Supplementary Note to Estimating Genetic Correlation from GWAS Summary Statistics

Brendan Bulik-Sullivan

November 22, 2014

Contents

1 Definitions

Let y_1 and y_2 denote phenotypes defined for individuals in a hypothetical population of infinite size (or more precisely, for individuals drawn from a distribution). Let g denote a set of additively-coded SNPs, and let g_1 and g_2 denote the best linear predictors of y_1 and y_2 that can be constructed (at infinite sample size) from the SNPs in S[1]. Then we can write

$$y_1 = g_1 + \epsilon_1;$$
$$y_2 = g_2 + \epsilon_2,$$

where ϵ_i denotes the residual, which is uncorrelated (in the population) with g_i by the definition of a projection. Note that so far this construction is applicable to arbitrary phenotypes[2].

Definition 1.1. *The narrow-sense (or additive)* ***heritability*** *of phenotype y_i explained by the SNPs in g, denoted $h^2_g(y_i)$ is defined*

$$h^2_g(y_i) := \mathrm{Cor}[g_i, y_i]^2, \tag{1.1}$$

where Cor *denotes the correlation between random variables, (alternatively, the correlation in a hypothetical population of infinite size), not the empirical correlation in some finite sample.*

Definition 1.2. *The [additive]* ***genetic covariance*** *between y_1 and y_2 among SNP set g, denoted $\rho_g(y_1, y_2)$ is defined*

$$\rho_g(y_1, y_2) := \frac{\mathrm{Cov}[g_1, g_2]}{\sqrt{\mathrm{Var}[y_1]\mathrm{Var}[y_2]}}. \tag{1.2}$$

Definition 1.3. *The [additive]* ***genetic correlation*** *between y_1 and y_2 among SNP set g, denoted $r_g(y_1, y_2)$ is defined*

$$r_g(y_1, y_2) := \frac{\rho_g}{\sqrt{h^2_g(y_1)h^2_g(y_2)}}. \tag{1.3}$$

Note that these definitions make sense even when either or both phenotypes are binary, and we refer to the specialization of definition 1.1 to a binary phenotype as the ***heritability of the observed phenotype***.

There are two challenges when dealing with binary phenotypes. The first is inferential: often studies of binary phenotypes will over-sample cases in order to increase power. Some work is required in order to obtain valid estimates of the parameters of a population with, say, 1% cases from an ascertained sample with 50% cases. Ascertainment is addressed in section **??**. The second challenge is definitional: the heritability of the observed phenotype depends strongly on the prevalence of the phenotype. For example, consider two liability threshold phenotypes y_1 and y_2, determined by the same underlying liability ψ, but with different thresholds. That is, $y_i := \mathbf{1}[\psi > \tau_i]$ for $i = 1, 2$. Suppose $h^2_g(\psi) = 1$, $\tau_1 = 0$ and $\tau_2 = 1.96$ (meaning the population prevalence of y_1 is 50% and the

[1] Formally, the g_i are constructed by projecting the phenotypes onto the vector space of functions $\{0, 1, 2\}^M \to \mathbb{R}$, where $M := |g|$. As a result g_i may account for a large proportion of the variance in phenotype, even if in truth the way in which the phenotype is determined from genotype and environmental factors is completely non-additive.

[2] Well, measurable finite-variance phenotypes. But this is no restriction at all on the genetic component, and hardly any restriction at all on the environmental component.

population prevalence of y_2 is 5%), then the heritability of the observed phenotype y_1 is 0.64, and the heritability of the observed phenotype y_2 is 0.23.

Sometimes it is desirable to compare the heritabilities or genetic covariances of phenotypes with different prevalences on an even footing, and this is the primary application of liability-scale heritability and liability scale genetic covariance. We note that one need not take the liability threshold model literally[3] in order for the conversion to the liability scale to be a useful procedure. One can view this conversion simply as a tool for comparing phenotypes with different prevalences that is inspired by – but not dependent on – the liability threshold model. An interpretation of LD Score regression under the liability threshold model is provided in section **??**.

There is no need to specify a scale when discussing genetic correlation. Genetic correlation on the observed scale is the same as genetic correlation on the liability scale is the same as genetic correlation in an ascertained sample.

2 Models

2.1 Quantitative Traits

Suppose we sample two cohorts for two phenotypes, y_1 and y_2, with sample sizes N_1 and N_2, such that N_s individuals are shared between cohorts and phenotyped for both traits. We model phenotypes as generated by the equations

$$y_1 = Y\beta + \delta;$$
$$y_2 = Z\gamma + \epsilon,$$

where Y and Z are matrices of genotypes normalized to mean zero and variance one[4], with dimensions $N_1 \times M$ and $N_2 \times M$, respectively; β and γ are $M \times 1$ vectors of per-normalized genotype effect sizes, and δ and ϵ are vectors of environmental or non-additive genetic effects, with dimensions $N_1 \times 1$ and $N_2 \times 1$, respectively. In this model, Y and Z are unobserved matrices of *all* SNPs, unlike Yang, *et al* [9], we model the effects of SNPs that are not genotyped as well as those that are.

