Estimating SNP-Based Heritability and Genetic Correlation in Case-Control Studies Directly and with Summary Statistics

Omer Weissbrod,[1,2,4,*] Jonathan Flint,[3] and Saharon Rosset[1,*]

Methods that estimate SNP-based heritability and genetic correlations from genome-wide association studies have proven to be powerful tools for investigating the genetic architecture of common diseases and exposing unexpected relationships between disorders. Many relevant studies employ a case-control design, yet most methods are primarily geared toward analyzing quantitative traits. Here we investigate the validity of three common methods for estimating SNP-based heritability and genetic correlation between diseases. We find that the phenotype-correlation-genotype-correlation (PCGC) approach is the only method that can estimate both quantities accurately in the presence of important non-genetic risk factors, such as age and sex. We extend PCGC to work with arbitrary genetic architectures and with summary statistics that take the case-control sampling into account, and we demonstrate that our new method, PCGC-s, accurately estimates both SNP-based heritability and genetic correlations and can be applied to large datasets without requiring individual-level genotypic or phenotypic information. Finally, we use PCGC-s to estimate the genetic correlation between schizophrenia and bipolar disorder and demonstrate that previous estimates are biased, partially due to incorrect handling of sex as a strong risk factor.

Introduction

Much of the theory underlying methods for estimating two key measures of disease genetic architecture, SNP-based heritability and genetic correlation, was designed for cohort studies of quantitative phenotypes. Consequently, when applied to studies of categorical traits, these methods may contain unacknowledged biases that may affect estimation accuracy.

The problem of accurately estimating SNP-based heritability and genetic correlation is usually translated into questions about variance and covariance components in properly defined mathematical models. A commonly held misconception states that variance components can be accurately calculated in case-control studies by virtue of applying a correction factor to results derived under a quantitative trait framework.[1–4] However, this is not true when risk factors (including risk variants) exert a strong influence on disease risk. In this paper we examine the validity of approaches for estimating heritability, genetic covariance, and correlation (covariance standardized to a [−1, 1] scale) in case-control studies of disease.

Broadly speaking, there are three common approaches for carrying out these tasks. The first is based on restricted maximum likelihood estimation (REML) in the linear mixed model (LMM)[5] framework and is implemented in some widely used tools.[6–8] This approach has been extensively applied to heritability estimation[1,5,7] and to genetic correlation estimation.[7,9–11]

The second approach is based on regression of phenotype correlations on genotype correlations and relies on less restrictive assumptions than the LMM approach. This approach was originally designed for quantitative phenotypes, in which case it is known as Haseman-Elston (HE) regression.[12,13] It has recently been adapted for case-control studies, in which case it is called PCGC.[14,15] A common misconception states that PCGC is the same as HE regression (up to a scaling factor), but this equivalence holds only in the absence of covariates.[14] Rather than being an extension of HE regression, PCGC was carefully derived from first principles to apply to all relevant situations in case-control studies. In this paper we extend PCGC to also estimate genetic correlation and to accommodate arbitrary genetic architectures.

The third approach is the family of linkage disequilibrium score regression (LDSC) methods, which estimate heritability and genetic correlation while accounting for LD,[2,16] and have recently been applied to numerous large-scale studies.[17–33] LDSC is attractive because it requires only publicly available summary statistics from genetic studies, thereby avoiding privacy and logistical concerns.[34] Other summary-statistics-based methods have also been proposed recently but we focus on LDSC, as alternative methods cannot be applied in the presence of LD[3] or are not directly designed for categorical phenotypes.[4,35,36]

Here we examine all three approaches under a common set of assumptions that is shared by all of them. We demonstrate that even when these assumptions hold, LDSC and REML yield biased estimates in the presence of covariates representing major risk factors such as sex and age, due to incorrect modeling of case-control

[1]Statistics Department, Tel Aviv University, Ramat Aviv 6997801, Israel; [2]Computer Science Department, Technion - Israel Institute of Technology, Haifa 3200003, Israel; [3]Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA 90405, USA
[4]Present address: Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
*Correspondence: oweissbrod@hsph.harvard.edu (O.W.), saharon@post.tau.ac.il (S.R.)
https://doi.org/10.1016/j.ajhg.2018.06.002.

ascertainment. In contrast, PCGC remains unbiased under all settings. We further develop a new version of PCGC, called PCGC-s, that can work with summary statistics that explicitly take the case-control sampling into account.

We demonstrate the value of PCGC-s by investigating the genetic correlation between schizophrenia and bipolar disorder and between type 1 diabetes and coronary artery disease. We demonstrate that the estimates of both quantities are severely biased under alternative methods, partially due to incorrect handling of sex—an important risk factor. Finally, we provide best practice recommendations depending on the available data and the trait characteristics.

Material and Methods

Underlying Mixed Effects Model

We adopt the theoretical framework of the liability threshold model,[37,38] which is the same model assumed by REML[1] and LDSC[2] for analysis of case-control studies. This model assumes that every individual i has a latent normally distributed liability value for trait t, a_t^i, such that case subjects of trait t are individuals whose liability exceeds a given cutoff.

We additionally assume that the liability of trait t can be decomposed into three terms corresponding to a covariates effect q_t^i, a genetic effect g_t^i, and an environmental effect e_t^i, $a_t^i = q_t^i + g_t^i + e_t^i$, such that the vectors of covariate effects $\boldsymbol{q}_t = [q_t^1, \ldots, q_t^n]^T$, of genetic effects $\boldsymbol{g}_t = [g_t^1, \ldots, g_t^n]^T$, and of environmental effects $\boldsymbol{e}_t = [e_t^1, \ldots, e_t^n]^T$ are given by:

$$\boldsymbol{q}_t = \boldsymbol{C}_t \beta_t$$

$$\boldsymbol{g}_t \sim \mathcal{N}\left(0; \sigma_{gt}^2 \boldsymbol{G}_t\right)$$

$$\boldsymbol{e}_t \sim \mathcal{N}\left(0; \left(1 - \sigma_{gt}^2\right)\boldsymbol{I}\right).$$

Here, $\boldsymbol{C}_t$ is a design matrix of covariates, $\boldsymbol{\beta}_t$ is a column vector of fixed effects, $\boldsymbol{G}_t$ is a matrix of genetic similarity coefficients (defined below), σ_{gt}^2 is a genetic variance parameter, and $\boldsymbol{I}$ is the identity matrix. The matrix $\boldsymbol{G}_t$ is typically given by $\boldsymbol{G}_t = \boldsymbol{X}_t \boldsymbol{W} \boldsymbol{X}_t^T$, where $\boldsymbol{X}_t$ is an $n \times m$ matrix of m standardized single-nucleotide polymorphisms (SNPs), and $\boldsymbol{W}$ is an $m \times m$ diagonal weighting matrix, which assigns different weights to different SNPs. This definition can accommodate any linear genetic architecture; it includes the standard model used by common REML software packages[6,7] and by LDSC[2,16] under the special case $\boldsymbol{W} = (1/m)\boldsymbol{I}$. However, it can also accommodate minor allele frequency (MAF)-dependent and LD-dependent architectures,[8,39] which correspond to a suitable choice of $\boldsymbol{W}$ (Supplemental Methods).

