-based study returns an unbiased estimate of a causal effect defined in an analogous way to Eq. **8**, namely flipping haplotypes only in gametes produced by parents heterozygous at the marker locus. Thus, in this context too, the family-based estimate can be interpreted as a LATE for these families.

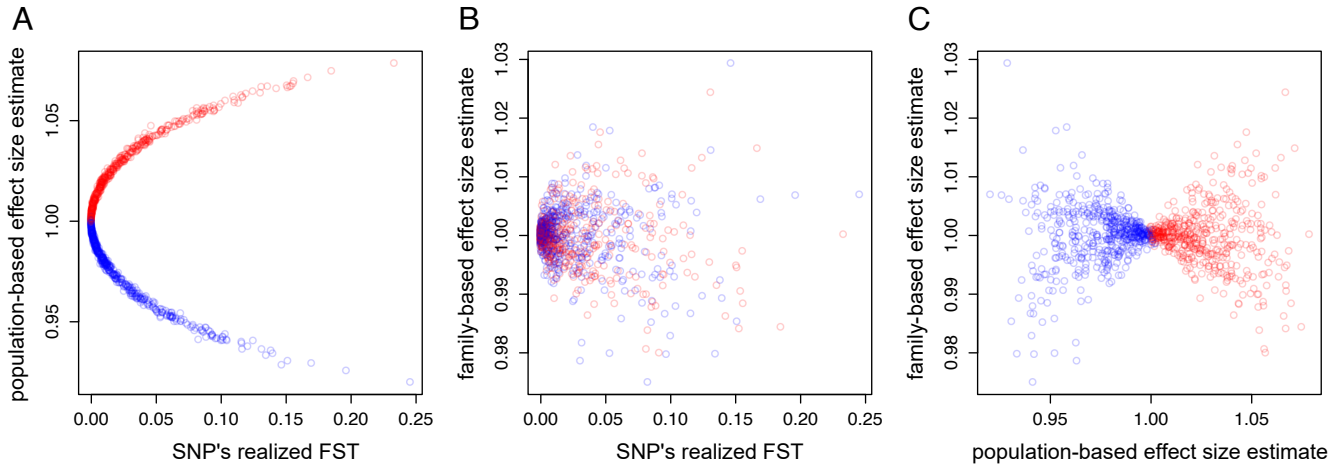

Fig. 1. Relationships between the effect-size estimates produced by population- and family-based association studies and F_{ST} in a simple model of population structure. There are two populations, red and blue, each inhabiting a distinct environment. At a biallelic locus, the focal allele has effect 1.1 on a trait of interest in the environment of the red population and effect 0.9 in the environment of the blue population. We simulated independent drift of allele frequencies at the locus in the two populations, and then conducted population- and family-based association studies of the trait in a large sample drawing equally from the two populations. (*A*) When drift results in some differentiation between the populations ($F_{ST} > 0$), the effect-size estimate from a population-based GWAS is systematically biased away from the true population-average effect of 1, with the direction of the bias reflecting whether the focal allele ended up at higher frequency in the red population (red points in *A*, *B*, and *C*) or the blue population (blue points in *A*, *B*, and *C*). (*B*) The estimate from a family-based GWAS depends instead on whether heterozygosity ends up being higher in the red population or the blue population; that is, on the population in which the focal allele at the locus moved closer to frequency 1/2. (*C*) Consequently, the family-based estimate differs from the population-based estimate in the presence of G×E. See *SI Appendix*, section 1.4 for mathematical details.

We can rewrite the family-based estimate as

$$\hat{\alpha}^{\text{fam}} \approx \alpha^{\text{flip}} + \left[\left(p_{A|M,Mm} - p_{A|M}\right) - \left(p_{A|m,Mm} - p_{A|m}\right)\right]\alpha. \quad [11]$$

Therefore, the family-based estimate at the marker locus will differ from the causal effect defined by the allele flip (and from the effect estimated in an unconfounded GWAS) if the A allele tends to co-occur on the same haplotype as the M allele more frequently in Mm heterozygotes than in the rest of the population. This phenomenon is mathematically analogous to the case of G×E considered above—compare Eqs. **6** and **11**—even though the sources of heterogeneity are distinct.

As a simple illustration of how LD differences between heterozygous and homozygous parents can influence the various association study designs, consider two populations that differ in both the degree of LD between, and the allele frequencies at, the marker and causal loci. In a sample drawn from across these two populations, the effect size estimated at the marker locus by a family-based association study will more strongly reflect the degree of LD in the population with greater heterozygosity at the marker locus, whereas an unconfounded population-based association study will instead reflect differences in the haplotype frequencies between the two populations (*SI Appendix*, section 1.6).