Now we introduce randomness: we model all of $Y, Z, \beta, \gamma, \delta$ and ϵ as random variables. Suppose that the $2M \times 1$ vector (β, γ) follows a distribution with mean zero and variance-covariance matrix

$$\mathrm{Var}[(\beta, \gamma)] = \frac{1}{M} \begin{pmatrix} h_1^2 I & \rho_g I \\ \rho_g I & h_2^2 I \end{pmatrix},$$

and the $2N \times 1$ vector (δ, ϵ) follows a distribution with mean zero and variance-covariance matrix

$$\mathrm{Var}[(\delta, \epsilon)] = \begin{pmatrix} (1-h_1^2) I & \rho_e I \\ \rho_e I & (1-h_2^2) I \end{pmatrix}.$$

[3] We also note that the liability threshold model is (trivially) completely general. Let y denote an arbitrary binary phenotype with prevalence K, and set $\tau := \Phi^{-}1(1-K)$, where Φ is the cdf of the standard normal distribution. We can construct a liability for y as follows: if individual i is a case, draw a liability ψ_i from a truncated standard normal with left truncation point τ. If individual i is a control, draw a liability ψ_i from a truncated standard normal with right truncation point τ. Then $y = \mathbf{1}[\psi > \tau]$.

[4] We ignore the distinction between normalizing and centering in the population and in the sample, since this introduces only $\mathcal{O}(1/N)$ error.

Finally suppose that each row (individual) of Y and Z represents an *i.i.d.* draw from a distribution with covariance matrix (LD matrix) R (except of course the N_s rows that are duplicated in Y and Z). We will write $\mathbb{E}[Y_{ij}Y_{ik}] = R_jk =: r_jk$. Note that since we assume normalized genotypes, R is both the covariance matrix and correlation matrix. Additionally, note that under this model, $\mathrm{Var}[y_1] = \mathrm{Var}[y_2] = 1$. Let $\rho := \rho_g + \rho_e$ denote the covariance between y_1 and y_2, which is also the correlation between y_1 and y_2, since both have variance one.

The assumption that all β is drawn with equal variance for all SNPs is, of course, not reasonable. We only make this assumption here for notational simplicity. In this paper, we use MAF- and LD-partitioned LD Score regression for estimation, as described in references [5, 2]. This technique minimizes confounding under models where $\mathrm{Var}[\beta]$ is correlated with MAF or LD Score.

Proposition 2.1. *Under this model, the expected genetic covariance between phenotypes is ρ_g, justifying our use of the notation ρ_g.*

Proof. Let X denote an $1 \times M$ vector of normalized, centered genotypes for an arbitrary individual. Under the model, the additive genetic component of y_1 for this individual is $\sum_j X_j\beta_j$, and the additive genetic component of y_2 for this individual is $\sum_j X_j\gamma_j$. Thus, the genetic covariance between y_1 and y_2 is

$$\mathrm{Cov}\left[\sum_{j=1}^{M} X_j\beta_j, \sum_{j=1}^{M} X_j\gamma_j\right],$$

We can simplify this covariance with some algebra:

$$\begin{aligned}
\mathrm{Cov}\left[\sum_{j=1}^{M} X_j\beta_j, \sum_{j=1}^{M} X_j\gamma_j\right] &= \mathbb{E}\left[\left(\sum_{j=1}^{M} X_j\beta_j\right)\left(\sum_{j=1}^{M} X_j\gamma_j\right)\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{M}\sum_{k=1}^{M} X_jX_j\beta_j\gamma_k\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{M} X_j^2\beta_j\gamma_j\right] \\
&= \sum_{j=1}^{M} \mathbb{E}[X_j^2]\mathbb{E}[\beta_j\gamma_j] \\
&= \rho_g.
\end{aligned}$$

□

2.2 Liability Threshold Model

Suppose unobserved liability is generated following the usual model for quantitative traits:

$$\psi_i = \sum_{j=1}^{M} X_{ij}\beta_j + \epsilon_i.$$

The observed binary phenotype has population prevalence K, and is generated from the unobserved liability via the liability threshold model:

$$y_i := \mathbf{1}[\psi_i > \tau],$$

where $\tau := \Phi^{-1}(1-K)$ is the liability threshold, and Φ denotes the cdf of the standard normal distribution.

