Expectation of the intercept from bivariate LD score regression in the presence of population stratification

Loic Yengo[a,1], Jian Yang[a,b] & Peter M. Visscher[a,b,1]

[a]Institute for Molecular Bioscience, The University of Queensland, Brisbane 4072, Australia; [b]Queensland Brain Institute, The University of Queensland, Brisbane 4072, Australia.

[1]To whom correspondence may be addressed.
Email: l.yengodimbou@uq.edu.au or peter.visscher@uq.edu.au.

Linkage disequilibrium (LD) score regression is an increasingly popular method used to quantify the level of confounding in genome-wide association studies (GWAS) or to estimate heritability and genetic correlation between traits. When applied to a pair of GWAS, the LD score regression (LDSC) methodology produces a statistic, referred to as the bivariate LDSC intercept, which deviation from 0 is classically interpreted as an indication of sample overlap between the two GWAS. Here we propose an extension of the theory underlying the bivariate LDSC methodology, which accounts for population stratification within and between GWAS. Our extended theory predicts an inflation of the bivariate LDSC intercept when sample sizes and heritability are large, even in the absence of sample overlap. We illustrate our theoretical results with simulations based on actual SNP genotypes and we propose a re-interpretation of previously published results in the light of our extended theory.

Initially introduced in Bulik-Sullivan *et al.* (2014)[1], the LD score regression (LDSC) methodology relies on a derivation for a particular SNP j of the expectation of its association χ^2-statistic (χ_j^2) as function of the LD score (ℓ_j), that is the sum of pairwise squared correlation between minor allele counts at SNP j versus all SNPs or versus SNPs in its vicinity, the heritability (h^2) of the trait plus a term indicating the level of confounding in the GWAS attributable to population stratification. The derivation proposed in Bulik-Sullivan *et al.* (2014) predicts that the latter term (a.k.a the LDSC intercept) increases linearly with the sample size (N) of the GWAS and the heritability of the trait, which raises a number of challenges in its interpretation as an indication of confounding when N is very large (N ∼400,000 for example). This problematic behavior of the LDSC intercept has been underlined in a recent publication by Loh *et al.* (2017)[2], which recommends the use of an alternative statistic to quantify the influence of population stratification on GWAS results.

The LDSC methodology was later extended in Bulik-Sullivan *et al.* (2015)[3] to analyse pairs of GWAS in order to quantify the genetic correlation between focus traits of each GWAS. The theory proposed in Bulik-Sullivan *et al.* (2015) introduced the bivariate LD score intercept, obtained from the regression of the product of association statistics (defined further below) onto LD scores, as a measure of sample overlap between the two GWAS. Compared to Bulik-Sullivan *et al.* (2014), the model underlying the bivariate LDSC methodology does not consider the presence of population stratification within and between studies, neither discusses how this might contribute to the expectation of the bivariate LDSC intercept. We therefore address this question here and propose an extension of the initial theory.

Theoretical results

Notations

We consider that GWASs of two traits $\mathbf{y}_1$ and $\mathbf{y}_2$ are performed in two cohorts: cohort 1 and cohort 2 respectively. Each SNP j is tested for association with each trait using a test statistic T_j defined as the ratio between estimated SNP effect from linear regression of $\mathbf{y}_1$ or $\mathbf{y}_2$ onto the minor allele count (MAC) over its estimated standard error. Let us denote T_{j1} and T_{j2} the test statistics for SNP j calculated for $\mathbf{y}_1$ in cohort 1 and for $\mathbf{y}_2$ in cohort 2 respectively. We propose hereafter an extension of the expectation of the product of T_{j1} and T_{j2}, initially proposed in Bulik-Sullivan *et al.* (2015), in the presence of population stratification within each cohort induced by genetic drift.

As Bulik-Sullivan et al. (2014), we assume that population stratification can be modelled in each cohort as a 50:50 mixture of two sub-populations deriving from a common ancestral population and thus having different allele frequencies spectra. We denote σ_{S1} and σ_{S2} as the mean phenotypic difference between sub-populations of cohort 1 and 2 respectively (environmental stratification). We introduce $F_{ST}^{(1)}$ and $F_{ST}^{(2)}$ as Wright's F_{ST} measures of allele frequency differences between sub-populations of cohort 1 and cohort 2 respectively (genetic stratification). For the sake of simplicity, we assume that the level of genetic and environmental stratification is similar between cohort 1 and cohort 2. Therefore $F_{ST}^{(1)} \approx F_{ST}^{(2)} = F_{ST}$ and $\sigma_{S1} \approx \sigma_{S2} = \sigma_S$. We finally denote h_1^2, h_2^2 and r_g, the heritabilities of $\mathbf{y}_1$ and $\mathbf{y}_2$ and their genetic correlation.

We consider the following model

$$\mathbf{y}_k = \mathbf{z}_k \beta_k + \mathbf{s}_k + \mathbf{e}_k \tag{1}$$

where $\mathbf{y}_k$ is the vector of phenotypes of N_k individuals from cohort k, $\mathbf{s}_k$ a N_k-dimensional vector which entries equal $\pm\sigma_S$ (mean of environmental fixed effects) depending on whether participants enrolled in cohort k are from one or the other sub-population, $\mathbf{z}_k$ a $N_k \times M$ matrix of scaled MAC, β_k a M dimensional vector of true SNP effect sizes on trait $\mathbf{y}_k$ and $\mathbf{e}_k$ a N_k-dimensional vector of residuals.

