Measuring missing heritability: Inferring the contribution of common variants

David Golan[a,1], Eric S. Lander[b,c,d,2], and Saharon Rosset[a,2]

[a]Department of Statistics and Operations Research, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel 69978; [b]Broad Institute of MIT and Harvard, Cambridge, MA 02142; [c]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and [d]Department of Systems Biology, Harvard Medical School, Boston, MA 02155

Contributed by Eric S. Lander, October 10, 2014 (sent for review June 15, 2014)

Genome-wide association studies (GWASs), also called common variant association studies (CVASs), have uncovered thousands of genetic variants associated with hundreds of diseases. However, the variants that reach statistical significance typically explain only a small fraction of the heritability. One explanation for the "missing heritability" is that there are many additional disease-associated common variants whose effects are too small to detect with current sample sizes. It therefore is useful to have methods to quantify the heritability due to common variation, without having to identify all causal variants. Recent studies applied restricted maximum likelihood (REML) estimation to case–control studies for diseases. Here, we show that REML considerably underestimates the fraction of heritability due to common variation in this setting. The degree of underestimation increases with the rarity of disease, the heritability of the disease, and the size of the sample. Instead, we develop a general framework for heritability estimation, called phenotype correlation–genotype correlation (PCGC) regression, which generalizes the well-known Haseman–Elston regression method. We show that PCGC regression yields unbiased estimates. Applying PCGC regression to six diseases, we estimate the proportion of the phenotypic variance due to common variants to range from 25% to 56% and the proportion of heritability due to common variants from 41% to 68% (mean 60%). These results suggest that common variants may explain at least half the heritability for many diseases. PCGC regression also is readily applicable to other settings, including analyzing extreme-phenotype studies and adjusting for covariates such as sex, age, and population structure.

genome-wide association studies | statistical genetics | heritability estimation

Comprehensive genomic studies have begun to uncover the genetic basis of common polygenic inherited diseases and traits, identifying thousands of loci and pointing to key biological pathways. However, the genetic variants implicated so far account for less than half the estimated heritability of most diseases and traits (1). Explaining the remainder of the heritability—often termed "missing heritability"—is of considerable biological interest and medical importance. This is our third article on exploring the mystery of missing heritability.

In our first paper (2), we noted that some of the apparently missing heritability may arise from a methodological issue. Specifically, we showed that the presence of genetic interactions among loci might substantially inflate estimates of the total (narrow-sense) heritability and thus overstate the extent of missing heritability. However, this likely is only a partial explanation.

In our second paper (3), we explored the design of association studies to discover genetic variants associated with the risk of a disease or trait. Specifically, our paper focused on rare variant association studies (RVASs), for which large-scale comprehensive efforts are just becoming feasible with advances in sequencing technology. The approach involves sequencing every human gene in a large case–control study to see whether the aggregate frequency (or "burden") of a set of rare variants differs between cases and controls. (Rare variants may be defined operationally as having frequency $\leq$0.5%.) We studied how the power of RVASs depends on various factors, such as the selection coefficient against null alleles, the type of rare variants to be aggregated (based, for example, on allele frequency and mutational type), and the population studied. We concluded that RVASs with adequate power to detect genetic effects of interest should involve at least 25,000 cases.

In this third paper, we turn our focus to common variant association studies (CVASs). (Such studies typically are referred to simply as genome-wide association studies, or GWASs, but we prefer the term CVAS to highlight the complementarity with RVAS.) CVAS involves testing millions of common genetic variants for correlation with disease in case–control studies. CVAS has the advantages that one can enumerate the complete set of common variants in a population; each variant is frequent enough to be tested individually; and variants may provide information about a nearby region (as the result of linkage disequilibrium). Whereas RVAS only now is becoming feasible, CVAS became practical with the advent of inexpensive large-scale genotyping arrays roughly a decade ago. CVASs have been performed for hundreds of diseases, involving a total of approximately 2 million samples. The fruits of these studies include the discovery of hundreds of loci for inflammatory bowel disease, schizophrenia, early heart disease, and type 2 diabetes (4).

Significance

Studies have identified thousands of common genetic variants associated with hundreds of diseases. Yet, these common variants typically account for a minority of the heritability, a problem known as "missing heritability." Geneticists recently proposed indirect methods for estimating the total heritability attributable to common variants, including those whose effects are too small to allow identification in current studies. Here, we show that these methods seriously underestimate the true heritability when applied to case–control studies of disease. We describe a method that provides unbiased estimates. Applying it to six diseases, we estimate that common variants explain an average of 60% of the heritability for these diseases. The framework also may be applied to case–control studies, extreme-phenotype studies, and other settings.

Whereas early association studies in the 1990s used loose thresholds for statistical significance (e.g., $P \leq 0.05$) and were notoriously irreproducible, CVAS imposes an extremely stringent threshold for statistical significance (on the order of

Author contributions: D.G., E.S.L., and S.R. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]Present address: Department of Genetics, Stanford University, Stanford, CA 94305.

[2]To whom correspondence may be addressed. Email: lander@broadinstitute.org or saharon@post.tau.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1419064111/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1419064111

$P \leq 5\times10^{-8}$) to reduce the number of false discoveries, which otherwise would be inflated when testing many hypotheses of which the vast majority are false (corresponding to non-associated SNPs). As a result, the discoveries have proven to be highly replicable. Although CVAS has been very effective at reliably identifying disease-associated loci, the genetic variants detected tend to have modest effects on disease risk and thus may be challenging to study biologically or clinically.

An open question has been how much of the heritability of a trait is attributable to common variants. A straightforward approach for inferring the heritability due to common variants is to add up the estimated heritability contributed by each of the genetic variants that have achieved clear genome-wide statistical significance. This calculation typically yields a relatively low proportion—e.g., 5–10% for height (5–8), which has estimated heritability of 80%. The obvious problem with the approach is that it provides only a lower bound that likely is a substantial underestimate because it ignores the many loci that have not yet reached genome-wide significance. Thus, there has been considerable interest in ways to estimate the total heritability attributable to common variants in an indirect manner that does not require definitively identifying the loci.

Visccher and coworkers [Yang et al. (9)] made major contributions to this program, focusing on the situation of quantitative phenotypes studied in a random population sample. The fundamental idea is to estimate the heritability due to common variants by studying the extent to which the phenotypic similarity across pairs of individuals in a sample is explained by their genotypic similarity at common variants. Rather than using simple correlation, they used a family of elegant statistical models, called linear mixed models (LMMs), and estimated the heritability using a technique called restricted maximum likelihood (REML) estimation. By applying REML to a study of height, they estimated that the common variants examined explained ∼50% of the heritability, which was considerably greater than the 5–10% obtained from summing the contributions of variants that have achieved statistical significance. Following this pioneering work, the REML approach has been applied to many other CVASs of quantitative phenotypes, yielding significant increases in estimated heritability explained by common variants.

Subsequently, the same group [Lee et al. (2011) (10)] sought to extend this approach to disease traits (or, more generally, any binary phenotypes) in case–control studies. Modifying the REML method for this setting, they applied it to three diseases studied by the Wellcome Trust Case Control Consortium (WTCCC): type 1 diabetes, Crohn's disease, and bipolar disorder. The estimated heritability explained by common variants was considerably higher than that obtained from summing the contributions of individual loci. Their approach has since been applied to numerous other disease phenotypes, with similar results (see, e.g., refs. 11–13).

Here, we reexamine the REML approach for disease traits in case–control studies. We identify a flaw in the underlying assumptions that creates a serious bias in the REML estimate for disease traits. Consequently, we show that the REML approach of Lee et al. (10) underestimates the heritability explained by common variants. The magnitude of the bias is affected by many factors, most notably the study size, the prevalence of the disease in the population, the proportion of cases in the study, the true underlying heritability, and the number of genotyped SNPs [some of these factors were noted independently by others (14, 15)]. For example, the simulation studies below show that for a disease with 0.1% prevalence in which common variants actually explain 50% of the heritability, the REML approach applied to a balanced case–control study of 4,000 individuals will yield, on average, an estimate as low as 30–35% (depending on the number of SNPs used). Importantly, this bias increases with study size, suggesting that the inaccuracy of REML estimates will increase as larger and larger CVASs are conducted.

Instead, we propose a general framework for heritability estimation, which we term phenotype correlation–genetic correlation (PCGC) regression, which produces unbiased estimates for case–control studies of disease traits. The approach generalizes traditional regression-based approaches used in genetics (16, 17). We show that PCGC regression yields substantially higher estimates of the heritability explained by common variants for several diseases, including Crohn's disease, bipolar disorder, type 1 diabetes, early-onset myocardial infarction (MI), schizophrenia, and multiple sclerosis (MS). The PCGC framework also is suitable for other settings involving phenotype-guided sampling, including selection of extreme phenotypes for quantitative traits.

Below, we begin by outlining a general model for a quantitative trait, and review methods for heritability estimation in this situation. We then turn to case–control studies of disease traits and describe current efforts for estimating heritability. We discuss the challenges induced by case–control sampling and use simulations to demonstrate that current methods result in serious downward-biased heritability estimates. To overcome this problem, we introduce an alternative approach, PCGC regression for heritability estimation, and demonstrate that it provides unbiased estimates and improved accuracy. We then use PCGC regression to estimate the heritability due to common SNPs in several case–control studies. Our results show that the fraction of heritability explained by common SNPs is larger than previously thought. We conclude by describing several extensions of our PCGC framework to other CVAS scenarios, such as accounting correctly for additional covariates and analyzing extreme-phenotype studies.

Results

General Model of Quantitative Traits. In the general case, a quantitative phenotype p depends on genotype g and environment e, according to a function ψ. For the ith individual, we have $p_i = \psi(g_i, e_i)$. Here, the genotype $g_i = (g_{i1}, g_{i2}, \ldots g_{in})$ is the diploid genotype at every variant site across the genome, where g_{ik} is the number of copies of a designated allele at the kth variant site, and f_j is the frequency of the designated allele. (We will assume that the variant sites are in linkage equilibrium and all alleles are biallelic, although these assumptions may be relaxed.)

This definition is completely general. It may be used with whole-genome sequence data, with the variant sites corresponding to every nucleotide in the human genome. Moreover, the function ψ allows for arbitrary gene–gene (GxG) and gene–environment (GxE) interactions.

For convenience, we will work below with "normalized" phenotypes p_i and genotypes g_{ik}, where the quantities p_i and g_{ik} have been centered to have mean 0 and standardized to have variance 1.

Additive Model for Quantitative Traits. Analyses of heritability typically assume a much simpler additive model. We do so here, assuming that our quantitative trait follows a simple additive model with no GxG or GxE interactions. We also assume there is no correlation in the environmental effects among individuals. We write

$$\begin{aligned} p_i &= g_i + e_i \\ g_i &= \sum_k u_k g_{ik}, \end{aligned} \qquad [1]$$

where u_k denotes the normalized effect of the kth variant. This model is illustrated in Fig. 1*A*.

Defining Heritability. Heritability quantifies how much of the variability of p is the result of variability in g. There are two types of heritability: broad-sense heritability H^2, which measures the full contribution of genes, and narrow-sense heritability h^2, which is meant to capture the "additive" contribution of genes (see

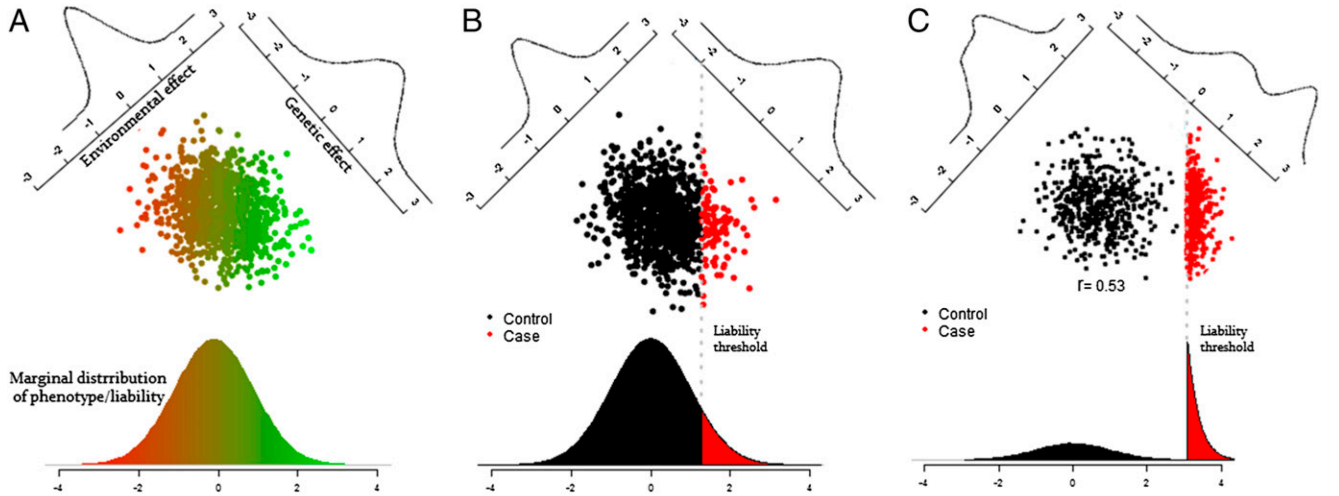


Fig. 1. Distributions of genetic effects, environmental effects, phenotypes, and liabilities in three study designs. In each of *A*, *B*, and *C*, a phenotype is assumed to depend on the sum of a genetic effect and an environmental effect. The scatterplot shows the joint distribution of the genetic and environmental effects, the upper left shows the marginal distributions of the environmental effect, the upper right shows the marginal distributions of the genetic effect, and the lower portion shows the marginal distribution of the phenotype. (*A*) Quantitative phenotype in a random sample of the population. (*B*) Disease phenotype in a random sample of the population. (*C*) Disease trait in a balanced case–control study. Disease phenotypes were simulated under a liability threshold model with disease prevalence of 10% (*B*) and 0.1% (*C*), with red points indicating affected individuals (liability above the threshold) and black points indicating unaffected individuals (liability below the threshold). In *C*, the marginal distributions of the genetic and environmental effects no longer are normally distributed, and there is an induced positive correlation between the genetic and environmental effects ($r = 0.53$).

ref. 2 for definitions). For our additive model, these definitions coincide, and we have

$$H^2 = h^2 = \sum_k u_k^2. \quad [2]$$

Estimating Heritability from Significantly Associated Variants. Suppose a certain subset S of variants has been shown definitively to be associated with our quantitative trait, based on a large CVAS with a stringent threshold for statistical significance. The heritability explained by these known loci may be estimated as

$$\hat{h}_S^2 = \sum_{k \in S} \hat{u}_k^2. \quad [3]$$

As noted above, h_S^2 provides a lower bound for the heritability. However, h_S^2 may greatly underestimate the actual heritability explained by common SNPs if the sum in Eq. **3** fails to include many additional variants associated with the trait because they have not reached statistical significance in the sample—for example, because of low effect sizes or lower minor allele frequencies. It is clear that most current disease studies have many false negatives, inasmuch as the number of loci identified has been continuing to grow with sample size.

Estimating the Aggregate Impact of All Variants. The challenge thus is to estimate the heritability attributable to all common variants associated with a disease, without actually identifying these variants. The basic idea follows from the classical notion, articulated by Galton (18) in the 19th century, that the heritability of a trait is captured by the extent to which genetic similarity between individuals predicts their phenotypic similarity. The challenge is to convert this notion into a mathematical procedure for estimating heritability.

Yang et al. (9) made the key observation that because we are interested in the value Σu_k^2 rather than each of the individual effect sizes u_k, the effect sizes may be regarded as "nuisance parameters." They adopted a "random effects" model in which the u_k are treated as identical and independently distributed random variables drawn from a distribution with mean 0 and variance σ_u^2. (Higher values of σ_u^2 imply larger effects, resulting in higher heritability, and vice versa.)

Under our additive model above with random effects, we have the following elegant relationship:

$$corr(p_i, p_j) = \mathbb{E}(p_i p_j) = h^2 G_{ij}, \quad [4]$$

where G_{ij} denotes the genetic correlation between individuals i and j, given by

$$G_{ij} = corr(g_i, g_j) = \frac{1}{m} \sum_{k=1}^{m} g_{ik} g_{jk}. \quad [5]$$

Eq. **4** provides an intuitive method for estimating heritability—regressing the empirical phenotypic correlations $(p_i p_j)$ onto the genetic correlations G_{ij}. The estimated slope of this regression is an unbiased estimator of the heritability. This procedure is known in the genetics literature as Haseman–Elston regression (16, 17). We refer to the general idea of regressing the phenotypic correlations onto the genetic correlations and using the slope for heritability estimation as PCGC and note that Haseman–Elston regression is a special case of PCGC regression.

To apply Eq. **4**, we need to know the genetic correlation G_{ij} between people. However, a problem shared by all heritability estimation methods is that these correlations are unknown and need to be estimated from the data. A classical approach used in human genetics is to use the expected kinship coefficient for individuals in a pedigree (for example, 50% for siblings, 3.125% for second cousins, or 0 for unrelated individuals). However, the genetic correlation between related individuals may vary considerably around the kinship coefficient, and unrelated individuals may have substantial cryptic sharing, resulting in nonzero correlations. For example, the genetic correlation between siblings varies around the expected value of 50% with an SD of 3.6% (19). Hence, more accurate estimates of G_{ij} may be obtained by directly examining partial genotype or complete sequence information. We return to the problem of estimating G_{ij} below.

Improving Heritability Estimates with REML. Although the PCGC regression approach for estimating heritability is easy to understand and implement, and produces unbiased estimators, it does not, in fact, make full use of all available information. In a sense, PCGC regression looks only at pairs of individuals at a time, whereas there is additional information in looking at trios of individuals simultaneously, or even at the entire cohort simultaneously. More precisely, the PCGC regression estimator is a moments-based estimator. In statistics, maximum likelihood estimators generally are preferred because they can extract more information and provide more precise estimates.

To use maximum-likelihood estimation, one is required to put forward an explicit probabilistic model of the data (by contrast, the PCGC regression approach requires only independence assumptions, namely that the genetic and environmental effects are independent, and that the environmental effects of different individuals are independent). The common additional assumptions are that the genetic component g and the environmental component e of the phenotype both follow a normal (Gaussian). In this case, the phenotype p is distributed normally, and the joint distribution of the phenotype vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ is given by

$$\boldsymbol{p} \sim N\left(0, Gh^2 + I\left(1-h^2\right)\right), \qquad [6]$$

where I denotes the identity matrix, and G is the genetic correlation matrix whose off-diagonal entries are the pairwise correlations G_{ij} and its diagonal is 1. This model is a special case of a more general statistical approach known as random effects models, and the problem of estimating heritability is a special case of the problem of estimating variance components.

Yang et al. (9) use this framework for the estimation of heritability, which allows them to apply well-established methods such as REML to obtain estimates of h^2. The resulting REML estimates are better than those based on the regression approach (i.e., they require fewer observations to reach the same accuracy of the estimate). For example, Yang et al. (9) use REML to estimate the narrow-sense heritability of height explained by common SNPs at 53.7 ± 10.0%. By contrast, the PCGC regression estimate (which in this case, reduces to standard Haseman–Elston regression) using the same data are 51.0 ± 13.5% (9).

Model for Disease Traits. From the description above, estimating the heritability due to common variants may be considered largely solved for the case of a quantitative trait. However, the primary focus of medical genetics is disease traits—which are binary (affected vs. unaffected) rather than quantitative. Disease traits pose further challenges.

Disease phenotypes traditionally have been modeled by a liability threshold model (illustrated in Fig. 1*B*; see ref. 2 for details). The model assumes the existence of a quantitative trait, called the "liability score" and denoted l. As above, we have $l = g + e$, where the genetic component g and the environmental component e both are normally distributed and uncorrelated with each other. Individuals are affected if and only if their liability score exceeds a threshold t. The value of t determines the prevalence of the disease in the population, so the liability threshold model can accommodate diseases of any frequency by adjusting the threshold parameter accordingly.

Heritability may be calculated based on either the unobserved liability scale (denoted h_l^2) or the observed binary disease phenotype (denoted h_o^2). Geneticists typically are interested in knowing h_l^2, but because the liability score is unobserved, they must estimate it indirectly based on h_o^2. Dempster and Lerner (20) discovered a surprisingly simple relationship between these two heritabilities:

$$h_l^2 = \frac{K(1-K)}{\varphi(t)^2} h_o^2, \qquad [7]$$

where K is the prevalence of the disease in the population, t is the threshold, and φ is the standard Gaussian density.

Adapting REML to Case–Control Studies of Disease Traits. Lee et al. (10) sought to use this framework to adapt the REML method to disease traits. In the case of a random sample from the population, they offer a simple recipe: (*i*) code the disease phenotype as a 0/1 variable, (*ii*) use the REML procedure for a quantitative trait to calculate the heritability of the 0/1 on this observed scale, and (*iii*) convert the resulting estimate to the liability scale as in Eq. **7**. There is an important caveat: although the liability is a continuous quantitative trait, the 0/1 variable itself is not and therefore does not actually fit the likelihood function assumed in the REML method (Eq. **6**). Although this approach yields unbiased estimates (*SI Appendix*, section 5.4.2), the resulting estimates no longer are maximum-likelihood estimates; thus, some of the favorable properties of maximum likelihood estimates no longer are guaranteed to hold.

Real disease studies, however, rarely involve a random sample from the population: the number of affected individuals captured in the sample would be too small. Instead, geneticists use case–control studies, in which cases are considerably oversampled relative to their prevalence in the population. Because of this oversampling of cases, several assumptions of the probabilistic model of REML are violated: (*i*) the marginal distributions of the genetic and environmental effects, as well as of the liability, no longer are normal; (*ii*) the multivariate distribution of these effects no longer is multivariate normal; and (*iii*) the genetic and environmental effects no longer are independent. We illustrate two of these problems below.

Problem 1: Nonnormality of the Liability. Whereas the liability score follows a normal distribution in the case of a random population sample, it does not when case–control sampling is used: the oversampling of cases inflates the right tail of the liability score distribution, resulting in a nonnormal distribution of the liability score in the study (Fig. 1*C*).

