Summary statistic analyses can mistake confounding bias for heritability

John B. Holmes*[1], Doug Speed †[2,3], and David J. Balding ‡[1,3]

[1]Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne, Australia.
[2]Aarhus Institute of Advanced Studies (AIAS), Aarhus University, Denmark.
[3]UCL Genetics Institute, University College London, United Kingdom.

June 5, 2019

Abstract

LD SCore regression (LDSC) has become a popular approach to estimate confounding bias, heritability and genetic correlation using only genome wide association study (GWAS) test statistics. SumHer is a newly-introduced alternative with similar aims. We show using theory and simulations that both approaches fail to adequately account for confounding bias, even when the assumed heritability model is correct. Consequently, these methods may estimate heritability poorly if there was inadequate adjustment for confounding in the original GWAS analysis. We also show that choice of summary statistic for use in LDSC or SumHer can have a large impact on resulting inferences. Further, covariate adjustments in the original GWAS can alter the target of heritability estimation, which can be problematic for test statistics from a meta-analysis of GWAS with different covariate adjustments.

Introduction

LD SCore regression (LDSC) uses genome-wide association test statistics to estimate confounding bias, the heritability tagged by SNPs (h^2_{SNP}), how h^2_{SNP} is distributed across the genome and the genetic correlation of pairs of traits (Bulik-Sullivan et al., 2015a,b; Finucane et al., 2015; Gazal et al., 2017). Its use of test statistics rather than individual genotype data means that it is effectively unlimited in sample size, and can make use of published studies that do not release the genotypes of participants. Moreover the test statistics can be obtained from a single GWAS or from a meta-analysis of multiple GWAS. These advantages have led to LDSC being very widely used.

*corresponding author, email: john.holmes@unimelb.edu.au
†email: doug@aias.au.dk
‡email: david.balding@unimelb.edu.au

LDSC regresses the test statistic at each SNP on an "LD score", defined as a sum of linkage disequilibrium (LD) coefficients over neighbouring SNPs. The regression slope and intercept are interpreted as, respectively, h^2_{SNP} and confounding bias not corrected in the GWAS analysis. SumHer (Speed and Balding, 2019) and S-LDSC (Finucane et al., 2015; Gazal et al., 2017) generalise LDSC by introducing weights into the LD score. The weights correspond to a heritability model that relates the expected heritability of a SNP to its properties known *a priori*. SumHer uses fixed, SNP-specific weights reflecting LD and minor allele fraction (MAF). In the most recent version of S-LDSC (Gazal et al., 2017), weights based on LD and MAF as well as functional annotations are estimated in the summary statistic analysis. HESS (Shi et al., 2016) and RSS (Zhu and Stephens, 2017) are other summary statistic methods that require more information than association test statistics.

Researchers using LDSC or SumHer usually have not performed the underlying GWAS analyses, but use test statistics obtained from public data repositories (Zheng et al., 2017) that may lack information needed to check the assumptions underlying these methods. Here, we examine the validity of these assumptions under a range of scenarios. We do not revisit the topic of the underlying heritability model (Speed and Balding, 2019), rather we will highlight problems that arise even when the simulation and analysis heritability models are the same.

We derive expected values of association statistics and show that confounding effects are SNP dependent, and correlated with LD score (Appendices 2-3), contravening a fundamental assumption of LDSC and SumHer. Thus a global adjustment term can fail to remove confounding effects, although a multiplicative adjustment can correct over-conservatism arising from use of genomic control (Speed and Balding, 2019).

We illustrate the magnitude of the problem through simulations. Our investigation covers two summary statistics and we show that inferences from LDSC or SumHer can be greatly impacted by this choice. Further, we show that the definition of h^2_{SNP} targeted by LDSC and SumHer varies with the covariates fitted in the GWAS analyses. This can be important in meta-analysis: a summary statistic heritability analysis based on studies with different covariate adjustments will merge estimates of different parameters.

A general model

We derive approximations for $\mathrm{E}[S_j]$ and perform simulation studies using a highly general phenotype model,

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I}), \tag{1}$$

where μ is an intercept, $\mathbf{Z}$ an $n \times m$ matrix of standardised SNP genotypes, $\boldsymbol{\alpha}$ a vector of SNP effect sizes, and $\boldsymbol{\Sigma}$ a diagonal matrix with jth entry σ_j^2. The $n \times p$ matrix $\mathbf{X}$ contains column-standardised

covariate values, while $\boldsymbol{\beta}$ is a vector of covariate effects. If $\mathrm{Cor}(\mathbf{X}, \mathbf{Z}) \neq \mathbf{0}$, the $\mathbf{X}$ are confounders for genetic association analysis. The most important example is population structure when both $\mathbf{X}$ and $\mathbf{Z}$ vary with, for example, geography or social strata.

The GCTA model (Yang et al., 2011a) is the special case of (1) in which σ_j^2 is constant over SNPs. This assumption also underlies LDSC. The LDAK model (Speed et al., 2017) is another special case of (1) in which $\sigma_j^2 \propto w_j(f_j(1-f_j))^{1+\alpha}$, where the w_j reflect local LD at SNP j, f_j is the MAF of SNP j and α is a parameter that reflects selection on the phenotype. The value $\alpha = -0.25$ has been shown to fit well over many human phenotypes (Speed et al., 2017). SumHer can be used with any heritability model, but here we implement SumHer with the LDAK values for σ_j^2 and $\alpha = -0.25$.

When (1) is used as a simulation model, usually only a subset of the available SNPs are assigned non-zero effects. When used as an analysis model, because the causal SNPs are unknown all available SNPs should be included in (1). This mismatch between simulation and analysis models arises as it is impossible in practice to limit analyses to causal SNPs.

Choice of test statistics

LDSC and SumHer both fit a linear regression to summary statistics $S_j, j = 1, \ldots, m$, obtained from a GWAS on n individuals:

$$\mathrm{E}[S_j] = \begin{cases} A + nh^2_{\mathrm{SNP}} \sum_i r_{ij}^2/m & \text{LDSC} \\ C\big(1 + n\sum_i r_{ij}^2 h_i^2\big), & \text{SumHer,} \end{cases} \tag{2}$$

where $h_j^2 \geq 0$ is interpreted as the expected heritability attributable uniquely to SNP j, with $h^2_{\mathrm{SNP}} = \sum_{j=1}^m h_j^2$. In (2), r_{ij}^2 is an estimated LD coefficient with $r_{ii} = 1$ (see Methods), while A and C are alternative adjustments for confounding effects not accounted for in the GWAS analysis. Estimates of C using SumHer were reported to be much lower than the corresponding estimates of A from LDSC (Speed and Balding, 2019), but this was due to the difference in heritability model rather than whether the confounding term was additive or multiplicative. SumHer found that many GWAS had over-corrected for confounding ($C < 1$), whereas LDSC analyses of the same data typically found $A > 1$, indicating a need for further confounding adjustment (Speed and Balding, 2019).