Under these assumptions, every individual i has an observed affection status indicator for trait t, y_t^i, such that $y_t^i = 1$ if and only if $a_t^i > \tau_t$, where $\tau_t = \Phi^{-1}(1 - K_t) + \mathrm{E}[\boldsymbol{C}_t]^T \boldsymbol{\beta}_t$ is the affection cutoff for trait t with prevalence K_t, and where $\Phi^{-1}(\cdot)$ is the inverse cumulative standard normal density.

For a pair of traits t_1, t_2, the concatenated liabilities vector follows a multivariate normal distribution,

$$\begin{bmatrix} \boldsymbol{a}_{t_1} \\ \boldsymbol{a}_{t_2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{C}_{t_1}\boldsymbol{\beta}_{t_1} \\ \boldsymbol{C}_{t_2}\boldsymbol{\beta}_{t_2} \end{bmatrix}, \begin{bmatrix} \sigma_{gt_1}^2 \boldsymbol{G}_{t_1} + \left(1 - \sigma_{gt_1}^2\right)\boldsymbol{I}_{t_1} & \rho_{t_1,t_2}\boldsymbol{G}_{t_1,t_2} \\ \rho_{t_1,t_2}\left(\boldsymbol{G}_{t_1,t_2}\right)^T & \sigma_{gt_2}^2 \boldsymbol{G}_{t_2} + \left(1 - \sigma_{gt_2}^2\right)\boldsymbol{I}_{t_2} \end{bmatrix} \right),$$

where $\boldsymbol{G}_{t_1,t_2} = \boldsymbol{X}_{t_1}\boldsymbol{W}\boldsymbol{X}_{t_2}^T$ is the matrix of between-study genetic similarity coefficients (i.e., the genetic similarity coefficients between each individual in study 1 and each individual in study 2), ρ_{t_1,t_2} is the genetic covariance, and $\boldsymbol{I}_{t_1}, \boldsymbol{I}_{t_2}$ are identity matrices of suitable dimensions.

The quantities we investigate in this paper are defined as follows:

(a) The SNP-based heritability of trait t, defined as $h_t^2 \triangleq \mathrm{var}(g_t^i)/\mathrm{var}(a_t^i)$.
(b) The SNP-based genetic covariance of two traits t_1, t_2, defined as $\rho_{t_1,t_2} \triangleq \mathrm{cov}(g_{t_1}^i, g_{t_2}^i)$.
(c) The SNP-based genetic correlation of two traits t_1, t_2, defined as $r_g \triangleq \rho_{t_1,t_2}/\sqrt{\mathrm{var}(g_{t_1}^i)\mathrm{var}(g_{t_2}^i)}$.

In the remainder of this article we use the shortened terms "heritability," "genetic covariance," and "genetic correlation" for brevity.

The Effect of Ignoring Covariates

The main contribution of PCGC-s over LDSC is its ability to account for covariates. Although it is rarely possible to measure all covariates affecting the trait of interest, covariates with a strong effect (such as the effect of sex on coronary artery disease) are often measured. This raises the question of whether omission of such important covariates affects heritability and genetic correlation estimates. We prove in the Supplemental Methods that if a method can provide unbiased estimates in settings with no covariates, it can also provide unbiased estimates in settings with covariates by simply ignoring these covariates, assuming that the covariate effects are (1) normally distributed and (2) uncorrelated with the genetic effect. The main idea behind the derivation is that the environmental effect represents the aggregated effect of unmeasured covariates and can thus absorb the effect of omitted covariates when these assumptions hold.

The assumption of normality approximately holds if a trait is influenced by a large number of covariates with small effects, owing to the central limit theorem. However, many traits are strongly influenced by a small number of non-normally distributed covariates, such as sex. Heritability estimates with omitted covariates can become inaccurate in the presence of such strong covariates. In contrast, genetic correlation is accurately estimated in the simulations even in the presence of strong non-normal covariates, suggesting that the errors in the estimation of genetic covariance and genetic variance approximately cancel out when dividing one by the other. However, this observation is currently unsupported by statistical theory.

The assumption that covariates are uncorrelated with the genetic effect is often violated when using heritable covariates, such as genetic principal components. This problem can be circumvented by regressing the omitted covariates out of the genotypes and correcting the individual-level affection cutoffs prior to parameter estimation or to computing summary statistics (Supplemental Methods). We caution that regression of covariates out of binary phenotypes as suggested in Bulik-Sullivan[40] can yield incorrect estimates in case-control studies, even for genetic correlation (as verified in the Results).

Marginal and Conditional Heritability

An important point often overlooked in heritability estimation is that covariates such as sex and age also contribute to the liability variance. Since the liability is non-identifiable, it is typically assumed to have a unit variance when conditioning on measured covariates. The liability is defined as $a_t^i = g_t^i + e_t^i + (\boldsymbol{C}_t^i)^T \boldsymbol{\beta}_t$, and thus its marginal variance is given by $\mathrm{var}[g_t^i + e_t^i] + \mathrm{var}[(\boldsymbol{C}_t^i)^T \boldsymbol{\beta}_t] = 1 + \mathrm{var}[(\boldsymbol{C}_t^i)^T \boldsymbol{\beta}_t]$ (assuming that covariates are uncorrelated with the genetic effects). Consequently, heritability is given by $\sigma_{gt}^2 / \left(1 + \mathrm{var}\left(\left(\boldsymbol{C}_t^i\right)^T \boldsymbol{\beta}_t\right)\right)$ (Supplemental Methods). Alternatively, one could assume that the marginal variance is 1, in which case the conditional variance is smaller than 1.

In practice, many studies define the genetic variance σ_{gt}^2 as the heritability, even in the presence of covariates. We therefore denote the former definition as marginal heritability and the latter definition as conditional heritability, because the latter definition uses the variance of the liability conditional on measured covariates.

In this paper we consider marginal heritability for two reasons: (1) this definition is arguably more natural, as different studies using different covariates are ultimately interested in estimating the same quantity; and (2) LDSC tends to severely underestimate the conditional heritability (as compared to less severe overestimation of marginal heritability). Therefore, we do not consider estimation of conditional heritability further in this paper.

PCGC-s with No Covariates

PCGC with no covariates estimates ρ_{t_1,t_2} by regressing standardized phenotypic correlations $\tilde{y}_{t_1}^i \tilde{y}_{t_2}^j$ on genetic similarity coefficients $G_{t_1,t_2}^{i,j}$ and then dividing the resulting estimator by a constant $f(t_1,t_2)$ (Supplemental Methods). This estimation encapsulates both genetic covariance and heritability, which for a trait t with no covariates is given by $\rho_{t,t}$.

The PCGC estimator can be computed without individual-level data by using the following two summary statistics:

$$z_t^k \triangleq \frac{1}{\sqrt{n_t}} \sum_{i=1}^{n_t} \tilde{y}_t^i X_t^{k,i}$$

$$\widehat{r}_t^{k,h} \triangleq \frac{1}{n_t} \sum_{i=1}^{n_t} X_t^{k,i} X_t^{h,i},$$

where n_t is the sample size of study t and $X_t^{k,i}$ is the value of the k^{th} variant of individual i in study t, after standardization. It is also possible to use logistic regression-based or other types of summary statistics, but this constitutes an approximation (Supplemental Methods).