Lee et al. (10) acknowledge the nonnormality induced by case–control sampling and propose the following analog to Eq. **7** to account for this issue in transforming REML estimates from the observed to the liability scale:

$$\hat{h}_l^2 = \frac{K^2(1-K)^2}{P(1-P)\varphi(t)^2} \hat{h}_o^2, \qquad [8]$$

where K and P are the prevalence of the disease in the population and the study respectively, and $\hat{h}_o^2$ refers to the REML estimate of heritability, when the phenotype is coded as 0/1, and treated as continuous. The same correction was derived by others in a Bayesian framework (21). Intuitively, the term $\frac{K^2(1-K)^2}{P(1-P)\varphi(t)^2}$ is a generalization of Eq. **7**, accounting for the nonnormality of the liability caused by the oversampling of cases. In a random sampling scheme, we have $K = P$, and Eq. **8** boils down to Eq. **7** as expected.

Problem 2: Case–Control Sampling Causes "Induced" GxE Interactions. Although Lee et al. (10) consider the issue of the nonnormality of the liability score induced by the case–control sampling, their analysis misses a subtle but important problem. It turns out that case–control sampling creates an induced positive correlation between the genetic and environmental effects for the samples in the study, as can be seen in Fig. 1*C*. Although there is no GxE interaction in the population, there is an obvious interaction between g and e under case–control sampling.

REML Underestimates the Heritability Explained. Just as it was demonstrated in ref. 2 that the presence of GxG interactions leads to underestimation of the fraction of heritability explained, we suspected that the presence of such induced GxE interactions might result in underestimation of the heritability. To test this idea, we simulated the entire generative process of a case–control study of a disease.

In each simulation run, we generated data for millions of individuals using a liability threshold model corresponding to the desired population prevalence K. For each individual, we generated genotypes at 10,000 independent loci, with minor allele frequencies between 0.05 and 0.5. The effect of each SNP was drawn from a normal distribution, as in Yang et al. (9). The liability was computed according to the polygenic model of Eq. **1**, with the variance of the genetic effect set to achieve the desired heritability. Each individual then was classified as affected or unaffected according to the appropriate threshold. Finally, all affected individuals were sampled for the study as cases, whereas unaffected individuals were chosen with a probability set to achieve the desired proportion of cases P in expectation. This process was repeated until 4,000 individuals (with the desired proportion of cases) were collected for the study. We note that the choice of simulating SNPs at linkage equilibrium was motivated by a theoretical result from Patterson et al. (22), which shows that the resulting distribution of correlation matrices is equivalent to the distribution obtained from a larger number of SNPs in linkage disequilibrium.

Our simulations confirm our expectation: heritability estimates obtained by applying REML and correcting using Eq. **8** are strongly downward biased (Fig. 2*A*). The magnitude of the bias increases when (*i*) the disease is rarer, (*ii*) the proportion of cases is closer to half, and (*iii*) the heritability is higher. Indeed, these circumstances each increase the induced GxE interaction.

To illustrate the magnitude of the bias, consider a situation representative of a balanced case–control study of a disease in which true underlying heritability due to common SNPs is 50%. If the prevalence is 0.1% (comparable to the frequencies of MS or Crohn's disease), then the REML method yields an estimated heritability of only 29.4% on average. The bias decreases for more common diseases. For prevalences of 0.5%, 1%, 5%, and 10%, the expected heritability estimate in a balanced case–control study is 34.9%, 37.9%, 43.7%, and 48.1%, respectively.

In addition to the factors described above, the bias of REML estimates depends on other factors—most importantly, it increases with the study size. To illustrate the effect of study size on the bias of REML estimates, we simulated case–control studies with a varying number of individuals (between 2,000 and 8,000) but kept all other parameters constant (we simulated 10,000 SNPs in linkage equilibrium, prevalence of disease in the population was set to 1%, average proportion of cases in the study was 30%, and the heritability was set to 50%). The average REML estimates of heritability decreased with study size, from 43.6% (SE: 0.7%) with 2,000 individuals to 35.3% (SE: 0.2%) with 8,000 individuals (Fig. 2*B*).

We note that Lee et al. (10) tested their method using simulations, which appeared to confirm that the REML method provides unbiased estimates. However, these simulations did not explicitly simulate genotypes. Instead, they proceeded as follows: they (*i*) generated individuals in batches of 100; (*ii*) assigned genetic correlations to all pairs of individuals, by assuming that individuals in the same batch have correlation 0.05 and individuals in separate batches have correlation 0; and (*iii*) simulated phenotypes, by generating liabilities according to Eq. **6** and comparing them to the threshold t. All affected individuals were retained for the simulation, together with an equal number of randomly selected unaffected individuals. The problem with this simulation scheme is that most batches contain few cases—indeed,

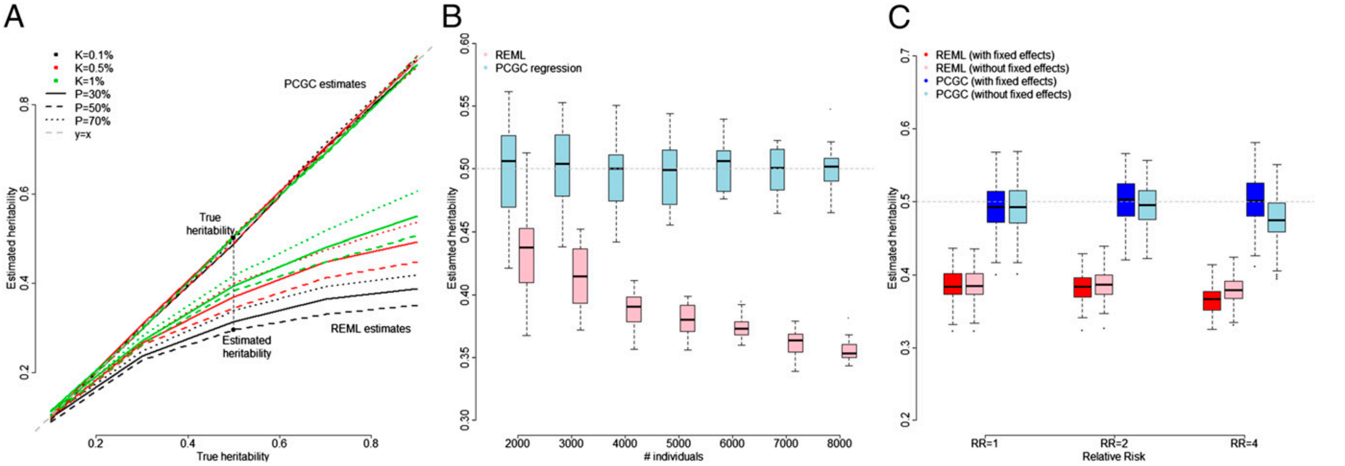


Fig. 2. Comparison of REML and PCGC regression. (*A*) REML yields biased estimates for case–control studies of diseases, whereas PCGC regression yields unbiased estimates. We simulated case–control studies for nine combinations of K (prevalence) and P (proportion of cases among overall samples), and for five values of h^2 (0.1, 0.3, 0.5, 0.7, and 0.9). For each combination of parameters, we show the average of 10 heritability estimates obtained by applying the REML method of Lee et al. (10) and PCGC regression to our simulated case–control data. REML produced biased estimates, whereas PCGC regression produced unbiased estimates for all scenarios. The bias of REML estimates increases as both the true heritability and overrepresentation of cases increase. To demonstrate the severity of the bias, consider the scenario of a disease with prevalence of 0.1% in a balanced case–control study (values typical for Crohn's disease or MS). When the true heritability is 50%, the estimated heritability would be 30% on average, as indicated by the black dots. (*B*) Heritability estimates for case–control studies with increasing sample size. Simulated case–control studies are as previously described, with the prevalence of the disease, the proportion of cases, and the heritability fixed at 1%, 30%, and 50%, respectively. The size of simulated studies ranged from 2,000 to 8,000. The bias of heritability estimates from REML increases with study size, whereas those from PCGC regression estimates remain unbiased. (*C*) Heritability estimation in the presence of fixed effects. We simulated case–control studies with an additional "sex" covariate, which either has no effect on the disease or increases the relative risk (RR) by twofold or fourfold. The prevalence of the disease in the population was 0.5%, the heritability was set to 50%, and the numbers of cases and controls were equal. Applying REML with or without accounting for the additional covariate resulted in underestimation of the heritability. Moreover, inclusion of the covariate as a fixed effect resulted in even lower estimates of heritability when the effect of the covariate on the phenotype was considerable. By contrast, PCGC regression correctly accounted for the presence of the covariate.

often zero or one for a disease with low frequencies. Because there were few cases within each batch, the genetic correlation between most case pairs, most control pairs, and most case–control pairs was exactly 0 (with a small number of nonzero correlations set at 0.05). In some simulation scenarios, up to 99.98% of the correlations were set to 0. In short, the simulation did not resemble actual genetic correlations. Our simulations above overcome this problem by explicitly simulating genotypes (*SI Appendix*, section 5).

PCGC Regression as a General Method for Estimating Heritability. Given the serious bias in REML estimates for case–control studies, we sought to develop an alternative approach based on PCGC regression. Because PCGC regression estimates are moment estimators, they (in contrast to maximum-likelihood estimators) do not require assuming an entire probabilistic model to obtain unbiased estimates; this is a useful feature in situations in which the actual probabilistic setup is complex.

PCGC regression is based on the simple idea that the heritability of a trait controls the strength of the relationship between genotype and phenotype. In the general case, the relationship among genetic correlation (G_{ij}), phenotypic correlation ($p_i p_j$), and the heritability due to common variants (h^2) may be expressed as

$$\mathbb{E}(p_i p_j) = f(h^2, G_{ij}), \qquad [9]$$

where the function f depends on (i) the design of the study and (ii) the properties of the phenotype. Given the function f, we can estimate h^2 by searching for the value that provides the best fit when across all pairs of individuals—for example, by minimizing the sum of squares between the actual and predicted phenotypic similarities:

$$\hat{h}^2 = \min_{0 \le h^2 \le 1} \sum_{i \ne j} \left[p_i p_j - f(h^2, G_{ij}) \right]^2. \qquad [10]$$

The simplest situation is one in which (i) the individuals comprise a random sample from the population and (ii) the phenotype is an additive polygenic quantitative trait, in which case we have $f(h^2, G_{ij}) = h^2 G_{ij}$. The value of h^2 may be estimated by linear regression; the estimate is the slope obtained by regressing the values of $p_i p_j$ onto the values of G_{ij}. This practice is known as Haseman–Elston regression.

PCGC regression may be extended to other study designs, although explicit expressions for f are harder to obtain. However, when focusing on studies of largely unrelated individuals, the values of G_{ij} are mostly small. Accordingly, f can be approximated by a Taylor series at $G_{ij} = 0$:

$$f(h^2, G_{ij}) = \frac{df}{dG_{ij}}(h^2, 0) G_{ij} + O\left(G_{ij}^2\right). \qquad [11]$$

Provided that g and e are normally distributed, we show (*SI Appendix*, section 1.3) that $\frac{df}{dG_{ij}}(h^2, 0)$ is linear in h^2, and thus

$$f(h^2, G_{ij}) = c h^2 G_{ij} + o\left(G_{ij}^2\right) \qquad [12]$$

for some constant c that depends on the properties of the phenotype and the study, but not on the heritability.

In this situation, we can use linear regression to estimate h^2. The only question is how to calculate the constant c. We provide step-by-step calculations for determining c under several relevant study designs (*SI Appendix*, sections 1–4).

As an example, consider the situation of case–control studies, as above. We show (*SI Appendix*, section 1.3) that the value of c is given by

$$c = \frac{P(1-P)\varphi(t)^2}{K^2(1-K)^2}, \qquad [13]$$

which is the reciprocal of the coefficient in Eq. **6**. [Notably, the REML correction is derived by Lee et al. (10) to correct for the problem of nonnormality (problem 1 above), whereas the use of regression in combination with this c addresses both the marginal nonnormality and the induced GxE correlations (problem 2) simultaneously.] As a test, we applied PCGC regression to the simulated case–control data. The results (Fig. 2*A*) confirmed that the approach indeed yields estimates that are unbiased and considerably more accurate than those achieved by the method of Lee et al. (10). We tested PCGC regression across many scenarios: simulating a wide range of disease prevalence (*SI Appendix*, Figs. S5 and S6), populations with cryptic relatedness (*SI Appendix*, Fig. S7); different numbers of SNPs (*SI Appendix*, Fig. S8), increasing study sizes (Fig. 2*B*), and alternative polygenic architectures (*SI Appendix*, Figs. S11 and S12); and estimating heritability in the presence of additional covariates (see below for more details, or see *SI Appendix*, section 5.6). In all scenarios, PCGC regression yielded unbiased estimates of heritability.

Estimating the Heritability Due to Common SNPs. So far, we have discussed PCGC regression as a general method for estimating heritability. Its input is a matrix G of genetic correlations and a vector p of phenotypes, and its output is an estimate of the heritability (the same generally is true for any correlation-based heritability estimation method). It is important to note that the interpretation of the resulting estimate depends heavily on the actual G used. Yang et al. (9) pioneered the approach of estimating G from genotyped common SNPs, and thus the result is an estimate of the heritability explained (or tagged) by common SNPs. If, for example, G were to be estimated using SNPs from only one chromosome, the result would be an estimate of the heritability explained by common SNPs on that chromosome (23).

The commonly used estimate of the genetic correlation is the empirical correlation, computed over the set A of variants genotyped or sequenced:

$$\hat{G}_{ij} = \frac{1}{|A|} \sum_{k \in A} g_{ik} g_{jk}. \qquad [14]$$

Because G typically is estimated by using only genotyped SNPs, the resulting estimate is interpreted as the heritability due to genotyped SNPs (sometimes referred to as "chip heritability," i.e., the heritability explained by the SNPs on the genotyping chip). Hence, it is expected to underestimate the heritability explained by all common SNPs, because the genotyped SNPs are in imperfect linkage disequilibrium (LD) with the ungenotyped common SNPs. Yang et al. (9) suggest a method for quantifying and correcting this underestimation. We adopt their correction when appropriate, because our focus is on methods for estimating heritability and not on estimating correlation matrices (*SI Appendix*, section 7). We address alternative approaches in *Discussion*.

Applying PCGC Regression to Case–Control Studies of Disease. We applied PCGC regression to six case-control studies of disease: the WTCCC studies of Crohn's disease, bipolar disorder, and type 1 diabetes investigated by Lee et al. (10); studies of MS and schizophrenia to which the same REML methodology was applied (11, 24, 25); and a study of early-onset MI (26). Where necessary, we applied stringent quality control, as suggested by Lee et al. (10), to mitigate batch effects (*SI Appendix*, section 10). We included sex as a covariate (*SI Appendix*, section 2) and removed the top 10 principal components of the correlation matrix to control for population structure (*SI Appendix*, section 8). To address the problem of imperfect LD between causal SNPs and genotyped SNPs (which, as noted above, results in underestimation of heritability), we applied the correction proposed by Yang et al. (9) (*SI Appendix*, section 7); we note that this correction also was applied by Lee et al. (10) for their REML estimates.

As expected, PCGC regression estimates were higher than REML estimates for all six diseases (Table 1). Specifically, the estimated heritability attributable to common variants increased from 39.3% to 47% for bipolar disorder, from 20.2% to 24.6% for Crohn's disease, from 14.6% to 16.3% for type 1 diabetes (excluding chromosome 6 from the analysis), from 38.2% to 42.1% for schizophrenia, and from 33.3% to 38.2% for MI. The most notable increase was for the least frequent disease: for MS, the estimated heritability attributable to common variants increased from 33.5% to 45.3%.

We can infer the proportion of the overall heritability attributable to common variants by dividing these values by published estimates of the total heritability (derived from population studies of the phenotypic correlations among relatives; Table 1). Mindful of the considerable uncertainty in these published estimates, we estimate that the proportion of the overall heritability attributable to common variants for these diseases ranges from 41% to 68% (mean 60%).

Extending PCGC Regression to Incorporate Covariates in Heritability Estimation. PCGC regression similarly can deal with other important situations. Although genetic and environmental effects often are assumed to represent the sum of many small effects and thus to be distributed normally, some specific effects may be very large. For example, men are considerably taller than women, the prevalence of Alzheimer's disease increases sharply with age, and lung cancer is more prevalent in smokers than nonsmokers. Such covariates (sex, age, smoking) also are referred to as "fixed effects" and must be accounted for in attempting to estimate the heritability due to common variants.

In the case of a randomly sampled continuous phenotype, REML methodology can be naturally extended to account for additional covariates, and this extension indeed is implemented in the popular GCTA software (27). In the scenario of case–control studies, Lee et al. (10) continue to treat the phenotype as quantitative, apply the extended REML approach to account for fixed effects, and use their correction (Eq. **6**) to transform the resulting estimates of heritability to the liability scale. However, the presence of fixed effects only aggravates the problems arising from case–control sampling, and as a result, the heritability estimates obtained in this manner are even more biased. The increased bias is seen in our simulations (Fig. 2*C* and *SI Appendix*, Fig. S13) and also is supported by theoretical arguments (28).

By contrast, PCGC regression can be extended readily to account for covariates while still yielding unbiased estimates. This is done by replacing the constant c from Eq. **7** with a set of constants c_{ij} that depend on the covariates for individuals i and j. Deriving the specific values of c_{ij} is more involved (*SI Appendix*, section 2). Briefly, we let z_i denote the covariates for individual i, which includes all relevant additional data (smoking habits, sex, age, and so on). Rather than having studywide parameters denoting (*i*) the fraction K of cases in the population, (*ii*) the liability threshold t, and (*iii*) the probability P that an individual in the study is affected, we have individual-specific parameters, where each individual has (*i*) an individual-specific probability K_i of being affected, conditional on her specific covariates; (*ii*) a corresponding individual-specific liability threshold t_i; and (*iii*) a corresponding individual-specific probability P_i of being affected, conditional on both her specific covariates and the fact that she was selected for the study.

Using these definitions, we show (*SI Appendix*, section 2.2) that

$$c_{ij} = \frac{\varphi(t_i)\varphi(t_j)\left[1-(P_i+P_j)\left(\frac{P-K}{P(1-K)}\right)+P_iP_j\left(\frac{P-K}{P(1-K)}\right)^2\right]}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}\left(K_i+(1-K_i)\frac{K(1-P)}{P(1-K)}\right)\left(K_j+(1-K_j)\frac{K(1-P)}{P(1-K)}\right)}. \quad [15]$$

Our simulations confirm that PCGC regression accounts correctly for additional covariates (Fig. 2*C*).

The relatively complex formulas for PCGC regression in the fixed-effects scenario shed light on a key difference between REML-based methods and PCGC regression. When no fixed effects are involved, both the REML method of Lee et al. (10) and PCGC regression appear rather similar: both may be viewed as a two-step procedure, in which the first step entails applying

Table 1. Estimates of phenotypic variance explained and proportion of heritability explained by common variants from REML and PCGC regression for six diseases: bipolar disorder, Crohn's disease, early-onset MI, MS, schizophrenia, and type 1 diabetes

		Phenotypic variance explained by common variants, % (SE)			
Phenotype	Prevalence, %	REML	PCGC	Estimated total heritability, %	Heritability explained by common variants, %
Bipolar disorder	0.5	39.3 (4.3)	47.0 (7.1)	71	66
Crohn's disease	0.1	20.2 (3.1)	24.6 (4.3)	50–60	41
MI	1	33.3 (6.1)	38.2 (9.1)	56	68
MS	0.1	33.5 (3.5)	45.3 (5.5)	25–75	60
Schizophrenia	1	38.2 (3.3)	42.1 (5)	64	66
Type 1 diabetes (w/o HLA)	0.5	14.5 (4.0)	16.3 (5.5)		
With HLA (*SI Appendix*, section 11)	0.5		51.3	72–88	58

SEs are given in parentheses. PCGC and REML estimates are corrected for estimated genetic correlations, as discussed in refs. 9 and 10, and include sex and the top 10 principal components as fixed effects (*SI Appendix*, sections 2 and 8, respectively). SEs for PCGC regression estimates are estimated using 100 jackknife iterations (*SI Appendix*, section 6), and REML SEs are produced by the GCTA software (27). For population prevalence and family-based estimate references, see *SI Appendix*, Table S8. The heritability explained by common variants was obtained by dividing (*i*) the PCGC estimate of the proportion of phenotypic variance explained by common variants by (*ii*) the reported heritability. Where a range of values is given for the heritability, we use the largest value; this provides a conservative estimate. For type 1 diabetes, the REML and PCGC analyses exclude chromosome 6, which includes the large effect of the MHC. Based on family studies, the MHC has been estimated to explain ∼50% of the heritability of type 1 diabetes, which would correspond to 36–44% of the phenotypic variance (39). Based on direct analysis of the relative risk of various MHC haplotypes, we obtain a conservative estimate that the MHC explains at least 35% of the phenotypic variance (*SI Appendix*, section 11); we use this value as the contribution of the MHC. w/o, without.

a simple recipe appropriate for quantitative traits to a disease trait coded as 0/1 and the second step entails applying a "correction" to the result, to account for issues overlooked in the first step.

Although Lee et al. (10) use the same approach when fixed effects are involved (i.e., they apply a standard REML procedure and an ex post facto correction), PCGC regression no longer may be viewed in this manner. Accounting for the fixed effects is an intrinsic part of the estimation process, rather than an ex post correction. We no longer regress the product of the phenotypes onto G_{ij} and divide the resulting slope by a single factor c to obtain liability-scale heritability estimates. Instead, we have a different c_{ij} for every pair of individuals, and we regress p_ip_j onto the product $c_{ij}G_{ij}$.

Thus, although REML-based methods and PCGC regression may appear similar, they actually are fundamentally different. PCGC regression constructs the estimates from first principles, which is why its estimates are unbiased in the presence of both case–control sampling and covariates.

Applying PCGC Regression to Extreme-Phenotype Studies of Quantitative Traits. PCGC regression also can be applied to other study designs in which unbiased REML estimates are not available because of the complexity of the situation. A particularly important design is an "extreme-phenotypes" study, in which only individuals with extremely high or low values of the phenotype are selected for genotyping. Lander and Botstein (29) showed that such designs are efficient for genetic mapping, because most of the power resides in individuals with extreme phenotypes.

Extreme-phenotype studies pose challenges similar to those of case–control studies for REML-based estimation of heritability, because the assumptions of the normality and GxE independence no longer hold (*SI Appendix*, Fig. S15). As a result, REML estimates will be biased.