In practice, S_j is often the Wald statistic T_j^2 from a classical simple linear regression (Yang et al., 2011b; Bulik-Sullivan et al., 2015a; Finucane et al., 2015), which can be inferred from p-values. Its null distribution is $F_{1,n-2}$, which converges to χ_1^2 as n increases. However, LDSC was proposed assuming that $S_j = n\hat{\alpha}_j^2$, where α_j is the effect of SNP j when both $\mathbf{Z}$ and $\mathbf{y}$ are standardised (Bulik-Sullivan et al., 2015a). Assuming that no covariates were fitted, S_j is $n/(n-1)$ times the standardised regression sum of squares $S\tilde{S}R_j = (n-1)SSR_j/SST$. When covariates are included this equality

no longer holds, but we check using simulation that $\mathrm{E}[S\tilde{S}R_j] \approx \mathrm{E}[n\hat{\alpha}_j^2]$, and where convenient we compute $\mathrm{E}[S\tilde{S}R_j]$ rather than $\mathrm{E}[n\hat{\alpha}_j^2]$.

Recently, GWAS test statistics have often been derived from a mixed regression model (Lippert et al., 2011; Loh et al., 2015) in which SNP j is tested while other SNPs are used to compute the variance structure of a random effect modelling cryptic kinship and population structure. We report the expectation of a general mixed model test statistic, finding that it depends on quantities not usually available from GWAS data repositories and that in general h^2_{SNP} is non-identifiable (Appendix 3.7).

Two definitions of h^2_{SNP}

We define $\sigma_y^2 = \sum_{j=1}^m \sigma_j^2 + \sigma_e^2$, the phenotypic variance after conditioning on covariates/confounders $\mathbf{X}$. However $\mathbf{X}$ may be unrecorded, or omitted from analysis, in which case σ_y^2 cannot be estimated, and only the total phenotypic variance $\sigma_y^2 + \sigma_c^2$ is available, where $\sigma_c^2 = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/(n-1)$ and $'$ denotes transpose. This leads to two definitions of the heritability of SNP j (Weissbrod et al., 2018):

$$h^2_{j,\mathrm{a}} = \sigma_j^2/\sigma_y^2, \tag{3}$$

$$h^2_{j,\mathrm{b}} = \sigma_j^2/(\sigma_y^2 + \sigma_c^2). \tag{4}$$

The conditional heritability, $h^2_{j,\mathrm{a}}$, is standard when the $\mathbf{X}$ are modelled as fixed effects (Mrode, 2014; Pirinen et al., 2013), while the marginal heritability, $h^2_{j,\mathrm{b}}$, is usually preferred for random-effect covariates (Heckerman et al., 2016; De Villemereuil et al., 2018). We use h_j^2 when there are no covariates or it is unimportant to distinguish $h^2_{j,\mathrm{a}}$ from $h^2_{j,\mathrm{b}}$. Henceforth we assume that the phenotype vector $\mathbf{y}$ is sample standardised, in which case $\sigma_j^2 = h_j^2$.

Methods

Data processing

We used genotypes from the eMERGE network (Verma et al., 2014), following the same quality control steps as Speed and Balding (2019). From the 25,875 individuals, we randomly selected 8000 to form the study population, simulated their phenotypes and computed GWAS summary statistics. The remaining 17,875 individuals were used as a reference panel to compute r^2 values for the summary statistic analyses.

We also generated three meta-analyses by dividing the study population randomly into two studies of size 4000, and calculating summary statistics for each study, both without and with covariate

adjustment. Each meta-analysis used within-study phenotype standardisation, and computed T_j^2 using inverse-variance weighting (Willer et al., 2010).

Of the SNPs remaining after quality control, 558,431 had non-zero LDAK weights (Speed et al., 2012) and only these SNPs contribute causal effects under the LDAK model and to SumHer analyses. We also restricted LDSC analyses, and simulations under the GCTA model (Yang et al., 2011a), to a set of 558,431 randomly-chosen SNPs.

Simulation of phenotypes and summary statistics

For 150 iterations, we randomly sampled 35,000 causal SNPs and, under each of the GCTA and LDAK models, we generated five phenotypes with different covariate and confounding effects such that $h^2_{\text{SNPa}} = 0.5$ in all cases: $\mathbf{y}_A$ (no covariates or confounding), $\mathbf{y}_B$ (covariate effect, no confounders), and $\mathbf{y}_{Ci}, i = 1, 2, 3$ (confounding). For $\mathbf{y}_B$, $\mathbf{X}$ has two columns, and the simulated effects were such that $\sigma_c^2 = \sigma_y^2/9$, so that $h^2_{\text{SNPb}} = 0.45$. To explore incomplete control of confounding, for all $\mathbf{y}_C$ phenotypes confounders correspond to a two-level hierarchical population structure. First, three subpopulations were identified using k-means clustering on the leading two principal components (PCs) of the SNP correlation matrix $\mathbf{Z}^T\mathbf{Z}/m$, restricted to SNPs with non-zero LDAK weight in order to minimise any effect of correlated SNPS.

Within each of these subpopulations, three sub-subpopulations were defined by k-means clustering on the two leading PCs computed only from subpopulation members. We assigned different phenotype means to the nine sub-subpopulations, while SNP effect sizes remained the same. For $\mathbf{y}_C$ phenotypes we consider both $\mathbf{X}$ corresponding to the three subpopulations (two columns), and $\mathbf{X}$ corresponding to all nine sub-subpopulations (eight columns). The h^2_{SNPb} values were 0.45 (C1), 0.475 (C2) and 0.49 (C3), corresponding to, respectively, 10%, 5% and 2% of the phenotypic variation due to confounding. $\mathbf{y}_A$ phenotypes and principal components were calculated using the LDAK software, while k-means clustering and the simulation of $\mathbf{y}_B$ and $\mathbf{y}_C$ phenotypes was undertaken in R (R Core Team, 2016).

For all phenotypes we compute T_j^2 and $n\hat{\alpha}_j^2$, both with and without adjusting for covariates $\mathbf{X}$. Based on these statistics we estimate h^2_{SNP} using SumHer for LDAK phenotypes (results in main text) while for GCTA phenotypes (Appendix 4.1) we used LDSC as implemented in the LDAK software (Speed et al., 2012). The two sets of results are broadly similar; we comment in the text on notable differences.

Large-effect SNPs

In part because of the problem of the unknown number of causal SNPs, but also due to model misspecification such as incomplete control of confounding, in many GWAS values of S_j arise that are extreme outliers under the assumed analysis model. Ideally, the solution would be to improve the analysis model, for example using a distribution with thicker tails than the Gaussian, or assigning an atom of prior probability at each SNP to a zero effect. However, because of computational advantages associated with model (1), in practice an ad-hoc data filtering approach is often adopted in which a SNP is removed if its estimated effect size is too large to be well-supported under the model. As we have control over confounding in our simulations, our main results do not use filtering. In Appendix 4.2, we consider the impact of filtering, where we follow Zheng et al. (2017) and exclude from analysis any SNP with $S_j > 80$. In the analysis of $\mathbf{y}_A$ and $\mathbf{y}_B$ simulations, no SNP was excluded, while for 32 C1, 2 C2, and 0 C3 LDAK and 44 C1, 14 C2, and 0 C3 GCTA simulations, at least one SNP was excluded for both T_j^2 and $n\hat{\alpha}_j^2$ when $\mathbf{X}$ was ignored in the GWAS analysis.