We denote $\mathbf{z}_{jk} = (z_{1j}, \ldots, z_{N_k j})'$ the j-th column of matrix $\mathbf{z}_k$. The i-th entry of $\mathbf{z}_{jk}$, denoted z_{ijk}, is defined as $z_{ijk} = (x_{ijk} - 2p_{jk})/\sqrt{2p_{jk}(1-p_{jk})}$ where x_{ijk} is the MAC at SNP j of individual i from cohort k and p_{jk} is minor allele frequency (MAF) of SNP j in cohort k. We assume that all sub-samples of cohorts 1 and 2 have derived from the same ancestral population and denote p_j as the MAF of SNP j in that ancestral population.

As in Bulik-Sullivan *et al.* (2015), we model the true genetic effects as

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0}_M \\ \mathbf{0}_M \end{bmatrix}, \frac{1}{M} \begin{bmatrix} h_1^2 \mathrm{I}_M & \rho_g \mathrm{I}_M \\ \rho_g \mathrm{I}_M & h_2^2 \mathrm{I}_M \end{bmatrix} \right) \tag{2}$$

and the residuals as

$$\boldsymbol{e} = \left[\begin{array}{c} \mathbf{e}_1 \\ \mathbf{e}_2 \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} \mathbf{0}_{N_1} \\ \mathbf{0}_{N_2} \end{array}\right], \left[\begin{array}{cc} (1-h_1^2)\mathrm{I}_{N_1} & \rho_e 1_{N_1} 1'_{N_2} \\ \rho_e 1_{N_2} 1'_{N_1} & (1-h_2^2)\mathrm{I}_{N_2} \end{array}\right]\right), \tag{3}$$

where $\rho_g = r_g\sqrt{h_1^2 h_2^2}$ is the genetic covariance between $\mathbf{y}_1$ and $\mathbf{y}_2$ and ρ_e is the covariance between error terms $\mathbf{e}_1$ and $\mathbf{e}_2$. We also define ρ as $\rho \equiv \rho_g + \rho_e$.

In cohort k, the least-squares estimate of the effect of SNP j on $\mathbf{y}_k$, denoted $\widehat{\beta}_{jk}$, is approximately equal to $\widehat{\beta}_{jk} \approx [\mathbf{z}'_{jk}\mathbf{y}_k]/N_k$ and has a sampling variance $\approx 1/N_k$. The accuracy of these approximations increases with the sample size N_k, that we assume here to be large, e.g. hundreds of thousands. Therefore the t-statistic T_{jk} can be approximated as $T_{jk} \approx [\mathbf{z}'_{jk}\mathbf{y}_k]/\sqrt{N_k}$.

Derivation of $\mathbb{E}[T_{j1}T_{j2}]$

We can express $\mathbb{E}[T_{j1}T_{j2}]$ as follows

$$\begin{aligned}
\mathbb{E}[T_{j1}T_{j2}] &= \frac{1}{\sqrt{N_1N_2}}\mathbb{E}\left[\left(\mathbf{z}'_{j1}\mathbf{z}_1\beta_1 + \mathbf{z}'_{j1}\mathbf{s}_1 + \mathbf{z}'_{j1}\mathbf{e}_1\right)\left(\mathbf{z}'_{j2}\mathbf{z}_2\beta_2 + \mathbf{z}'_{j2}\mathbf{s}_2 + \mathbf{z}'_{j2}\mathbf{e}_2\right)\right] \qquad (4)\\
&= \frac{1}{\sqrt{N_1N_2}}\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{z}_1\beta_1)(\mathbf{z}'_{j2}\mathbf{z}_2\beta_2)\right] + \frac{1}{\sqrt{N_1N_2}}\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{e}_1)(\mathbf{z}'_{j2}\mathbf{e}_2)\right]\\
&\quad + \frac{1}{\sqrt{N_1N_2}}\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{e}_1)(\mathbf{z}'_{j2}\mathbf{z}_2\beta_2)\right] + \frac{1}{\sqrt{N_1N_2}}\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{e}_1)(\mathbf{z}'_{j2}\mathbf{s}_2)\right]\\
&\quad + \frac{1}{\sqrt{N_1N_2}}\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{z}_1\beta_1)(\mathbf{z}'_{j2}\mathbf{e}_2)\right] + \frac{1}{\sqrt{N_1N_2}}\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{s}_1)(\mathbf{z}'_{j2}\mathbf{e}_2)\right]\\
&\quad + \frac{1}{\sqrt{N_1N_2}}\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{s}_1)(\mathbf{z}'_{j2}\mathbf{z}_2\beta_2)\right] + \frac{1}{\sqrt{N_1N_2}}\mathbb{E}\left[(\mathbf{z}'_{j2}\mathbf{s}_2)(\mathbf{z}'_{j1}\mathbf{z}_1\beta_1)\right]\\
&\quad + \frac{1}{\sqrt{N_1N_2}}\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{s}_1)(\mathbf{z}'_{j2}\mathbf{s}_2)\right]. \qquad (5)
\end{aligned}$$