Our PCGC regression framework can deal with extreme-phenotype studies in a fashion analogous to that of case–control studies. For example, when sampling individuals whose phenotypes are either below the Kth quantile, or above the $(1-K)$th quantile, the value of c is given by (*SI Appendix*, section 4)

$$c=\left(1+\frac{\varphi(t)t}{K}\right)^2, \qquad [16]$$

where $t=\Phi^{-1}(1-K)$ is the upper threshold for inclusion on the liability scale, and φ is the standard Gaussian density, as before. We derive similar expressions for studies in which different quantiles are used for including extremely high or extremely low phenotypes in the study, as well as several other generalizations (*SI Appendix*, section 4). Simulations confirm that the PCGC estimates are unbiased (*SI Appendix*, Fig. S16).

To study the potential benefit of extreme-phenotype studies, we used simulations to measure the improvement in the accuracy of heritability estimates obtained by extreme-phenotype sampling (*SI Appendix*, section 4). We find, for example, that the accuracy obtained by randomly sampling N individuals from the population can be achieved by sampling approximately $N/8$ individuals from each of the top and bottom deciles (*SI Appendix*, Fig. S17). We observe similar accuracy benefits when comparing the SEs of heritability estimates of HDL levels in GWASs based on random-sampling vs. GWASs based on extreme-phenotype sampling (*SI Appendix*, section 10.3).

Discussion

The genetic architecture of most common traits and diseases is complex, involving contributions from genetic variants at many genes and, potentially, interactions among them. These genetic variants likely span the range of allele frequencies, from common variants in the population to rare variants present at extremely low frequencies.

CVASs (typically called GWASs) of traits and diseases already have uncovered numerous statistically significant associations with common variants at individual genetic loci. However, statistical analyses suggest that many more associated variants lurk below the surface—falling short of statistical significance because of inadequate sample sizes. Accordingly, it would be valuable to have reliable methods to infer the overall contribution of common variants without the need to identify each individual locus.

REML Methods for Estimating Heritability. Visscher and coworkers (9) pioneered the random-effects approach for using CVAS data to estimate the narrow-sense heritability of a quantitative trait due to common variants. By modeling effect sizes as random variables, this method elegantly circumvents the need to estimate the effect size of each SNP. Instead, it focuses on estimating the overall heritability due to the entire set of common variants. Applying this REML methodology to a wide range of quantitative phenotypes indicates that for many phenotypes, common genetic variants account for a substantial portion of the overall heritability—much more than the portion explained by the individual loci that so far have attained genome-wide significance.

Visscher and coworkers (10) subsequently sought to extend the REML methodology to disease phenotypes. Disease studies typically involve case–control designs, wherein cases are considerably oversampled relative to their frequency in the population; this sampling design violates various assumes of the REML framework. Lee et al. (10) attempted to account for these issues by applying a post hoc correction. However, as demonstrated by our extensive simulations (and by actual genetic studies of six diseases), their method yields strongly downwardly biased estimates of the heritability due to common variants in a variety of interesting and relevant scenarios. The bias depends on properties of the disease, including the prevalence in the population and the true underlying heritability. Most troublingly, the bias increases with study size and with the proportion of cases in the sample. (The bias also depends on the number of SNPs actually genotyped, although this is of secondary importance.) We conclude that REML methods do not provide a suitable framework for estimating heritability for disease phenotypes.

PCGC Regression. To solve this problem, we developed PCGC regression, which provides a powerful framework for estimating the heritability due to common variants in a wide range of scenarios. Extensive simulations show that PCGC regression yields heritability estimates that are unbiased (as expected mathematically) and more accurate than the REML approach, when applied to case–control studies.

PCGC regression is a general framework for heritability estimation. In the case of an unascertained quantitative phenotype, it boils down to the well-known regression method of Haseman and Elston. When dealing with case–control studies, PCGC regession allows unbiased estimation of heritability, even in the presence of covariates. As such, it subsumes several recent anecdotal observations (14, 15), provides a theoretical foundation, and, importantly, allows for the incorporation of covariates. In addition to case–control studies, the PCGC regression framework can readily accommodate other complex study designs. One important application is extreme-phenotype studies for quantitative traits, which may be more cost-effective than random sampling studies for the purposes of identifying causal loci and estimating heritability.

One limitation of PCGC regression is that it is based on a first-order approximation of the relationship between phenotypic similarity and genetic correlation. This approximation is expected to be accurate when individuals in the study largely are unrelated. However, in populations with a high degree of cryptic relatedness, this might not be the case. In such cases, PCGC regression can be extended by using second-order, or higher-order, approximations (*SI Appendix*, section 1.4). We note, however, that for the WTCCC population, using first- or second-order approximations yielded very little difference.

Design of Simulations. Our paper highlights a critical issue concerning the design of genetic simulations. Lee et al. (10) did not observe the serious bias inherent in REML methods for case–control studies because they used a highly incomplete simulation. In particular, they did not explicitly generate genotypes and phenotypes for each individual, but rather arbitrarily assigned genetic-correlation values to pairs of individuals. This shortcut eliminated critical correlations between genotype and phenotype. By contrast, our simulations use a generative approach: they explicitly (*i*) assign effect sizes to each variant, (*ii*) generate genotypes for each individual, and (*iii*) ascertain cases and controls based on phenotype, by selecting or rejecting individuals. These simulations readily revealed the large downward bias in REML estimates.

One issue with the generative approach is the considerable running time of each simulation, which greatly limits the possible number of individuals and SNPs that can be simulated, as well as the prevalence of the disease simulated. To overcome this problem, we developed a dynamic programming approach, allowing direct sampling of genotypes of cases (*SI Appendix*, section 5.8). We implemented this approach in a software package called simCC, which is freely available from our website. Simulations using simCC are considerably faster; therefore, we could expand our simulations from 10,000 SNPs to 100,000 SNPs, with qualitatively similar results (*SI Appendix*, Fig. S18).

We note that our simulations fail to be fully realistic because they do not model linkage disequilibrium among variants. Although the mathematical result by ref. 22 means that this limitation has little impact on the results in the context of heritability estimation, it is interesting to ask how linkage disequilibrium might be included for the benefit of other endeavors. Several authors have used an approach that takes actual genotypes from an existing study (e.g., WTCCC), assigns effect sizes, and then calculates phenotypes (e.g., refs. 13, 23, 30). However, as currently implemented, this approach has the serious flaw that it does not impose selection to obtain cases and controls and, accordingly, cannot yield a realistic correlation structure for a case–control study (such as the striking correlation effects in Fig. 1*C*). One could solve this problem by using a vastly larger collection of actual genotypes, such that one could obtain cases by imposing stringent selection (i.e., discarding 90–99% of samples, depending on the disease frequency); however, current datasets are too small for his approach. Alternatively, it may be possible to use programs such as HapGen (31) to simulate realistic genotypes from a smaller sample of actual genotypes. Whether this approach is feasible at the required scale is a topic for further study.

Improved Estimates of Genetic Correlation. We demonstrate that PCGC regression yields unbiased heritability estimates, given knowledge of the genetic correlation matrix G. However, G is unknown, and estimates of G are used by most heritability estimation methods. As pointed out by Yang et al. (9), replacing the true value of G with a noisy estimate results in underestimation of the heritability. This effect is not unique to heritability estimation and is known as "diluted" regression (or "errors-in-variables"): when regressing a dependent variable y onto a noisy measurement of an explanatory variable x, the estimated slope is attenuated. It is important to stress that this bias is not a result of the heritability estimation method, but rather to the use of noisy estimates of genetic correlations instead of the true correlations.

Yang et al. (9) overcome this problem by quantifying the bias via simulations, and correcting for it post hoc. This eliminates the bias but increases the variance of the estimate. We adopt their method when estimating the heritability due to common SNPs using real data.

Several recent papers (13, 21, 30, 32–35) focused on improving the estimation of the genetic correlation matrices, by accounting for sparsity (33, 34), LD and LD-dependent genetic architecture (30), prior information regarding effect sizes of different SNPs (35), or the relationship between minor allele frequency and effect size (30). As better methods for estimating the genetic correlation matrix are developed, they may be used in PCGC regression in place of the correction method of Yang et al. (9). More accurate estimates of the genetic correlation matrix are expected to increase the estimated proportion of heritability due to common SNPs even further.

Deviating from the Probabilistic Assumptions of the Model. A key assumption throughout most works addressing the problem of heritability estimation in general, and estimating heritability using case–control GWAS in particular, is that the genetic and environmental effects follow a normal distribution. The assumption of normality typically is justified by the central limit theorem, as both the genetic and environmental effects are assumed to be the sum of many small contributions, resulting in an approximately normal distribution.

These assumptions might prove invalid when some covariates have a considerable effect on the phenotype: for example, the effect of sex on height or coronary artery disease. When the problematic covariate is known and observed, it can easily be accounted for directly by including it as a fixed effect in the model. However, not all important covariates may be known or observed. When the phenotype is quantitative, normality assumptions can be tested; therefore, situations in which the normality assumption is invalid can, at the very least, be identified. On the other hand, for disease phenotypes, the liability is unobserved, and it is unclear how to test the normality assumptions, as well as what the effect of deviating from these assumptions would be. This question remains to be addressed in future work.

We assume throughout that the underlying genetic architecture is additive. However, one common speculation is that GxG interactions play a considerable role in human disease (36). In this case, one still might attempt to estimate the additive heritability explained by common SNPs by means described here. However, the proper interpretation of the result might depend on the exact genetic architecture.

Another key assumption is that the selection of individuals for the study is guided only by their phenotypes. Although this often is the situation with case–control designs, other, more complicated selection schemes exist. Two important examples are covariate-driven sampling [in which the selection is guided by both the phenotype and by a risk factor of interest, e.g., type 2 diabetes patients with a low body mass index (BMI) and controls with a high BMI (37)] and case–control matching (in which, for each case in the study, an effort is made to recruit a control with similar characteristics). Such designs pose an interesting challenge for future research.

Role of Common Variants in Common Disease. Beyond the methodological results in this paper, our key biological finding is that the heritability of disease traits attributable to common genetic variants is even higher than current estimates. For the diseases analyzed above, the heritability estimated by PCGC regression is 9–34% (mean 19%) higher than the estimates produced by REML.

The phenotypic variance attributable to common variants is 25%, 38%, 42.1%, 45%, and 47% for Crohn's disease, early-onset MI, schizophrenia, MS, and bipolar disorder, respectively. For type 1 diabetes, our estimate is 16.3% when we exclude the large effects due to the major histocompatibility complex (MHC), and 51.3% when we include it. [The contribution of the MHC can be estimated from family studies (38, 39), as well as by considering the contribution of the individual common variants as fixed effects (*SI Appendix*, section 11).]

We can estimate the proportion of heritability explained by common variants by dividing (*i*) the proportion of phenotypic variance explained by common variants by (*ii*) the proportion of the phenotypic variance explained by additive genetic factors—that is, the total heritability—based on phenotypic similarity among relatives (e.g., monozygotic and dizygotic twins). We note that estimates of the total heritability involve substantial uncertainty [due to small study size and potential artifacts resulting from the underlying genetic architecture (2)] and may vary considerably across studies; where multiple estimates had been reported, we used the largest value. The proportions are 41%, 58%, 60%, 64%, 66%, 66%, and 68% for Crohn's disease, type 1 diabetes, MS, schizophrenia, bipolar disorder, and early-onset MI, respectively, with a mean value of 60% (Table 1).

Our improved estimates still may underestimate the true proportion of heritability explained by common variants for three reasons. First, we calculated the proportion of heritability explained by using the largest estimate of the total heritability, when multiple estimates had been reported; this yields a conservative estimate. Second, as noted above, uncertainty about the genetic correlation matrix decreases the heritability explained. Improved estimates of the genetic correlations likely will increase the estimated heritability explained by common SNPs. Third, our analysis assumes that the contributions of variants are drawn from a uniform prior distribution, regardless of biological context. Instead, variants might be categorized based on biological annotation (e.g., those within or near coding regions, regulatory elements, and so on). A distinct correlation matrix might be estimated for each set and the matrices used simultaneously for heritability estimation, a task that can be accommodated readily by PCGC regression (*SI Appendix*, section 3). A refined correlation structure should provide a better model of the genetic architecture of disease, and thus would be expected to yield higher estimates of the heritability explained by common variants, as well as to provide useful scientific insights.

Our results suggest that larger CVASs will identify many additional common variants related to common diseases, although many additional common variants likely still will have effect sizes that fall below the limits of detection given practically achievable sample sizes. Still, common variants clearly will not explain all heritability. As discussed in the first two papers in this series (2, 3), rare genetic variants and genetic interactions likely will make important contributions as well. Fortunately, advances in DNA sequencing technology should make it possible in the coming years to carry out comprehensive studies of both common and rare genetic variants in tens (and possibly hundreds) of thousands of cases and controls, resulting in a fuller picture of the genetic architecture of common diseases.

ACKNOWLEDGMENTS. We are grateful to the Wellcome Trust Case Control Consortium for making data on multiple diseases available (www.wtccc.org.uk), the ANZgene Consortium for sharing the MS data, Sekar Kathiresan for sharing the MI data, and Dan Rader for sharing the HDL data. D.G. was supported by a Colton fellowship and a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. E.S.L. was supported in part by NIH HG003067 and by funds from the Broad Institute. S.R. was supported in part by Israeli Science Foundation Grant 1487/12.

1. Maher B (2008) Personal genomes: The case of the missing heritability. *Nature* 456(7218):18–21.
2. Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* 109(4):1193–1198.
3. Zuk O, et al. (2014) Searching for missing heritability: Designing rare variant association studies. *Proc Natl Acad Sci USA* 111(4):E455–E464.
4. Welter D, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42(Database issue, D1):D1001–D1006.
5. Visscher PM (2008) Sizing up human height variation. *Nat Genet* 40(5):489–490.
6. Weedon MN, et al.; Diabetes Genetics Initiative; Wellcome Trust Case Control Consortium; Cambridge GEM Consortium (2008) Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* 40(5):575–583.
7. Lettre G, et al.; Diabetes Genetics Initiative; FUSION; KORA; Prostate, Lung Colorectal and Ovarian Cancer Screening Trial; Nurses' Health Study; SardiNIA (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* 40(5):584–591.
8. Gudbjartsson DF, et al. (2008) Many sequence variants affecting diversity of adult human height. *Nat Genet* 40(5):609–615.
9. Yang J, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42(7):565–569.
10. Lee SH, Wray NR, Goddard ME, Visscher PM (2011) Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* 88(3):294–305.
11. Lee SH, et al.; ANZGene Consortium; International Endogene Consortium; Genetic and Environmental Risk for Alzheimer's disease Consortium (2013) Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Hum Mol Genet* 22(4):832–841.
12. Do CB, et al. (2011) Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genet* 7(6):e1002141.
13. Zaitlen N, et al. (2013) Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet* 9(5):e1003520.
14. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* 46(2):100–106.
15. Chen G-B (2014) Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. *Front Genet* 5:107.
16. Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 2(1):3–19.
17. Elston RC, Buxbaum S, Jacobs KB, Olson JM (2000) Haseman and Elston revisited. *Genet Epidemiol* 19(1):1–17.
18. Galton F (1886) Regression towards mediocrity in hereditary stature. *J Royal Anthr Inst* 15:246–263.
19. Visscher PM, et al. (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* 2(3):e41.
20. Dempster ER, Lerner IM (1950) Heritability of threshold characters. *Genetics* 35(2):212–236.
21. Zhou X, Carbonetto P, Stephens M (2013) Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet* 9(2):e1003264.
22. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12):e190.
23. Lee SH, et al.; Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ); International Schizophrenia Consortium (ISC); Molecular Genetics of Schizophrenia Collaboration (MGS) (2012) Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* 44(3):247–250.
24. Bahlo M, et al.; Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nat Genet* 41(7):824–828.
25. Ripke S, et al.; Multicenter Genetic Studies of Schizophrenia Consortium; Psychosis Endophenotypes International Consortium; Wellcome Trust Case Control Consortium 2 (2013) Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* 45(10):1150–1159.
26. Kathiresan S, et al.; Myocardial Infarction Genetics Consortium; Wellcome Trust Case Control Consortium (2009) Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet* 41(3):334–341.
27. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet* 88(1):76–82.
28. Burton PR (2003) Correcting for nonrandom ascertainment in generalized linear mixed models (GLMMs), fitted using Gibbs sampling. *Genet Epidemiol* 24(1):24–35.
29. Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121(1):185–199.
30. Speed D, Hemani G, Johnson MR, Balding DJ (2012) Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* 91(6):1011–1021.
31. Spencer CCA, Su Z, Donnelly P, Marchini J (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* 5(5):e1000477.
32. Crossett A, Lee AB, Klei L, Devlin B, Roeder K (2013) Refining genetically inferred relationships using treelet covariance smoothing. *Ann Appl Stat* 7(2):669–690.
33. Golan D, Rosset S (2011) Accurate estimation of heritability in genome wide studies using random effects models. *Bioinformatics* 27(13):i317–i323.
34. Guan Y, Stephens M (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann Appl Stat* 5(3):1780–1815.
35. Zhang Z, et al. (2014) Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLoS One* 9(3):e93017.
36. Manolio TA, et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461(7265):747–753.
37. Voight BF, et al.; MAGIC investigators; GIANT Consortium (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 42(7):579–589.
38. Wei Z, et al. (2009) From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet* 5(10):e1000678.
39. Risch N (1987) Assessing the role of HLA-linked and unlinked determinants of disease. *Am J Hum Genet* 40(1):1–14.

Measuring missing heritability: Inferring the Contribution of Common Variants

David Golan[1] Eric S. Lander[2] Saharon Rosset[1]

June 2014

1. Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv, Israel.

2. Broad Institute of Harvard and MIT, Cambridge, MA 02142;
Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and
Department of Systems Biology, Harvard Medical School, Boston, MA 02155.

Note: This supplementary text includes a complete and self-contained description of our methods, repeating the materials from the manuscript where appropriate.

Contents

1 PCGC regression

1.1 Liability threshold model - notations

Denote K the prevalence of a condition in the population and P the prevalence in the study.

Under the liability threshold model, we assume that each individual i has an unknown liability $l_i = g_i + e_i$ where g_i is a genetic random effect, which can be correlated across individuals, and e_i is the environmental random effect, which is assumed to be independent of each other and of the genetic effects. Both effects are assumed to follow a Gaussian distribution with variances σ_g^2 and $1 - \sigma_g^2$ respectively. A person is then assumed to be a case if her liability exceeds a threshold $t = \Phi^{-1}(1-K)$, i.e. the phenotype y_i is given by $y_i = \mathbb{I}\{l_i > t\}$. This definition guarantees that the prevalence in the population is indeed K.

1.2 Selection probabilities

When the study is observational, the probability of being included in the study is independent of the phenotype. However, in a case-control study, the proportion of cases is usually greatly ascertained. To model this fact, we define a random indicator variable s_i indicating whether individual i was selected to the study.

Denote $P_{\text{case}}, P_{\text{control}}$ the probabilities that a case and a control would be selected for the study. It follows that:

$$\frac{KP_{\text{case}}}{(1-K)P_{\text{control}}} = \frac{P}{1-P}.$$

solving for P_{control} we get:

$$P_{\text{control}} = \frac{K(1-P)}{P(1-K)} P_{\text{case}}.$$

Assuming $P_{\text{case}} = 1$ yields $P_{\text{control}} = \frac{K(1-P)}{P(1-K)}$. This assumption is commonly referred to as the "full ascertainment" [21] assumption. We note, however, that this assumption only serves to simplify the mathematical notation, and that relaxing it does not alter any of the following results, as the additional P_{case} term cancels out in all of the final equations.

For a random individual with unknown phenotype the selection probability is:

$$KP_{\text{case}} + (1-K)P_{\text{control}} = \frac{K}{P} P_{\text{case}} = \frac{K}{P}.$$

1.3 Heritability estimation

Next, consider a pair of individuals in the study, whose genetic effects are correlated and denote by ρ the correlation. To account for the design of the study, we introduce an additional variable, $\mathcal{S}$, which indicates that both individuals were selected for the study.

Denote by Z_{ij} the product of the standardized phenotypes:

$$Z_{ij} = \frac{(y_i - P)(y_j - P)}{P(1-P)}.$$

The variable Z_{ij} can obtain three values:

$$Z_{ij} = \begin{cases} \frac{1-P}{P} & y_i = y_j = 1 \\ -1 & y_i \neq y_j \\ \frac{P}{1-P} & y_i = y_j = 0 \end{cases}.$$

We write down the expected value of Z_{ij}, conditional on $\mathcal{S}$ (the individuals are part of the study) and given ρ:

$$\mathbb{E}[Z_{ij} \mid \mathcal{S}=1;\rho] = \frac{1-P}{P}\mathbb{P}(y_i = y_j = 1 \mid \mathcal{S}=1;\rho) -$$

$$\mathbb{P}(y_i \neq y_j \mid \mathcal{S}=1;\rho) + \frac{P}{1-P}\mathbb{P}(y_i = y_j = 0 \mid \mathcal{S}=1;\rho).$$

We apply Bayes' theorem to the first of the three summands on the right:

$$\mathbb{P}(y_i = y_j = 1 \mid \mathcal{S}=1;\rho) = \frac{\mathbb{P}(\mathcal{S}=1 \mid y_i = y_j = 1;\rho)\mathbb{P}(y_i = y_j = 1;\rho)}{\mathbb{P}(\mathcal{S}=1;\rho)}.$$

Under the full ascertainment assumption $\mathbb{P}(\mathcal{S}=1 \mid y_i = y_j = 1;\rho) = 1$, and so

$$\mathbb{P}(y_i = y_j = 1 \mid \mathcal{S}=1;\rho) = \frac{\mathbb{P}(y_i = y_j = 1;\rho)}{\mathbb{P}(\mathcal{S}=1;\rho)}.$$

Similarly:

$$\mathbb{P}(y_i = y_j = 0 \mid \mathcal{S}=1;\rho) = \frac{\mathbb{P}(\mathcal{S}=1 \mid y_i = y_j = 0;\rho)\mathbb{P}(y_i = y_j = 0;\rho)}{\mathbb{P}(\mathcal{S}=1;\rho)},$$

and since a control is selected to the study with probability $\frac{K(1-P)}{P(1-K)}$, this boils down to:

$$\Big(\frac{K(1-P)}{P(1-K)}\Big)^2 \frac{\mathbb{P}(y_i = y_j = 0;\rho)}{\mathbb{P}(\mathcal{S}=1;\rho)}.$$

For the case of $y_i \neq y_j$, one individual is a case, and is automatically selected, while the other is a control and is selected with probability $\frac{K(1-P)}{P(1-K)}$. Hence:

$$\mathbb{P}(y_i \neq y_j \mid \mathcal{S}=1;\rho) = \frac{K(1-P)}{P(1-K)} \frac{\mathbb{P}(y_i \neq y_j;\rho)}{\mathbb{P}(\mathcal{S}=1;\rho)},$$

Using these results we get:

$$\mathbb{E}[Z_{ij} \mid \mathcal{S}=1;\rho] =$$

$$\frac{\frac{1-P}{P}\mathbb{P}(y_i = y_j = 1;\rho) - \frac{K(1-P)}{P(1-K)}\mathbb{P}(y_i \neq y_j;\rho) + \frac{P}{1-P}\left(\frac{K(1-P)}{P(1-K)}\right)^2 \mathbb{P}(y_i = y_j = 0;\rho)}{\mathbb{P}(\mathcal{S}=1;\rho)}.$$

Denote the numerator by $A(\rho)$ and the denominator by $B(\rho)$. We wish to approximate the latter equation using a Taylor series around $\rho = 0$. Such an approximation would take the form:

$$\mathbb{E}[Z_{ij} \mid \mathcal{S}=1;\rho] = \frac{A(0)}{B(0)} + \frac{A'(0)B(0) - B'(0)A(0)}{B(0)^2}\rho + \mathcal{O}(\rho^2).$$

We later discuss a second order approximation as well.