PGSs. To study the influence of G×E in applications of PGSs, we focus on the general case of the relationship between a PGS for trait X and the value of trait Y. As special cases, X and Y could be the same trait or X and Y could be the same trait but in a different set of environments.

An investigator has access to effect-size estimates $\hat{\beta}_l$ at a set of genotyped loci $l \in \Lambda$ from a prior GWAS on trait X. The sample in which this GWAS was carried out does not overlap with the sample used in the investigator's subsequent analysis. Given individual i's genotype g_{il}, $l \in \Lambda$, the investigator calculates a trait-X PGS for the individual:

$$PGS_i = \sum_{l\in\Lambda} \hat{\beta}_l g_{il}. \quad [12]$$

If individual i is in family f, their value for trait Y is given by

$$Y_i = Y^* + \sum_{l\in L} \left(\alpha_l + \alpha_{fl} + \alpha_{il}\right) g_{il} + \epsilon_f + \epsilon_i, \quad [13]$$

which is the multilocus version of Eq. **1**, with the set of loci L causally underlying variation in trait Y. We note that, in laying out this model, we have specified the relationship between the trait-X-associated loci and trait Y, rather than between traits X and Y directly. While the latter is usually the biological motivation for PGS studies, in practice, the researcher does not know whether the effects on trait Y of the loci in the trait-X GWAS are mediated through trait X itself (or due to horizontal pleiotropy or confounding; see, e.g., refs. 6, 13, and 38–41).

We are interested in the causal relationship between the PGS for trait X and the value of trait Y, both in terms of a hypothetical manipulation and in terms of estimates produced by standard least-squares regression approaches. We consider population-based designs that regress Y on PGS and family-based designs that regress Y on PGS controlling for the maternal and paternal PGSs.

Because a PGS is constructed using estimated rather than true effect sizes, it reflects the signal of causal effects partially (and noisily) captured by the genotyped loci on which the PGS is based. Additionally, the effect-size estimates may be biased due to confounding in the GWAS. The noise and bias in the original GWAS used to construct the PGS weights are implicit in our $\hat{\beta}$s (see ref. 13 for a fuller treatment). For simplicity, we assume that the loci included in the PGS are a subset of the loci that causally affect the trait of interest ($\Lambda \subseteq L$) and that the causal loci are in Hardy–Weinberg and linkage equilibrium in the sample. As we show, even under these simplifying assumptions, interpreting the effects of PGSs in the presence of G×E is not straightforward. Moreover, our results extend naturally to the case where the PGS

loci are not themselves causal but instead tag nearby causal loci (a point that we return to below).

Interpreting PGSs in terms of hypothetical manipulations. As in the single locus case, we can again consider two thought experiments to understand causal effects in the PGS case, the first involving randomization and the second a genetic manipulation. First, we imagine randomly assigning genotypes at the loci in Λ to our sampled individuals, independent of their environments or genetic backgrounds. Under this randomization, least-squares regression of trait Y on the trait-X PGS returns, in expectation, a coefficient

$$\delta^{\text{rand}} = \frac{\sum_{l\in\Lambda} H_l \hat{\beta}_l \alpha_l}{\sum_{l\in\Lambda} H_l \hat{\beta}_l^2}, \quad [14]$$

where $H_l = 2p_l(1-p_l)$ is the heterozygosity at locus l under the assumption of Hardy–Weinberg equilibrium. Under our assumption that the loci in Λ are causal for trait X, the numerator in Eq. **14** can be interpreted as the genic covariance between the trait-X PGS and trait Y, while the denominator is the genic variance of the trait-X PGS.

It is more complicated to define the causal effect of the PGS in terms of hypothetical genetic manipulations, in which we imagine experimentally flipping alleles at the PGS loci: With many loci contributing to the PGS, a desired change in the value of the trait-X PGS can be achieved via many different manipulations, and these different manipulations will, in general, not have the same average effect on trait Y. Nonetheless, in *SI Appendix*, section 2, we describe two manipulation strategies that respect the sample variance structure of the PGS and return an effect of the PGS, δ^{flip}, that is the same as that given by the randomization thought experiment described above, δ^{rand}.

When the PGS alleles do not themselves causally affect the trait, but instead tag ungenotyped causal alleles, as in the single marker locus case above, we can imagine the investigator flipping haplotypes of a given length around each PGS locus. Under this definition, the α_l terms in Eq. **14** would be replaced by expressions for the average effects of the haplotype swaps (i.e., Eq. **8**, assuming that the flipped haplotype is long enough at each marker locus to include the causal locus).

Spelling out these manipulations makes it clear that the estimand (Eq. **14**) is in fact a complicated object. In so far as the manipulations are ad hoc and may be a somewhat contrived basis for a causal interpretation, a more straightforward perspective may be to view the individual SNP effects as causal, and to think of the slope of a PGS as a function of these causal effects. The question then becomes whether it is possible to estimate this function in an unbiased way.