If we assume independence between $\mathbf{e}_k$ and β_k, between $\mathbf{e}_k$ and $\mathbf{s}_k$, and that $\mathbb{E}[\mathbf{e}_k|\mathbf{z}_k] = \mathbf{0}$ and $\mathbb{E}[\beta_k|\mathbf{z}_k] = \mathbf{0}$, which are classical assumptions made in Bulik-Sullivan *et al.* (2014), then

$$\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{e}_1)(\mathbf{z}'_{j2}\mathbf{z}_2\beta_2)\right] = \mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{e}_1)(\mathbf{z}'_{j2}\mathbf{s}_2)\right] = \mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{z}_1\beta_1)(\mathbf{z}'_{j2}\mathbf{e}_2)\right] = 0 \tag{6}$$

and

$$\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{s}_1)(\mathbf{z}'_{j2}\mathbf{e}_2)\right] = \mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{s}_1)(\mathbf{z}'_{j2}\mathbf{z}_2\beta_2)\right] = \mathbb{E}\left[(\mathbf{z}'_{j2}\mathbf{s}_2)(\mathbf{z}'_{j1}\mathbf{z}_1\beta_1)\right] = 0. \tag{7}$$

This therefore leads to simplify $\mathbb{E}[T_{j1}T_{j2}]$ as

$$\mathbb{E}[T_{j1}T_{j2}] = \frac{1}{\sqrt{N_1N_2}}\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{z}_1\beta)(\mathbf{z}'_{j2}\mathbf{z}_2\beta)\right] + \frac{1}{\sqrt{N_1N_2}}\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{s}_1)(\mathbf{z}'_{j2}\mathbf{s}_2)\right] + \frac{1}{\sqrt{N_1N_2}}\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{e}_1)(\mathbf{z}'_{j2}\mathbf{e}_2)\right]. \tag{8}$$

Equation (8) shows a strong similarity with equation (1) from Bulik-Sullivan *et al.* (2015) but one can already notice the inclusion of the second term on the right side of the equation, representing the contribution of population stratification within and between cohorts. We now further simplify equation (8).

Let us start with the first term of the right side of equation (8). We can rewrite it as

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{z}_{j1}'\mathbf{z}_1\beta_1)(\mathbf{z}_{j2}'\mathbf{z}_2\beta_2)\right] &= \mathbb{E}\left[\beta_1'\left(\mathbf{z}_1'\mathbf{z}_{j1}\mathbf{z}_{j2}'\mathbf{z}_2\right)\beta_2\right] = \mathbb{E}\left[\mathbb{E}\left[\beta_1'\left(\mathbf{z}_1'\mathbf{z}_{j1}\mathbf{z}_{j2}'\mathbf{z}_2\right)\beta_2|\mathbf{z}_1,\mathbf{z}_2\right]\right] \\
&= \frac{\rho_g}{M}\mathbb{E}\left[\mathbb{E}\left[\operatorname{tr}\left(\mathbf{z}_1'\mathbf{z}_{j1}\mathbf{z}_{j2}'\mathbf{z}_2\right)|\mathbf{z}_1,\mathbf{z}_2\right]\right] \\
&= \frac{\rho_g}{M}\mathbb{E}\left[\operatorname{tr}\left(\mathbf{z}_1'\mathbf{z}_{j1}\mathbf{z}_{j2}'\mathbf{z}_2\right)\right] \\
&= \frac{\rho_g}{M}\sum_{k=1}^{N_1}\sum_{h=1}^{N_2}\sum_{q=1}^{M}\mathbb{E}\left[z_{kq}^{(1)}z_{hq}^{(2)}z_{kj}^{(1)}z_{hj}^{(2)}\right]. \qquad (9)
\end{aligned}$$

Denote $\mathcal{O}_s$ as the set of samples overlapping cohort 1 and cohort 2. We use the simplified notation "$i \in \mathcal{O}_s$" (or "$i \notin \mathcal{O}_s$") to indicate that individual i belongs (or does not belong) to $\mathcal{O}_s$. We can therefore write