Results

For derivations of the expectations given below, see Appendix 3. Simulation results reported here are for SumHer analyses of LDAK phenotypes (see Methods); corresponding results using LDSC analyses of GCTA phenotypes are broadly similar (Appendix, Figures S1-4).

No confounding ($\mathbf{Cor}(\mathbf{X}, \mathbf{Z}) = \mathbf{0}$)

For a single GWAS with no covariate/confounder effects

$$\mathrm{E}[T_j^2] \approx c_j \left(1 + n \sum_i r_{ij}^2 h_i^2\right) \tag{5}$$

$$\mathrm{E}[(n-1)\hat{\alpha}_j^2] = \mathrm{E}[S\tilde{S}R_j] = \frac{\mathrm{E}[SSR_j]}{\mathrm{E}[SST]} \approx n\frac{\sigma_y^2 + n\sum_i r_{ij}^2\sigma_i^2}{(n-1)\sigma_y^2} \approx 1 + n\sum_i r_{ij}^2 h_i^2, \tag{6}$$

where $c_j = 1/(1 - \sum_i r_{ij}^2 h_i^2)$. For complex traits h_i^2 is typically small, so that c_j slightly exceeds 1 for many j. Both (5) and (6) are special cases of the SumHer model (2), but the deviation of C from unity in (5) is not due to confounding.

SumHer estimates of h^2_{SNP} based on GWAS summary statistics in the absence of covariate/confounder effects are centred close to the true value of 0.5 for both statistics (Figure 1(a)), so that for our simulations the deviation of the c_j from 1 appears to be negligible. The mean estimate of h^2_{SNP} does not noticeably change when A or C is estimated rather than fixed at the true values ($A = C = 1$), but

the variance increases due to uncertainty arising from the additional parameter estimation.

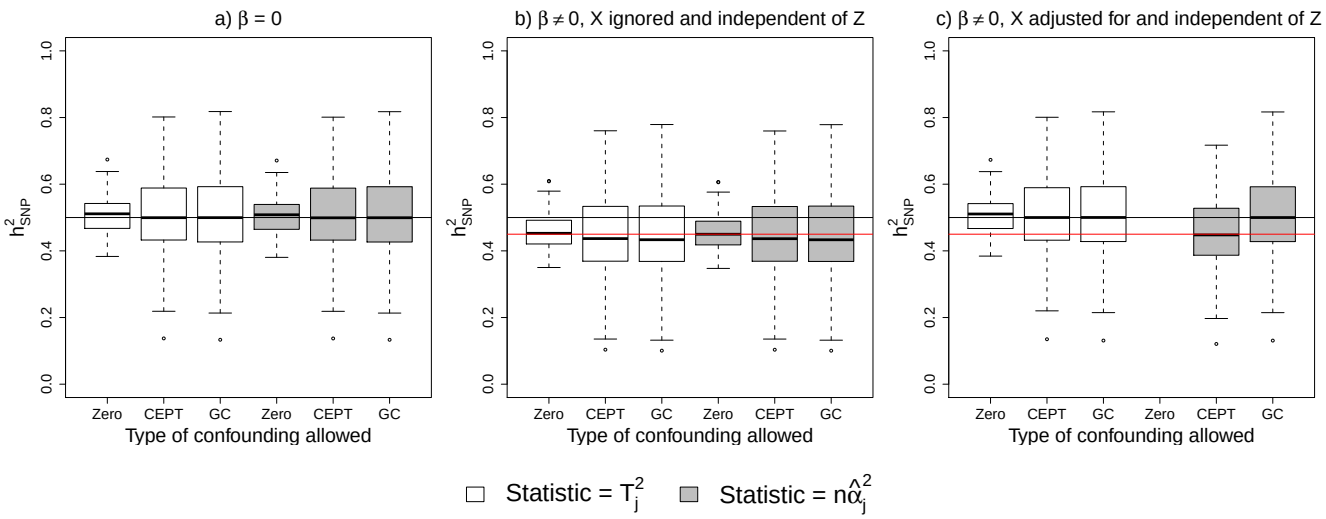

Figure 1: Estimates of h^2_{SNP} obtained by analysis of summary statistics from simulated GWAS. The black and red horizontal lines indicate the values of h^2_{SNPa} and h^2_{SNPb}, the SNP heritability without and with conditioning on covariates. Zero, CEPT and GC refer to no, A and C confounding terms being estimated in the analysis. Phenotypes are simulated under the LDAK model and the analyses performed using SumHer. (a) Phenotypes with no covariate effects. (b) Phenotypes with covariate effects but $\mathbf{X}$ ignored in the analysis. (c) Phenotypes with covariate effects and $\mathbf{X}$ adjusted for in the analysis. The "Zero" estimates when $S_j = n\hat{\alpha}_j^2$ are all negative and are not shown.

When covariates affect $\mathbf{y}$ but $\mathbf{X}$ is ignored in the GWAS analysis, $h^2_{j,\text{b}}$ is estimated rather than $h^2_{j,\text{a}}$ because now $\text{E}[SST] \approx \sigma_y^2 + \sigma_c^2$ rather than σ_y^2. Again, the average estimate of h^2_{SNP} changes little when A or C is estimated rather than fixed at 1 (Figure 1(b)). When $\mathbf{X}$ is included in the GWAS analysis,

$$\text{E}[T_j^2] = \text{E}[\text{E}[T_j^2|\mathbf{X}]] \approx \frac{\sigma_y^2 + (n-p)\sum_i r_{ij}^2\sigma_i^2}{\sigma_y^2 - \sum_i r_{ij}^2\sigma_i^2} = c_j\left(1 + (n-p)\sum_i r_{ij}^2 h^2_{i,\text{a}}\right), \tag{7}$$

which is the same as (5) but with $n-p$ in place of n and $h_i^2 = h^2_{i,\text{a}}$. Further,

$$\text{E}[S\tilde{S}R_j] \approx C\left(1 + (n-p)\sum_i r_{ij}^2 h^2_{i,\text{a}}\right) = A + (n-p)\sum_i r_{ij}^2 h^2_{i,\text{b}}, \tag{8}$$

where $A = C = \sigma_y^2/(\sigma_y^2 + \sigma_c^2) = h^2_{\text{SNPb}}/h^2_{\text{SNPa}}$. Figure 1(c) reflects properties evident from (7) and (8): only h^2_{SNPa} can be estimated from T_j^2, whereas either h^2_{SNPa} or h^2_{SNPb} can be estimated from $n\hat{\alpha}_j^2$, according to whether C or A is fitted.