Note that with $\rho = 0$, the phenotypes of the two individuals are i.i.d. and so $A(0) = 0$. Therefore, the Taylor approximation takes the form:

$$\mathbb{E}[Z_{ij} \mid \mathcal{S}=1;\rho] = \frac{A'(0)}{B(0)}\rho + \mathcal{O}(\rho^2).$$

Similarly, with $\rho = 0$ the events of being included in the study are i.i.d. for both individuals, so $B(0) = \frac{K^2}{P^2}$. All that remains is to find $A'(0)$.

We are interested in computing the probabilities of the three possible combinations of phenotypes:

$$\mathbb{P}(y_i = y_j = 1; \rho, \sigma_g^2) = \int_t^\infty \int_t^\infty f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2,$$

$$\mathbb{P}(y_i \neq y_j; \rho, \sigma_g^2) = 2\int_{-\infty}^t \int_t^\infty f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2,$$

and

$$\mathbb{P}(y_i = y_j = 0; \rho, \sigma_g^2) = \int_{-\infty}^t \int_{-\infty}^t f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2,$$

where f_{ρ,σ_g^2} is the multivariate Gaussian density, namely:

$$f_{\rho,\sigma_g^2}(l_1, l_2) = \frac{1}{2\pi} |\Sigma|^{-\frac{1}{2}} e^{-\frac{(l_1,l_2)\Sigma^{-1}(l_1,l_2)^\intercal}{2}},$$

with Σ denoting the covariance matrix of the liabilities, given explicitly by:

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \sigma_g^2 + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (1 - \sigma_g^2) = \begin{pmatrix} 1 & \rho\sigma_g^2 \\ \rho\sigma_g^2 & 1 \end{pmatrix}.$$

Note that this correlation structure implies that assuming normality of g and e is enough to guarantee that the resulting first-order PCGC-regression formula would be linear in σ_g^2. Both ρ and σ_g^2 only appear as the product $\rho\sigma_g^2$. Hence, differentiating w.r.t. to ρ results in an expression of the form $f(\rho\sigma_g^2)\sigma_g^2$, for some function f, and setting $\rho = 0$ yields $f(0)\sigma_g^2$, so the first-order approximation is of the form $f(0)\sigma_g^2\rho$.

The determinant of Σ is $|\Sigma| = 1 - \rho^2\sigma_g^4$ and its inverse is $\Sigma^{-1} = \frac{1}{1-\rho^2\sigma_g^4} \begin{pmatrix} 1 & -\rho\sigma_g^2 \\ -\rho\sigma_g^2 & 1 \end{pmatrix}$, and so the density function f_{ρ,σ_g^2} can be written as:

$$f_{\rho,\sigma_g^2}(l_1, l_2) = \frac{1}{2\pi\sqrt{1 - \rho^2\sigma_g^4}} e^{-\frac{l_1^2 + l_2^2 - 2l_1 l_2 \rho\sigma_g^2}{2(1-\rho^2\sigma_g^4)}}.$$

Deriving $A(\rho)$ requires deriving each of the three double integrals w.r.t. ρ:

$$\frac{d}{d\rho} \int_t^\infty \int_t^\infty f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2 = \int_t^\infty \int_t^\infty \frac{d}{d\rho} f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2.$$

Setting $\rho = 0$ in the last expression yields:

$$\int_t^\infty \int_t^\infty l_1 l_2 \sigma_g^2 \frac{1}{2\pi} e^{-\frac{l_1^2 + l_2^2}{2}} = \sigma_g^2 \Big[\int_t^\infty l \frac{1}{\sqrt{2\pi}} e^{-\frac{l^2}{2}} dl \Big]^2 = \sigma_g^2 \varphi(t)^2.$$

Explanation: we differentiate and set $\rho = 0$. By the chain rule, the derivative of any expression with ρ^2 is 0 with $\rho = 0$, and obviously the derivative of any expression which does not depend on ρ is 0. The only expression whose derivative is therefore not 0 at $\rho = 0$ is $-2l_1 l_2 \rho\sigma_g^2$ in the numerator of the exponent. The denominator of the exponent is 2 at $\rho = 0$, and so the derivative is $l_1 l_2 \sigma_g^2 \frac{1}{2\pi} e^{-\frac{l_1^2 + l_2^2}{2}}$.

Similarly:

$$\frac{d}{d\rho} \int_{-\infty}^t \int_t^\infty f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2 = -\sigma_g^2 \varphi(t)^2,$$

and

$$\frac{d}{d\rho} \int_{-\infty}^t \int_{-\infty}^t f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2 = \sigma_g^2 \varphi(t)^2.$$

Using these results we can write down $A'(0)$:

$$A'(0) = \Big[\frac{1-P}{P} + 2\frac{K(1-P)}{P(1-K)} + \frac{P}{1-P}\Big(\frac{K(1-P)}{P(1-K)}\Big)^2\Big]\sigma_g^2\varphi(t)^2 = \frac{1-P}{P(1-K)^2}\sigma_g^2\varphi(t)^2,$$

and so:

$$\mathbb{E}[Z_{ij} \mid \mathcal{S} = 1; \rho] \approx \frac{A'(0)}{B(0)}\rho = \frac{\frac{1-P}{P(1-K)^2}\sigma_g^2\varphi(t)^2}{\frac{K^2}{P^2}}\rho = \frac{P(1-P)}{K^2(1-K)^2}\sigma_g^2\varphi(t)^2\rho.$$

Hence, when the error of the approximation is small, the slope obtained by regressing Z_{ij} on G_{ij} is an unbiased estimator of $\frac{P(1-P)}{K^2(1-K)^2}\sigma_g^2\varphi(t)^2$, thus dividing it by $\frac{P(1-P)}{K^2(1-K)^2}\varphi(t)^2$ yields an unbiased estimator of σ_g^2 - the liability scale heritability.

1.4 Second order approximation

While the first order approximation yields very satisfactory results in our simulations, one can obtain a better estimator using a better approximation. The second term of the Taylor series takes the form:

$$\frac{A''(0)B(0) - 2B'(0)A'(0)}{B(0)^2}\rho^2.$$

We have already derived:

$$B(0) = \frac{K^2}{P^2},$$

and

$$A'(0) = \frac{1-P}{P(1-K)^2}\sigma_g^2\varphi(t)^2.$$

Now, $B(\rho)$ is the probability that two individuals with genetic correlation ρ are included in the study:

$$B(\rho) = P(y_i = y_j = 1; \rho, \sigma_g^2) + \frac{K(1-P)}{P(1-K)}P(y_i \neq y_j; \rho, \sigma_g^2) + \Big(\frac{K(1-P)}{P(1-K)}\Big)^2 P(y_i = y_j = 0; \rho, \sigma_g^2).$$

Using the previous derivatives of the double integrals yields:

$$B'(0) = \Big[1 - 2\frac{K(1-P)}{P(1-K)} + \Big(\frac{K(1-P)}{P(1-K)}\Big)^2\Big]\sigma_g^2\varphi(t)^2 =$$

$$\Big[1 - \frac{K(1-P)}{P(1-K)}\Big]^2\sigma_g^2\varphi(t)^2 = \Big[\frac{K-P}{P(1-K)}\Big]^2\sigma_g^2\varphi(t)^2.$$

Computing $A''(0)$ requires computing the second derivative of the two-dimensional density at $\rho = 0:$

$$\int_t^\infty\int_t^\infty \frac{d^2}{d\rho^2}f_{\rho,\sigma_g^2}(l_1, l_2) = \int_t^\infty\int_t^\infty \sigma_g^4(l_1^2 - 1)(l_2^2 - 1)f_{0,\sigma_g^2}(l_1, l_2) = \sigma_g^4\varphi(t)^2t^2,$$

so

$$A''(0) = \frac{1-P}{P(1-K)^2}\sigma_g^4\varphi(t)^2t^2.$$

Hence:

$$\frac{A''(0)B(0) - 2B'(0)A'(0)}{B(0)^2} = \frac{\frac{1-P}{P(1-K)^2}\sigma_g^4\varphi(t)^2t^2\frac{K^2}{P^2} - 2\Big[\frac{K-P}{P(1-K)}\Big]^2\sigma_g^2\varphi(t)^2\frac{1-P}{P(1-K)^2}\sigma_g^2\varphi(t)^2}{\frac{K^4}{P^4}}$$

$$= \frac{P}{K^4}\sigma_g^4\varphi(t)^2\frac{1-P}{(1-K)^2}\Big[t^2K^2 - 2\Big[\frac{K-P}{(1-K)}\Big]^2\varphi(t)^2\Big],$$

and the second order approximation can be written as:

$$\mathbb{E}[Z_{ij} \mid \mathcal{S}=1;\rho] = \frac{P(1-P)}{K^2(1-K)^2}\sigma_g^2\varphi(t)^2\rho$$

$$+\frac{P}{K^4}\varphi(t)^2\frac{1-P}{(1-K)^2}\Big[t^2K^2 - 2\Big[\frac{K-P}{(1-K)}\Big]^2\varphi(t)^2\Big]\sigma_g^4\frac{\rho^2}{2!} + \mathcal{O}(\rho^3).$$

Since K and P are assumed to be known, the estimation problem boils down to a single-variable non-linear regression in σ_g^2. Higher order approximations can be derived similarly.

2 Dealing with fixed effects

2.1 Extending the liability threshold model

It is often desired to include fixed effects in the analysis of a complex phenotype. Such fixed effects might include external information such as sex, diet and exposure to environmental risks, but can also be genetic variants with known effects or estimates of population structure such as projections of several top principal components.

Since the liability threshold model is in fact a probit model, these effects can be included in the usual manner:

$$l_i = x_i^\intercal\beta + g_i + e_i,$$

where x_i is a vector of the values of the relevant covariates and β is a vector of their respective effect sizes.

An individual is a case if $l_i > t$, as before. However, an equivalent formulation would be to subtract the fixed effects from the threshold, rather than adding them to the liability:

$$t_i = t - x_i^\intercal\beta,$$

thus keeping the previous formulation of the liability as a sum of genetic and environmental effects.

2.2 Heritability estimation with known fixed effects

Assume the fixed effects are known, and so the t_i's are known. We define:

$$K_i = P(y_i = 1; t_i),$$

and:

$$P_i = P(y_i = 1 \mid s_i = 1; t_i),$$

to be the probability of the i'th individual being a case when accounting for fixed effects through the adjustment of the threshold t_i.

We redefine:

$$Z_{ij} = \frac{(y_i - P_i)(y_j - P_j)}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}},$$

so now Z_{ij} can obtain four possible values:

$$Z_{ij} = \begin{cases} \frac{(1-P_i)(1-P_j)}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}} & y_i = y_j = 1 \\ \frac{-P_i(1-P_j)}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}} & y_i = 0, y_j = 1 \\ \frac{-P_j(1-P_i)}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}} & y_i = 1, y_j = 0 \\ \frac{P_iP_j}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}} & y_i = y_j = 0 \end{cases}.$$

We now wish to derive the same first order approximation before, while conditioning on the fixed effects. Repeating the same steps as before while conditioning on t_i, t_j the expression for $A(\rho)$ is now:

$$\begin{aligned} A(\rho; t_i, t_j) \quad = \quad & \frac{(1-P_i)(1-P_j)}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}}\mathbb{P}(y_i = y_j = 1; \rho, t_i, t_j) + \\ & \frac{K(1-P)}{P(1-K)}\frac{-P_i(1-P_j)}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}}\mathbb{P}(y_i = 0, y_j = 1; \rho, t_i, t_j) + \\ & \frac{K(1-P)}{P(1-K)}\frac{-P_j(1-P_i)}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}}\mathbb{P}(y_i = 1, y_j = 0; \rho, t_i, t_j) + \\ & \Big(\frac{K(1-P)}{P(1-K)}\Big)^2\frac{P_iP_j}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}}\mathbb{P}(y_i = 0, y_j = 0; \rho, t_i, t_j). \end{aligned}$$

Each phenotype was standardized to have mean 0, conditional on the relevant fixed effects. Additionally, with $\rho = 0$ the phenotypes are independent and so the expected value of Z at $\rho = 0$ is 0. An immediate result is that $A(0) = 0$. To see this consider the Taylor expansion of

$$\mathbb{E}[Z_{ij} \mid \mathcal{S} = 1; \rho, t_i, t_j] = \frac{A(0; t_i, t_j)}{B(0; t_i, t_j)} + \sum_{i=1}^{\infty} c_i\rho^i,$$

for some constants $\{c_i\}_{i=1}^{\infty}$. Setting $\rho = 0$ yields:

$$\mathbb{E}[Z_{ij} \mid \mathcal{S} = 1; 0, t_i, t_j] = \frac{A(0; t_i, t_j)}{B(0; t_i, t_j)},$$

but on the other hand:

$$\mathbb{E}[Z_{ij} \mid \mathcal{S} = 1; 0, t_i, t_j] = \mathbb{E}\Big[\frac{(y_i - P_i)(y_j - P_j)}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}}\Big] =$$

$$\mathbb{E}\Big[\frac{y_i - P_i}{\sqrt{P_i(1-P_i)}}\Big]\mathbb{E}\Big[\frac{y_j - P_j}{\sqrt{P_j(1-P_j)}}\Big] = 0.$$

Therefore, $\frac{A(0;t_i,t_j)}{B(0;t_i,t_j)} = 0$ and so $A(0; t_i, t_j) = 0$. We conclude that the first order Taylor approximation is again of the form:

$$\mathbb{E}[Z_{ij} \mid \mathcal{S} = 1; \rho, t_i, t_j] = \frac{A'(0; t_i, t_j)}{B(0; t_i, t_j)}\rho + \mathcal{O}(\rho^2).$$

To derive an explicit expression for $A'(0; t_i, t_j)$ we need to differentiate the double integrals as before. When both i and j are cases, with thresholds t_i, t_j respectively, the double integral takes the form:

$$\int_{t_i}^{\infty}\int_{t_j}^{\infty} f_{\rho,\sigma_g^2}(l_1, l_2)dl_1dl_2,$$

differentiating it w.r.t. ρ and setting $\rho = 0$ yields:

$$\varphi(t_i)\varphi(t_j)\sigma_g^2,$$

and differentiating the other double integrals yields similar results:

$$\frac{d}{d\rho}\int_{-\infty}^{t_i}\int_{-\infty}^{t_j} f_{\rho,\sigma_g^2}(l_1,l_2)dl_1dl_2\mid_{\rho=0}=\varphi(t_i)\varphi(t_j)\sigma_g^2,$$

and:

$$\begin{aligned}\frac{d}{d\rho}\int_{t_i}^{\infty}\int_{-\infty}^{t_j} f_{\rho,\sigma_g^2}(l_1,l_2)dl_1dl_2\mid_{\rho=0} &= \\ \frac{d}{d\rho}\int_{t_j}^{\infty}\int_{-\infty}^{t_i} f_{\rho,\sigma_g^2}(l_1,l_2)dl_1dl_2\mid_{\rho=0} &= -\varphi(t_i)\varphi(t_j)\sigma_g^2,\end{aligned}$$

We differentiate $A(\rho;t_i,t_j)$ w.r.t. ρ, and using the previous results and some algebra we get:

$$A'(0;t_i,t_j)=\frac{\varphi(t_i)\varphi(t_j)\sigma_g^2}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}}\Big[1-(P_i+P_j)(1-\frac{K(1-P)}{P(1-K)})+P_iP_j\Big(1-\frac{K(1-P)}{P(1-K)}\Big)^2\Big]=$$

$$\frac{\varphi(t_i)\varphi(t_j)\sigma_g^2}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}}\Big[1-(P_i+P_j)\Big(\frac{P-K}{P(1-K)}\Big)+P_iP_j\Big(\frac{P-K}{P(1-K)}\Big)^2\Big].$$

As a sanity check, set $P_i=P_j=P$ and $t_i=t_j=t$ to get:

$$\frac{\varphi(t)^2\sigma_g^2}{P(1-P)}\Big[1-2P\Big(\frac{P-K}{P(1-K)}\Big)+P^2\Big(\frac{P-K}{P(1-K)}\Big)^2\Big]=\frac{\varphi(t)^2\sigma_g^2}{P(1-P)}\Big(\frac{1-P}{1-K}\Big)^2=\frac{\varphi(t)^2\sigma_g^2(1-P)}{P(1-K)^2},$$

which is the expression derived for the no fixed effects case.

Moreover, the the probability of inclusion of individual i in the study, conditional on the fixed effects is now:

$$K_i+(1-K_i)\frac{K(1-P)}{P(1-K)}.$$

and so:

$$B(0;t_i,t_j)=\Big(K_i+(1-K_i)\frac{K(1-P)}{P(1-K)}\Big)\Big(K_j+(1-K_j)\frac{K(1-P)}{P(1-K)}\Big).$$

Plugging the derived expressions for $A'(0;t_i,t_j),B(0;t_i,t_j)$, we conclude that regressing Z_{ij} on

$$\frac{\varphi(t_i)\varphi(t_j)\Big[1-(P_i+P_j)\Big(\frac{P-K}{P(1-K)}\Big)+P_iP_j\Big(\frac{P-K}{P(1-K)}\Big)^2\Big]}{\sqrt{P_i(1-P_i)}\sqrt{P_j(1-P_j)}\Big(K_i+(1-K_i)\frac{K(1-P)}{P(1-K)}\Big)\Big(K_j+(1-K_j)\frac{K(1-P)}{P(1-K)}\Big)}G_{ij},$$

yields an estimator of heritability on the liability scale.

2.3 Dealing with unknown fixed effects

More often than not, the effects of relevant fixed effects are unknown and must be estimated from the data. However, estimating effect sizes under ascertainment in case-control studies is notoriously problematic. Specifically, under the threshold (probit) model, ignoring the ascertainment yields biased estimators.

A special exception is the case of logistic regression. In their seminal paper, Prentice and Pyke (1979) [15] proved that using a logistic regression to estimate fixed effects from ascertained data yields consistent estimators of these effects in the (unascertained) population, and that the ascertainment only biases the intercept.

We therefore suggest a two-step procedure for estimating heritability. First, we estimate the fixed effects using a logistic regression model. We then correct the effect of the ascertainment, and obtain the individual-specific thresholds. Lastly, we plug the thresholds into the estimation scheme described above.

More elaborately, by Bayes' theorem:

$$P_i = \frac{\mathbb{P}(s_i = 1 \mid y_i = 1; x_i) K_i}{\mathbb{P}(s_i = 1; x_i)},$$

by the complete ascertainment assumption $\mathbb{P}(s_i = 1 \mid y_i = 1, x_i) = 1$, and according to the selection scheme:

$$\mathbb{P}(s_i = 1; x_i) = K_i + \frac{K(1-P)}{P(1-K)}\big(1 - K_i\big).$$

We can thus solve for K_i and express it is a function of P_i:

$$K_i = \frac{\frac{K(1-P)}{P(1-K)} P_i}{1 + \frac{K(1-P)}{P(1-K)} P_i - P_i}$$

We then use logistic regression to obtain $\hat{P}_i$ - a consistent estimator of P_i, and use this estimate to obtain an estimate of K_i, which is in turn used to estimate the threshold:

$$\hat{t}_i = \Phi^{-1}\big(1 - \hat{K}_i\big),$$

and the estimates of the individual-wise thresholds are used for estimating the liability-scale heritability.

2.3.1 Estimating the added variance due to fixed effects

Lastly, the presence of fixed effects increases the variance of the liability, so σ_g^2 no longer equals h^2. The appropriate definition of heritability is now:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2 + \sigma_t^2} = \frac{\sigma_g^2}{1 + \sigma_t^2},$$

where σ_t^2 is the variance of the thresholds in the population, and so the estimate of σ_g^2 can be transformed to an estimate of the heritability simply by dividing it by $1 + \sigma_t^2$. Therefore, obtaining an estimate of the heritability in the presence of fixed-effects requires an estimate of σ_t^2 in the population. To estimate σ_t^2 we use the law of total variance:

$$\sigma_t^2 = V(t) = V(\mathbb{E}(t \mid y)) + \mathbb{E}(V(t \mid y))$$

where y is the phenotype of the individual. Furthermore:

$$V(\mathbb{E}(t \mid y)) = K(1-K)(\mathbb{E}(t \mid y = 1) - \mathbb{E}(t \mid y = 0))^2,$$

and:

$$\mathbb{E}(V(t \mid y)) = KV(t \mid y = 1) + (1-K)V(t \mid y = 0).$$

Once we condition on a specific phenotype, both the expected and variance of T can be estimated from the data, as they are no longer affected by enrichment of cases in the data. Specifically, we use the estimated thresholds $\hat{t}_i$ to estimate the expected value and the variance of the threshold for cases and controls, and plug the estimates into the equations above to obtain $\hat{\sigma}_t^2$.