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{z}_{j1}'\mathbf{z}_1\beta_1)(\mathbf{z}_{j2}'\mathbf{z}_2\beta_2)\right] &= \frac{\rho_g}{M}\sum_{(k,h)\in\mathcal{O}_s}\sum_{q=1}^{M}\mathbb{E}\left[z_{kq}^{(1)}z_{hq}^{(2)}z_{kj}^{(1)}z_{hj}^{(2)}\right] \\
&\quad+\frac{\rho_g}{M}\sum_{(k,h)\notin\mathcal{O}_s}\sum_{q=1}^{M}\mathbb{E}\left[z_{kq}^{(1)}z_{hq}^{(2)}z_{kj}^{(1)}z_{hj}^{(2)}\right] \\
&= \frac{N_s^2\rho_g}{M}\sum_{q=1}^{M}\mathbb{E}\left[\frac{1}{N_s^2}\sum_{(k,h)\in\mathcal{O}_s}z_{kj}^{(1)}z_{kq}^{(1)}z_{hj}^{(2)}z_{hq}^{(2)}\right] \\
&\quad+\frac{\rho_g}{M}\sum_{(k,h)\notin\mathcal{O}_s}\sum_{q=1}^{M}\mathbb{E}\left[z_{kj}^{(1)}z_{kq}^{(1)}\right]\mathbb{E}\left[z_{hj}^{(2)}z_{hq}^{(2)}\right] \\
&= \frac{N_s^2\rho_g}{M}\sum_{q=1}^{M}\mathbb{E}\left[\left(\frac{1}{N_s}\sum_{k\in\mathcal{O}_s}z_{kj}^{(1)}z_{kq}^{(1)}\right)\left(\frac{1}{N_s}\sum_{h\in\mathcal{O}_s}z_{hj}^{(2)}z_{hq}^{(2)}\right)\right] \\
&\quad+\frac{\rho_g}{M}\sum_{(k,h)\notin\mathcal{O}_s}\sum_{q=1}^{M}\mathbb{E}\left[z_{kj}^{(1)}z_{kq}^{(1)}\right]\mathbb{E}\left[z_{hj}^{(2)}z_{hq}^{(2)}\right] \\
&= \frac{N_s^2\rho_g}{M}\mathbb{E}\left[\sum_{q=1}^{M}\hat{r}_{jq}^2\right]+\frac{(N_1N_2-N_s^2)\rho_g}{M}\sum_{q=1}^{M}\mathbb{E}\left[z_{kj}^{(1)}z_{kq}^{(1)}\right]\mathbb{E}\left[z_{hj}^{(2)}z_{hq}^{(2)}\right]
\end{aligned}$$

(10)

where $\hat{r}_{jq}^2$ is the squared sample correlation between MAC at SNP j and MAC at SNP q.

Besides, $\mathbb{E}\left[z_{kq}^{(1)} z_{kj}^{(1)}\right] = \mathbb{E}[\hat{r}_{jq}]$ is the expectation of the sample correlation in cohort 1 between MAC at SNP j and MAC at SNP q, which we assumed to be the same as the expectation of the sample correlation in cohort 2, i.e. $\mathbb{E}\left[z_{kq}^{(1)} z_{kj}^{(1)}\right] = \mathbb{E}\left[z_{hj}^{(2)} z_{hq}^{(2)}\right] = \mathbb{E}[\hat{r}_{jq}]$.

Therefore $\mathbb{E}\left[z_{kq}^{(1)} z_{kj}^{(1)}\right] \times \mathbb{E}\left[z_{hq}^{(2)} z_{hj}^{(2)}\right] = \mathbb{E}[\hat{r}_{jq}]^2 = \mathbb{E}[\hat{r}_{jq}^2] - \mathbb{V}(\hat{r}_{jq})$. From equation (2.3) in Bulik-Sullivan *et al.* (2014) we can therefore deduce that

$$\mathbb{E}\left[z_{kq}^{(1)} z_{kj}^{(1)}\right] \times \mathbb{E}\left[z_{hq}^{(2)} z_{hj}^{(2)}\right] \approx \underbrace{r_{jq}^2 + F_{ST}^2 + (1 - F_{ST}^2)/N\, r_{jq}^2}_{\mathbb{E}[\hat{r}_{jq}^2]} - \underbrace{(1 - F_{ST}^2)/N}_{\mathbb{V}(\hat{r}_{jq})} = r_{jq}^2 + F_{ST}^2, \tag{11}$$

where, in the equation above, $N = N_1$ or $N = N_2$ indifferently. We can therefore rewrite equation (10) as

$$\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{z}_1\beta_1)(\mathbf{z}'_{j2}\mathbf{z}_2\beta_2)\right] = \frac{N_s^2\rho_g}{M}\mathbb{E}\left[\hat{\ell}_j\right] + \frac{(N_1N_2 - N_s^2)\rho_g}{M}\ell_j + (N_1N_2 - N_s^2)\rho_g F_{ST}^2, \tag{12}$$

where $\hat{\ell}_j = \sum_{q=1}^{M} \hat{r}_{jq}^2$ is the sample LD score of SNP j calculated only from samples in $\mathcal{O}_s$ and $\ell_j = \sum_{q=1}^{M} r_{jq}^2$ is theoretical true LD score.

If we assume that $\mathcal{O}_s$ is a random sample of cohort 1 and cohort 2, then the population structure within $\mathcal{O}_s$ is expected to be similar to that within cohort 1 and within cohort 2. Bulik-Sullivan *et al.* (2014) derived an approximation of the expectation of the sample LD score $\mathbb{E}[\hat{\ell}_j] = \mathbb{E}\left[\sum_{q=1}^{M} \hat{r}_{jq}^2\right]$ (equation 2.4 of their supplementary note) that we rewrite here as

$$\mathbb{E}[\hat{\ell}_j] \approx \ell_j + MF_{ST}^2 + \frac{M(1 - F_{ST}^2)}{N_s} \approx \ell_j + MF_{ST}^2 + \frac{M}{N_s}. \tag{13}$$