Figure 2 shows that h^2_{SNPa} and h^2_{SNPb} are estimated with no apparent bias if, respectively, both studies did and did not adjust for covariates in a two-GWAS meta-analysis. Again, the inclusion

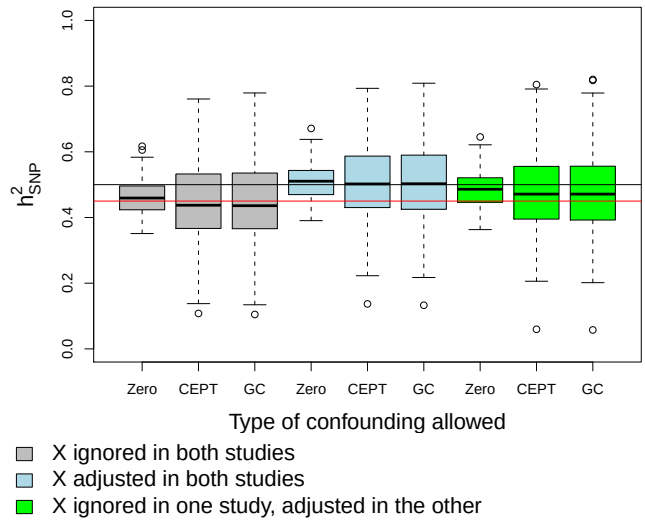

Figure 2: Estimates of h^2_{SNP} obtained from SumHer analysis of summary statistics calculated from a meta-analysis of two GWAS. The black and red horizontal lines indicate the values of h^2_{SNPa} and h^2_{SNPb}. Zero, CEPT and GC refer to no, A and C confounding terms in the analysis model.

of A or C terms has little effect on the mean estimates, as there is no confounding. When there is a mismatch in covariate adjustments between the two GWAS, the estimate of h^2_{SNP} is intermediate between h^2_{SNPa} and h^2_{SNPb} (Figure 2, green bars). In practice many meta-analyses do combine studies with different covariate adjustments, which may not adversely affect association tests but does affect heritability analyses. Examples include the meta-analyses of height (Lango et al., 2010) and blood pressure (The International Consortium for Blood Pressure Genome-Wide Association Studies et al., 2011) re-analysed using LDSC (Bulik-Sullivan et al., 2015a), and those of psychiatric traits (Okbay et al., 2016) and Type 2 diabetes (Scott et al., 2017).

Confounding ($\mathbf{Cor}(\mathbf{X}, \mathbf{Z}) \neq \mathbf{0}$)

When confounder $\mathbf{X}$ is ignored in the GWAS analysis:

$$\mathrm{E}[T_j^2] \approx c_j \left(1 + \frac{n a_j}{\sigma_y^2 + \sigma_c^2} + n \sum_i r_{ij}^2 h_{i,\text{b}}^2 \right) = c_j \mathrm{E}[S\tilde{S}R_j], \tag{9}$$

where $1/c_j = 1 - a_j/(\sigma_y^2 + \sigma_c^2) - \sum_{i \neq j} r_{ij}^2 h_{i,\text{b}}^2$ and $a_j = \left(\sum_{k=1}^p \mathrm{Cor}(\mathbf{Z}_j, \mathbf{X}_k)\boldsymbol{\beta}_k\right)^2$ with $\mathbf{X}_k$ denoting column k of $\mathbf{X}$. Assuming $c_j \approx 1$, only $h_{j,\text{b}}^2$ is estimable, and (9) includes an additive constant resembling A in (2). However, this term is SNP-dependent, and for it to correspond to A in (2) we require a_j to be independent of LD score, which typically does not hold (see Appendix 2). Instead,

we expect the estimate of h^2_{SNP} to be inflated by an amount that depends on both $\text{Cov}(\mathbf{X}, \mathbf{Z})$ and $\sigma_c^2/(\sigma_y^2+\sigma_c^2)$. Replacing (9) with the SumHer regression model leads to similar difficulties.

When $\mathbf{X}$ is included in the GWAS analysis, the estimated SNP effect, $\hat{\alpha}_j$, can be obtained from the linear regression of the residuals of $\mathbf{y}|\mathbf{X}$ on the residuals of $\mathbf{Z}|\mathbf{X}$ (Frisch and Waugh, 1933), and

$$\text{E}[T_j^2|\mathbf{Z}, \mathbf{X}] \approx \frac{b\sigma_e^2 + n\sum_i(\hat{r}_{ij} - \hat{\boldsymbol{\gamma}}_i'\hat{\boldsymbol{\Sigma}}^2_{\mathbf{X}}\hat{\boldsymbol{\gamma}}_j)^2\sigma_i^2/\bar{R}_j^2}{b\sigma_e^2 + \sum_{i\neq j}\bar{R}_i^2\sigma_i^2 - \sum_{i\neq j}(\hat{r}_{ij} - \hat{\boldsymbol{\gamma}}_i'\hat{\boldsymbol{\Sigma}}^2_{\mathbf{X}}\hat{\boldsymbol{\gamma}}_j)^2\sigma_i^2/\bar{R}_j^2}, \tag{10}$$

where $\boldsymbol{\Sigma}^2_{\mathbf{X}} = \text{Var}[\mathbf{X}]$ and, from the regression of $\mathbf{Z}_j$ on $\mathbf{X}$, $\bar{R}_j^2$ is one minus the coefficient of determination and $\hat{\boldsymbol{\gamma}}_j$ is the vector of estimated coefficients, while $b = (n-p-2)/(n-2)$. Further,

$$\text{E}[S\tilde{S}R_j|\mathbf{Z}] \approx \frac{\sigma_y^2}{\sigma_y^2 + \sigma_c^2}(b(1-h^2_{\text{SNPa}}) + n\sum_i(\hat{r}_{ij} - \hat{\boldsymbol{\gamma}}_i'\hat{\boldsymbol{\Sigma}}^2_{\mathbf{X}}\hat{\boldsymbol{\gamma}}_j)^2 h^2_{i,\text{a}}/\bar{R}_j^2). \tag{11}$$

Now, $S\tilde{S}R_j = (n-1-\mathbf{Z}_j'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_j)\hat{\alpha}_j^2$ and, unlike when $\text{Cor}(\mathbf{X}, \mathbf{Z}) = \mathbf{0}$, the term multiplying $\hat{\alpha}_j^2$ varies over SNPs.