3 PCGC regression with multiple correlation matrices

It is often desired to estimate the heritability due to several genomic regions simultaneously, or more generally to include more than one variance component in the estimation. PCGC can be easily extended to deal with such situations. Denote G_i the correlation matrix for the i'th variance component (out of p), and denote $\sigma^2_{g_i}$ the narrow-sense heritability due to this component. Consider a pair of individuals j, k and denote $\rho_i = G_{i_{jk}}$, the correlation between these individuals in the i'th variance component. Using these notations, Σ is given explicitly by:

$$\Sigma = \begin{pmatrix} 1 & \sum_{i=1}^p \rho_i \sigma^2_{g_i} \\ \sum_{i=1}^p \rho_i \sigma^2_{g_i} & 1 \end{pmatrix}.$$

The determinant of Σ is $|\Sigma| = 1 - (\sum_{i=1}^p \rho_i \sigma^2_{g_i})^2$ and its inverse is

$$\Sigma^{-1} = \frac{1}{1 - (\sum_{i=1}^p \rho_i \sigma^2_{g_i})^2} \begin{pmatrix} 1 & -\sum_{i=1}^p \rho_i \sigma^2_{g_i} \\ -\sum_{i=1}^p \rho_i \sigma^2_{g_i} & 1 \end{pmatrix},$$

and so the density function $f_{\rho_1,\ldots,\rho_p,\sigma^2_{g_1},\ldots,\sigma^2_{g_p}}$ can be written as:

$$f_{\rho_1,\ldots,\rho_p,\sigma^2_{g_1},\ldots,\sigma^2_{g_p}}(l_1, l_2) = \frac{1}{2\pi\sqrt{1 - (\sum_{i=1}^p \rho_i \sigma^2_{g_i})^2}} e^{-\frac{l_1^2 + l_2^2 - 2l_1 l_2 (\sum_{i=1}^p \rho_i \sigma^2_{g_i})}{2(1 - (\sum_{i=1}^p \rho_i \sigma^2_{g_i})^2)}}.$$

We now proceed as before, but instead of using a first-order Taylor approximation of a univariate function, our function is multivariate. The first-order Taylor approximation would therefore involve the gradient, which requires differentiating the double integrals w.r.t. each of the ρ_i's:

$$\frac{d}{d\rho_i} f_{\rho_1,\ldots,\rho_p,\sigma^2_{g_1},\ldots,\sigma^2_{g_p}}(l_1, l_2) \mid_{\rho_i = 0} = l_1 l_2 \sigma^2_{g_i} \frac{1}{2\pi} e^{-\frac{l_1^2 + l_2^2}{2}}.$$

Computing the integrals in the same manner as before and plugging the results into the approximation we get:

$$\mathbb{E}[Z_{ij} \mid \mathcal{S} = 1; \rho] = \frac{P(1-P)}{K^2(1-K)^2} \varphi(t)^2 \sum_{i=1}^p \sigma^2_{g_i} \rho_i + \sum_{i=1}^p \mathcal{O}(\rho_i^2).$$

Hence, to incorporate multiple variance components, we simply run a multiple regression of the Z_{ij} variable on the respective entries of the correlation matrices $G_1, \ldots G_p$ and transform the obtained coefficients as just as in the case of the univariate PCGC.

4 Extreme phenotype studies

The same regression idea can be used to estimate heritability in extreme phenotype studies. To illustrate this point, assume we are genotyping only individuals with extreme values of the phenotype. More specifically, assume we are genotyping the K_1 individuals with the lowest values of the phenotype, and the K_2 individuals with the highest values of the phenotype. In other words, we take individuals whose phenotype's Z-score is below $t_1 = Z_{K_1}$ or above Z_{1-K_2}, where Z_X is the X'th quantile of the standard normal distribution.

Using the previous notations we now write:

$$\mathbb{E}\big[y_i y_j \mid \mathcal{S} = 1; \rho, \sigma^2_g\big] = \frac{\mathbb{E}\big[y_i y_j; \mathcal{S} = 1, \rho, \sigma^2_g\big]}{\mathbb{P}(\mathcal{S} = 1; \rho, \sigma^2_g)},$$

and denote the numerator and denominator as usual by $A(\rho)$ and $B(\rho)$ respectively. As before, these expressions have no closed form solution. we approximate the expectation on the left using a Taylor series around $\rho = 0$. The two terms of the Taylor series take the form:

$$\frac{A(0)}{B(0)} + \frac{A'(0)B(0) + B'(0)A(0)}{B(0)^2} \rho.$$

We start with B:

$$B(\rho) = \int_{-\infty}^{t_1}\int_{-\infty}^{t_1} f_{\rho,\sigma_g^2}(y_i,y_j)dy_idy_j + 2\int_{-\infty}^{t_1}\int_{t_2}^{\infty} f_{\rho,\sigma_g^2}(y_i,y_j)dy_idy_j +$$

$$\int_{t_2}^{\infty}\int_{t_2}^{\infty} f_{\rho,\sigma_g^2}(y_i,y_j)dy_idy_j.$$

Clearly, for $\rho = 0$ we have $B(0) = (K_1+K_2)^2$. Using previous results we get:

$$B'(0) = \sigma_g^2\Big[\varphi(t_2)-\varphi(t_1)\Big]^2.$$

Moving to A, we have:

$$A(\rho) = \int_{-\infty}^{t_1}\int_{-\infty}^{t_1} y_iy_jf_{\rho,\sigma_g^2}(y_i,y_j)dy_idy_j + 2\int_{-\infty}^{t_1}\int_{t_2}^{\infty} y_iy_jf_{\rho,\sigma_g^2}(y_i,y_j)dy_idy_j +$$

$$\int_{t_2}^{\infty}\int_{t_2}^{\infty} y_iy_jf_{\rho,\sigma_g^2}(y_i,y_j)dy_idy_j.$$

Using previous results again we get:

$$A'(0) = \Big[-\varphi(t_1)t_1+\Phi(t_1)\Big]^2 + 2\Big[-\varphi(t_1)t_1+\Phi(t_1)\Big]\Big[\varphi(t_2)t_2+(1-\Phi(t_2))\Big] + \Big[\varphi(t_2)t_2+(1-\Phi(t_2))\Big]^2$$

$$= \Big[-\varphi(t_1)t_1+K_1+\varphi(t_2)t_2+K_2\Big]^2\sigma_g^2.$$

Setting $\rho = 0$, the phenotypes are now i.i.d and so $A(0) = \mathbb{E}[y, s=1]^2 = \Big[\varphi(t_2)-\varphi(t_1)\Big]^2$.
It thus follows that:

$$\mathbb{E}\big[y_iy_j \mid \mathcal{S}=1;\rho,\sigma_g^2\big] = \frac{A(0)}{B(0)} + \frac{A'(0)B(0)-B'(0)A(0)}{B(0)^2}\rho + \mathcal{O}(\rho^2) =$$

$$\Big(\frac{\varphi(t_2)-\varphi(t_1)}{K_1+K_2}\Big)^2 + \frac{\Big[-\varphi(t_1)t_1+K_1+\varphi(t_2)t_2+K_2\Big]^2(K_1+K_2)^2-\Big[\varphi(t_2)-\varphi(t_1)\Big]^4}{(K_1+K_2)^4}\sigma_g^2\rho + \mathcal{O}(\rho^2)$$

One common complication is that one of the tails may be over-sampled (for example if two studies of different sizes are merged, where each study focused on a different tail of the distribution). Formally, we sample P_1 of the bottom K_1 and P_2 of the top K_2. We can then write:

$$A(\rho) = P_1^2\int_{-\infty}^{t_1}\int_{-\infty}^{t_1} y_iy_jf_{\rho,\sigma_g^2}(y_i,y_j)dy_idy_j + 2P_1P_2\int_{-\infty}^{t_1}\int_{t_2}^{\infty} y_iy_jf_{\rho,\sigma_g^2}(y_i,y_j)dy_idy_j +$$

$$+P_2^2\int_{t}^{\infty}\int_{t}^{\infty} y_iy_jf_{\rho,\sigma_g^2}(y_i,y_j)dy_idy_j,$$

and similarly:

$$B(\rho) = P_1^2\int_{-\infty}^{t_1}\int_{-\infty}^{t_1} f_{\rho,\sigma_g^2}(y_i,y_j)dy_idy_j + 2P_1P_2\int_{-\infty}^{t_1}\int_{t_2}^{\infty} f_{\rho,\sigma_g^2}(y_i,y_j)dy_idy_j +$$

$$P_2^2\int_{t_2}^{\infty}\int_{t_2}^{\infty} f_{\rho,\sigma_g^2}(y_i,y_j)dy_idy_j,$$

so

$$B(0) = (K_1P_1 + P_2K_2)^2$$

and

$$B'(0) = \sigma_g^2\Big[P_2\varphi(t_2) - P_1\varphi(t_1)\Big]^2.$$

Similarly:

$$A'(0) = \Big[P_1(-\varphi(t_1)t_1 + K_1) + P_2(\varphi(t_2)t_2 + K_2)\Big]^2,$$

and

$$A(0) = \Big[P_2\varphi(t_2) - P_1\varphi(t_1)\Big]^2.$$

Hence:

$$\mathbb{E}\big[y_iy_j \mid \mathcal{S} = 1; \rho, \sigma_g^2\big] = fracA(0)B(0) + \frac{A'(0)B(0) - B'(0)A(0)}{B(0)^2}\rho + \mathcal{O}(\rho^2) =$$

$$\Big(\frac{R\varphi(t_2) - \varphi(t_1)}{K_1 + RK_2}\Big)^2 + \frac{\Big[-\varphi(t_1)t_1 + K_1 + R\varphi(t_2)t_2 + RK_2\Big]^2(K_1 + RK_2)^2 - \Big[R\varphi(t_2) - \varphi(t_1)\Big]^4}{(K_1 + K_2)^4} + \mathcal{O}(\rho^2),$$

where $R = \frac{P_2}{P_1}$.

We note that the phenotype is assumed to have mean 0 and variance 1 *in the population*. Typically, the mean and variance of the phenotype are well-known, otherwise it criteria for inclusion in the study are hard to establish. In this case standardization is straightforward.

5 Simulations

5.1 Description of the Lee et al. simulation scheme

Lee et al. perform simulations of ascertained case-control studies in the following manner:

Phenotypes are simulated in blocks of 100 individuals. Given the true genetic variance σ_g^2, liabilities for each block are sampled from a multivariate Gaussian distribution with mean 0 and covariance matrix $\Sigma = G\sigma_g^2 + I(1 - \sigma_g^2)$, where G is given by:

$$G_{ij} = \begin{cases} 1 & i = j \\ 0.05 & i \neq j \end{cases},$$

that is – individuals within each block have a genetic correlation of 0.05 while individuals in different blocks are perfectly unrelated. Individuals with liabilities higher than the threshold t are considered cases.

Cases are always included in the study. Controls are included in the study with probability $\frac{K}{1-K}$. Blocks are generated until 100 individuals are accumulated, and this process is repeated 100 times, so in total $10,000$ cases and controls are accumulated.

5.2 Discussion of problems with the simulations in Lee et al.

There are several key aspects in which the simulations of Lee et al. differ from the true generative process of the data.

First, the resulting correlation matrix is highly degenerate, with most of the correlations being 0 – as many as 99.99%. Since the underlying idea behind heritability estimation from unrelated individuals is leveraging the minor – but non-zero – correlations among unrelated individuals, restricting the simulations to largely degenerate correlation structures is highly unrealistic and counterproductive.

Second, in reality, correlations between individuals span a wide range of values. while in Lee et al.'s simulations the correlations are either 0 or 0.05. Hence negative correlations are impossible, and the expected correlation is larger than 0, contrary to the unrelated individuals assumption.

Thirdly, the selection procedure, in which cases are up-sampled compared to their prevalence in the population, results in cases being more similar genetically than controls in real ascertained studies. This is a well known phenomenon which has recently attracted considerable attention in the context of GWAS [29]. This is hardly captured by the simulation process of Lee et al. For example, cases from different blocks always have 0 genetic correlation between them in Lee et al.'s simulations, while in a realistic scenario they should have positive genetic correlation, whose magnitude depends on the heritability of the phenotype and the number of causative loci.

Lastly, for small enough prevalence, all blocks in a typical simulation would contain only a single individual. In this case the genetic effects would be completely interchangeable with the environmental effects since $G = I$ and so heritability would be impossible to estimate. In reality, however, a very small prevalence should yield very closely related individuals, thus generating an intuitively easier estimation problem.

5.3 Simulations using a generative model

To see if these problems have a major impact, we ran simulations using the full generative model. This was done as follows:

1. The MAFs of 10,000 SNPs were randomly sampled from $U[0.05, 0.5]$.
2. SNP effect sizes were randomly sampled from $N(0, \frac{\sigma_g^2}{m})$.
3. For each individual, we:
   (a) Randomly generated a genotype using the MAFs, and normalized it (according to Yang et al.'s model [25]).
   (b) Used the genotype and the effect sizes, to compute the genetic effect.
   (c) Sampled an environmental effect from $N(0, 1 - \sigma_g^2)$.
   (d) Computed liability and phenotype.
   (e) If the phenotype was a case, the individual was automatically included in the study. Otherwise the individual was included in the study with probability $\frac{K(1-P)}{P(1-K)}$.
4. Step (2) was repeated until enough individuals were accumulated (4,000).
5. The genotypes of all included individuals were used to compute $G = \frac{ZZ^\intercal}{m}$. where Z is the matrix of normalized genotypes.
6. We used REML, as implemented in the GCTA software [26], to estimate heritability with G as the genetic correlation matrix, and a $0 - 1$ vector of the phenotypes as input.

We ran ten repetitions of this simulation for all combinations of $\sigma_g^2 \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $P \in \{0.1, 0.3, 0.5\}$ and $K \in \{0.001, 0.005, 0.01\}$ (we also simulated data for $P = 0.7$, which is displayed in the main text). The average heritabiltiy estimates for all combinations of are given in figure 1, displaying a clear downwards bias of the heritabilty estiamtes obtained by applying REML and correcting for the effects of ascertainment as described in [11]. We note that simulated sets with less ascertainment (high K, low P) are less biased, most notably $P = 0.1$ and $K = 0.01$ - the yellow line in figure 1.

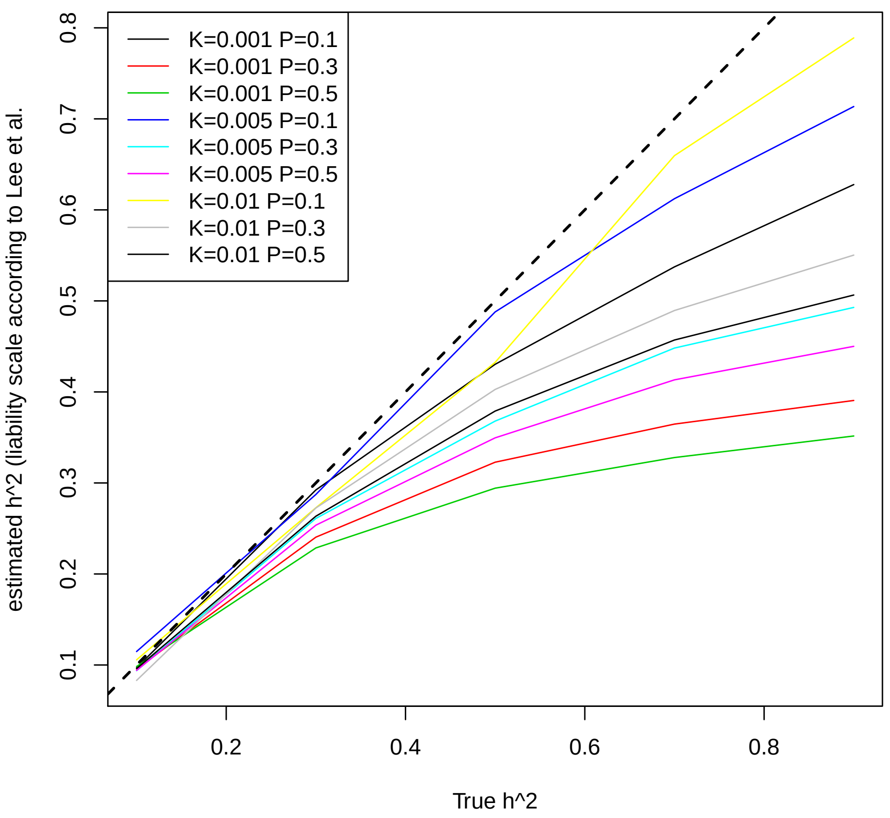


Figure 1: Estimated heritability on the liability scale using REML for various values of σ_g^2, K and P, demonstrating the non-linearity of $\hat{h_g^2}$ (as a function of σ_g^2), even after correcting the estimate as per Lee et al. The dashed line is $y = x$, i.e. the true heritability.

We then applied PCGC to the same set of simulated data. The results are portrayed in figure 2. As can clearly be seen from the figure, PCGC estimators are unbiased. Moreover, as can be seen in 4, PCGC estimates are not only unbiased, but also considerably more accurate than REML estimates when comparing their mean square errors (MSEs). Figure 3 gives the MSEs for corresponding to figure 2A in the main text.

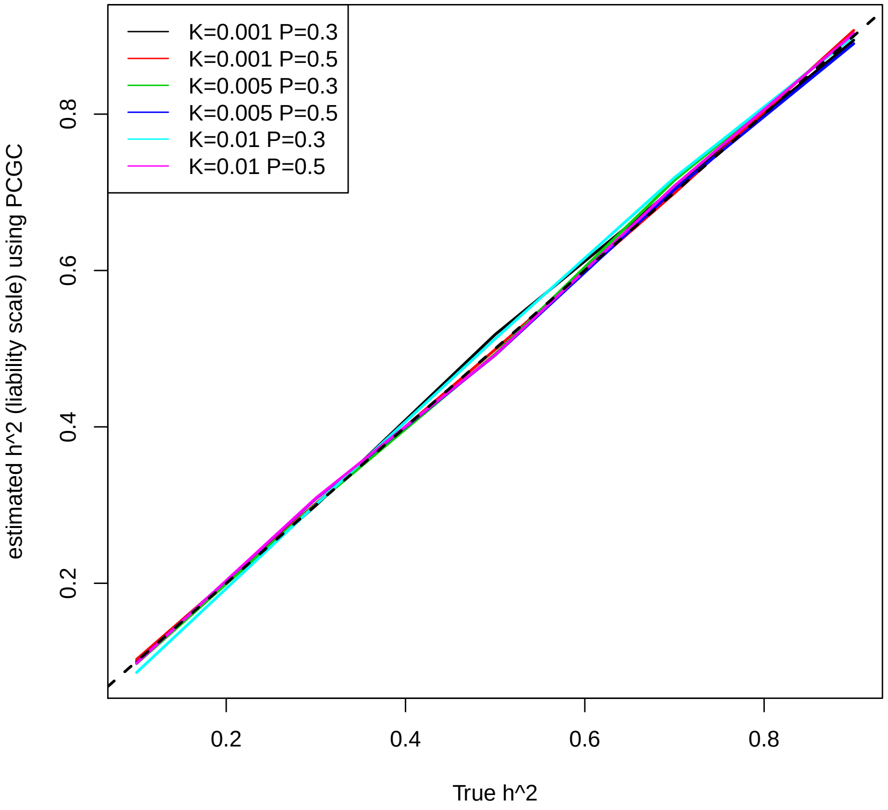


Figure 2: Corrected average estimates obtained by PCGC using the same simulations as in figures 1-2. The dashed line is $y = x$, indicating the estimators are indeed unbiased.

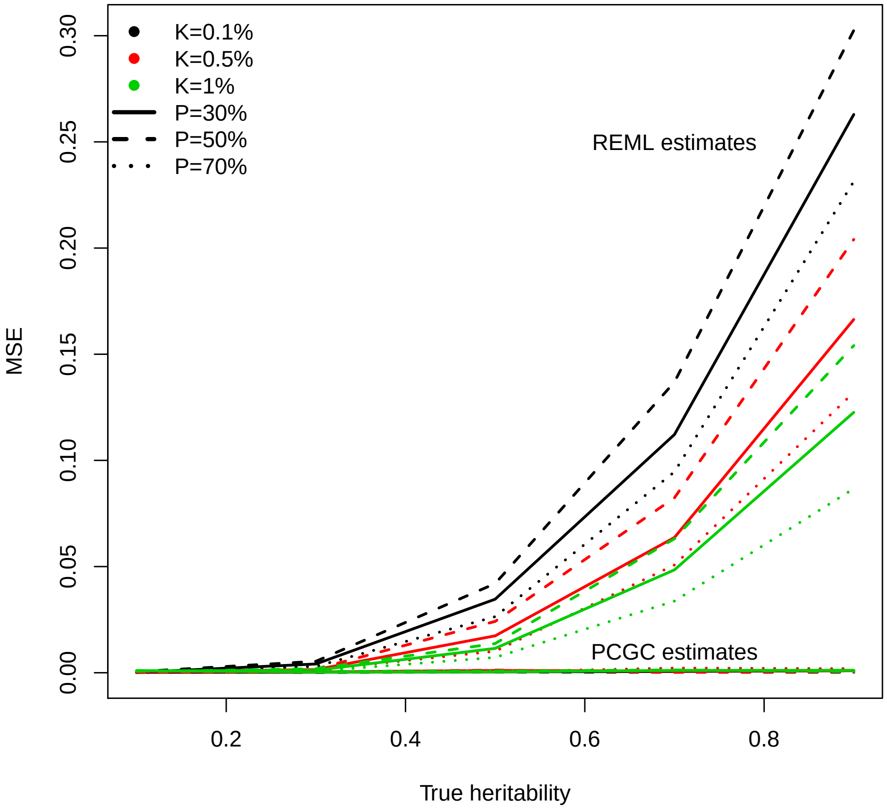
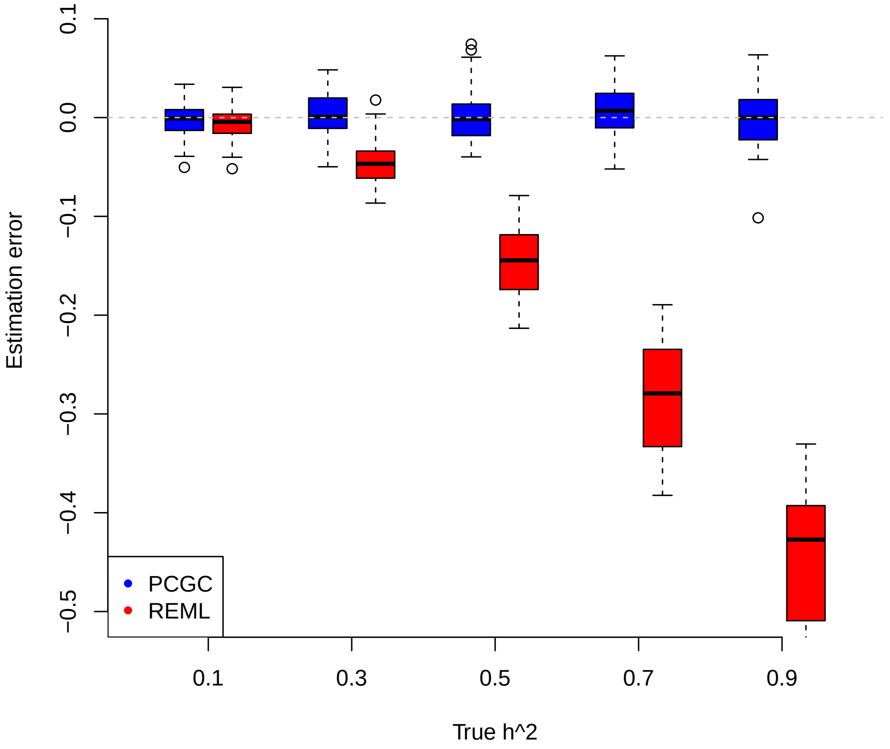


Figure 3: MSEs estimated from the simulations of main-text figure 2A, indicating that PCGC regression estimates are not only unbiased, but also considerably more accurate than REML estimates.