If we combine equations (12) and (13) we get

$$\begin{aligned}\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{z}_1\beta_1)(\mathbf{z}'_{j2}\mathbf{z}_2\beta_2)\right] &\approx \frac{N_1N_2\rho_g}{M}\ell_j + N_s^2\rho_g F_{ST}^2 + N_s\rho_g + (N_1N_2 - N_s^2)\rho_g F_{ST}^2 \\ &= \frac{N_1N_2\rho_g}{M}\ell_j + N_s\rho_g + N_1N_2\rho_g F_{ST}^2. \end{aligned} \tag{14}$$

From Bulik-Sullivan *et al.* (2015) we know that $\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{e}_1)(\mathbf{z}'_{j2}\mathbf{e}_2)\right] = N_s\rho_e$. Therefore from combining equations (8) and (14) we have

$$\mathbb{E}[T_{j1}T_{j2}] \approx \frac{\sqrt{N_1N_2}\rho_g}{M}\ell_j + \frac{N_s\rho}{\sqrt{N_1N_2}} + \rho_g F_{ST}^2\sqrt{N_1N_2} + \frac{\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{s}_1)(\mathbf{z}'_{j2}\mathbf{s}_2)\right]}{\sqrt{N_1N_2}}. \tag{15}$$

The last term to derive, $\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{s}_1)(\mathbf{z}'_{j2}\mathbf{s}_2)\right]$, can be rewritten as in Bulik-Sullivan *et al.* (2014) (eq. 2.11) as

$$\mathbb{E}\left[(\mathbf{z}'_{j1}\mathbf{s}_1)(\mathbf{z}'_{j2}\mathbf{s}_2)\right] = N_s^2\sigma_S^2 F_{ST}. \tag{16}$$

By combining equations (15) and (16), we obtain the final expression

$$\boxed{\mathbb{E}[T_{j1}T_{j2}] \approx \frac{\sqrt{N_1N_2}\rho_g}{M}\ell_j + \frac{N_s\rho}{\sqrt{N_1N_2}} + \rho_g F_{ST}^2\sqrt{N_1N_2} + \frac{N_s^2F_{ST}\sigma_S^2}{\sqrt{N_1N_2}}} \tag{17}$$

As a special case we can obtain the univariate LD score regression equation by assuming that cohort 1 is the same as cohort 2, i.e. $N_1 = N_2 = N_s = N$ and $\rho = 1$. We hence have

$$\mathbb{E}[\chi_j^2] \approx \frac{Nh^2}{M}\ell_j + 1 + NF_{ST}\left(h^2F_{ST} + \sigma_S^2\right) \text{ (same as eq. 2.12 in Ref}^1\text{)}. \tag{18}$$

Another interesting special case is the absence of sample overlap ($N_s = 0$), which leads to

$$\mathbb{E}[T_{j1}T_{j2}] \approx \frac{\sqrt{N_1N_2}\rho_g}{M}\ell_j + \rho_g F_{ST}^2\sqrt{N_1N_2}, \tag{19}$$

i.e. a non-zero intercept is expected even in the absence of sample overlap. Note in this case that the intercept is proportional to the geometric mean of the sample sizes of both GWAS ($\sqrt{N_1N_2}$) and the genetic covariance between the traits. It is therefore expected to increase with large sample sizes and more heritable traits. As a numerical example, although F_{ST}^2 is small in general, values of $h^2F_{ST}^2\sqrt{N_1N_2}$ can be as large as ~0.17, which would be indicative of sample overlap if we take for example $h^2 = 0.5$, $F_{ST} = 0.001$, $N_1 = 450,000$ (sample size of UK Biobank) and $N_2 = 250,000$ (sample size of Wood *et al.*, 2014).

Simulations

We performed a simulation to quantify the inflation of the bivariate LDSC intercept created when the sample size of each GWAS and the heritability are large. We used for our simulations genotypes at 1,123,348 HapMap 3 SNPs (Online methods) from 348,502 unrelated (genetic relationship < 0.05) participants of the UK Biobank (UKB) with European ancestry (Online methods). To mimic independent GWAS, we randomly split our dataset in two sub-samples of equal size ($N_1 = N_2 = 174,251$), and simulated traits from 10,000 causal variants (randomly sampled among HapMap 3 SNPs) and with an heritability varying from 0.1, 0.2,. . ., up to 0.9. Each trait was simulated with same SNPs effect sizes in each sub-sample so that the genetic correlation is expected to be $r_g = 1$. For each simulation replicate, we performed a GWAS of each simulated trait in each sub-sample separately, then used GWAS summary statistics to perform a bivariate LD score regression. LD score regression was performed using the LDSC software v1.0.0 and using LD scores from European samples of the 1,000 genomes reference panel. We performed

100 simulation replicates for each value of expected heritability.

We present the results our this simulation in Figure 1. Overall, we found an inflation of the bivariate LDSC intercept, which increases with the heritability of the trait. For example, with $h^2 = 0.9$ we observed bivariate LDSC intercepts as large as ∼0.1 (s.e. 0.02), which under the theory developed in Bulik-Sullivan *et al.* (2015), would falsely indicate a potential overlap of $\sim 0.1\sqrt{N_1 N_2} \approx 8,712$ participants between the two sub-samples of the UKB. We also observed an inflation of the univariate LDSC intercept (Figure 1, panel **b**), which is expected under Bulik-Sullivan *et al.* (2014) theory.