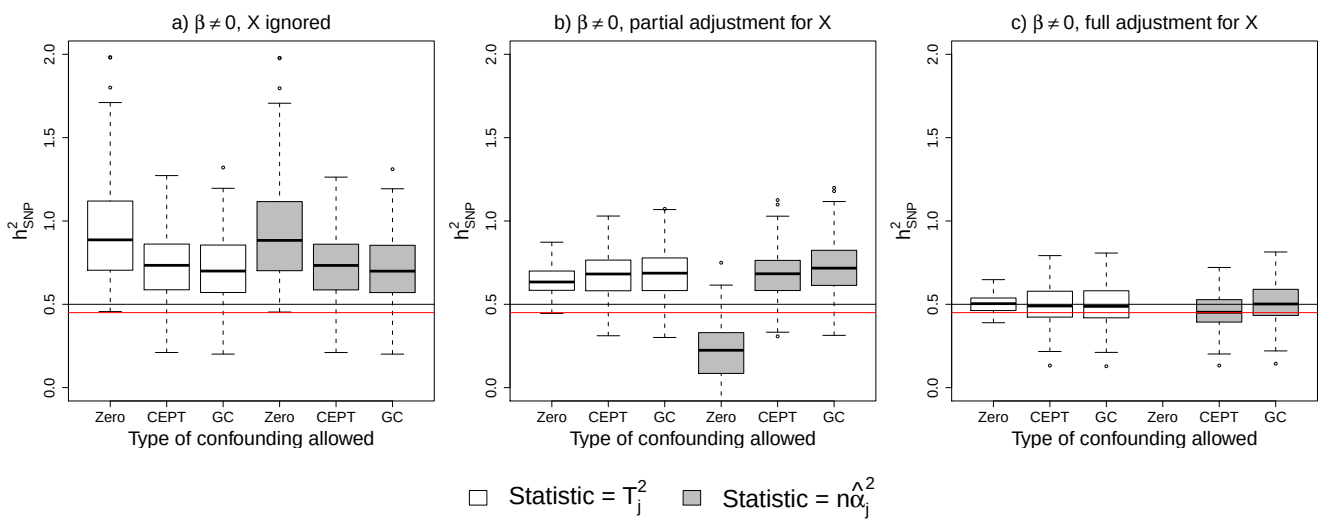

Figure 3: Similar to Figure 1, but here GWAS phenotypes are subject to confounding: phenotype means differ among three subpopulations that each consist of three sub-subpopulations. Subpopulations were constructed by applying k-means clustering to principal components of the SNPs with non-zero LDAK weight. Estimates of h^2_{SNP} from a GWAS with (a) no covariate adjustment, (b) adjustment for the three subpopulations but not the sub-subpopulations, (c) full covariate adjustment.

As expected, ignoring confounders results in inflated estimates of h^2_{SNP} (Figures 3(a) and 4(a, b)). The values of h^2_{SNPb}, are 0.45 for $C1$, 0.475 for $C2$, and 0.49 for $C3$ phenotypes, yet the average estimates are in the reverse order ($C1 > C2 > C3$) because of the inadequately-corrected confounding (Figure 4(a,b)).

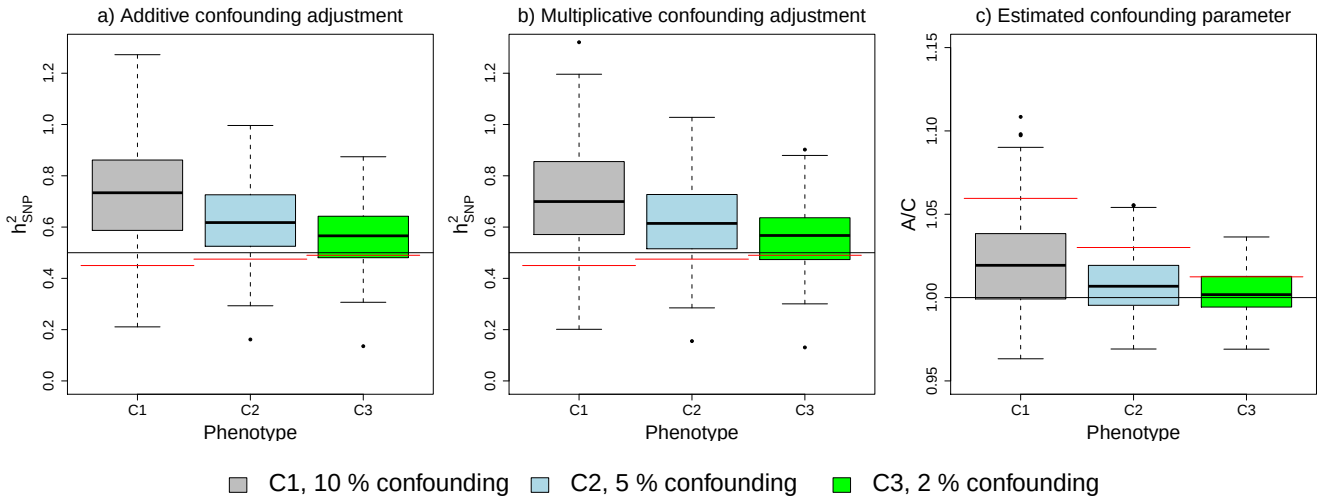

Figure 4: Estimating h^2_{SNP} and confounding parameters from phenotypes when $h^2_{\text{SNPa}} = 0.5$ and 10% (C1), 5% (C2) and 2% (C3) of phenotypic variance due to confounding. See Figure 3 for details of the confounding. The black lines in (a,b) indicate the simulated value of h^2_{SNPa} and the red lines the simulated value of h^2_{SNPb}, while the box plot shows the distribution of estimates when applying the confounding adjustment indicated in the plot heading. In (c), the black line corresponds to $A/C = 1$ corresponds to zero confounding bias and the red line the mean level of confounding, estimated as $\bar{S}_j - 1 - n\overline{\sum_i r^2_{ij} h^2_{j,\text{b}}}$. Note that the y-axis differs between (a,b) and (c).

The A/C estimates are consistently too low (Figure 4(c)) because the positive association between a_j and LD score leads to some of the confounding being misinterpreted as heritability. Comparing $C1$, $C2$ and $C3$ phenotypes, we find that the bias in h^2_{SNP} is, like a_j, a function of $\sigma^2_c/(\sigma^2_y+\sigma^2_c)$, the proportion of phenotypic variance due to confounding. Figure 4(c) shows average A/C estimates > 1 in the presence of confounding, but this does not always hold, and estimates $A < 1$ have been reported, such as in GWAS of rheumatoid arthritis (Bulik-Sullivan et al., 2015a), age at first birth and number of children (Barban et al., 2016), body mass index and hip-waist ratio (Speed and Balding, 2019), which could be due to confounding of the type considered here. The level of bias in h^2_{SNP} was higher in the LDAK simulations than for the GCTA simulations (Appendix 4.1).

Our finding of inadequate adjustment for confounding is concordant with the results of two recent analyses of stratified populations (DeVlaming et al., 2017; Berg et al., 2018), but not with Lee et al. (2018) who considered confounding by parental genotype. This is because parental average genotype generates an additive genetic effect, which inflates the slope but not the intercept of the summary statistic regression.

Partial covariate adjustment in the GWAS analysis reduced but did not eliminate bias in h^2_{SNP} estimates (Figure 3(b)) and led to divergence of the estimates based on T^2_j and $n\hat{\alpha}^2_j$. Full covariate adjustment did lead to unbiased estimates of h^2_{SNPa} when $S_j = T^2_j$ whether A or C or neither was fitted (Figure 3(c)). When $S_j = n\hat{\alpha}^2$, fitting A or C led to estimates of h^2_{SNPb} and h^2_{SNPa}, respectively.