5.3.1 Distribution of estimates around the mean

Apart from being unbiased and having lower MSEs, PCGC estimates also display an equal or lower variance to the REML estimates. This can be seen in figure 4.



x

Figure 4: Distribution of estimation errors for various values of h^2. Estimates for different values of K and P are grouped together.

5.4 Additional simulations

5.4.1 Weak ascertainment

To test the robustness of our method to the different values of P and K, we ran a wide range of simulations, with less extreme ascertainment. The results are given in figure 5, showing how the downward-bias of the REML estimates is smaller for less ascertained studies, and disappears completely when the studies are non-ascertained. in both cases the PCGC yields unbiased estimates.

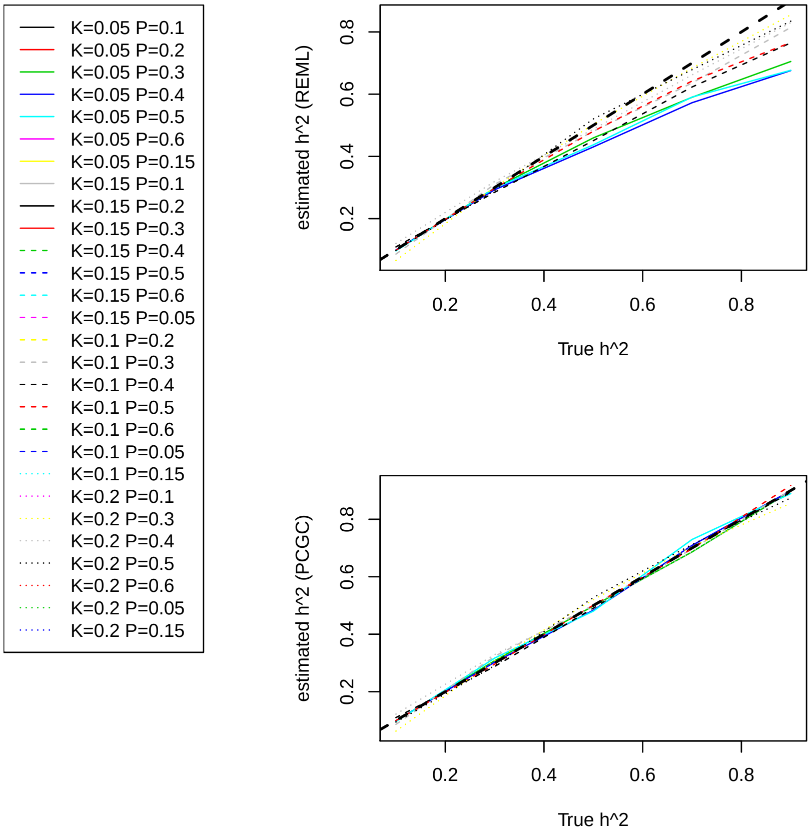


Figure 5: Comparing corrected REML estimates (top) to PCGC estimates (bottom) for studies with low-to-intermediate ascertainment.

5.4.2 non-ascertained simulations

We also applied our method to non-ascertained simulations, where both the REML and PCGC methods produce unbiased estimators, as can be seen in figure 6.

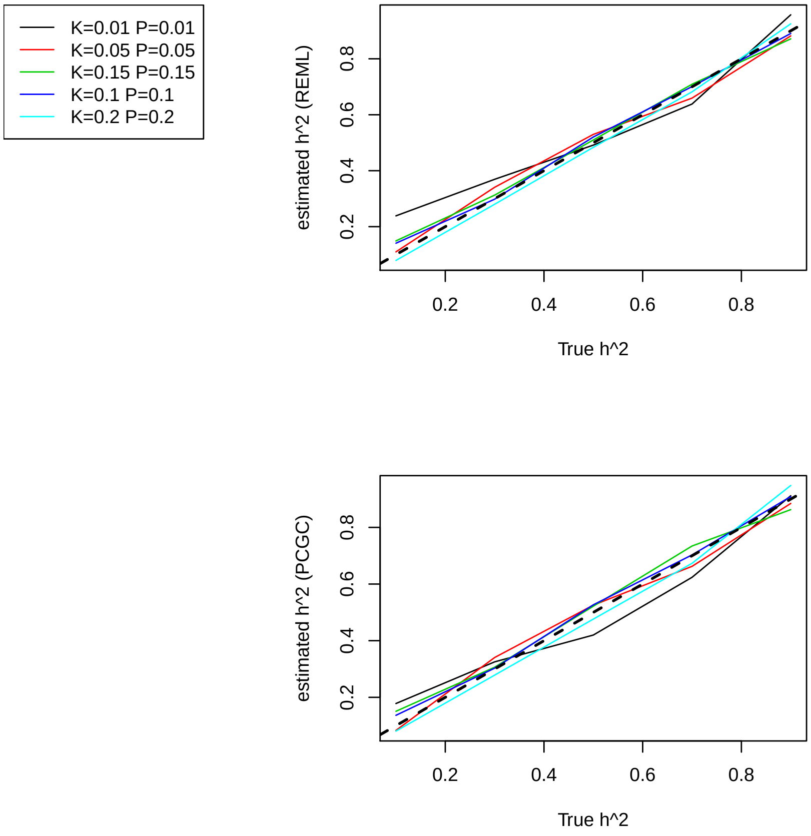


Figure 6: Comparing corrected REML estimates (top) to PCGC estimates (bottom) for studies with no ascertainment. Both methods yield unbiased estimates, except for low values of heritability, where the positivity constraint introduces a mild positive bias.

5.4.3 Simulations with population structure

To study the robustness of our method to the presence of subtle population structure, we modified our simulations to introduce population structure similar to the simulations in Lee et al. For every batch of 100 individuals, we randomly selected 5% of the SNPs and fixed them for the entire batch. By doing so, the expected correlation within each batch is 0.05, as in Lee et al. while the actual realized correlations are still realistic. The process of phenotype generation and individual selection then proceeds as described earlier. We ran simulations with $\sigma_g^2 \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $P \in \{0.1, 0.3, 0.5\}$ and $K \in \{0.001, 0.005, 0.01\}$. The resulting PCGC estimates were unbiased for all these scenarios, as can be seen in figure 7.

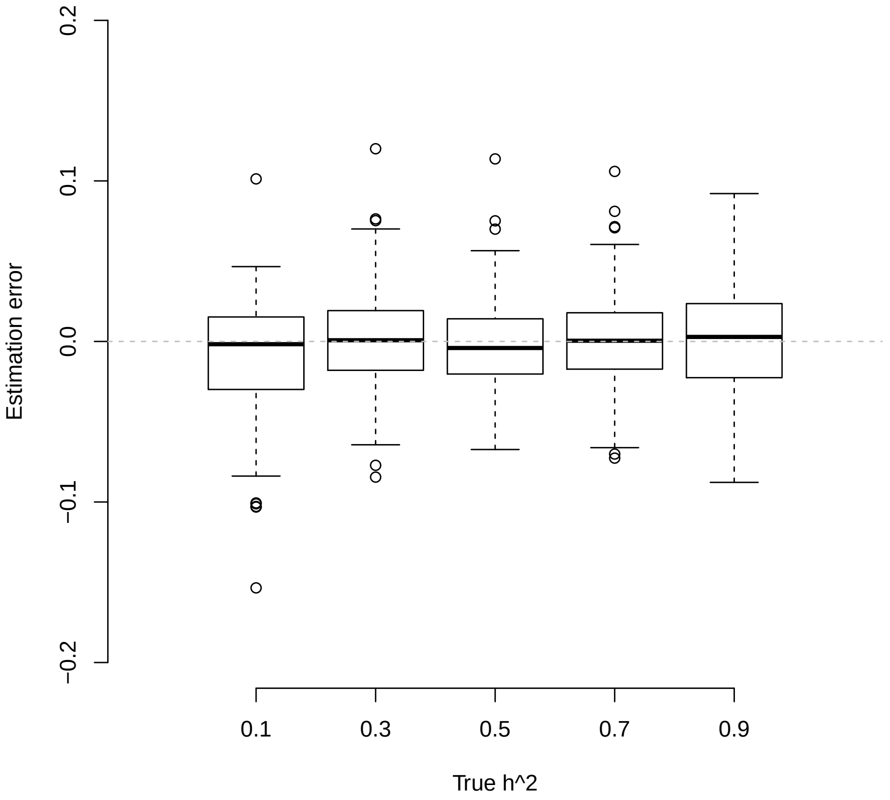


Figure 7: Estimation errors for simulations generated with mild population structure. PCGC estimates are still unbiased in the presence of such structure.

5.4.4 Simulations with different numbers of SNPs

The number of SNPs used in the simulations determines the variance of the realized genetic correlations between individuals. Since PCGC relies on a first-order approximation around zero correlation, high variance of genetic correlations might decrease the accuracy of PCGC. To study the robustness of PCGC regression to such changes, we reran our initial simulations using 1,000 or 100 SNPs. PCGC still produced unbiased estimates. As expected, when the number of SNPs was smallest (100) and the heritability was highest (0.9), the results displayed considerably higher variance. The results are summarized in figure 8. Observing the results in figure 8, one might suspect that PCGC regression estimates are slightly biased downwards when the number of SNPs is small and the heritability is high. To further test this point we re-ran 80 simulations setting the number of SNPs to 100, and setting either $h^2 = 0.7$ or $h^2 = 0.9$. The average estimate for $h^2 = 0.7$ was 0.699 (s.e.= 0.03) and for $h^2 = 0.9$ it was 0.899 (s.e.= 0.026). We thus conclude that even in this extreme setup, PCGC regression estimates are unbiased.

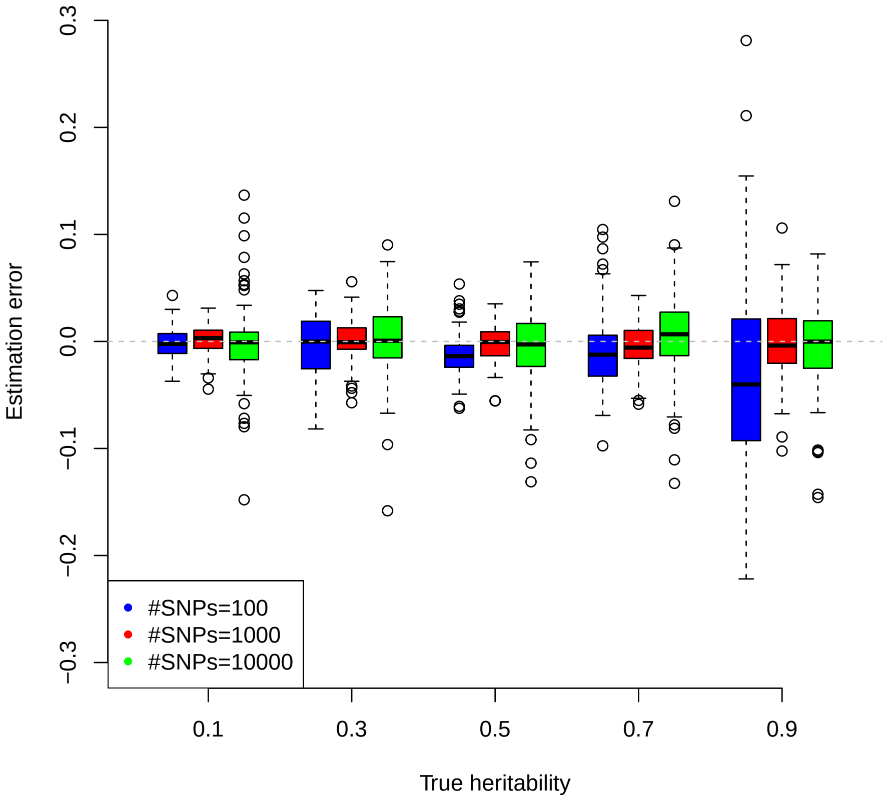


Figure 8: Comparing PCGC estimates for various simulation scenarios with different number of SNPs used to simulate the phentoypes and compute the genetic correlations.

5.4.5 Second order simulations

To study the accuracy of PCGC, we implemented a version of PCGC which utilizes the second order approximation as detailed before. We then compared the first- and second-order estimates of heritability on our initial set of simulated data. The resulting second order estimates were almost identical to the results obtained by the first order approximation (figure 9).

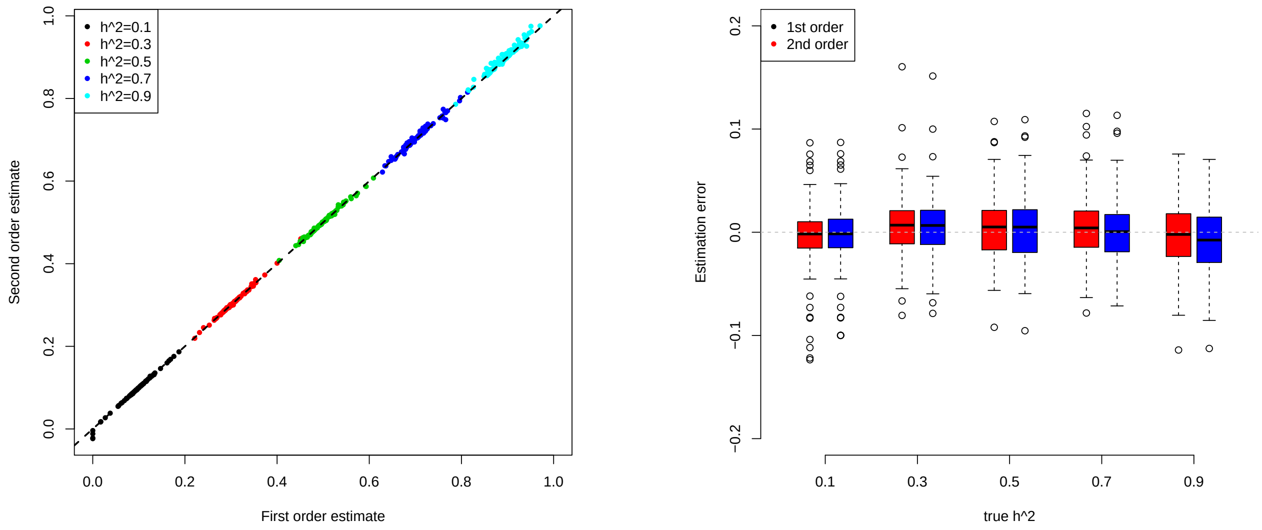


Figure 9: Comparing the first and second order PCGC heritability estimates. Left: Plotting the 1st vs. 2nd order estimates for each simulation run. Right: Comparing estimation errors across different simulation setups.

5.4.6 Simulations with changing study size and number of SNPs

Following [27], we ran simulations with an increasing number of samples in the study, while fixing all the other parameters: $K = 1\%$, $P = 30\%$, $h^2 = 50\%$. Our simulations (Figure 10) demonstrated that the bias of REML estimates does indeed increase with sample size, as suggested by [27]. In contrast, PCGC-regression estimates remain unbiased. We also conducted simulations with varying number of SNPs. Here, too, the our results agreed with the results of [27], and demonstrated that the bias decreases as the number of SNPs increase. Additionally, our results seem to agree with the observation of [27] that the magnitude on the bias depends on the ratio of the number of samples to the (effective) number of SNPs. Importantly, the effective number of genotyped SNPs, i.e. the number of SNPs after accounting for LD, which determines the behavior of the genetic correlation matrix [14], is not expected to increase indefintely, contrary to the increasing size of GWAS. We therefore find that the dependence of the bias on the number of SNPs is of secondary importance.

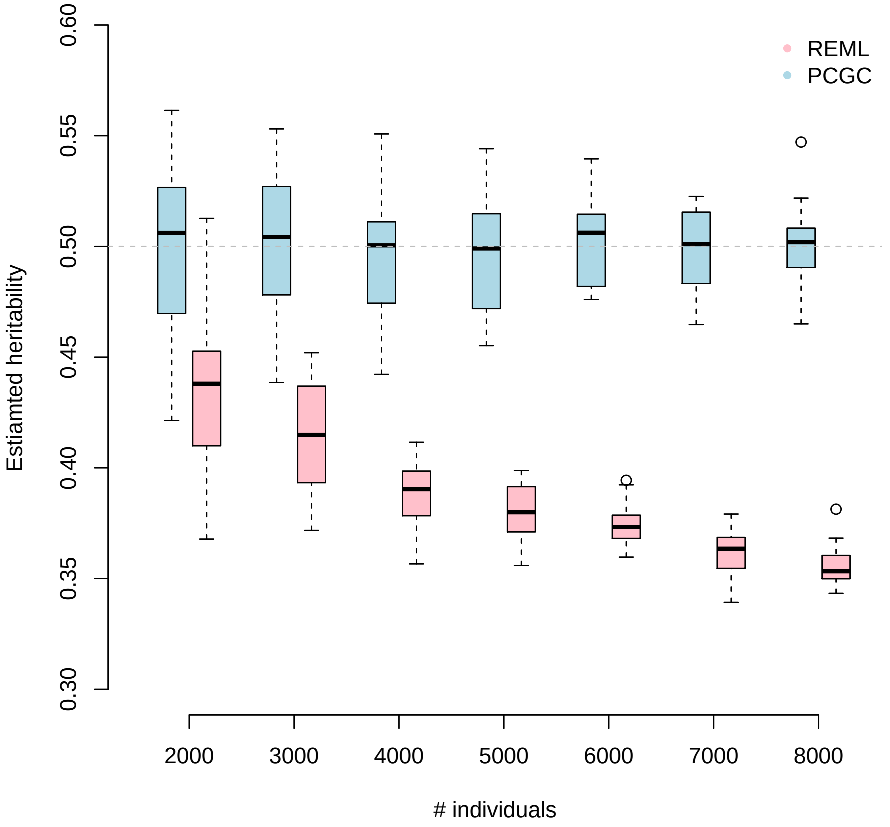


Figure 10: Performance of REML and PCGC regression with fixed parameters and increasing study size. We simulated case-control studies with $K = 1\%$,$P = 30\%$ and $h^2 = 50\%$, with study sizes increasing from $2,000$ to $8,000$ in steps of $1,000$. We then estimated the heritability, for 20 simulations, using either REML or PCGC regression. The results demonstrate that the bias of REML estimates increases with the size of the study, while PCGC regression estimates remain unbiased

5.5 Simulations with different levels of polygenicity

Following [30], we ran simulations using the same parameters as in the previous section (with $n = 4,000$), but where not all SNPs were causal. Instead, only 10% or 1% had non-zero effects and the rest of the SNPs had zero effect on the phenotype. Our simulations demonstrated that PCGC regression still yields unbiased estimates in this setup (Figure 11), and that these estimates are more accurate than REML estimates when comparing MSE (Figure 12).

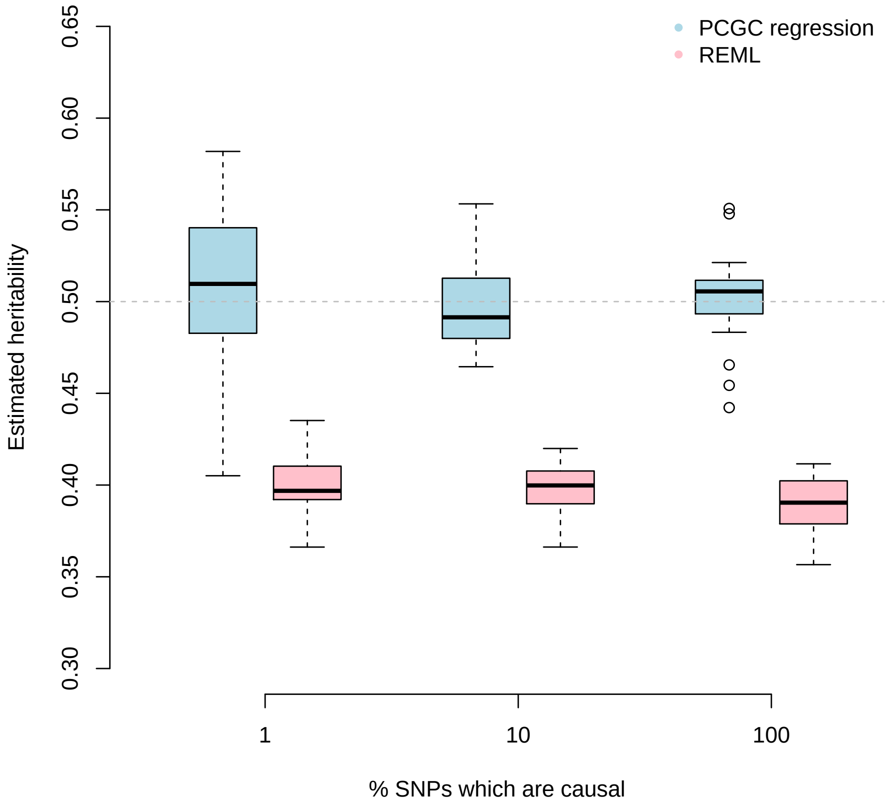


Figure 11: Performance of REML and PCGC regression with varying degrees of polygenicity. We simulated case-control studies with $K = 1\%$,$P = 30\%$, $h^2 = 50\%$, and $n = 4,000$. We simulated $10,000$ SNPs in linkage equilibrium, but the set of causal SNPs was either 1%, 10% or 100% of these SNPs. We then estimated the heritability, for 30 simulations, using either REML or PCGC regression. The results demonstrate that PCGC regression estimates remain unbiased in these setups.