Using these quantities and denoting S_{t_1,t_2} as the set of all pairs of indices i, j that refer to the same individual shared between the two studies, the PCGC estimator can be written as:

$$\widehat{\rho}_{t_1,t_2}^{\text{pcgc}-\text{s}} \triangleq \frac{1}{f(t_1,t_2)} \frac{\frac{\sqrt{n_{t_1} n_{t_2}}}{m} \sum_{k=1}^{m} z_{t_1}^k z_{t_2}^k - \sum_{(i,j)\in S_{t_1,t_2}} G_{t_1,t_2}^{i,j} \left(\tilde{y}_{t_1}^i \tilde{y}_{t_2}^j\right)}{\frac{n_{t_1} n_{t_2}}{m^2} \sum_{k,h=1}^{m} \widehat{r}_{t_1}^{k,h} \widehat{r}_{t_2}^{k,h} - \sum_{(i,j)\in S_{t_1,t_2}} \left(G_{t_1,t_2}^{i,j}\right)^2},$$

where m is the number of variants and $f(t_1,t_2)$ is given by: $f(t_1,t_2) = \frac{\sqrt{P_{t_1}(1-P_{t_1})P_{t_2}(1-P_{t_2})}\phi(\tau_{t_1})\phi(\tau_{t_2})}{K_{t_1}(1-K_{t_1})K_{t_2}(1-K_{t_2})}.$

Here, K_t and P_t are the prevalence of trait t and the case-control proportion of study t, respectively, $\tau_t = \Phi^{-1}(1-K_t)$ is the liability cutoff, and $\phi(\bullet)$, $\Phi(\bullet)$ are the density and cumulative distribution of the standard normal distribution, respectively.

The resulting estimator approximately coincides with the LDSC estimator if there are no overlapping individuals (i.e., individuals that are included in both studies) and the in-sample LD estimates in both studies are the same as in the reference population used by LDSC.[40] The extension to estimating multiple variance components or for using MAF and LD-dependent genetic architectures is straightforward (Supplemental Methods).

The second term of the numerator and of the denominator can be computed by research groups with access to the genotypes and phenotypes of overlapping individuals, which often consist of control cohorts, or can be approximated via the approximation $G_{t_1,t_2}^{i,j} \approx 1.0$ for overlapping individuals, as done implicitly in LDSC.[2] However, we caution that even minor deviations (which can occur for example by regressing principal components out of genotypes) can affect the approximation (Supplemental Methods).

A particularly convenient property of $\widehat{\rho}_{t_1,t_2}^{\text{pcgc}-\text{s}}$ in the absence of covariates is that when estimating the genetic correlation, all terms dependent on the trait prevalence vanish. This is convenient because the true trait prevalence is often not known with certainty.

PCGC-s with Covariates

In the presence of covariates, PCGC estimates ρ_{t_1,t_2} by regressing $\tilde{y}_{t_1}^i \tilde{y}_{t_2}^j$ on $G_{t_1,t_2}^{i,j} Q_{t_1,t_2}^{i,j}$, where $Q_{t_1,t_2}^{i,j}$ is a quantity that depends on the covariates of individuals i and j, and so the regression constant is different for every pair of individuals.[14] The corresponding PCGC-s estimator is given by:

$$\widehat{\rho}_{t_1,t_2}^{\text{pcgc}-\text{covar}-\text{s}} \triangleq \frac{\frac{1}{m} \sum_{k=1}^{m} z_{t_1}^{k,\text{covar}} z_{t_2}^{k,\text{covar}} - \sum_{(i,j)\in S_{t_1,t_2}} \tilde{y}_{t_1}^i \tilde{y}_{t_2}^j G_{t_1,t_2}^{i,j} Q_{t_1,t_2}^{i,j}}{\frac{1}{m^2} \sum_{k,h=1}^{m} \widehat{r}_{t_1}^{k,h,\text{covar}} \widehat{r}_{t_2}^{k,h,\text{covar}} - \sum_{(i,j)\in S_{t_1,t_2}} \left(G_{t_1,t_2}^{i,j} Q_{t_1,t_2}^{i,j}\right)^2}.$$

The above quantities are defined as follows:

$$z_t^{k,\text{covar}} \triangleq \sum_{i=1}^{n_t} \tilde{y}_t^i X_t^{k,i} \sum_{\mathrm{a}=0}^{1} \mathrm{u}_{\mathrm{t,a}}^i$$

$$\widehat{r}_t^{k,h,\text{covar}} \triangleq \sum_{i=1}^{n_t} X_t^{k,i} X_t^{h,i} \sum_{\mathrm{a,b}=0}^{1} \mathrm{u}_{\mathrm{t,a}}^{\mathrm{i}} \mathrm{u}_{\mathrm{t,b}}^{\mathrm{i}}$$

$$Q_{t_1,t_2}^{i,j} \triangleq \sum_{a,b=0}^{1} u_{t_1,a}^i u_{t_2,b}^j, \text{where } \mathrm{u}_{\mathrm{t,0}}^i,\ \mathrm{u}_{\mathrm{t,1}}^j \text{are given by :}$$

$$\mathrm{u}_{\mathrm{t,0}}^{\mathrm{i}} \triangleq \frac{\phi\left(\tau_t^i\right)}{\sqrt{P_t^i(1-P_t^i)}\left(K_t^i + (1-K_t^i)\frac{K_t(1-P_t)}{P_t(1-K_t)}\right)} \frac{K_t(1-P_t)}{P_t(1-K_t)} P_t^i$$

$$u_{t,1}^{i} \triangleq \frac{\phi\left(\tau_t^i\right)}{\sqrt{P_t^i\left(1-P_t^i\right)\left(K_t^i+\left(1-K_t^i\right)\frac{K_t(1-P_t)}{P_t(1-K_t)}\right)}}\left(1-P_t^i\right).$$

Here, K_t^i is the probability of individual i being a case conditional on her covariates, P_t^i is the probability of individual i being a case conditional on her covariates and on being ascertained into the study, and $\tau_t^i = \Phi^{-1}(1-K_t^i)$ is the liability cutoff of individual i conditional on her covariates.

The full derivation, extensions for multiple variance components and for MAF and LD-dependent architectures, and an approximation that requires a single summary statistic instead of using $\hat{r}_t^{k,h,\text{covar}}$ (which requires a number of statistics equal to the number of pairs of variants) are provided in the Supplemental Methods.

As in the case of no covariates, the second term of the numerator and denominator can be computed by research groups with access to overlapping individuals, which often consist of control cohorts. Third parties with no access to overlapping individuals can approximate the terms on the right-hand sides of the numerator and the denominator given appropriate summary statistics (Supplemental Methods).

Results

We are interested in estimating the following quantities (see Material and Methods for exact definitions): (a) heritability, the fraction of liability variance explained by genetics; (b) genetic covariance, the covariance between the genetic components of two traits on the liability scale; and (c) genetic correlation, the genetic covariance standardized to a $[-1,1]$ scale.