Estimates of the effect of the PGS. Family-based designs for estimating the effect of the PGS rely on the regression

$$Y = \mu + \delta PGS + \gamma(PGS_m + PGS_p) + \epsilon, \quad [15]$$

where PGS is the PGS of the offspring and PGS_m and PGS_p are the maternal and paternal PGSs respectively. The estimate $\hat{\delta}^{\text{fam}}$ produced by ordinary least squares (OLS) estimation of this regression has been interpreted as an estimate of the direct effect of the PGS on the phenotype, since the parental PGSs effectively control for population structure and for indirect effects of genotypes of relatives on the phenotype of the offspring.

Since we can write

$$PGS = \left(PGS_m + PGS_p\right)/2 + \varsigma, \quad [16]$$

with ς the segregation deviation of the offspring's PGS from the midparent value, the estimate $\hat{\delta}^{\text{fam}}$ produced by OLS estimation of Eq. **15** is the same, in expectation, as that produced by OLS estimation of the regression

$$Y = \tilde{\mu} + \delta\varsigma + \tilde{\epsilon}, \quad [17]$$

by the Frisch–Waugh–Lovell theorem (pp. 35 and 36 in ref. 42; *SI Appendix*, section 3.1).

In the presence of G×E interactions, and under the simplifying assumptions laid out above concerning the sets of PGS and causal loci Λ and L, the family-based estimate is, in expectation,

$$\hat{\delta}^{\text{fam}} = \frac{\text{Cov}(Y,\varsigma)}{\text{Var}(\varsigma)} = \frac{\sum_{l\in\Lambda} H_l \hat{\beta}_l \left(\alpha_l + \mathbb{E}\left[\alpha_l^f \mid h_l\right]\right)}{\sum_{l\in\Lambda} H_l \hat{\beta}_l^2}, \quad [18]$$

where h_l denotes that a parent is heterozygous at locus l (*SI Appendix*, section 3.2).

As this expression makes clear, the estimate from a family-based GWAS will differ systematically from the randomization or allele-flipping effect of the PGS if parents heterozygous at any loci in the PGS are nonrandomly distributed across environments ($\mathbb{E}\left[\alpha_l^f \mid h_l\right] \neq 0$).

Without information about parental genotypes, a population-based study of the effect of the PGS effect is based on the regression

$$Y = \mu + \delta PGS + \epsilon. \quad [19]$$

In that setting, from Eq. **16**, OLS estimation of this regression produces

$$\hat{\delta}^{\text{pop}} = \frac{\text{Cov}(Y, (PGS_m + PGS_p)/2 + \varsigma)}{\text{Var}(PGS)} \quad [20]$$

$$= \frac{V_{PGS} - V_\varsigma}{V_{PGS}} \cdot \hat{\delta}^{\text{between-fam}} + \frac{V_\varsigma}{V_{PGS}} \cdot \hat{\delta}^{\text{fam}} \quad [21]$$

in expectation, where V_{PGS} is the population variance of PGS values, V_ς is the contribution of within-family segregation to the PGS variance, and $\hat{\delta}^{\text{between-fam}} = \text{Cov}(Y, (PGS_m + PGS_p)/2)/(V_{PGS} - V_\varsigma)$ is the "between-family" slope one would obtain by regressing offspring phenotypes on midparent PGSs. The between-family slope absorbs direct genetic effects as well as environmental and genetic confounding.

Even in the absence of genetic or environmental confounding, the estimated effect of the PGS is a mixture of between- and within-family effects. Thus, while the family-based estimate $\hat{\delta}^{\text{fam}}$ can be seen as a component of the population-level estimate $\hat{\delta}^{\text{pop}}$, $\hat{\delta}^{\text{pop}}$ additionally reflects the G×E effect in the children of parents homozygous at PGS loci.

Summary. In the presence of G×E, the family-based estimate of the causal effect of the PGS, $\hat{\delta}^{\text{fam}}$, can be biased away from the causal effect defined by genotype–environment randomization or by experimental manipulation of genotypes if, for any of the loci in the PGS, heterozygotes and homozygotes are distributed differently across environments.

At any given locus, there exists a set of individuals—the offspring of heterozygotes—for whom a family-based GWAS provides an unbiased estimate of an average causal effect, i.e., a LATE (cf. Eq. **6**). However, this set of individuals will be different for each SNP, and so, in general, $\hat{\delta}^{\text{fam}}$ does not have a causal interpretation either as an ATE or a LATE. Instead,

the family-based estimate $\hat{\delta}^{\text{fam}}$ is a strangely weighted average across loci and across families (Eq. **18**), and the fact that it is a weighted sum of per-SNP LATEs, each of which has a separate causal interpretation, does not in itself endow $\hat{\delta}^{\text{fam}}$ with a causal interpretation. In general, in the presence of an interaction between genetic effects and familial environments, and nonrandom distribution of family genotypes across interacting environments, family-based estimates $\hat{\delta}^{\text{fam}}$ do not have internal validity.