These results indicate that although confounding adjustment can mask some of the heritability signal, which is intuitively why (10) and (11) differ from (7) and (8), the differences appear negligible in this case. However, for populations with much stronger stratification adjusted for in the GWAS stage, the distinction between (10), (11) and (7), (8) was reported to be important for estimating h^2_{SNP} (Luo et al., 2018).

Sample overlap and mis-specification of the heritability model

Our focus has been on heritability analyses when the underlying GWAS failed to account for all covariates or confounders. Here we briefly discuss some other possible issues that complicate summary statistic heritability analysis.

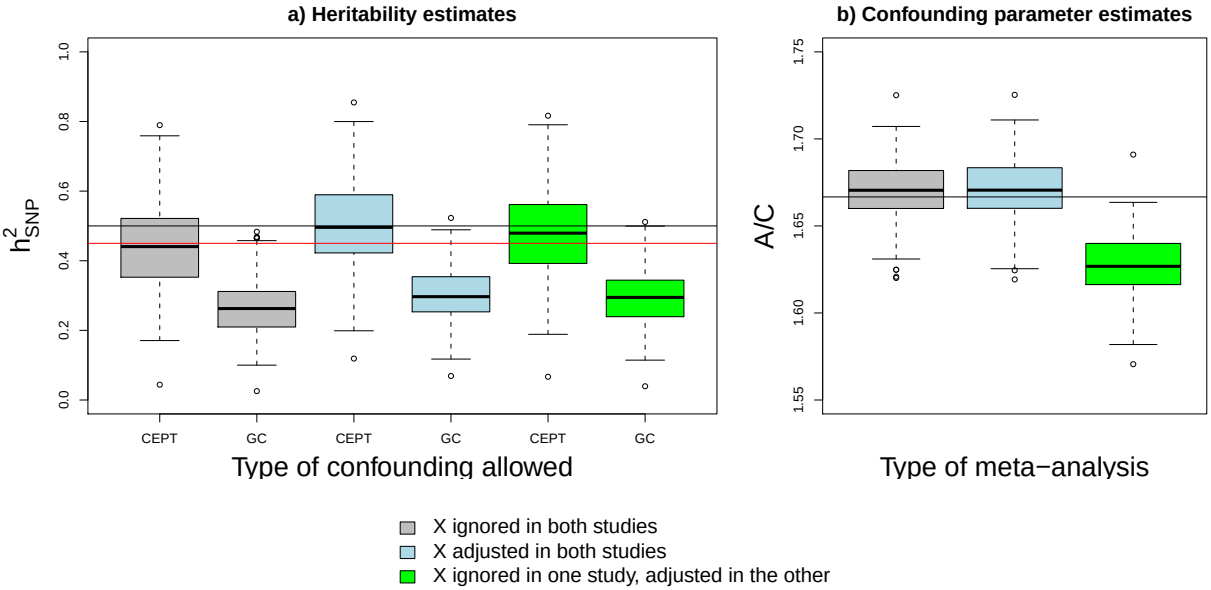

Figure 5: Similar to Figure 2, but now the meta-analysis includes overlapping 4000 individuals out of a total of 12,000. The black and red horizontal lines in (a,c) indicate the values of h^2_{SNPa} and h^2_{SNPb}. The black horizontal line in (b,d) indicates the expected intercept. CEPT and GC refer to A and C confounding terms in the analysis. Plot (a) gives estimates of h^2_{SNP} obtained under either an A or C view of the confounding parameter, while plot (b) gives estimates of the confounding parameter. Phenotypes were simulated in accordance to a LDAK model.

First, consider a meta-analysis in which some individuals have been included in multiple studies, but all non-genetic effects were accounted for in the component studies. In this case, the expected test statistic will be additively inflated (see Appendix 3.6 and Figure 5), and estimation of h^2_{SNP} remains unbiased. Furthermore, if the level of overlap is known, then the inflation in the expected statistic can be predicted in advance. In contrast, the population structure in our simulations implies some relatedness among all individuals, leading to a relationship between confounding and LD that cannot

be modelled using A or C. The example of non-trivial intercepts in Lee et al. (2018) appropriately correcting for confounding is similar to sample overlap, as it was based on twins, so can be viewed as two dependent samples, each consisting of unrelated individuals.