We are concerned with the three following questions:

1. Can quantities (a)–(c) be estimated reliably given genotypic and phenotypic data?
2. Can quantities (a)–(c) be estimated reliably given summary statistics via LDSC?
3. Can quantities (a)–(c) be estimated reliably given summary statistics via an alternative method?

The answers to questions 1 and 2 are summarized in Table S1. Briefly, PCGC is the only method that can estimate all quantities of interest under all investigated settings. REML provides inconsistent estimates of quantities (a) and (b) and empirically provides consistent estimates of quantity (c). LDSC can provide consistent estimates of quantities (a) and (b) in the absence of covariates and provides consistent estimates of quantity (c) when no covariates are included in the analysis. To answer question 3, we present a reformulation of PCGC called PCGC-s that can estimate quantities (a)–(c) reliably using only summary statistics, both with and without covariates (Supplemental Methods).

Simulation Studies

We conducted simulation studies to investigate the behavior of the evaluated methods in case-control studies; such simulations require first obtaining a very large pool with hundreds of thousands of individuals, and then sampling a small fraction of case subjects according to the trait prevalence.[1,14,16,41]

Our simulations were based on the liability threshold model. Briefly, for every non-genetic covariate k we sampled two independent effect for two different traits (t_1 and t_2), denoted as $\beta_{t_1}^k, \beta_{t_2}^k$, from a normal distribution. In addition, for every SNP j we sampled two correlated effect sizes, $b_{t_1}^j, b_{t_2}^j$. A subset of these effects were equal to zero (determined according to the desired trait polygenicity). Every other pair of effects was sampled independently from a bivariate normal distribution, whose covariance matrix was determined according to the desired heritabilities and genetic correlation of the two traits. In most simulations the variance of all pairs of effects was the same, though we also evaluated MAF and LD-dependent architectures, as described below.

For every individual i, we generated a vector of uniformly spaced SNPs whose LD decays exponentially with distance, denoted as $\boldsymbol{x}^i$, and a vector of independent covariates denoted as $\boldsymbol{c}^i$. Finally, for every individual i and for every trait $t \in \{t_1, t_2\}$, we generated (1) a normally distributed environmental effect e_t^i; (2) a liability given by $a_t^i = (\boldsymbol{x}^i)^T \boldsymbol{b}_t + (\boldsymbol{c}^i)^T \beta_t + e_t^i$; and (3) a case/control label, where case subjects are individuals with $a_t^i > \tau_t$, with τ_t being the empirical $1-K_t$ percentile of the liabilities in the population, and K_t is the prevalence of trait t. We kept on sampling individuals until obtaining the desired number of case and control subjects. The full simulation details are provided in the Supplemental Methods.

Our simulations span a wide range of scenarios, with various levels of prevalence, heritability, genetic correlation, sample sizes, number of SNPs, number of covariates, LD patterns, fraction of shared controls, and trait polygenicity. In each experiment we varied one or more of the above parameters while keeping the others fixed. The default simulation parameters used 1% prevalence, 50% heritability, and 50% genetic correlation, with each study having 2,000 case subjects, 1,000 unique and 1,000 overlapping control subjects, and 10,000 SNPs whose LD decays exponentially with distance, and with a correlation of between 25% and 90% between consecutive SNPs (consequently, the correlation between every pair of SNPs separated by at least 25 SNPs is <0.001 in all settings). In most simulations all SNPs influenced the phenotype, though we verified that relaxing this assumption does not affect the results (see details below). 100 simulations were conducted for each unique combination of settings.

The examined methods included (1) PCGC-s; (2) PCGC-s-LD, which is an approximate version of PCGC-s that uses external LD estimates (but uses data about overlapping individuals; Supplemental Methods); (3) LDSC with omitted covariates (LDSC-omit); and (4) REML, using the implementation in GCTA[6] (exact execution details are provided in the Supplemental Methods). Note that PCGC-s is exactly equivalent to PCGC when all required summary

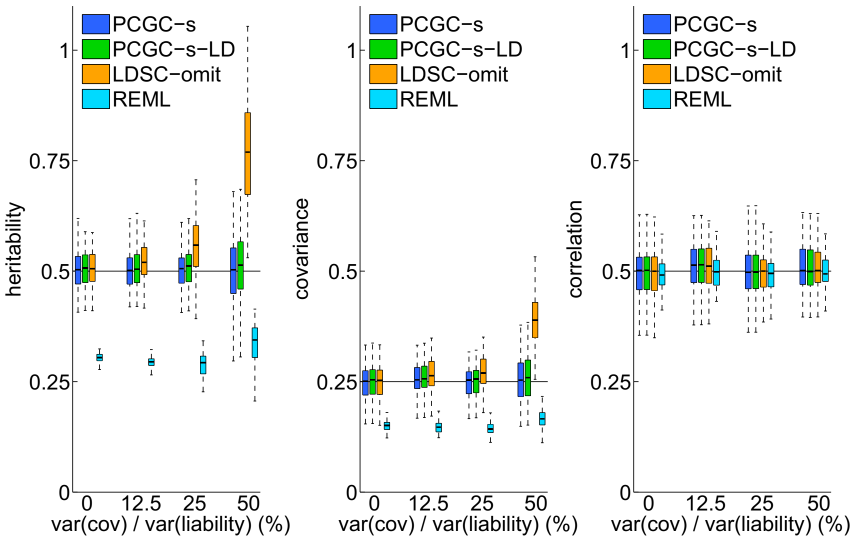

Figure 1. The Effect of Covariate Strength
PCGC-s and PCGC-s-LD estimate all parameters accurately under all settings; LDSC-omit estimates of heritability and genetic covariance become increasingly inaccurate as the covariates strength increases; REML misestimates heritability and genetic covariance under all settings. All methods estimate genetic correlation accurately. The black horizontal lines indicate the true parameter values. 100 experiments were performed for each unique combination of settings, and each study included 2,000 case subjects and 2,000 control subjects.

statistics are provided. LDSC-omit refers to LDSC that does not include any covariates in the analysis and was used because explicit inclusion of covariates can lead to highly biased estimates, as demonstrated below. In most simulations LDSC-omit was based on our own implementation, to avoid confounding the analysis by implementation details. Specifically, our implementation of LDSC-omit used a predetermined intercept and did not weight summary statistics by their posterior variance, similarly to PCGC-s-LD (see Discussion for elaboration on these issues). In additional simulations described below, we demonstrated that when using the ldsc software instead of our own implementation, LDSC-omit became less accurate.

Our first experiment examined the impact of covariate effect magnitude on the estimation of heritability, genetic covariance, and genetic correlation. We simulated datasets with five binary covariates that explained various fractions of the liability variance, where the first covariate accounted for 95% of the aggregated covariates effect. All methods estimated correlation well, but PCGC-s and PCGC-s-LD were the only methods that estimated the two other quantities accurately (Figure 1). Both PCGC-s and PCGC-s-LD estimated heritability significantly more accurately than LDSC-omit ($p < 2.1 \times 10^{-2}$, $p < 1.7 \times 10^{-6}$, $p < 6.5 \times 10^{-24}$ for covariates explaining 12.5%, 25%, and 50% of the liability variance, respectively; binomial test for PCGC-s-LD; PCGC-s results were effectively the same). The accuracy of LDSC-omit improved as effect sizes became smaller; LDSC-omit and PCGC give very similar estimates in the absence of covariates, as expected from theory (Supplemental Methods). REML consistently underestimated heritability despite using the correction for case-control ascertainment implemented in GCTA.[1] We note that the extent of under-estimation by REML is not fixed with a known ratio but depends on various unknown parameters.[14] We also obtained similar results when ignoring the contribution of covariates to the liability variance (Material and Methods, Figure S1).