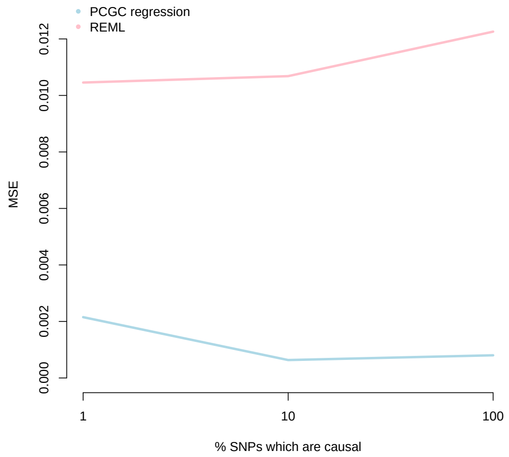


Figure 12: Comparison of the MSE of REML and PCGC regression with varying degrees of polygenicity (see Figure 11 for details).

5.6 Simulations with fixed effects

When addressing the problem of heritability estimation in the presence of fixed effects, it is important to distinguish two different tasks: (i) estimating the actual heritability, namely $\frac{\sigma_g^2}{\sigma_l^2}$, and (ii) estimating the actual parameter σ_g^2. While the former is our focus of interest, the latter can serve to demonstrate the We therefore tested PCGC-regression's ability to handle fixed effects in three scenarios as detailed below.

Normally distributed fixed effect: Liabilities were generated by:

$$l_i = \beta x_i + g_i + e_i,$$

with x_i $N(0,1)$, so the overall variance of the liability is $1+\beta^2$. The heritability was set to 50%, so $\sigma_g^2 = \frac{1}{2}(1+\beta^2)$. The threshold was adjusted so that the prevalence of the disease in the population was 1%, and each simulated study contained 30% cases on average, and a study size of 4,000. The heritability was set to 50%, with $K = 1\%$, and 10,000 SNPs in linkage equilbrium were used in each simulation.

It is important to note that in this setup, not accounting for the fixed effect is not expected to bias the estimated heritability, since the fixed effect and the environmental effect are interchangeable. Consider for example a fixed effect with variance of 1. In this case, the overall variance of the liability is 2, and so, to achieve heritability of 50%, we need to set $\sigma_g^2 = 1$. This is equivalent to setting $\sigma_g^2 = \sigma_e^2 = 0.5$, without any fixed effects. Our simulations demonstrate that PCGC regression provides an unbiased estimate of both the heritability and σ_g^2 in the presence of fixed effects (Table 1).

β^2 (%)	True σ_g^2 (%)	Mean estimated σ_g^2 (s.e.) (%)	Mean estimated h^2(%)
1	50.5	51 (3.4)	50.5 (3.4)
5	52.5	51.8 (3)	49.4 (2.9)
10	55	55.4 (2.9)	50.1 (2.5)
25	62.5	62.4 (2.9)	50.1 (2.8)

Table 1: Estimated values of σ_g^2 and h^2, averaged over 20 simulations, for various values of β. Our simulations demonstrate that PCGC regression estimates σ_g^2 correctly in the presence of fixed effects.

Non-normally distributed continuous fixed effect: We simulated data using a similar scheme as before, but with x_i following Student's-T distribution with 3 degrees of freedom. We then estimated the heritability using REML and PCGC regression, each time with and without including the fixed effect in the analysis. The results are shown in Figure 13. Our simulations demonstrate that PCGC regression successfully accounts for the fixed effects. Without accounting for the fixed effect, PCGC regression yields biased estimates, and the magnitude of the bias increases with the variance of the fixed effect. Similarly, the bias of REML estimates increases with the variance of the fixed effect. When fixed effects are included, the bias of REML estimates is slightly mitigated.

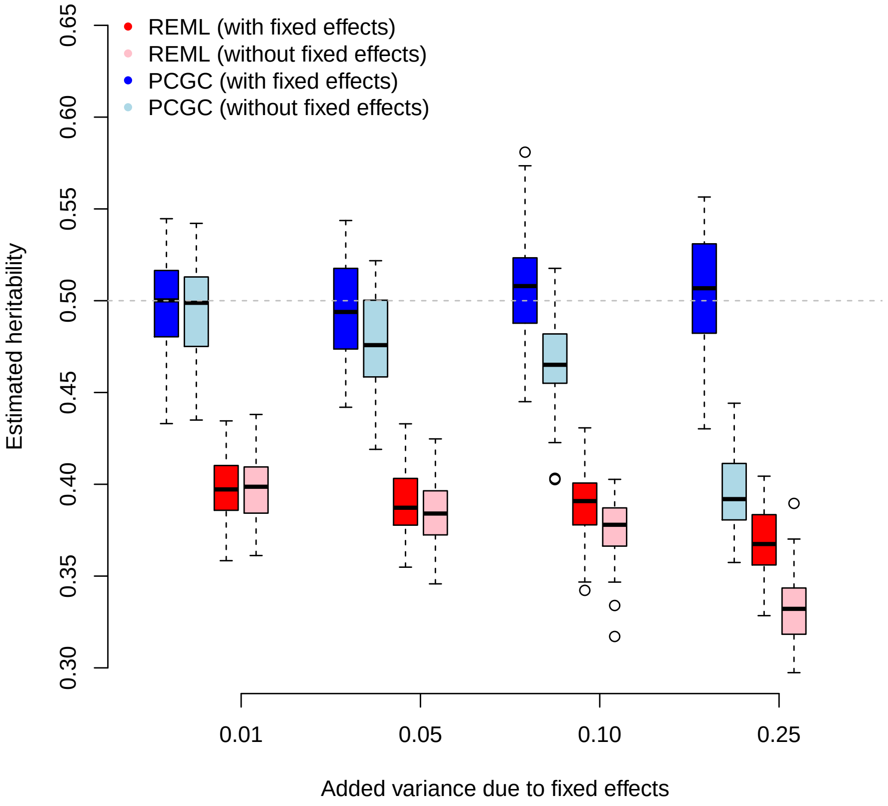


Figure 13: Estimation of heritability in the presence of a T-distributed continuous fixed effect. PCGC regression yields unbiased estimates when accounting for the fixed effect, while all other methods yield biased estimates. The bias increases as the effect of the fixed-effect on the liability increases.

Dichotomous fixed effect: for each simulated individual we generated a 0-1 ”sex" variable with probability 0.5 for each value. We fixed the heritability to 0.5 and the proportion of cases in the study to 0.3, and defined the risk of individuals with sex = 0 to 0.005. The risk of individuals with sex = 1 was either $0.005, 0.01$ or 0.02, i.e. the relative risk (RR) was $1, 2$ or 4. The thresholds were adjusted to depend on the sex accordingly. Phenotypes were generated after accounting for the fixed effects by way of changing the threshold as described earlier, and individuals were selected for the study based only on their phenotypes (i.e. independently of their sex). We then simulated 100 sets of genotypes/phenotypes/sex for each of the three possible RR values. We then estimated the heritability using REML and PCGC regression, each time with and without including the fixed effect in the analysis. The results are displayed in figure 14.

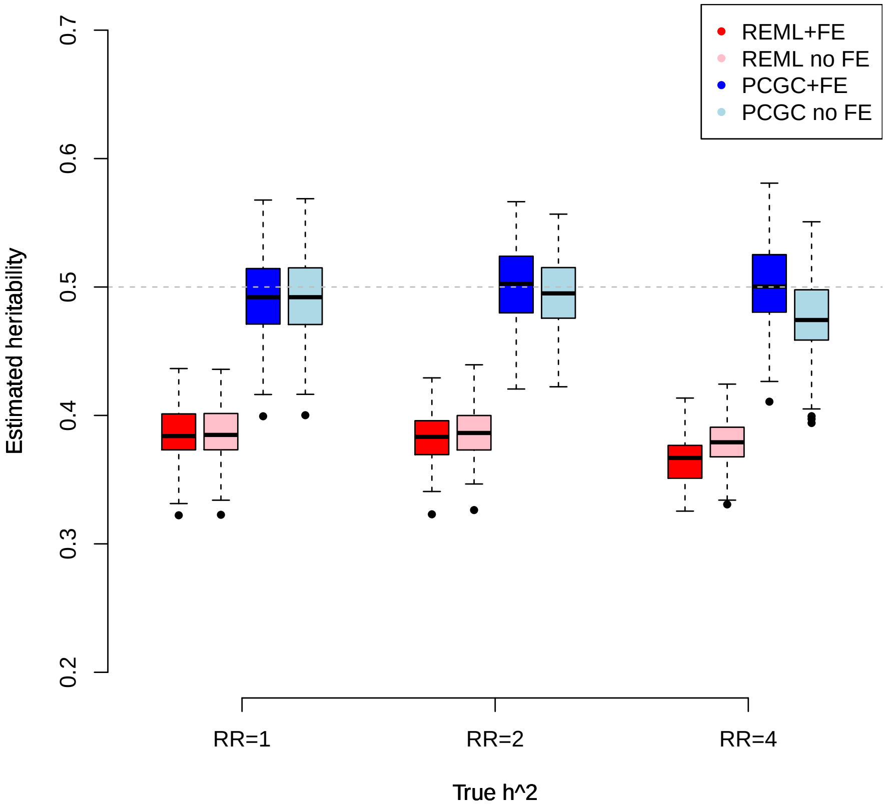


Figure 14: Estimation of heritability with and without accounting for fixed effects using REML and PCGC. The true underlying heritability is 0.5 and the only method which is consistently unbiased is PCGC with fixed effects.

5.7 Extreme phenotype simulations

The effects of extreme phenotype sampling on the marginal and joint distributions of the genetic and environmental effects are portrayed in figure 15. We simulated extreme phenotype studies with parameters similar to the real data at hand (see real data results below). Individuals at the bottom 25% of the phenotype distribution were included with probability 25%, while individuals at the top 10% were included with probability 1 (i.e. $K_1 = 0.1, K_2 = 0.25, R = 0.25$). Genotypes and phenotypes were selected until 400 individuals were accumulated. The phenotypes were generated using $\sigma_g^2 \in 0.3, 0.5, 0.7$. For each heritability value we generated 100 simulations. Figure 16 demonstrates that PCGC estimates are unbiased in these setups.

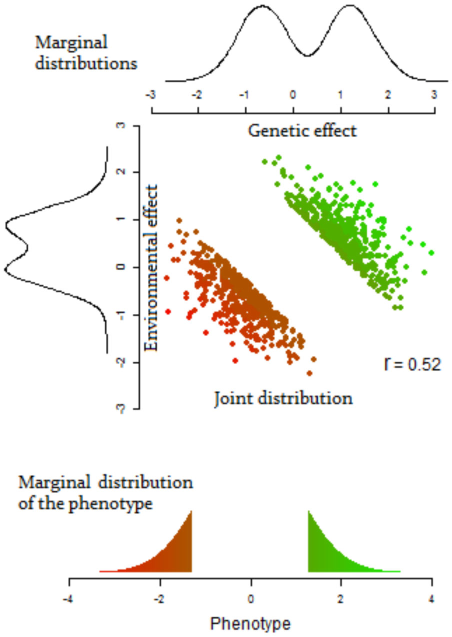


Figure 15: Quantitative trait in an extreme-phenotype study. A quantitative trait follows the same distribution as in figure 1A of the main text, but only the top and bottom deciles are included in the study. The top and left panels show the marginal distributions of the genetic and environmental effects, respectively. The middle panel shows the joint distribution of the genetic and environmental effects, and the bottom panel shows the marginal distribution of the phenotype in the study. Similarly to case-control studies, oversampling of extreme values of the phenotype results in non-normality of the genetic and environmental effects, and an induced positive correlation between g and e.

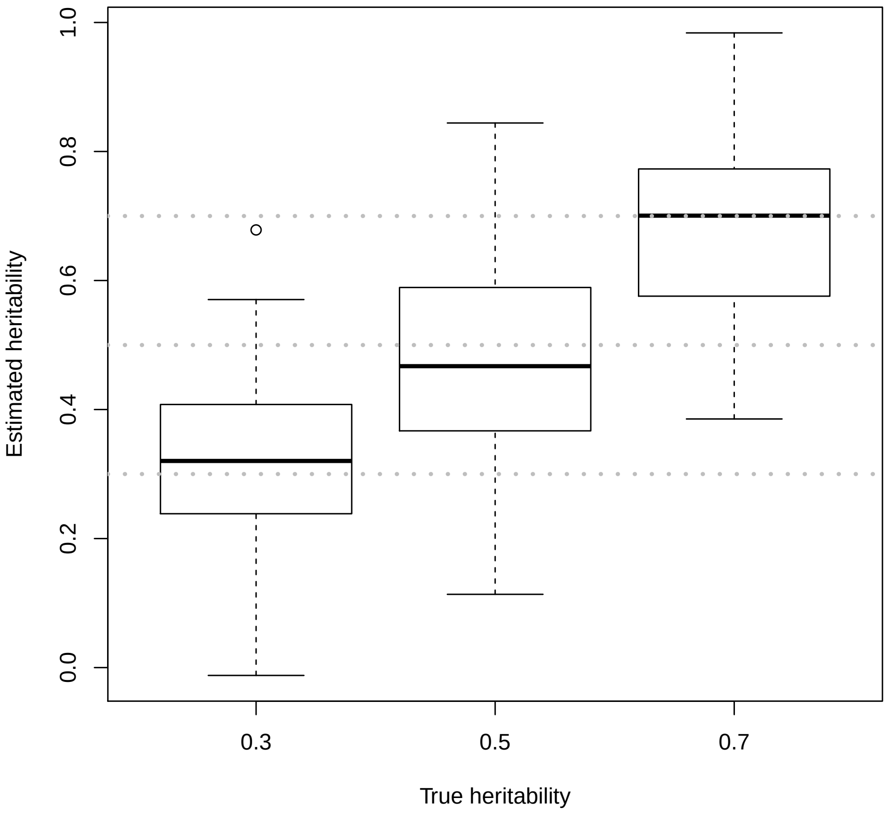


Figure 16: Estimation of heritability in extreme phenotype studies. Extreme phenotype studies were simulated using $K_1 = 0.1, K_2 = 0.25, R = 0.25, n = 400$ and $\sigma_g^2 \in 0.3, 0.5, 0.7$, and estimated using the appropriate PCGC variant. The results demonstrate that PCGC yields unbiased estimates in this setup as well.

We then compared the variance of the heritability estimates of extreme phenotype studies, to the variance of random sampling studies. To do so we simulated extreme phenotype studies of the top and bottom deciles ($K_1 = K_2 = 0.1, R = 1$) with $n = 500$, and random sampling studies ($K_1 = K_2 = 0.5, R = 1$) with $n \in 500, 1000, 2000$. In all setups σ_g^2 was set to 0.5. Our results suggest that in this setup, random sampling requires roughly four times more individuals to achieve estimation error similar to sampling of the top and bottom deciles. The results are illustrated in Figure 17.

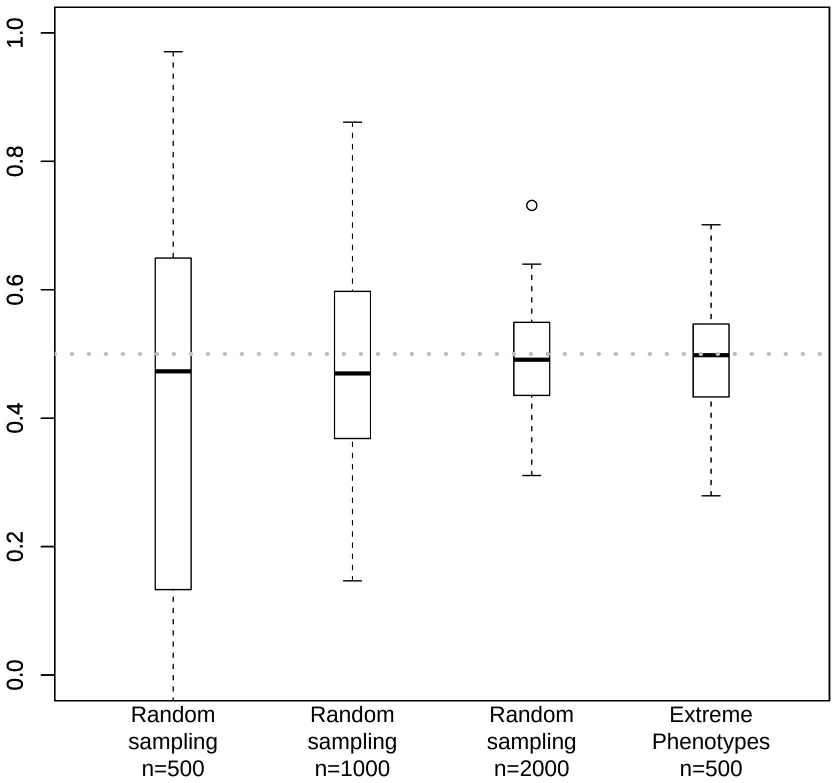


Figure 17: Comparison of variance of extreme-phenotype and random sampling studies. We simulated extreme phenotype studies using $K_1 = K_2 = 0.1, R = 1, n = 500$, and compared them to random sampling studies with $n \in 500, 1000, 2000$. Our results suggest that random sampling requires roughly 4 times more individuals to achieve similar variance of the estimators as in the exterme phenotypes design. In all simulations σ_g^2 was set to 0.5.

5.8 Simulation of case-control data using dynamic programming

We are interested in sampling from the conditional distribution of $(g_1, g_2, ..., g_n) \mid p = 1$. We use the chain rule to write break the joint probability into a serial product of conditional probabilities:

$$P(g_1, g_2, ..., g_n \mid p = 1) = P(g_1 \mid p = 1) \prod_{i=2}^{n} P(g_i \mid g_1, ..., g_{i-1}, p = 1)$$

By Bayes' theorem, any conditional probability can be written as:

$$P(g_i \mid g_1, ..., g_{i-1}, p = 1) = \frac{P(p = 1 \mid g_1, ..., g_i) P(g_i \mid g_1, ..., g_{i-1})}{P(p = 1 \mid g_1, ..., g_{i-1})}.$$

The term $P(g_i \mid g_1, ..., g_{i-1})$ does not involve conditioning on the phenotype, and therefore is straightforward to compute, especially under linkage equilibrium, where $P(g_i \mid g_1, ..., g_{i-1}) = P(g_i)$. To compute the term $P(p = 1 \mid g_1, ..., g_i)$, we note that

$$l = \sum_{j=1}^{m} u_j g_j + e = \sum_{j=1}^{i} u_j g_j + \sum_{j=i+1}^{m} u_j g_j + e,$$

since the genotypes $g_1, ..., g_i$ are known, and the effect sizes $u_1 ..., u_m$ are known, $\sum_{j=1}^{i} u_j g_j$ is known. On the other hand $g_{i+1}, ..., g_m$ are unknown, so $\sum_{j=i+1}^{m} u_j g_j \dot{\sim} N(0, \sum_{j=i+1}^{m} u_j^2)$, and so:

$$l \dot{\sim} N(\sum_{j=1}^{i} u_j g_j, 1 - \sum_{j=1}^{i} u_i^2).$$

Therefore:

$$P(p = 1 \mid g_1, ..., g_i) = 1 - \Phi\Big(\frac{t - \sum_{j=1}^{i} u_j g_j}{\sqrt{1 - \sum_{j=1}^{i} u_j^2}}\Big).$$

Using these formulae, the probability of observing each possible genotype at locus i is calculated, and the genotype is sampled according to these probabilities.

We note that computing the sampling probabilities involved computing two probabilities: $P(p = 1 \mid g_1, ..., g_i)$ conditions on the actual genotypes, and therefore is unaffected by LD structure, while $P(g_i \mid g_1, ..., g_{i-1})$ does not condition on the phenotype, and is therefore unaffected by the case-control sampling scheme. Hence, any method that allows computing $P(g_i \mid g_1, ..., g_{i-1})$ using realistic LD structure (e.g., methods based on hidden Markov models), can be modified to generate case samples with realistic LD.

We implemented our method (called simCC) in an R script which is freely available on our website. To validate simCC, we re-ran the same simulation set as in figure 4 in the main text, resulting in very similar results (Figure 18, left panel). In addition, the improved running time allowed us to extend the simulations to $100,000$ SNPs. Each simulation ran in 2-3 hours instead of several days (Figure 18, right panel). For all scenarios studied PCGC-regression estimates remained unbiased.