These conclusions extend to the case where the PGS loci are not causal themselves, but instead tag nearby causal loci. In this case, the within-family estimate of the PGS effect will reflect the level of LD in heterozygotes for each marker (Eq. **11**, assuming no G×E at causal loci). If the patterns of LD differ between heterozygous and homozygous parents, the within-family estimate is not, in general, an unbiased estimate of the average effect of the PGS defined in terms of haplotype manipulation, nor can it be interpreted as a LATE for any subsample.

We note further that, while the within-family PGS slope (Eq. **18**) provides an estimate of the slope in an unconfounded population-based PGS analysis (Eq. **21**), it does not in general provide an unbiased estimate of this slope: in the presence of interactions, the two are not in general equal in expectation.

Under what conditions is the family-based regression coefficient on the PGS an internally valid estimate of an average causal effect? In this section and the next, we address two more specific questions about family-based PGS studies in the presence of heterogeneous effects: 1) Does there exist a narrower set of assumptions in which the estimates obtained from family-based PGS studies are unbiased estimates of ATEs or LATEs? 2) What proportion of the population variance in the trait can be attributed to the slope of the PGS estimated from family-based studies?

The within-family regression slope of the PGS, $\hat{\delta}^{\text{fam}}$, has been described as an estimate of the direct causal effect of the PGS. As our results show, when genetic effects interact with the familial environment or when LD patterns differ between parents that are homozygous versus heterozygous for PGS loci, this claim does not hold. The within-family regression slope of the PGS has also been described more specifically as "a weighted average over the direct effects of the [PGS] for the individuals in the population," with this average taken "over any heterogeneity of effects across individuals that may exist" (SI Appendix, section 7.1 in ref. 43). While, by the same token, this claim cannot be generally true, here we consider under what restricted conditions it is valid.

There are two special conditions that jointly allow us to recover an interpretation of the family-based regression coefficient $\hat{\delta}^{\text{fam}}$ as a weighted average of direct PGS causal effects. The first is that the G×E effects α_l^f are, for each family f, a fixed multiple C_f of the population-average effect α_l across all causal loci, such that a family's environment either amplifies ($C_f > 0$) or dampens ($C_f < 0$) the genetic contribution to the trait by a constant factor across loci. The second is that the effect-size estimates $\hat{\beta}_l$ used to construct the PGS differ in expectation from the population-average causal effects β_l only by a multiplicative factor B that is constant across loci (allowing also for uncorrelated noise).

If these two conditions both hold, each family is characterized by a well-defined PGS slope $\delta_f = (1 + C_f)/B$, such that any genetic manipulation that increases the PGS of an offspring in family f by 1 unit—no matter which loci are flipped and in what direction—increases the offspring's trait value by δ_f in expectation. The regression coefficient $\hat{\delta}^{\text{fam}}$ can then be interpreted as an estimate of an average of these family-specific slopes δ_f.

Even under these strict conditions, however, this average does not weight each family equally, and therefore does not return the average PGS slope in the sample. Instead, each family is weighted proportional to its segregation variance for the PGS, with families in which parents are heterozygous at more of the PGS loci upweighted in the calculation of $\hat{\delta}^{\text{fam}}$ (*SI Appendix*, section 3.3). The reason is that these families contribute more within-family variation in the PGS, and it is within-family PGS variation upon which family-based estimation of δ depends. Thus, while the estimate of δ produced by the within-family PGS regression is not an ATE across the sample of genotyped families, it is, under the specific conditions laid out above, a LATE in an "effective sample" (44) of families weighted according to the parents' PGS segregation variance. This is a specific example of a more general phenomenon, known in statistics and econometrics, where, in estimation of a coefficient of interest, controlling for additional covariates (in our case, the parental PGSs) causes OLS to upweight some observations and downweight others (e.g., refs. 44, 45, and pp. 76 and 79 in ref. 42; see *SI Appendix*, section 3.3).

While these parallels to regression results from other fields are enlightening, in practice, the assumptions of a fixed multiplicative G×E effect across loci and a fixed multiplicative bias in the effect-size estimates used to construct the PGS are unlikely to be met. When considering the PGS as a "treatment," we rely on a sum of small estimated effects. Because the causal loci tagged by the PGS loci will often be pleiotropic in their effects, the property of a single G×E multiplier C_f—although a useful heuristic—is unlikely to hold for the trait of interest in the family-based study. Moreover, it is worth noting that the estimated weights used to build the PGS are not just noisy; for some traits at least, they are biased to varying degrees across loci because of confounding.