The next experiment examined the implications of having normal versus non-normal covariate effects, by considering three settings: (1) a single binary covariate, (2) a single normally distributed covariate, and (3) 20 equally strong binary covariates. In all settings the covariates jointly explained 40% of the liability variance. Setting 1 encodes a non-normal aggregated effect, whereas settings 2 and 3 encode a normal and an approximately normal effect (owing to the central limit theorem), respectively. In setting 1, LDSC-omit was substantially less accurate than PCGC-s ($p < 3.21 \times 10^{-19}$; binomial test) and PCGC-s-LD ($p < 2.73 \times 10^{-20}$; binomial test), because its underlying model is violated in the presence of strong non-normally distributed covariates (Figure 2, Material and Methods). The bias of LDSC-omit decreased when decreasing the magnitude of the covariate effects, similarly to the results shown in Figure 1.

In additional experiments, we simulated data with one strong and four weak binary covariates as in the first experiment, where the covariates jointly explained 25% of the liability variance, and verified that the results remained similar under various levels of heritability (Figure S2), genetic correlation (Figure S3), prevalence (Figure S4), LD (Figure S5), fraction of shared controls (Figure S6), numbers of covariates (Figure S7), sample sizes (Figure S8), numbers of simulated causal SNPs (Figure S9), and trait polygenicity (Figure S10). We also explored running LDSC-omit using the ldsc software (Figure S11) and using logistic regression-based summary statistics (Figures S12 and S13).

We also examined the effect of using LDSC without omitting covariates, by regressing measured covariates out of the phenotypes and genotypes prior to computing summary statistics, as previously recommended.[16,40] Our results demonstrate that LDSC estimates are severely down-biased in this setting, with an average bias of more than 10% in heritability and covariance estimation, and of more than 5% in correlation estimation, under realistic settings (Figures S14 and S15).

Next, we performed a set of experiments with a MAF and LD-dependent genetic architectures. Specifically, we simulated phenotypes according to the LDAK model,[8] which

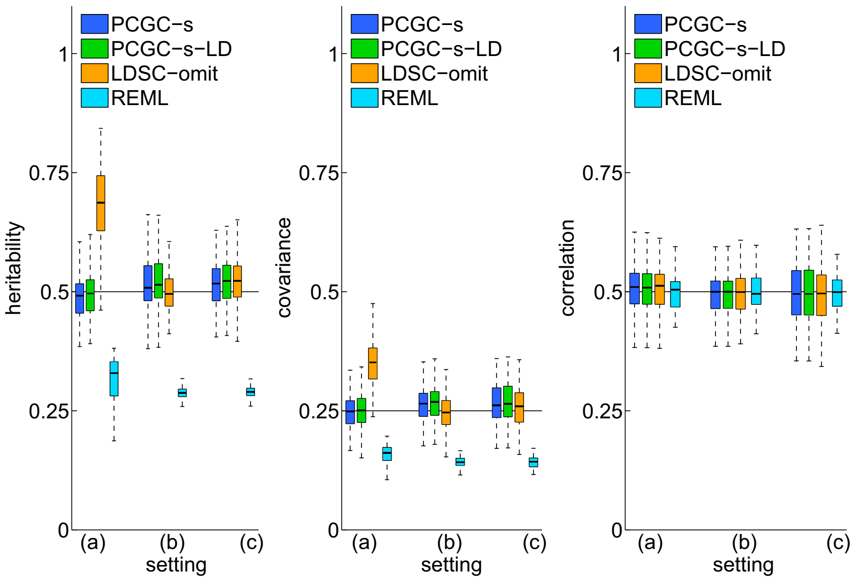

Figure 2. The Effect of the Covariate Effects Distribution
Setting (a) includes a single binary covariate; setting (b) includes a single normally distributed covariate; setting (c) includes 20 binary variables with equal strength, yielding an approximately normal aggregated effect owing to the central limit theorem. PCGC-s and PCGC-s-LD are the only methods that accurately estimate heritability and genetic covariance in setting (a), where the covariates effects distribution is far from normal. 100 experiments were performed for each unique combination of settings, and each study included 2,000 case subjects and 2,000 control subjects.

replaces the standard genetic similarity matrix $\boldsymbol{G}_t = \boldsymbol{X}_t\boldsymbol{X}_t^T/m$ with the matrix $\boldsymbol{G}_t = \boldsymbol{X}_t\boldsymbol{W}\boldsymbol{X}_t^T/M$, where $\boldsymbol{W} = \mathrm{diag}[(p^j(1-p^j))^{0.75}w^j]$, p^j is the MAF of SNP j, w^j minimizes the L_2 norm of $(1-\sum_{k=1}^{m}(r^{k,j})^2 w^k)$, and $M = \sum_k W_{kk}$. All methods yielded biased estimates of heritability and genetic covariance when using the incorrect genetic similarity matrix $\boldsymbol{G}_t = \boldsymbol{X}_t\boldsymbol{X}_t^T/m$ (Figure S16). However, PCGC-s and PCGC-s-LD became unbiased when using the correct genetic similarity matrix, whereas the other methods remained biased even when using the correct genetic similarity matrix (Figure S17). Interestingly, all methods yielded empirically unbiased estimates of genetic correlation even when using an incorrect model, suggesting that the approximation errors cancel themselves out, similarly to the patterns observed when not correctly modeling case-control ascertainment. A numerical summary of all the results reported in this section is provided in Table S2.

Finally, we note that PCGC-s-LD is highly computationally efficient. Since PCGC-s-LD uses only summary statistics, it can perform estimation for data with millions of variants and hundreds of thousands of individuals in less than 1 hr (results not shown).

Estimating the Genetic Architecture of Schizophrenia and Bipolar Disorder

To demonstrate the behavior of the methods on real data, we studied the heritability and genetic correlation of schizophrenia (SCZ)[42] and bipolar disorder (BP).[43] To prevent confounding due to population stratification,[44] we restricted the analysis to two highly concordant Swedish datasets consisting of 1,745 SCZ-affected case subjects, 1,268 BP-affected case subjects, and 6,293 control subjects, 2,566 of which are shared between the studies[42,43] (Supplemental Methods). The covariates included 10 principal components and sex, which is a major risk factor for both diseases.