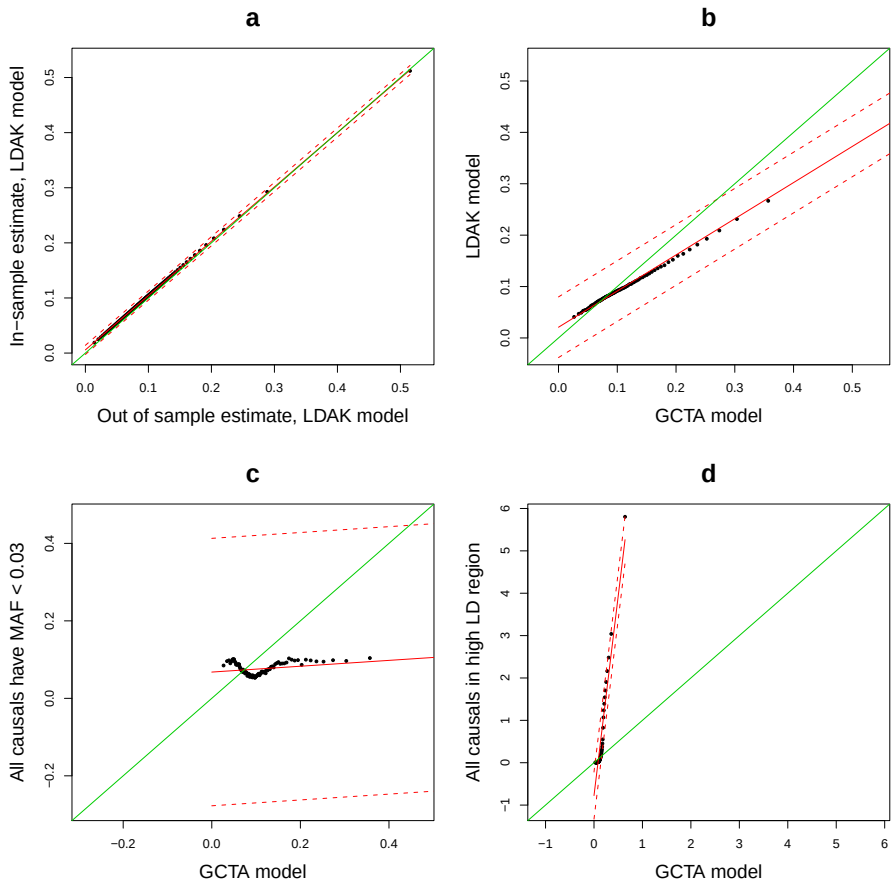

Figure 6: Comparing the slope of $\mathrm{E}[S_j]$, $n\sum_i r_{ij}^2 h_j^2$, for a range of heritability models. In (a), the slopes given by assuming a LDAK model where r_{ij}^2 is estimated by an external reference panel (x-axis) and using sample genotypes (y-axis) are compared. (b) compares the slopes assuming a GCTA (x-axis) and LDAK (y-axis) model. (c) compares the slopes assuming a GCTA model (x-axis) and a heritability model where only variants with MAF < 0.03 are causal (y-axis). (d) compares the slopes assuming a GCTA (x-axis) and a heritability model where only SNPs where $\sum_i r_{ij}^2 > 13$ are causal (y-axis). In (c) and (d), causal variant effects are assumed $a_j \sim \mathcal{N}(0, \sigma_g^2/m_{\text{causal}})$. where m_{causal} is the number of causal variants. The green line in the plots is the 45 degree line, while the red lines correspond to the the line of best fit (solid) and associated 95 % prediction intervals. The dots are percentile bins.

Figure 6 (a) shows that use of a well-chosen reference panel (which is the case for our simulations) has minimal impact on parameter estimates compared with computing LD coefficients from the GWAS genotypes (In-sample estimates). In Figure 6 (b-d), we compare the relationship of the predictor $\sum_i r_{ij}^2 h_i^2$ with summary statistics, under some extreme scenarios to illustrate sensitivity to mis-specification of the heritability model. If the true heritability model was LDAK, but the summary

statistic analysis assumed a GCTA model, h^2_{SNP} would be under-estimated, and the confounding parameter A over-estimated (Figure 6 (b), see also Speed and Balding (2019)). If all causal variants have low MAF (Figure 6 (c)), then h^2_{SNP} is again under-estimated. If causal variants are all in regions of high LD (Figure 6 (d)), then h^2_{SNP} is over-estimated and A is under-estimated.

Discussion

We have shown theoretically, and illustrated using simulation, that GWAS confounding bias is in general SNP dependent and correlated with LD. Therefore if the original GWAS analysis did not avoid confounding effects, heritability analysis of GWAS summary statistics using regression on an LD-related predictor, as implemented in LDSC (Bulik-Sullivan et al., 2015a) and SumHer (Speed and Balding, 2019), can fail to adequately account for confounding bias, leading to poor h^2_{SNP} estimation. To avoid this, we propose that such analysis proceeds in practice by first fitting a model with A or C unconstrained, and testing for deviation of its value from one. If A/C is significantly different from one subsequent heritability analysis would be unreliable. By this criterion, heritability analysis should proceed for only 7 out of 24 LDSC analysis of GWAS reported in Table 1 Bulik-Sullivan et al. (2015a), and 3/25 LDSC analyses and 17/25 SumHer analyses of GWAS reported in Supplementary Table 2 of Speed and Balding (2019).

When covariates were fitted in the GWAS analysis, we found very different results according to whether S_j was chosen to be the Wald statistic T_j^2 or the statistic $n\hat{\alpha}_j^2$ used to justify LDSC (Bulik-Sullivan et al., 2015a). Further, the estimable definition of h^2_{SNP} varies with the covariate adjustment performed in the original association analysis. The statistic $S\tilde{S}R$, closely related to $n\hat{\alpha}_j^2$, can be used to estimate h^2_{SNPb} regardless of the (non-confounder) covariates fitted, and hence a valid meta-analysis of h^2_{SNP} estimates is possible. However, $S\tilde{S}R$ is often not available in published GWAS results, and like T_j^2 it is subject to SNP-dependent confounding that can bias estimates of h^2_{SNP}. We have also shown that mis-specification of the heritability model can cause a regression intercept different from one, and biased h^2_{SNP} estimates, even if the underlying GWAS correctly accounted for confounders.

One source of error in published GWAS test statistics is genomic control, which applies a common multiplicative adjustment, derived under an assumption of sparse causal effects. It tends to over-adjust for highly-polygenic traits, which is corrected using SumHer (Speed and Balding, 2019).

We have only considered quantitative phenotypes, and we have not examined in detail the question of the validity of h^2_{SNP} analyses based on mixed model association statistics. In the appendix, we show that the exact expected value of such statistics contain terms not usually obtainable from public databases. Further h^2_{SNP} is non-identifiable without further assumptions, which if ignored

could lead to severely biased estimates.

Our findings accord with those of others (DeVlaming et al., 2017; Berg et al., 2018), and some statements/results in the original LDSC paper (Bulik-Sullivan et al., 2015a). For example, Bulik-Sullivan et al. (2015a) found that some heritability was inferred in simulations of confounding-only phenotypes, which was attributed to linked selection generating the correlation between confounding effect and LD score. Further, their interpretation of the intercept was based on average results from distinct populations, not replicate samples from the same population. This ignores the structure, and hence confounding, that is specific to a population. Supplementary Table 4C in Bulik-Sullivan et al. (2015a) shows that in the presence of confounding only ($h^2_{\text{SNP}} = 0$), one population (cou3) has higher LDSC h^2_{SNP} estimates for all three traits (0.144, 0.254, 0.229) than for any of the other 18 trait/population combinations (average: 0.030). Most importantly, the claim that LD score is not associated with confounding (Bulik-Sullivan et al., 2015a) was based on marginalising over the confounding term a_j, which is inappropriate as a_j and LD score are both SNP specific.

Acknowledgements

DS is funded by the European Unions Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement number 754513, by Aarhus University Research Foundation (AUFF) and the Independent Research Fund Denmark under Project 7025-00094B. DB is funded by the Australian Research Council under grant DP190103188. The eMERGE Network was initiated and funded by NHGRI through the following grants: U01HG006828 (Cincinnati Childrens Hospital Medical Center/Boston Childrens Hospital); U01HG006830 (Childrens Hospital of Philadelphia); U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006382 (Geisinger Clinic); U01HG006375 (Group Health Cooperative); U01HG006379 (Mayo Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University Medical Center); and U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center). Access to eMERGE Network data was granted under dbGaP Project 14422, 'Comprehensive testing of SNP-based prediction models'.

Declaration of interests

The author declare there are no conflicts of interest.

References

Brendan K. Bulik-Sullivan, Po-Ru Loh, Hilary Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47:291–295, 2015a.

B. K. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P.-R. Loh, ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium, Laramie Duncan, J. R. B. Perry, Nick Patterson, Elise B. Robinson, M. J. Daly, A. L. Price, and B. M. Neale. An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11):1236–1241, 2015b.

Hilary K. Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh Farh, Stephan Ripke, Felix R. Day, ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, The RACI Consortium, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R. B. Perry, Yukinori Okada, Soumya Raychaudhuri, Mark Daly, Nick Patterson, Benjamin M. Neale, and Alkes L. Price. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235, 2015.