What is the contribution of within-family PGS variation to phenotypic variation? The coefficient estimated in a population-based GWAS for the effect of a PGS on a trait can often reflect substantial environmental and genetic confounding, whereas the coefficient estimated in a family-based GWAS removes all environmental and most genetic confounding (2, 13). Motivated by these considerations, studies often compare the within-family and population-based PGS slopes, treating their ratio, $\hat{\delta}_{\text{fam}}/\hat{\delta}_{\text{pop}}$, or its square, $\hat{\delta}^2_{\text{fam}}/\hat{\delta}^2_{\text{pop}}$, as an estimate of the proportion of PGS association or variance that is due to direct causal effects of (or tagged by) the PGS loci (26, 35, 43).

As can be seen in Eq. **21**, the population-based estimate of the PGS slope is composed of the within-family slope $\hat{\delta}_{\text{fam}}$, weighted by the proportion of PGS variance that comes from within families (V_ς/V_{PGS}), and a between-family slope $\hat{\delta}_{\text{between-fam}}$, weighted by the proportion of PGS variance that comes from between families ($[V_{PGS} - V_\varsigma]/V_{PGS}$). The between-family slope can be thought of as a combination of the true direct effects between families, $\delta_{\text{between-fam}}$, and an extra term due to confounding. Treating the fraction $\hat{\delta}_{\text{fam}}/\hat{\delta}_{\text{pop}}$ as the fraction of the population-level association that is due to direct effects implicitly assumes that $\delta_{\text{between-fam}} = \hat{\delta}_{\text{fam}}$ in expectation. As we have shown, this interpretation is not valid in general. Therefore, taking ratios of slopes obtained in family-based versus standard GWASs may yield biased estimates of the proportional contribution of direct genetic effects.

The estimate $\hat{\delta}_{\text{fam}}$ is estimated from—and therefore only strictly applies to—the within-family PGS variance V_ς. Attributing the explanatory value of $\hat{\delta}_{\text{fam}}$ to the PGS variance among individuals V_{PGS} requires an extrapolation from within to between families. As we have shown, the within-family slope

generally lacks internal validity, and so, in the presence of heterogeneity, this extrapolation can lead to a biased estimate of the variance explained by direct effects. Similar biases may also affect other methods that use genetic segregation from parents in order to partition the population phenotypic variation into genetic and nongenetic components (46, 47).

The population variance of the PGS can be decomposed into contributions from within and between families:

$$V_{PGS} = \mathrm{Var}(PGS) = \mathrm{Var}\left(\frac{PGS_m + PGS_p}{2} + \varsigma\right) = \mathrm{Var}\left(\frac{PGS_m + PGS_p}{2}\right) + V_\varsigma, \quad [22]$$

noting that the segregation deviation ς is uncorrelated with $(PGS_m + PGS_p)/2$. If we focus on the within-family variance V_ς, and consider the proportion of the overall phenotypic variance explained by fitted values $\varsigma\hat{\delta}_{\mathrm{fam}}$ based on it, we obtain

$$\frac{\hat{\delta}^2_{\mathrm{fam}} V_\varsigma}{V_\mathrm{P}}. \quad [23]$$

This approach is coherent and the fraction is interpretable as the contribution to phenotypic variance arising from Mendelian segregation within families at the PGS loci. The interpretation is valid because families are weighted the same in the calculation of $\hat{\delta}_{\mathrm{fam}}$ as they are in V_ς; the interpretation of Eq. **23** therefore does not require an extrapolation from families with greater segregation variance (more heterozygous parents) to families with lower segregation variance (less heterozygous parents). The quantity in Eq. **23** can also be measured via a hypothetical experimental manipulation, as laid out in *SI Appendix*, section 3.4.

These results establish a solid basis for causal statements about the proportion of phenotypic variance explained by within-family PGS variance. In the absence of assortative mating or other sources of long-range signed linkage disequilibrium, the within-family variance of the PGS is expected to be half the sample-wide variance ($V_\varsigma = V_{PGS}/2$)—less if alleles with the same directional effect on the trait are in positive LD, as is the case with assortative mating, and more if they are in negative LD, as arises under various kinds of selection. Therefore, our results indicate that only half (and potentially less) of the phenotypic variance attributed to the PGS based on within-family estimates of its effect can potentially be estimated in an unbiased manner, whereas the rest relies on an extrapolation based on the assumption of no heterogeneity of genetic effects across families.

A caveat is that we have not considered the multiple covariates often included in family studies beyond parental genotypes, which will further complicate the interpretation of causal effects (48). Clearly, further work is required to define precisely what is being estimated in practice in family-based GWASs.