The PCGC-s heritability estimates for SCZ and BP were 39.2% and 41.7%, respectively. The estimated genetic correlation was 42.4%, which is substantially lower than previous estimates of 68% using REML[10] and 79% using LDSC.[2] We further verified that when omitting covariates, the PCGC-s estimates increased to 60%, suggesting that incorrect treatment of non-genetic risk factors can lead to inflated estimates. When invoking LDSC on the same data using the ldsc software, the estimated correlation could not be computed when omitting covariates due to negative estimated heritabilities and was 15.8% when regressing the covariates out of the phenotypes (Table 1). We also estimated the heritabilities and genetic correlation under the LDAK model[8] and obtained very similar estimates (Table S3). Namely, the heritability estimates for SCZ and BP were 40.0% and 46.1%, respectively, and the estimated genetic correlation was 43.8%. Overall, we conclude that improper handling of covariates and of sample overlap in case-control studies can lead to substantially biased estimates and to incorrect conclusions regarding the genetic architecture of genetic diseases.

Estimating the Genetic Architecture of Type 1 Diabetes and Coronary Artery Disease

To further evaluate PCGC-s, we studied the correlation between type 1 diabetes (T1D) and coronary artery disease (CAD), using data from the Wellcome Trust Case Control Consortium 1 (WTCCC1).[45] It is known that T1D is associated with an increased risk for CAD,[46] but the role of genetics in this association is not clear. We chose to explore this example because of the expected impact of covariates on the result: T1D is very strongly affected by SNPs in the major histocompatibility complex (MHC) region, and sex is a major risk factor for CAD. We thus modeled the effects of these risk factors as fixed rather than random and investigated the implications of inclusion and exclusion of these covariates. The analysis details are provided in the Supplemental Methods.

The results demonstrated the existence of a positive genetic correlation between T1D and CAD and corroborated

Table 1. Results of Real Data Analysis of Psychiatric Disorders

Covariates		SCZ		BP		Correlation
		$\widehat{\sigma}_g^2$	$\widehat{h}^2$	$\widehat{\sigma}_g^2$	$\widehat{h}^2$	
Omitted	PCGC-s	0.127 (0.059)	0.127 (0.059)	0.259 (0.044)	0.259 (0.044)	0.561 (0.149)
	PCGC-s-LD	0.139 (0.047)	0.139 (0.047)	0.282 (0.057)	0.282 (0.057)	0.602 (0.178)
	LDSC-omit	0.467 (0.101)	0.467 (0.101)	0.293 (0.109)	0.293 (0.109)	0.451 (0.190)
	LDSC-omit +intercept	0.467 (0.101)	0.467 (0.101)	0.293 (0.109)	0.293 (0.109)	–
Included	PCGC-s	0.399 (0.068)	0.392 (0.062)	0.426 (0.051)	0.417 (0.045)	0.437 (0.077)
	PCGC-s-LD	0.438 (0.059)	0.430 (0.049)	0.465 (0.059)	0.455 (0.058)	0.424 (0.084)
	LDSC	0.412 (0.084)	0.405 (0.070)	0.356 (0.105)	0.348 (0.103)	0.527 (0.176)
	LDSC+intercept	0.412 (0.084)	0.405 (0.077)	0.356 (0.105)	0.349 (0.093)	0.158 (0.112)

Shown are the estimated values of the genetic variance σ_g^2 (also termed the conditional heritability in this paper), the marginal heritability h^2 (which is equal to σ_g^2 when no covariates are present and smaller than σ_g^2 in the presence of covariates) and the genetic correlation. Standard errors were computed via a block jackknife of 200 blocks of consecutive SNPs. LDSC+intercept is the LDSC estimator when fitting an intercept from the data.[2] LDSC-omit is different from PCGC-s-LD with omitted covariates because of differences in the predetermined intercept value due to normalization (Supplemental Methods). LDSC results were computed using the ldsc software. Values marked with "–" could not be computed because of negative or illegal parameter estimates.

the simulation studies (Table 2, Table S4). As expected, inclusion of covariates had a minor effect on PCGC-s estimates, decreasing the heritability estimate for T1D from 23.7% to 18.3% and for CAD from 40.5% to 39.9%, and slightly increasing the genetic correlation estimate from 18.1% to 19.2%. The LDSC heritability estimates for T1D and CAD when omitting covariates (35% and 58.8%, respectively) were greater than those of PCGC-s (consistent with our simulation results) and the correlation estimate was also greater (28.4%). LDSC heritability estimates were nonsensical (non-positive or greater than one) when including covariates or fitting an intercept rather than using a predetermined one. REML estimation of genetic correlation using gcta failed to converge.

We conclude that accounting for covariates can substantially affect heritability and genetic correlation estimates. However, we caution that the results are sensitive to preprocessing of the data (Tables S5–S7, Supplemental Methods; see Discussion). We also present genetic correlation estimates between all phenotypes included in the WTCCC1 study, confirming some well-known significant correlations, such as between hypertension and coronary artery disease; and others that have been tentatively suggested in the literature, such as between rheumatoid arthritis and coronary artery disease[47,48] (Table S8).

Discussion

Our major conclusions regarding the existing approaches can be summarized as follows. (1) REML severely misestimates heritability and genetic covariance in case-control studies under all settings (as has been pointed out previously[7,14,41]). In settings without binary covariates, REML accurately estimates genetic correlation, but it can become slightly biased in the presence of such covariates. (2) LDSC estimates are accurate in the absence of covariates but can become biased in the presence of binary covariates with strong effects. Importantly, regressing covariates out of phenotypes prior to running LDSC can lead to a very severe bias and should always be avoided. We further caution that the software implementation of LDSC can lead to different estimates than those of PCGC-s even in the absence of covariates due to different data preprocessing procedures, as discussed below. (3) PCGC accurately estimates all quantities of interest directly or with summary statistics. (4) Standard summary statistics cannot be used to estimate genetic correlation for traits with binary non-genetic risk factors; we propose here a novel formulation of privacy-preserving summary statistics which can be used for this task.

Another potentially problematic aspect of genetic correlation estimation is analysis of cohorts from ancestrally divergent populations. Our preliminary analysis demonstrated that analysis of such cohorts can lead to inflated and unstable genetic correlation estimates for all methods, even when using a large number of PCs as covariates (results not shown). We therefore opted to focus our analysis on two Swedish cohorts. Previous estimates of the genetic correlation between schizophrenia and bipolar disorder were based on cohorts from divergent European populations, which may be another reason for the large difference between our estimates and previous ones.[2,10,44]

When comparing different methods, it is important to distinguish between the underlying mathematics and the software implementation. Even though PCGC-s and LDSC are roughly equivalent in the absence of covariates, the software implementation of PCGC-s is careful to perform case-control-aware data preprocessing (e.g., avoiding in-sample SNP standardization, and avoid assuming that the diagonal of the genetic similarity matrix is exactly 1.0; Supplemental Methods). This can lead to major differences between the estimates of the software implementations in real data analysis. We therefore recommend that