Steven Gazal, Hilary K. Finucane, Nicholas A Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoeh, Brendan K. Bulik-Sullivan, Benjamin M. Neale, Alexander Gusev, and Alkes L. Price. Linkage disequilibrium dependent architecture of human complex traits reveals action of negative selection. *Nature Genetics*, 49:1421–1427, 2017.

Doug Speed and David J. Balding. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature Genetics*, 51:277–84, 2019.

Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, 99: 139–153, 2016.

Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Annals of Applied Statistics*, 11:1561–1592, 2017.

Jie Zheng, A. Mezut Erzurumluoglu, Benjamin L. Elsworth, John P. Kemp, Laurence Howe, Philip C. Haycock, Gibran Hemani, Katherine Tansey, Charles Laurin, Early Genetics and Lifecourse Epidemiology (EAGLE) Consortium, Beate St Pourcain, Nicole M. Warrington, Hilary K. Finucane,

Alkes L. Price, B. K. Bulik-Sullivan, Verneri Anttila, Lavinia Paternoster, Tom R. Gaunt, David M. Evans, and Benjamin M. Neale. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*, 33(2):272–279, 2017.

J. Yang, S. Hong Lee, M.E. Goddard, and P. M. Visscher. Gcta: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88:76–82, 2011a.

Doug Speed, Na Cai, the UCLEB Consortium, Michael R. Johnson, Sergey Nejentsev, and David J. Balding. Reevaluation of SNP heritability in complex human traits. *Nature Genetics*, 49(7): 986–992, 2017.

Jian Yang, Michael N. Weedon, Shaun Purcell, Guilaume Lettre, Karol Estrada, Cristen J. Willer, Albert V. Smith, Erik Ingelsson, Jeffrey R. O'Connell, Massimo Mangino, Reedik Mägi, Pamela A. Madden, Andrew C. Heath, Dale R. Nyholt, Nicholas G. Martin, Grant W. Montgomery, Timothy M. Frayling, Joel N. Hirschhorn, Mark I. McCarthy, Michael E. Goddard, Peter M. Visscher, and the GIANT Consortium. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19:807–812, 2011b.

Christoph Lippert, Jennifer Listgarten, Ting Liu, Carl M. Kadie, Robert I. Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 10: 833–837, 2011.

Po-Ru Loh, George Tucker, Brendan K. Bulik-Sullivan, Bjarni J. Vilhjálmsson, Hilary K. Finucane, Rany M. Salem, Daniel I. Chasman, Paul M. Ridker, Benjamin M. Neale, Bonnie Berger, Nick Patterson, and Alkes L. Price. Efficient bayesian mixed model analysis increases association power in large cohorts. *Nature Genetics*, 47:284–290, 2015.

Omer Weissbrod, Jonathan Flint, and Saharon Rosset. Estimating snp-based heritability and genetic correlation in case-control studies directly and with summary statistics. *The American Journal of Human Genetics*, 103:89–99, 2018.

R.A. Mrode. *Linear models for the prediction of animal breeding values.* CABI publishing, Third edition, 2014.

Matti Pirinen, Peter Donnelly, and Chris C.A. Spencer. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics*, 7(1):369–390, 2013.

D. Heckerman, Deepti Gurdasani, Carl Kadie, Cristina Pomilla, Tommy Carstensen, Hilary Martin, Kenneth Ekoru, Rebecca N. Nsubuga, Gerald Ssenyomo, Anatoli Kamali, Pontiano Kaleebu, Christian Widmer, and Manjinder S. Sandhu. Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National academy of Sciences of the United States of America*, 113(27):7377–7382, 2016.

Pierre De Villemereuil, Michael B. Morrissey, Shinichi Nagakawa, and Holger Schielzeth. Fixed effect variance and the estimation of repeatabilities and heritabilities: Issues and solutions. *Journal of Evolutionary Biology*, 31(4):621–632, 2018.

Shefali S. Verma, Mariza de Andrade, Gerard Tromp, Helena Kuivaniemi, Elizabeth Pugh, Bahram Namjou-Khales, Shubhabrata Mukherjee, Gail P. Jarvik, Leah C. Kottyan, Amber Burt, Yuki Bradford, Gretta D. Armstrong, Kimberly Derr, Dana C. Crawford, Jonathan L. Haines, Rongling Li, David Crosslin, and Marylyn D. Ritchie. Imputation and quality control steps for combining multiple genome-wide datasets. *Frontiers in Genetics*, 5:1–15, 2014.

Cristen J. Willer, Yun Li, and Goncalo R. Abecasis. Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 2010.

D. Speed, G. Hemani, M. R. Johnson, and D. J. Balding. Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics*, 91:1011–1021, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

Allen H. Lango, K. Estrada, G. Lettre, S.I. Berndt, M.N. Weedon, F. Rivadeneira, C.J. Willer, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

The International Consortium for Blood Pressure Genome-Wide Association Studies et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478 (7367):103–109, 2011.

Aysu Okbay, Bart M.L Baselmans, Jan-Emmanuel De Neve, Patrick Turley, Michel G. Nivard, Mark Alan Fontana, S. Fleur W. Meddens, et al. Genetic variants associated with subjective well-being, depressive symptoms and neuroticism identified through genome-wide analyses. *Nature Genetics*, 48:624–633, 2016.

Robert A Scott, Laura J Scott, Reedik Mägi, Letizia Marullo, Kyle J Gaulton, Marika Kaakinen, et al. An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes*, 66 (11):2888–2902, 2017.

Ragnar Frisch and Fredrick V. Waugh. Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401, 1933.

Nicola Barban, Rick Jansen, Ronald de Vlaming, Ahmad Vaez, Jornt J. Mandemakers, et al. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nature Genetics*, 48(12): 1462–1472, 2016.

Ronald DeVlaming, Magnus Johannesson, Patrik K.E. Magnusson, M. Arfan Ikram, and Peter M. Visscher. Equivalence of LD-Score Regression and Individual-Level-Data Methods. *Biorχiv*, 2017. doi: 10.1101/211821.

Jermey J. Berg, Arbel Harpak, Nasa Sinnott-Armstrong, Anja Moltke Jørgensen, Hakhamanesh Mostafavi, Yair Field, Evan A. Boyle, Xinjun Zhang, Fernando Racimo, Jonathan K. Pritchard, and Graham Coop. Reduced signal for polygenic adaptation of height in UK Biobank. *Biorχiv*, 2018. doi: 10.1101/354951.

James J. Lee, Matt McGue, William G. Iacomo, and Carson C. Chow. The accuracy of LD score regression as an estimator of confounding and genetic correlations in genome-wide association studies. *Genetic Epidemiology*, 42:7, 2018.

Yang Luo, Xinyi Li, Xin Wang, Steven Gazal, Josep Maria Mercader, 23 and Me Research Team, SIGMA Type 2 Diabetes Consortium, Benjamin M. Neale, Jose C. Florez, Adam Auton, Alkes L. Price, Hilary K. Finucane, and Soumya Raychaudhuri. Estimating heritability of complex traits in admixed populations with summary statistics. *Biorχiv*, 2018. doi: 10.1101/503144.