Discussion

While the random inheritance of parental alleles by children offers a natural experiment (49), the genetic variation that this experiment generates—based on which one hopes to establish "treatment" effects of alleles on phenotypes of interest—is contributed only by heterozygous parents. In other words, we do not observe alternative allelic treatments among the offspring of homozygotes and therefore cannot learn about treatment effects on their phenotypes. In contrast, the effect of an allele across families in a population reflects inheritance from both homozygous and heterozygous parents. The consequence is that if there is heterogeneity in effects between the children of homozygous and heterozygous parents, family studies will generally result in a biased estimate of the average effect of an allele in a population. In the case of the effect size estimated by a family GWAS for a single locus, the estimate can nonetheless be viewed as a LATE for the children of heterozygotes, and thus has internal validity for a well-defined subset of families. The same does not hold for a PGS, however, which has no general interpretation as a LATE at the population level in the presence of effect-size heterogeneity.

That so many fields of study focus on reliably estimating ATEs reflects the fact that heterogeneity in causal effects is ubiquitous. Genetics is no exception, with many well-studied sources of heterogeneity, including G×G and G×E interactions, and for marker loci, varying patterns of LD. For the specific issues with family studies raised here to be a problem, however, requires two conditions to be met: i) Genetic effects must be heterogeneous across environments and ii) genotypes must be nonrandomly distributed across the relevant environments. Given extensive evidence for heterogeneity in the effects of alleles due to G×E (and G×G) in settings where experimental manipulations are possible, the first condition seems highly plausible. In turn, the evidence of environmental and genetic confounding in human GWASs for an increasing number of traits shows that SNP genotypes are nonrandomly distributed across environmental and genetic backgrounds. Thus, it seems important to consider what happens to estimates produced by family-based GWASs in this setting.

In doing so, we have focused primarily on G×E as the source of heterogeneity in genetic effects. As noted above, however, the effect of an allele could also systematically differ across families (α_f) because it is involved in epistatic interactions with alleles at other loci in the genome (i.e., because of G×G). By analogy to our G×E model above, epistatic interactions would lead to biases in family-based GWASs if parents who are heterozygous at the focal study locus tend to have systematically different genotypes at loci that interact epistatically with the focal locus, relative to the population distribution of such genetic backgrounds. We have also ignored parent–offspring interactions in family-based studies. Following the same logic, interactions between alleles of parents and offspring will result in family GWAS estimates that are the average effect of the focal allele in an offspring conditional on the genetic background of a heterozygous parent. Thus, a nonrandom distribution of genetic backgrounds in heterozygous parents is also a potential source of bias.

In addition to interactions with the environment or other loci, differences between homozygotes and heterozygotes in the degree of LD between marker loci and the causal loci that they tag can also give rise to variation in effect sizes at the marker loci. As we have shown (Eq. **11**), these differences can also lead to biases in family-based estimates of average effect sizes. One plausible way that such differences in LD can arise is when GWASs draw samples from across ancestry gradients; for example, from across the north–south gradient in haplotype diversity within Europeans (50). Evidence that changes in LD patterns contribute to differences in the prediction accuracy of PGSs over even short genetic distances (51–53) supports the view that sampling families from across ancestry gradients may be an important source of heterogeneity.

To provide an interpretation of the effect size being estimated at a tag—rather than causal—locus, we proposed a thought experiment in which a haplotype of a given length is experimentally swapped. Under the assumption of tight linkage between the causal and marker locus and no LD between causal loci, the effect

estimated in family studies corresponds to the average effect of this manipulation in the offspring of heterozygotes at the marker locus. In the presence of LD between causal loci, the family-based effect-size estimate at the marker locus absorbs the effects of any causal loci on its chromosome that are in LD with it, complicating even the single-locus interpretation (13). More generally, we have not considered how LD among causal loci will complicate causal interpretation of PGSs. As our results show, even in the simpler case of uncorrelated causal loci, articulating a causal interpretation of family-based studies is not straightforward.

Further complications will arise if there is selection bias on the participants in the population-based or family-based GWAS (54), as the effects estimated may then be biased relative to those of a representative sample. An additional issue specific to family studies is that the key assumption of allele randomization by Mendelian segregation will be violated when the selection of families in the study occurs based on a heritable phenotype of the offspring.

Ultimately, how large the biases that we have uncovered are in practice is an empirical question. For single loci, effect-size estimates obtained in family-GWASs carried out in relatively homogenous populations may be largely unaffected. It is less clear, however, how these biases compound in analyses of PGSs. Moreover, the growing recognition of the utility of family studies in removing confounding will make them ever more essential as GWAS and PGS analyses are applied to samples drawn from genetically and environmentally heterogeneous groups. Yet it is precisely such samples in which the biases that we have highlighted could become substantially more important.