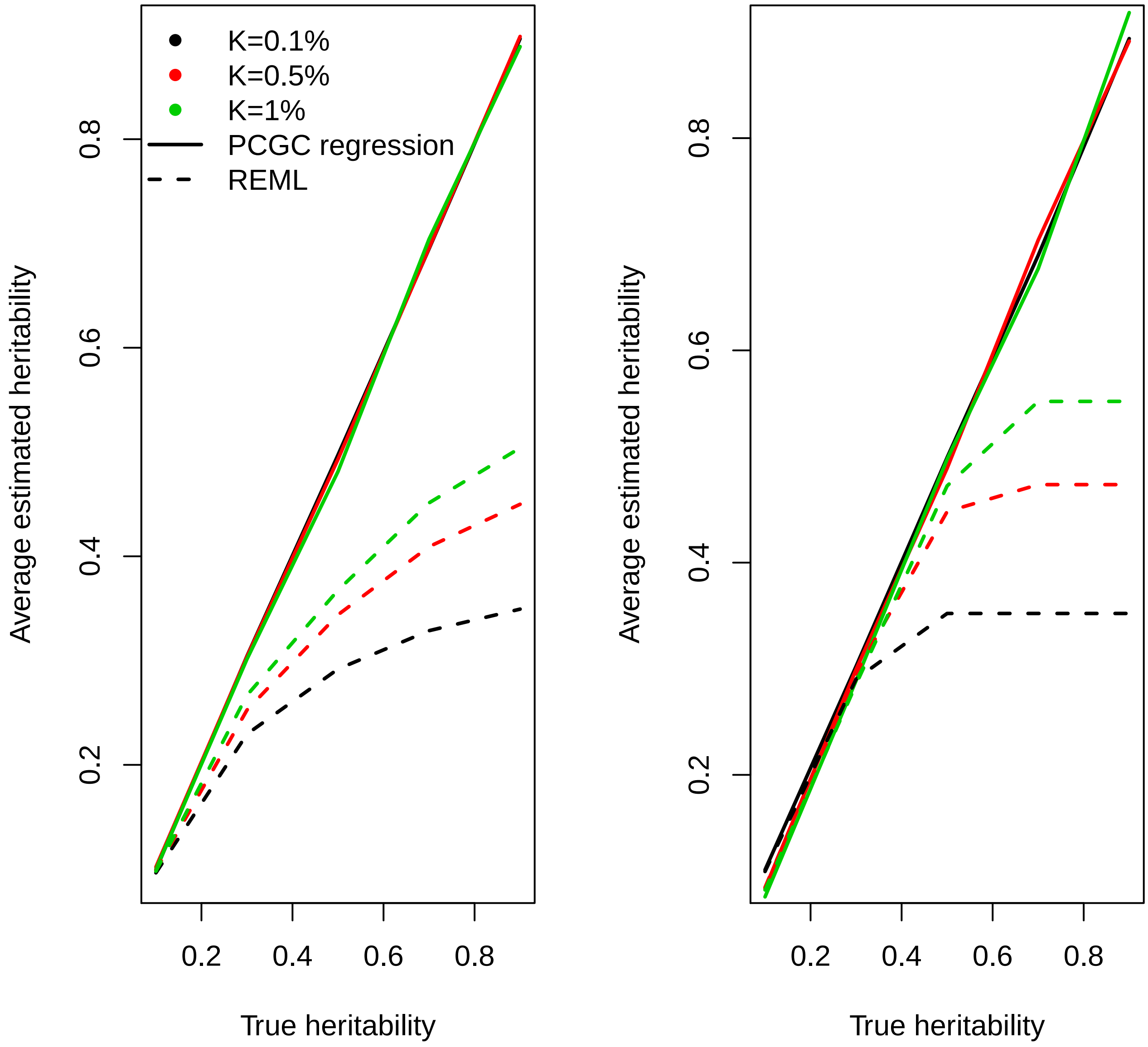


Figure 18: Comparison of PCGC regression and REML estimation for simulated case-control studies using simCC. Left panel - simulations using $10,000$ SNPs. Right panel - simulations using $100,000$ SNPs. All simulated studies were balanced ($P = 50\%$).

6 Inference

Standard inference for linear regression is not applicable in our case due to the breaking of several of the basic assumptions used in linear regression inference. For example, the errors are non-normal and not independent. We therefore use the jackknife [5] to estimate the variance of our estimates. Since our method is three to four orders of magnitude faster than REML estimation, applying such a procedure is not considerably time-demanding. More specifically, we use the jackknife to estimate the standard deviation of the estimate, and use this standard deviation to construct a confidence interval (CI). At each jackknife iteration we remove one individual from the data, and remove all relevant entries to the genetic correlation matrix. We constructed 95% confidence intervals for our simulations and, these CIs covered the true value of the heritability in 98% of the simulations, indicating the jackknife is conservative in this case, and the true standard error of the estimates is usually smaller than the jackknife estimates.

7 Accounting for imperfect linkage-disequilibrium

Yang et al. [25] suggest that the genetic correlation matrix, estimated based on the genotyped data, is in fact not the true underlying correlation matrix, as the latter should be computed only using causal SNPs. Since causal SNPs are unknown, and often not genotyped, the estimated genetic correlation is a noisy and biased estimate of the true genetic correlation. To correct for this effect, Yang et al. suggest using a different correlation matrix

$$G^* = \beta(G - I) + I$$

where G is the estimated genetic correlation matrix, and β is a constant correcting the bias between the true and estimated genetic correlations, effectively accounting for the imperfect LD between causal and genotyped SNPs. Using extensive simulations using real data and realistic assumptions they estimate β at roughly 0.875. This value of β corresponds to a scenario where the causal SNPs can of any minor allele frequency ($c = 0$ using their notation). It is easy to show that using G^* in REML estimation yields an estimate of heritability which is $\frac{1}{\beta}\hat{h}^2$, where $\hat{h}^2$ is the heritability estimate obtained using G. In the case of PCGC, we regress the product of normalized phenotypes Z_{ij} on G_{ij}. Hence, plugging in $G^*_{ij} = \beta G_{ij}$ as the covariate has the exact same effect of increasing the estimated heritability by a factor of $\frac{1}{\beta}$. Table 2 provides the heritability estimates without this correction.

Phenotype	Prevlance (%)	REML (s.e.) (%)	PCGC (s.e.) (%)
BD	0.5	34.4 (3.7)	41.1 (6.3)
CD	0.1	17.7 (2.7)	21.5 (3.7)
MI	1	29.1 (5.3)	33.4 (7.9)
MS	0.1	29.3 (3.1)	39.7 (4.9)
SCH	1	33.4 (2.9)	37.2 (4.4)
T1D	0.5	16.5 (3.6)	18.7 (4.8)

Table 2: Estimated heritability due to common SNPs using REML and PCGC regression, without adjusting the results for imperfect LD.

8 Controlling for structure

Following [25, 11], and many others, we included 10 principal components (PCs) as fixed effects in the analysis. The idea of controlling for structure using PCs was first discussed in [16] and [17] review both PC-based approaches and mixed-model approaches.

In addition to including PCs for the purpose of controlling for the confounding effect of population structure, it is important to note that PCGC-regression uses a two-step approach to handle fixed effects, wherein fixed-effects are fit in the first step and only then the heritability is estimated. Since PCs are taken care of in the first step, we removed the top 10 PCs from the genetic correlation matrix by defining:

$$G' = G - \sum_{i=1}^{10} \lambda_i v_i v_i^\intercal$$

where λ_i, v_i are the i'th eigenvalue and eigenvector of G, respectively, so as to refrain from accounting for them twice. We then use G' for estimating the heritability, as well as include the top 10 principal components as fixed effects.

9 Additional Tables

	REML	Adjusted REML	PCGC	Adjested PCGC
All autosomes	16.5 (3.6)	18.9 (4.1)	18.7 (4.8)	21.4 (5.5)
without chr 6	12.7 (3.5)	14.6 (4)	14.2 (4.8)	16.3 (5.5)
chr 6 only	4.1 (1.1)	4.7 (1.3)	3.2 (1.5)	3.7 (1.7)

Table 3: Breakdown of T1D heritability estimates to chromosome 6 and the rest of the autosomal chromosomes.

Phenotype	Prevlance (%)	REML (s.e.) (%)	PCGC (s.e.) (%)
BD	0.5	40.1 (4.2)	47.7 (7.5)
CD	0.1	20.1 (3.1)	24 (4.9)
MI	1	33.3 (6.1)	38.2 (9.4)
MS	0.1	34.7 (3.5)	46.6 (6.3)
SCH	1	38.7 (3.3)	42.5 (5.0)
T1D	0.5	18.9 (4.1)	21.4 (5.5)

Table 4: Estimated heritability due to common SNPs using REML and PCGC regression, without including sex as a fixed effect.

10 Quality Control

Lee et al. (2011) stress the importance of applying a stringent quality control (QC) process to genotype data to avoid detecting spurious heritability due to genotyping differences between cases and controls or between different control groups. In all analyses, we excluded sex chromosomes, SNPs with MAF<5% and all individual pairs with an estimated genetic correlation >0.05 based on the correlation matrix. The last step is done to ensure individuals in the study are not closely related. Some additional QC steps or relevant information are provided below, and the final results of all QC procedures are summarized in Table 5.

Phenotype	# Samples	# SNPs
CD	3,639	282,081
BD	3,548	275,370
T1D	3,759	292,223
MS	3,606	286,559
MI	3,361	620,797
SCH	6,731	616,728

Table 5: The number of samples and number of SNPs that passed QC ad were used for estimating the heritability due to common SNPs.

10.1 MS, MI and SCH data

All three datasets were provided after QC, and so we only needed to apply the steps described above. The MS data originated from the ANZGene consortium, and was previously desribed in [10]. For MI heritability estimation we used only the Italian Atherosclerosis Thrombosis and Vascular Biology (ATVB) study, which was part of a larger MI study [8]. The ATVB was the largest part of the study, and we focused on it to mitigate population structure issues. We also note that when inspecting the age of individuals in the study it seemed that the ATVB study had the best matching of cases and controls by age, while in other sections of the study the matching was problematic. Since the early-onset MI phenotype is defined by age, inclusion of older controls is problematic and would result in under-estimation of heritability, as the age variable can (falsely) explain a significant share of the case/control difference. For SCH heritability estimation, we merged the sw5 and sw6 cohorts from [18], which together compose the largest cohort genotyped using the same genotyping platform. As an additional QC step, we removed SNPs which failed to pass QC in one of the cohorts.

10.2 WTCCC data

The WTCCC data [3] was supplied without any prior QC. In addition to the QC steps described above, we performed additional QC steps as described in Lee et al. (2011): we removed SNPs with missing rate >1% and SNPs which displayed a significantly different missing rate between cases and controls (p-value<0.05). We also removed SNPs which deviated from Hardy-Weinberg (HW) equilibrium in the control groups (p-value <0.05). Additionally we removed SNPs which displayed a significant difference in frequency between the two control groups. We removed all the individuals appearing in the WTCCC exclusion lists, which include duplicate samples, first or second degree relatives, individuals which are not of European descent and others. In addition we removed individuals with missing rate >1%.

10.3 Extreme-HDL levels data

We obtained genotype data of individuals with extremely high or lower levels of HDL. Genotypes were supplied after QC. In addition we removed SNPs with MAF below 1% (but we note that results are similar for other MAF thresholds). Since HDL levels behave differently for males and females, we focused on male samples only. Additionally we focused on individuals of Caucasian ancestry. One individual was indicated to be an extreme outlier using PCA, and was therefore removed as well. HDL levels were centered and scaled based on population-specific parameters from NHANES III, 1988-94.

After QC, our sample included 452 individuals, sampled from the top 10% and bottom 25% of the HDL distribution. Individuals with high HDL levels were over-represented in the study, composing 57.8% of the study, resulting in $K_1 = 0.1, K_2 = 0.25, R = 0.287$. Using these parameters we estimated the heritabity at 45% (s.e.= 39%, using the jackknife). To make sure that the estimated heritability is not inflated due to additional population structure, we removed the four top principal components from the genetic correlation matrix. We note that the number was chosen using the method of [14] instead of simply removing 10 PCs due to the small sample size.

We note that this estimate is not adjusted using the usual β correction of Yang et al. (which accounts for unobserved genotypes in a standard Affymetrix 500K chip) because the study used the Affymetrix 1M chip.

Our simulations suggested that estimating heritability using extreme phenotype studies can greatly reduce the estimation error. To test this in practice, we compared our estimate and its standard error (s.e.) to three studies which used similar methods to estimate the heritability of HDL levels. Zaitlen et al. [28] used a random sample of 38,167 individuals to estimate the narrow-sense heritability of HDL at 45% (s.e.= 1.7%). Vattikuti et al. [22] used a random sample of 8,451 individuals, and estimated the narrow-sense heritability at 48% (s.e.= 11%). Lastly, Browning and Browning [2] used direct pair-wise IBD estimation in 5,402 individuals to estimate the narrow-sense heritability at 46% (s.e.= 17%). First, it is interesting to note that all four estimates are largely in agreement. Second, one can compare the standard errors of the various methods. All methods utilize information about pairs of individuals, and so the variance of the estimates scales as $O(\frac{1}{n})$ instead of the usual $O(\frac{1}{\sqrt{n}})$. Once accounting for sample size, heritability estimation using extreme phentype studies with PCGC regression is considerably more accurate – it requires less than one fourth of the study size to reach a smaller standard error as the other study designs. This is inline with the general spirit of our simulation results.

11 Estimating the contribution of common variants in the MHC region to the heritability of type-1 diabetes

Based on family studies of type-1 diabetes ("top down" estimates) [19], the MHC region is estimated to account for roughly 50% of the heritability, corresponding to 36-44% of the phenotypic variance.

Roughly similar values can be obtained by considering the contribution of the common haplotypes and genotypes at the MHC locus ("bottom up" estimates). Specifically, we used published population frequencies and odds-ratios for HLA haplotypes significantly associated with type-1 diabetes [6] (Table 6). We considered the top 10 most strongly associated haplotypes. (i) When we treat each haplotype as an independent fixed-effect covariate, and use the method outlined in section 2.3.1 above to compute the variance (σ_t^2) and heritability ($\frac{1}{1+\sigma_t^2}$), these 10 haplotypes explain 25.8% of the phenotypic variance. (ii) When we perform a joint analysis of the haplotypes using a multi-allelic model (i.e., a single fixed effect with multiple levels, one level per allele), the estimated fraction of phenotypic variance explained increases to 28.8%. (iii) We then estimated the contribution of the well-known interaction between HLA-DR3 and HLA-DR4, whereby heterozygotes are at much higher risk than would be predicted by the additive effects of the separate haplotypes; we used data on genetic interactions from the supplementary of [6] to estimate the heritability due to this interaction at 3.8%. The combined estimate is thus 32.6% of the phenotypic variance.

Haplotype (DRB1/DQA1/DQB1)	Frequency per individual (%)	Relative risk[1]	Phenotypic variance explained (%)	Heritability explained (%)[2]
0301 0501 0201	12.5	3.64	2.3	2.6
0401 0301 0302	4.5	8.40	2.8	3.2
0402 0301 0302	1.0	3.60	0.2	0.3
0404 0301 0302	3.2	1.60	0.1	0.1
0405 0301 0302	0.2	11.40	0.2	0.2
0701 0201 0303	4.3	0.02	5.1	5.8
1104 0501 0301	2.3	0.07	1.5	1.7
1303 0501 0301	1.0	0.08	0.6	0.7
1401 0101 0503	2.1	0.02	2.6	3.0
1501 0102 0602	12.0	0.03	10.4	11.9
(i) Total, treating haplotypes as binary fixed effects			25.8	29.3
(ii) Total, treating haplotypes as multi-allelic system[3]			28.8	32.7
(iii) Additional contribution of HLA-DR3/4 interaction[4]	2.1		3.8	4.3
Total (ii)+(iii)			32.6	37

Table 6: Analysis of the fraction of phenotypic variance and heritability of type-1 diabetes explained by common HLA alleles. Data are from [6].

[1] Studies report odds ratios, but for low prevalence diseases the OR and RR are very close.

[2] Taking the upper bound of family-based estimates of heritability (88%), meaning this is a conservative estimate.

[3] This is done by defining a different threshold for each allele, and using population frequencies to estimate the added variance.

[4] DR3/4 is defined as the four highly associated allele pairs from Supp. table 3 of [6]: DR0301 and each of DR0401/2/4/5. We use the frequencies and odds ratios reported and treat this locus as a multi-allelic locus. Notably, using data in a coarser resolution from [24] suggests that the proportion of variance explained by DR3/4 interaction is slightly higher: 5.8%.

We note that this analysis of haplotype effects based on the data in Erlich et al. (2008) yields a conservative estimate, because the data provides a very fine-scale breakdown of the various allele combinations. As a result, many allele-combinations have relatively few hits, and are therefore not statistically significant and cannot be included in the analysis. Repeating our analysis using two other studies of two different populations [24, 1], which study the HLA region at a much coarser resolution, the fraction of phenotypic variance explained by common alleles is 38.7% and 30%, respectively (Table 7).

Study	% phenotypic variance explained	% heritability explained
Erlich et al. (2008) [6]	32.6	37
Wolf et al. (1983) [24]	38.7	44
Svejgaard et al. (1986) [1]	30	34.1

Table 7: Estimated phenotypic variance and heritability explained by common MHC alleles from frequencies and odds-ratios of three studies of three different populations.

In summary, we conclude that common variants at the MHC explain at least 35% of the phenotypic variance. We use this estimate in Table 1 in the main text.

12 Population values references

Phenotype	Prevlance (%)	Reference	Estimated total heritability (%)	Reference
BD	0.5	[11]	71	[4]
CD	0.1	[11]	50-60	[20]
MI	1	[8]	56	[13]
MS	0.1	[10]	25-75	[23]
SCH	1	[18]	64	[12]
T1D	0.5	[11]	72-88	[7, 9]

Table 8: Disease prevalences and estimated total heritabilities for all six phenotypes analysed, with their corresponding references.

References

[1] HLA associations in insulin-dependent diabetes: search for heterogeneity in different groups of patients from a homogeneous population. *Tissue antigens*, 28:237–244, 1986.

[2] Sharon R Browning and Brian L Browning. Identity-by-descent-based heritability analysis in the northern finland birth cohort. *Human genetics*, 132(2):129–138, 2013.

[3] Paul R Burton, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[4] Jack Edvardsen, Svenn Torgersen, Espen Røysamb, Sissel Lygren, Ingunn Skre, Sidsel Onstad, and Per Anders Øien. Heritability of bipolar spectrum disorders. unity or heterogeneity? *Journal of affective disorders*, 2008.

[5] Bradley Efron and Robert J Tibshirani. An introduction to the bootstrap (chapman & hall/crc monographs on statistics & applied probability). 1994.

[6] Henry Erlich, Ana Maria Valdes, Janelle Noble, Joyce A Carlson, Mike Varney, Pat Concannon, Josyf C Mychaleckyj, John A Todd, Persia Bonella, Anna Lisa Fear, Eva Lavant, Anthony Louey, and Priscilla Moonsamy. HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families. *Diabetes*, 57:1084–1092, 2008.

[7] Valma Hyttinen, Jaakko Kaprio, Leena Kinnunen, Markku Koskenvuo, and Jaakko Tuomilehto. Genetic liability of type 1 diabetes and the onset age among 22,650 young finnish twin pairs a nationwide follow-up study. *Diabetes*, 52(4):1052–1055, 2003.

[8] Sekar Kathiresan, Benjamin F Voight, Shaun Purcell, Kiran Musunuru, Diego Ardissino, Pier M Mannucci, Sonia Anand, James C Engert, Nilesh J Samani, Heribert Schunkert, et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature genetics*, 41(3):334–341, 2009.

[9] Kirsten O Kyvik, Anders Green, and Henning Beck-Nielsen. Concordance rates of insulin dependent diabetes mellitus: a population based study of young danish twins. *Bmj*, 311(7010):913–917, 1995.

[10] S Hong Lee, Denise Harold, Dale R Nyholt, Michael E Goddard, Krina T Zondervan, Julie Williams, Grant W Montgomery, Naomi R Wray, and Peter M Visscher. Estimation and partitioning of polygenic variation captured by common snps for alzheimer's disease, multiple sclerosis and endometriosis. *Human molecular genetics*, 22(4):832–841, 2013.

[11] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305, 2011.

[12] Paul Lichtenstein, Benjamin H Yip, Camilla Björk, Yudi Pawitan, Tyrone D Cannon, Patrick F Sullivan, and Christina M Hultman. Common genetic determinants of schizophrenia and bipolar disorder in swedish families: a population-based study. *The Lancet*, 373(9659):234–239, 2009.

[13] JJ Nora, RH Lortscher, RD Spangler, AH Nora, and WJ Kimberling. Genetic–epidemiologic study of early-onset ischemic heart disease. *Circulation*, 61(3):503–508, 1980.

[14] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

[15] Ross L Prentice and Ronald Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

[16] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[17] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

[18] Stephan Ripke, Colm O'Dushlaine, Kimberly Chambert, Jennifer L Moran, Anna K Kähler, Susanne Akterin, Sarah E Bergen, Ann L Collins, James J Crowley, Menachem Fromer, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature genetics*, 45(10):1150–1159, 2013.

[19] N Risch. Assessing the role of HLA-linked and unlinked determinants of disease. *American journal of human genetics*, 40:1–14, 1987.

[20] J Sofaer. Crohn's disease: the genetic contribution. *Gut*, 34(7):869–871, 1993.

[21] Jon Stene. Assumptions for different ascertainment models in human genetics. *Biometrics*, pages 523–527, 1977.

[22] Shashaank Vattikuti, Juen Guo, and Carson C Chow. Heritability and genetic correlations explained by common snps for metabolic syndrome traits. *PLoS genetics*, 8(3):e1002637, 2012.

[23] Corey T Watson, Giulio Disanto, Felix Breden, Gavin Giovannoni, and Sreeram V Ramagopalan. Estimating the proportion of variation in susceptibility to multiple sclerosis captured by common snps. *Scientific reports*, 2, 2012.

[24] E Wolf, K M Spencer, and A G Cudworth. The genetic susceptibility to type 1 (insulin-dependent) diabetes: analysis of the HLA-DR association. *Diabetologia*, 24:224–230, 1983.

[25] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

[26] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88(1):76, 2011.

[27] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics*, 46(2):100–106, 2014.

[28] Noah Zaitlen, Peter Kraft, Nick Patterson, Bogdan Pasaniuc, Gaurav Bhatia, Samuela Pollack, and Alkes L Price. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS genetics*, 9(5):e1003520, 2013.

[29] Noah Zaitlen, Bogdan Paşaniuc, Nick Patterson, Samuela Pollack, Benjamin Voight, Leif Groop, David Altshuler, Brian E Henderson, Laurence N Kolonel, Loic Le Marchand, et al. Analysis of case–control association studies with known risk variants. *Bioinformatics*, 28(13):1729–1737, 2012.

[30] X Zhou, P Carbonetto, and M Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics*, 9(2):e1003264, 2013.