Under the assumptions made in this simulation (i.e. $N_s = 0$, $\sigma_S^2 = 0$, $N_1 = N_2 = N$ and $\rho_g = h^2$), Equation (17) predicts a affine relationship between univariate LDSC intercepts (I_u) within each cohort and bivariate LDSC intercept (I_b): $I_b = I_u - 1$. We validated this prediction in our simulated data as shown on Figure 2.

Empirical results: GWAS of height and body mass index (BMI)

We used summary statistics from published GWAS of height[4] (Wood *et al.*, 2014; median $N \approx 252,083$) and BMI[5] (Locke *et al.*, 2015; median $N \approx 233,692$). We also performed a GWAS of height and BMI in 348,502 unrelated participants of the UK Biobank (UKB) of European ancestry. We then estimated the bivariate LDSC intercept obtained from the comparison of Wood *et al.* (2014) with GWAS of height in UKB as well as from the comparison of Locke *et al.* (2015) with GWAS of BMI in UKB. We found in the first case a bivariate LDSC intercept ∼0.15 (s.e. 0.04) for height and ∼0.01 (0.01) for BMI. We previously reported similar observations using test statistics from linear mixed model association analyses in the UKB (Yengo *et al.*, 2018)[6]. Under Bulik-Sullivan *et al.* (2015) theory, these estimates suggest a significant overlap of $\sim 0.15 \times \sqrt{252,083 \times 348,502} = 44,460$ participants between the Wood *et al.* (2014) study and UKB but no signigicant overlap between partcipants of the Locke *et al.* (2015) study and UKB. Given that the same cohorts are included in the Locke *et al.* (2015) study and in the Wood *et al.* (2014) GWAS, these two conclusions are therefore contradictory or inconsistent with the Bulik-Sullivan *et al.* (2015) theory.

To further illustrate this contradiction, we performed in the 348,502 unrelated UKB participants a GWAS of height in females ($N_1 = 188,465$) and males ($N_2 = 160,037$) separately. Although no true sample overlap is to be expected, we nonetheless found a significant bivariate LDSC intercept of ~ 0.1 (s.e. 0.02), which also suggests a significant sample overlap under Bulik-Sullivan *et al.* (2015) theory.

In the light of the theoretical extension proposed in this note, we believe that these observations can be explained by the large sample sizes considered here, by the difference of heritability between height and BMI and the amount of trait variance explained by population stratification.