Conclusion

To date, one use of family studies has been to provide rigorous evidence that the PGS contains some signal of a genetic causal effect. While this use is valid, one might argue that, given that it is well established that almost any trait is partially heritable and that genetic correlations are widespread, it should not be surprising that there are nonzero genetic effects on a trait (55, 56). The focus on family studies providing ATEs reflects the desire for richer quantitative answers. Our results indicate that, in a general setting, a number of claims about what can be learned from family-based GWASs need to be refined.

Our findings further establish that family studies can be interpreted as providing an unbiased estimate of the contribution that the segregation of PGS alleles within families makes to the overall phenotypic variation (Eq. ***23***). Since, under standard quantitative genetics models, the within-family variance is a sizable contribution to the population variation, even this more limited use of family-based GWASs may yield important insights. Moreover, in practice, the biases that arise from family studies due to various sources of heterogeneity are likely smaller than the effects of confounding on population-based GWAS estimates. In that sense, estimates from family studies are more interpretable as causal effects than those from population-wide studies, and discrepancies between family- and population-based GWASs still offer a useful heuristic for identifying and dissecting how confounding affects PGSs.

Data, Materials, and Software Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. We thank David Blei, Cyrus Samii, and Matthew Stephens for helpful discussions, and Vince Buffalo, Doc Edge, Arbel Harpak, Cyrus Samii, Matthew Stephens, and Eli Weinstein for comments on an earlier draft of the manuscript. Funding was provided by the NIH (NIH R35 GM136290 awarded to G.C. and R01 HG011432 coawarded to M.P.) and a Branco Weiss fellowship to C.V.

Author affiliations: [a]Department of Ecology & Evolution, University of Chicago, Chicago, IL 60637; [b]Department of Biological Sciences, Columbia University, New York, NY 10027; [c]Department of Systems Biology, Columbia University, New York, NY 10032; and [d]Center for Population Biology and Department of Evolution and Ecology, University of California, Davis, CA 95616

1. B. J. Vilhjálmsson, M. Nordborg, The nature of confounding in genome-wide association studies. *Nat. Rev. Genet.* **14**, 1–2 (2013).
2. A. I. Young, S. Benonisdottir, M. Przeworski, A. Kong, Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**, 1396–1400 (2019).
3. J. J. Berg *et al*., Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725 (2019).
4. M. Sohail *et al*., Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* **8**, e39702 (2019).
5. H. Mostafavi *et al*., Variable prediction accuracy of polygenic scores within an ancestry group. *eLife* **9**, e48376 (2020).
6. R. Border *et al*., Cross-trait assortative mating is widespread and inflates genetic correlation estimates. *Science* **378**, 754–761 (2022a).
7. R. Border *et al*., Assortative mating biases marker-based heritability estimators. *Nat. Commun.* **13**, 660 (2022b).
8. R. S. Spielman, R. E. McGinnis, W. J. Ewens, Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516 (1993).
9. D. B. Allison, Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* **60**, 676–690 (1997).
10. G. R. Abecasis, L. R. Cardon, W. O. C. Cookson, A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**, 279–292 (2000).
11. L. J. Howe *et al*., Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nat. Genet.* **54**, 581–592 (2022).
12. A. I. Young *et al*., Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nat. Genet.* **54**, 897–905 (2022).
13. C. Veller, G. M. Coop, Interpreting population-and family-based genome-wide association studies in the presence of confounding. *PLoS Biol.* **22**, e3002511 (2024).
14. G. Davey Smith, S. Ebrahim, 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
15. J. W. Madole, K. P. Harden, Building causal knowledge in behavior genetics. *Behav. Brain Sci.* **46**, e182 (2023a).
16. J. Neyman, On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat. Sci.* **5**, 465–472 (1990). Translated and edited by D. M. Dabrowska and T. P., Speed from the Polish original, which appeared in Roczniki Nauk Rolniczych Tom X. *Ann. Agric. Sci.* 1–51 (1923).
17. D. B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688 (1974).
18. P. W. Holland, Statistics and causal inference. *J. Am. Stat. Assoc.* **81**, 945–960 (1986).
19. D. B. Rubin, Causal inference using potential outcomes: Design, modeling, decisions. *J. Am. Stat. Assoc.* **100**, 322–331 (2005).
20. G. W. Imbens, J. D. Angrist, Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–475 (1994).
21. G. W. Imbens, Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *J. Econ. Lit.* **48**, 399–423 (2010).
22. M. G. Findley, K. Kikuta, M. Denly, External validity. *Annu. Rev. Polit. Sci.* **24**, 365–393 (2021).
23. T. T. Morris, N. M. Davies, G. Hemani, G. Davey Smith, Population phenomena inflate genetic associations of complex social traits. *Sci. Adv.* **6**, eaay0328 (2020).
24. J. Becker *et al*., Resource profile and user guide of the Polygenic Index Repository. *Nat. Hum. Behav.* **5**, 1744–1758 (2021).
25. J. W. Madole, K. P. Harden, Causal complexity in human research: On the shared challenges of behavior genetics, medical genetics, and environmentally oriented social science. *Behav. Brain Sci.* **46**, e206 (2023b).
26. M. N. Meyer *et al*., Wrestling with social and behavioral genomics: risks, potential benefits, and ethical responsibility. *Hastings Cent. Rep.* **53**, S2–S49 (2023).
27. M. Lynch, B. Walsh, *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, Massachusetts, 1998).
28. S. H. Barcellos, L. S. Carvalho, P. Turley, Education can reduce health differences related to genetic risk of obesity. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E9765–E9772 (2018).
29. A. I. Young, F. L. Wauthier, P. Donnelly, Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nat. Genet.* **50**, 1608–1614 (2018b).
30. C. Zhu *et al*., Amplification is the primary mode of gene-by-sex interaction in complex human traits. *Cell Genomics* **3**, 100297 (2023).
31. K. Hou *et al*., Calibrated prediction intervals for polygenic scores across diverse contexts. *Nat. Genet.* **56**, 1386–1396 (2024).
32. A. Durvasula, A. Price, Distinct explanations underlie gene-environment interactions in the UK Biobank. medRxiv [Preprint] (2023). https://doi.org/10.1101/2023.09.22.23295969 (Accessed 21 January 2024).
33. B. Brumpton *et al*., Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nat. Commun.* **11**, 3519 (2020).