Table 2. Results of Real Data Analysis of T1D and CAD

Covariates		T1D		CAD		
		$\hat{\sigma}_g^2$	$\hat{h}^2$	$\hat{\sigma}_g^2$	$\hat{h}^2$	Correlation
Omitted	PCGC-s	0.237 (0.044)	0.237 (0.044)	0.405 (0.063)	0.405 (0.063)	0.181 (0.115)
	PCGC-s-LD	0.245 (0.045)	0.245 (0.045)	0.420 (0.065)	0.420 (0.065)	0.181 (0.115)
	LDSC-omit	0.350 (0.046)	0.350 (0.046)	0.588 (0.066)	0.588 (0.066)	0.284 (0.074)
	LDSC-omit +intercept	0.013 (0.105)	0.013 (0.105)	0.020 (0.109)	0.020 (0.109)	–
Included	PCGC-s	0.241 (0.066)	0.183 (0.050)	0.435 (0.070)	0.399 (0.062)	0.192 (0.139)
	PCGC-s-LD	0.250 (0.069)	0.190 (0.052)	0.451 (0.065)	0.413 (0.060)	0.191 (0.139)
	LDSC	−1.75 (0.038)	–	−0.33 (0.058)	–	–
	LDSC+intercept	−0.03 (0.046)	–	−0.07 (0.09)	–	–

The table fields are the same as in Table 1. LDSC results are based on our own implementation to provide a detailed comparison with PCGC-s that is not confounded by implementation details. Results using the ldsc software are provided in Table S4.

researchers use our software implementation of PCGC-s for analysis of case-control studies regardless of the presence of covariates, because PCGC-s is careful to preprocess case-control data correctly.

An important issue often raised in the context of heritability estimation regards the validity of the assumed model. One specific concern concerns the use of allele frequency and LD-dependent architectures.[8,39,49,50] While important, this concern is not directly related to our results, as PCGC-s can accommodate arbitrary linear genetic architectures (Supplemental Methods). Additional concerns include the difference between "SNP heritability" and "narrow sense heritability" which assumes that all causal SNPs are measured[7,51] and the potentially larger difference between narrow sense heritability and the true genetic heritability in the presence of non-additive effects.[52] These concerns are well founded and should certainly be addressed in practice. However, they are not directly related to our study, which focuses on the performance of different methods when the model assumed by these methods (liability threshold model and additivity) holds. We believe our conclusion, that commonly used methods can be biased under their own modeling assumptions, is of major interest even given the concerns about the validity of the assumptions themselves.

An important question that has been debated recently concerns the relationship between causal effect sizes and MAF and LD patterns.[8,39,49] PCGC-s can be readily modified to use any linear genetic architecture and can thus accommodate different genetic architectures. Several researchers recently advocated comparing between different models via the data likelihood,[8] but unfortunately exact likelihood estimation is infeasible in ascertained case-control studies. The determination of genetic architectures under case-control studies is therefore a potential line of future work.

Finally, the LDSC framework includes several techniques not considered in this work: estimation of the contribution of functional annotations to the liability variance,[53] improved estimation by weighting of summary statistic,[16] and fitting an intercept from the data rather than using a predetermined one.[16] The first technique can be readily adapted into the PCGC-s framework (Supplemental Methods). We do not recommend using the other techniques in case-control studies, as the derivations underlying these techniques assume an additive phenotype with genotype-environment independence. Adapting these procedures into case-control studies under a formal theoretical framework remains a potential avenue for future work.

Supplemental Data

Supplemental Data include 17 figures, 8 tables, and Supplemental Methods (mathematical derivations) and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.06.002.

Acknowledgments

This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from https://www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under awards 076113, 085475, and 090355. We wish to thank the Swedish Bipolar Collection (SWEBIC, PI Mikael Landén) for making data available. The authors thank Noah Zaitlen, Joel Mefford, and Na Cai for useful discussions.

This work was supported by the Israeli Science Foundation grant 1804/16. This collaboration started at the Computational Genomics Summer Institute funded by NIH grant GM112625. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of Interests

The authors declare no competing interests.

Received: February 2, 2018
Accepted: June 6, 2018
Published: July 5, 2018

Web Resources

PCGC-s, https://github.com/omerwe/PCGCs

References

1. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.
2. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., Duncan, L., Perry, J.R., Patterson, N., Robinson, E.B., et al.; ReproGen Consortium; Psychiatric Genomics Consortium; and Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3 (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.
3. Palla, L., and Dudbridge, F. (2015). A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. Am. J. Hum. Genet. *97*, 250–259.
4. Zhou, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. Ann. Appl. Stat. *11*, 2027–2051.
5. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
6. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
7. Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., de Candia, T.R., Lee, S.H., Wray, N.R., Kendler, K.S., et al.; Schizophrenia Working Group of Psychiatric Genomics Consortium (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat. Genet. *47*, 1385–1392.
8. Speed, D., Cai, N., Johnson, M.R., Nejentsev, S., Balding, D.J.; and UCLEB Consortium (2017). Reevaluation of SNP heritability in complex human traits. Nat. Genet. *49*, 986–992.
9. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M., and Wray, N.R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics *28*, 2540–2542.
10. Lee, S.H., Ripke, S., Neale, B.M., Faraone, S.V., Purcell, S.M., Perlis, R.H., Mowry, B.J., Thapar, A., Goddard, M.E., Witte, J.S., et al.; Cross-Disorder Group of the Psychiatric Genomics Consortium; and International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat. Genet. *45*, 984–994.
11. Chen, G.B., Lee, S.H., Brion, M.J., Montgomery, G.W., Wray, N.R., Radford-Smith, G.L., Visscher, P.M.; and International IBD Genetics Consortium (2014). Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. Hum. Mol. Genet. *23*, 4710–4720.
12. Haseman, J.K., and Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. *2*, 3–19.
13. Chen, G.-B. (2014). Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. Front. Genet. *5*, 107.
14. Golan, D., Lander, E.S., and Rosset, S. (2014). Measuring missing heritability: inferring the contribution of common variants. Proc. Natl. Acad. Sci. USA *111*, E5272–E5281.
15. Bonnet, A. (2018). Heritability estimation of diseases in case-control studies. Electron. J. Stat. *12*, 1662–1716.
16. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M.; and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
17. Robinson, E.B., St Pourcain, B., Anttila, V., Kosmicki, J.A., Bulik-Sullivan, B., Grove, J., Maller, J., Samocha, K.E., Sanders, S.J., Ripke, S., et al.; iPSYCH-SSI-Broad Autism Group (2016). Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. Nat. Genet. *48*, 552–555.
18. The Brainstorm Consortium, Anttila, V., Bulik-Sullivan, B., Finucane, H.K., Walters, R.K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G.J., Gormley, P., et al. (2018). Analysis of shared heritability in common disorders of the brain. Science *360*, eaap8757.
19. Lo, M.-T., Hinds, D.A., Tung, J.Y., Franz, C., Fan, C.-C., Wang, Y., Smeland, O.B., Schork, A., Holland, D., Kauppi, K., et al. (2017). Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. Nat. Genet. *49*, 152–156.
20. Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., Yi, Q., Li, C., Li, X., Shen, J., et al. (2017). Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. Nat. Genet. *49*, 1576–1583.
21. Sniekers, S., Stringer, S., Watanabe, K., Jansen, P.R., Coleman, J.R.I., Krapohl, E., Taskesen, E., Hammerschlag, A.R., Okbay, A., Zabaneh, D., et al. (2017). Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. Nat. Genet. *49*, 1107–1112.
22. Hobbs, B.D., de Jong, K., Lamontagne, M., Bossé, Y., Shrine, N., Artigas, M.S., Wain, L.V., Hall, I.P., Jackson, V.E., Wyss, A.B., et al.; COPDGene Investigators; ECLIPSE Investigators; LifeLines Investigators; SPIROMICS Research Group; International COPD Genetics Network Investigators; UK BiLEVE Investigators; and International COPD Genetics Consortium (2017). Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. Nat. Genet. *49*, 426–432.
23. Luciano, M., Hagenaars, S.P., Davies, G., Hill, W.D., Clarke, T.-K., Shirali, M., Harris, S.E., Marioni, R.E., Liewald, D.C., Fawns-Ritchie, C., et al. (2018). Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. Nat. Genet. *50*, 6–11.
24. Lane, J.M., Liang, J., Vlasac, I., Anderson, S.G., Bechtold, D.A., Bowden, J., Emsley, R., Gill, S., Little, M.A., Luik, A.I., et al. (2017). Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics

with neuropsychiatric and metabolic traits. Nat. Genet. *49*, 274–281.