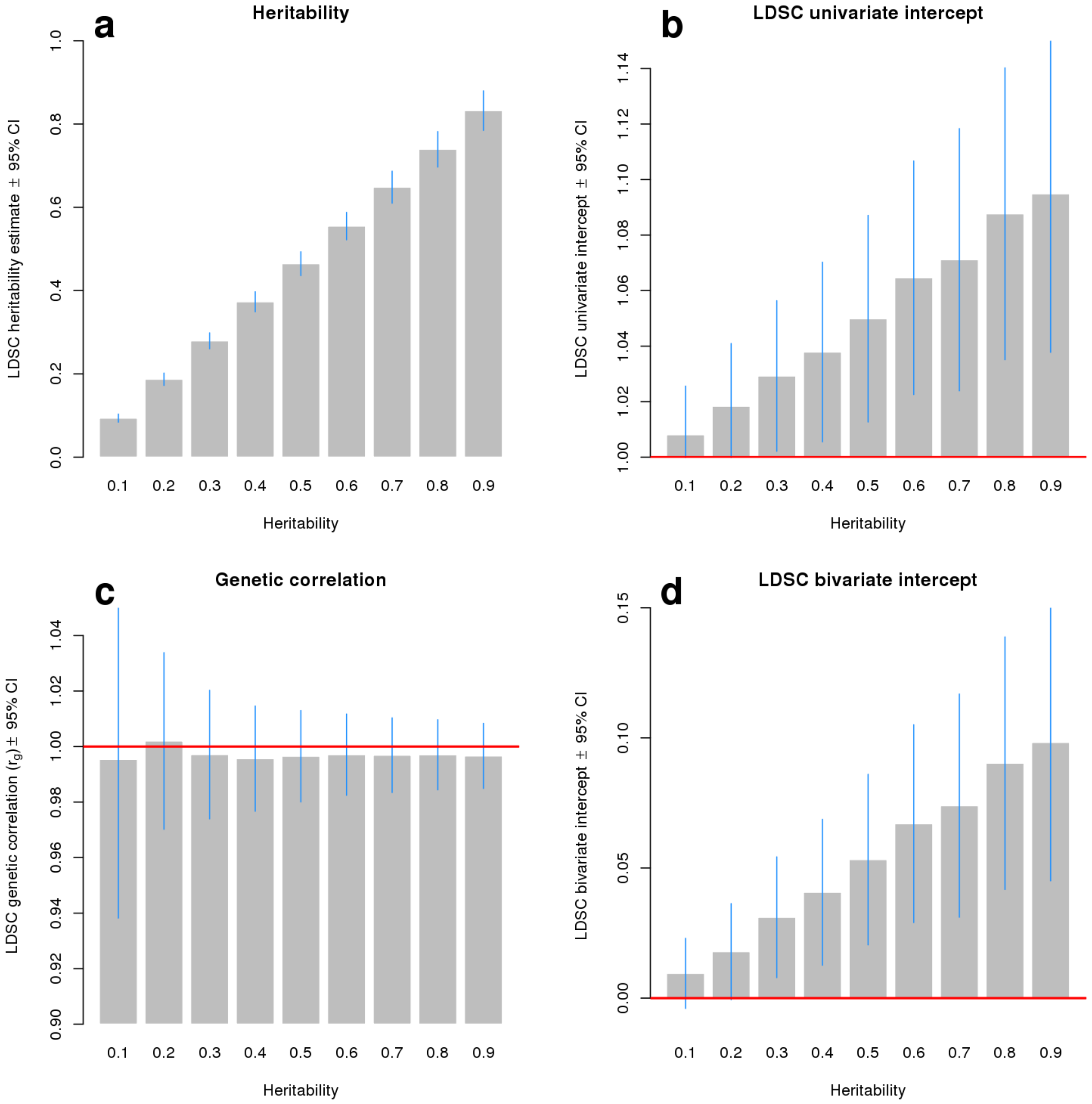

Figure 1: Statistics from the LD score regression applied to 900 simulated GWAS. Panels **a** and **b** show univariate LD score regression intercepts and estimates of heritability respectively obtained from analyzing summary statistics from each sub-sample separately, then averaged between the two independent sub-samples of participants of the UK Biobank (UKB). Panels **c** and **d** show estimates of genetic correlations (expected to be $r_g = 1$) between the two sub-samples and bivariate LD score regression intercepts respectively, indicating sample overlap between the two sub-samples of UKB, in particular when the underlying heritability is > 0.5.