34. J. J. Lee, C. C. Chow, The causal meaning of Fisher's average effect. *Genet. Res.* **95**, 89–109 (2013).
35. A. Kong *et al*., The nature of nurture: Effects of parental genotypes. *Science* **359**, 424–428 (2018).
36. S. Trejo, B. W. Domingue, Genetic nature or genetic nurture? Introducing social genetic parameters to quantify bias in polygenic score analyses *Biodemography Soc. Biol.* **64**, 187–215 (2018).
37. J. Fletcher, Y. Wu, T. Li, Q. Lu, Interpreting polygenic score effects in sibling analysis. *PLoS One* **19**, e0282212 (2024).
38. N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, J. W. Smoller, Pleiotropy in complex traits: Challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).
39. J. K. Pickrell, Fulfilling the promise of Mendelian randomization. bioRxiv [Preprint] (2015). https://doi.org/10.1101/018150 (Accessed 21 January 2024).
40. M. Verbanck, C.-Y. Chen, B. Neale, R. Do, Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
41. J. Morrison, N. Knoblauch, J. H. Marcus, M. Stephens, X. He, Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nat. Genet.* **52**, 740–747 (2020).
42. J. D. Angrist, J.-S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press, Princeton, NJ, 2009).
43. A. Okbay *et al*., Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat. Genet.* **54**, 437–449 (2022).
44. P. M. Aronow, C. Samii, Does regression produce representative estimates of causal effects? *Am. J. Polit. Sci.* **60**, 250–267 (2016).
45. J. D. Angrist, A. B. Krueger, "Empirical strategies in labor economics" in *Handbook of Labor Economics, volume 3, chapter 23*, O. C. Ashenfelter, D. Card, Eds. (Elsevier, Amsterdam, 1999), pp. 1277–1366.
46. P. M. Visscher *et al*., Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.* **2**, e41 (2006).
47. A. I. Young *et al*., Relatedness disequilibrium regression estimates heritability without environmental bias. *Nat. Genet.* **50**, 1304–1310 (2018a).
48. G. W. Imbens, J. M. Wooldridge, Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* **47**, 5–86 (2009).
49. R. A. Fisher, Statistical methods in genetics. *Heredity* **6**, 1–12 (1952).
50. A. Auton *et al*., Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res.* **19**, 795–803 (2009).
51. Y. Wang *et al*., Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).
52. F. Privé *et al*., Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* **109**, 12–23 (2022).
53. Y. Ding *et al*., Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* **618**, 774–781 (2023).
54. S. Benonisdottir, A. Kong, Studying the genetics of participation using footprints left on the ascertained genotypes. *Nat. Genet.* **55**, 1413–1420 (2023).
55. N. H. Barton, P. D. Keightley, Understanding quantitative genetic variation. *Nat. Rev. Genet.* **3**, 11–21 (2002).
56. E. Turkheimer, Three laws of behavior genetics and what they mean. *Curr. Dir. Psychol. Sci.* **9**, 160–164 (2000).