25. Ji, S.-G., Juran, B.D., Mucha, S., Folseraas, T., Jostins, L., Melum, E., Kumasaka, N., Atkinson, E.J., Schlicht, E.M., Liu, J.Z., et al.; UK-PSC Consortium; International IBD Genetics Consortium; and International PSC Study Group (2017). Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. Nat. Genet. *49*, 269–273.
26. Day, F.R., Thompson, D.J., Helgason, H., Chasman, D.I., Finucane, H., Sulem, P., Ruth, K.S., Whalen, S., Sarkar, A.K., Albrecht, E., et al.; LifeLines Cohort Study; InterAct Consortium; kConFab/AOCS Investigators; Endometrial Cancer Association Consortium; Ovarian Cancer Association Consortium; and PRACTICAL consortium (2017). Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. Nat. Genet. *49*, 834–841.
27. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.
28. McKay, J.D., Hung, R.J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D.C., Caporaso, N.E., Johansson, M., Xiao, X., Li, Y., et al.; SpiroMeta Consortium (2017). Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat. Genet. *49*, 1126–1132.
29. Pardiñas, A.F., Holmans, P., Pocklington, A.J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S.E., Bishop, S., Cameron, D., Hamshere, M.L., et al.; GERAD1 Consortium; CRESTAR Consortium; GERAD1 Consortium; CRESTAR Consortium; GERAD1 Consortium; and CRESTAR Consortium (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat. Genet. *50*, 381–389.
30. Hammerschlag, A.R., Stringer, S., de Leeuw, C.A., Sniekers, S., Taskesen, E., Watanabe, K., Blanken, T.F., Dekker, K., Te Lindert, B.H.W., Wassing, R., et al. (2017). Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and metabolic traits. Nat. Genet. *49*, 1584–1592.
31. Ferreira, M.A., Vonk, J.M., Baurecht, H., Marenholz, I., Tian, C., Hoffman, J.D., Helmer, Q., Tillander, A., Ullemar, V., van Dongen, J., et al.; 23andMe Research Team; AAGC collaborators; BIOS consortium; and LifeLines Cohort Study (2017). Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. Nat. Genet. *49*, 1752–1757.
32. Kemp, J.P., Morris, J.A., Medina-Gomez, C., Forgetta, V., Warrington, N.M., Youlten, S.E., Zheng, J., Gregson, C.L., Grundberg, E., Trajanoska, K., et al. (2017). Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. Nat. Genet. *49*, 1468–1475.
33. Sanchez-Roige, S., Fontanillas, P., Elson, S.L., Pandit, A., Schmidt, E.M., Foerster, J.R., Abecasis, G.R., Gray, J.C., de Wit, H., Davis, L.K., et al.; 23andMe Research Team (2018). Genome-wide association study of delay discounting in 23,217 adult research participants of European ancestry. Nat. Neurosci. *21*, 16–18.
34. Pasaniuc, B., and Price, A.L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. *18*, 117–127.
35. Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am. J. Hum. Genet. *99*, 139–153.
36. Zhu, X., and Stephens, M. (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. Ann. Appl. Stat. *11*, 1561–1592.
37. Dempster, E.R., and Lerner, I.M. (1950). Heritability of Threshold Characters. Genetics *35*, 212–236.
38. Falconer, D.S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann. Hum. Genet. *29*, 51–76.
39. Gazal, S., Finucane, H.K., and Price, A.L. (2018). Reconciling S-LDSC and LDAK functional enrichment estimates. bioRxiv. https://www.biorxiv.org/content/early/2018/01/30/256412.
40. Bulik-Sullivan, B. (2015). Relationship between LD Score and Haseman-Elston Regression. bioRxiv, 018283. https://www.biorxiv.org/content/early/2015/04/20/018283.
41. Hayeck, T.J., Zaitlen, N.A., Loh, P.R., Vilhjalmsson, B., Pollack, S., Gusev, A., Yang, J., Chen, G.B., Goddard, M.E., Visscher, P.M., et al. (2015). Mixed model with correction for case-control ascertainment increases association power. Am. J. Hum. Genet. *96*, 720–730.
42. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
43. Charney, A.W., Ruderfer, D.M., Stahl, E.A., Moran, J.L., Chambert, K., Belliveau, R.A., Forty, L., Gordon-Smith, K., Di Florio, A., Lee, P.H., et al. (2017). Evidence for genetic heterogeneity between clinical subtypes of bipolar disorder. Transl. Psychiatry *7*, e993–e993.
44. Bhatia, G., Gusev, A., Loh, P.-R., Finucane, H.K., Vilhjalmsson, B.J., Ripke, S., Purcell, S., Stahl, E., Daly, M., de Candia, T.R., et al. (2016). Subtle stratification confounds estimates of heritability from rare variants. bioRxiv, 048181. https://www.biorxiv.org/content/early/2016/04/12/048181.
45. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
46. Orchard, T.J., Costacou, T., Kretowski, A., and Nesto, R.W. (2006). Type 1 diabetes and coronary artery disease. Diabetes Care *29*, 2528–2538.
47. Goodson, N. (2002). Coronary artery disease and rheumatoid arthritis. Curr. Opin. Rheumatol. *14*, 115–120.
48. Maradit-Kremers, H., Nicola, P.J., Crowson, C.S., Ballman, K.V., and Gabriel, S.E. (2005). Cardiovascular death in rheumatoid arthritis: a population-based study. Arthritis Rheum. *52*, 722–732.
49. Yang, J., Zeng, J., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2017). Concepts, estimation and interpretation of SNP-based heritability. Nat. Genet. *49*, 1304–1310.
50. Speed, D., and Balding, D. (2018). Better estimation of SNP heritability from summary statistics provides a new understanding of the genetic architecture of complex traits. bioRxiv. https://www.biorxiv.org/content/early/2018/03/19/284976.
51. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A., Lee, S.H., Robinson, M.R., Perry, J.R., Nolte, I.M., van Vliet-Ostaptchouk, J.V., et al.; LifeLines Cohort Study (2015). Genetic

variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. *47*, 1114–1120.

52. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA *109*, 1193–1198.

53. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al.; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; and RACI Consortium (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.