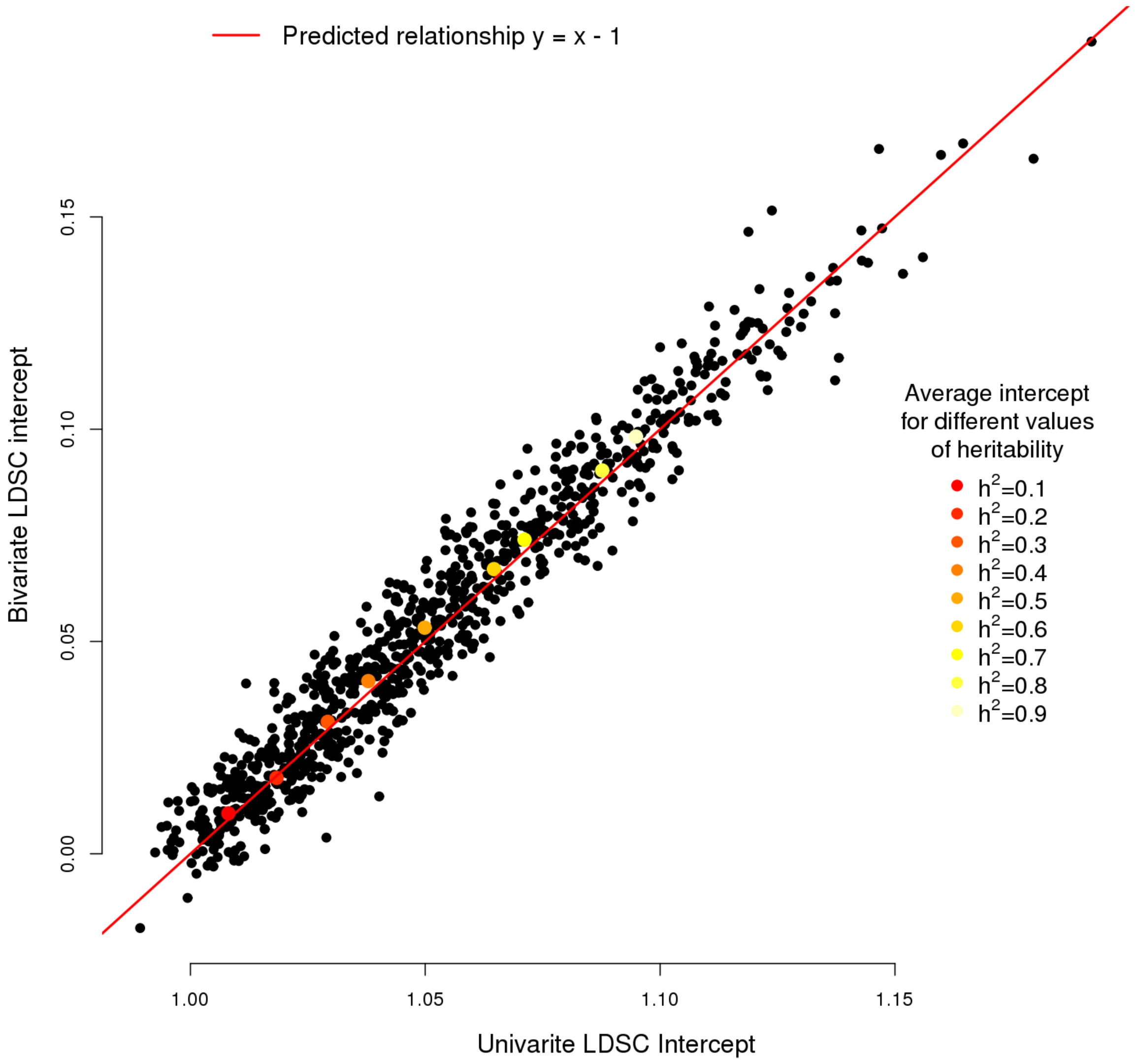

Figure 2: Relationship between univariate (x-axis) and bivariate (y-axis) LD score regression intercepts under the assumption that the same trait is analysed in both cohort (i.e. the genetic correlation $r_g = 1$), that population stratification does not explain any phenotypic variance ($\sigma_S^2 = 0$) and in the absence of sample overlap ($N_s = 0$). Each black dot corresponds one simulation replicate as described in our simulation study. Colored dots represent the mean of 100 simulations replicates obtained with a fixed value of heritability (h^2).

Discussion

We have developed in this note an extension of the theory underlying the bivariate LD score regression methodology in the presence of population stratification within each GWAS. Beyond the Bulik-Sullivan *et al.* (2015) theory, our results show that a non-zero bivariate LDSC intercept does not always indicate sample overlap but may also reflect patterns of population stratification within each study that are shared between studies.

Our extended theory thus explains and predicts a series of puzzling observations that the initial theory does not. For example, we can explain inconsistent detection of sample overlap from GWAS of traits with different heritabilities. Other theoretical extensions of the LDSC methodology have been previously proposed. We may for instance refer to the works of Lu *et al.* (2017)[7] (GNOVA method) who generalized the "sample overlap" term (N_s) by replacing it with the sum of genetic relatedness coefficients between partcipants of the two studies. This extension therefore predicts an inflation of the bivariate LDSC intercept if relatives span both studies, which is more general than the restriction to actual sample overlap. Another contribution by Lee *et al.* (2018)[8] is also worth mentioning here as it provides a rigorous mathematical framework that not only refines our understanding of the LD score regression methodology but also helps clarifying its interpretation. These two examples, among many others, both illustrate the effervescence of researches driven by the LD score regression methodology.

In conclusion, our findings improve the interpretation of results from bivariate LDSC analyses and may further have implications in other methodologies which use output statistics from LD score regression in their inference.

Online methods

UK Biobank data and summary statistics

We used genotypic and phenotypic data (height and body mass index) measured in participants of the UK Biobank. Samples and SNPs selection have been described in a previous publication[6]. We also analysed publicly available GWAS summary statistics from the Wood *et al.* (2014) and Locke *et al.* (2015) GWAS of height and BM respectively. Summary statistics from these studies were downloaded from the following link: `https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files`.

GWAS simulation

We simulated 900 GWAS (9 values of heritabiliy × 100 simulation replicates) according to the following steps: 1) For each simulation replicate and each value of heritability, we randomly sampled $M = 10,000$ SNPs as causal variants. 2) For each of the 348,502 UKB unrelated participants, we then simulated a quantitive trait y using the following equation

$$y_i = \sum_{j=1}^{M} \left[(x_{ij} - 2p_j) \left\{ 2p_j(1-p_j) \right\}^{-1/2} \right] \beta_j + e_i, \tag{20}$$

where x_j is the MAC of individual i at SNP j, p_j is the MAF of SNP j and β_j and e_i are independent normally distributed terms such as

$$\beta_j \sim \mathcal{N}(0, h^2/M) \text{ and } e_i \sim \mathcal{N}(0, 1-h^2). \tag{21}$$

3) Once the trait is simulated we randomly split the cohort into 2 equally sized sub-cohorts and performed SNP-trait association analyses using PLINK[9] in each sub-cohort.

Acknowledgements
This research was supported by the Australian Research Council (DP130102666 and DP160103860), the Australian National Health and Medical Research Council (1078037, 1078901, 1103418, 1107258, 1127440 and 1113400). This research has been conducted using the UK Biobank Resource under project 12505. We thank Tim Frayling, Joel Hirschhorn and members of GIANT consortium for fruitful discussions.

References

1. Bulik-Sullivan B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291295 (2014).

2. Loh, P.-R. *et al.* Mixed model association for biobank-scale data sets. *bioRxiv* 194944 (2018). doi:10.1101/194944.

3. Bulik-Sullivan B.K. *et al.* An Atlas of Genetic Correlations across Human Diseases and Traits. *Nat. Genet.* **47**, 12361241 (2015).

4. Wood A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173-1186 (2014).

5. Locke A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).

6. Yengo L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in 700,000 individuals of European ancestry. *Biorxiv* (2018).

7. Lu Q. *et al.* A Powerful Approach to Estimating Annotation-Stratified Genetic Covariance via GWAS Summary Statistics. *Am. J. Hum. Genet.* **101**, 939-964 (2017).

8. Lee J.J. *et al.* The accuracy of LD Score regression as an estimator of confounding and genetic correlations in genome-wide association studies. *Biorxiv* (2018).

9. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007).