3 Derivations

3.1 Non-Ascertained Samples

3.1.1 Genetic Covariance

Suppose we directly genotype SNP j. We estimate effect sizes using least-squares:

$$\hat{\beta}_j := \frac{1}{N_1} Y_j^\mathsf{T} y_1;$$
$$\hat{\gamma}_j := \frac{1}{N_2} Z_j^\mathsf{T} y_2,$$

where Y_j and Z_j denote the genotypes of all individuals at SNP j and have dimensions $N_1 \times 1$ and $N_2 \times 1$, respectively.

Proposition 3.1. *Under the model described in 2.1,*

$$\mathbb{E}[\hat{\beta}_j \hat{\gamma}_j] = \frac{\rho_g}{M} \ell_j + \frac{N_s \rho}{N_1 N_2}. \tag{3.1}$$

Proof. By the law of total expectation,

$$\mathbb{E}[\hat{\beta}_j \hat{\gamma}_j] = \mathbb{E}[\mathbb{E}[\hat{\beta}_j \hat{\gamma}_j \,|\, Y, Z]]$$

First,

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_j \hat{\gamma}_j \,|\, Y, Z] &= \frac{1}{N_1 N_2} \mathbb{E}[Y_j^\mathsf{T} y_1 y_2^\mathsf{T} Z_j] \\
&= \frac{1}{N_1 N_2} Y_j^\mathsf{T} \mathbb{E}[(Y\beta + \delta)(Z\gamma + \epsilon)] Z_j \\
&= \frac{1}{N_1 N_2} Y_j^\mathsf{T} \left(Y \mathbb{E}[\beta \gamma^\mathsf{T}] Z + \mathbb{E}[\delta^\mathsf{T} \epsilon] \right) Z_j \\
&= \frac{1}{N_1 N_2} \left(\frac{\rho_g}{M} Y_j^\mathsf{T} Y Z_j^\mathsf{T} Z + \rho_e Y_j^\mathsf{T} Z_j \right).
\end{aligned}$$

To remove the conditioning on Y and Z, we need only compute

$$\frac{1}{N_1 N_2} \mathbb{E}[Y_j^\mathsf{T} Z_j] = \frac{N_s}{N_1 N_2},$$

and

$$\frac{1}{N_1 N_2} \mathbb{E}[Y_j^\mathsf{T} Y Z_j^\mathsf{T} Z] = \ell_j + \frac{M N_s}{N_1 N_2}.$$

Thus,

$$\mathbb{E}[\hat{\beta}_j\hat{\gamma}_j] = \frac{\rho_g}{M}\ell_j + \frac{N_s\rho}{N_1N_2}.$$

Note that this expression does not contain any terms in which both the quantities N_s and ℓ_j appear. In the special case where there are no overlapping samples shared between the two cohorts, this expression simplifies to

$$\mathbb{E}[\hat{\beta}_j\hat{\gamma}_j] = \frac{\rho_g}{M}\ell_j.$$

□

3.1.2 Regression Weights

We can improve the efficiency of the LD Score regression by computing the conditional variance $\mathrm{Var}[\hat{\beta}\hat{\gamma} \,|\, \ell_j]$ and weighting the regression by the reciprocal of this variance. In order to compute this conditional variance, we need further assumptions: in addition to the assumptions from 2.1, assume that the phenotypes follow a multivariate normal distribution[5].

If phenotypes are normally distributed, then by the central limit theorem, $\hat{\beta}$ and $\hat{\gamma}$ are jointly normally distributed with expectation zero. Thus,

$$\begin{aligned}
\mathrm{Var}[\hat{\beta}_j\hat{\gamma}_j \,|\, Y, Z] &= \mathbb{E}[\hat{\beta}^2\hat{\gamma}^2] \\
&= \mathrm{Var}[\hat{\beta}]\mathrm{Var}[\hat{\gamma}] + 2\mathbb{E}[\hat{\beta}\hat{\gamma}]^2 \\
&= \left(\frac{h_1^2\ell_j}{M} + \frac{1}{N_1}\right)\left(\frac{h_2^2\ell_j}{M} + \frac{1}{N_2}\right) + 2\left(\frac{\rho_g\ell_j}{M} + \frac{\rho N_s}{N_1N_2}\right)^2. \qquad (3.2)
\end{aligned}$$

Note that we only assume normality in order to compute regression weights. If (quantitative) phenotypes are not normally distributed, this will not affect the expectation of the LD Score regression estimates (see 3.1.1, which makes no distributional assumptions about β and γ beyond first and second moments), but will increase the standard error, because in this case the regression weights will not be optimal for correcting for heteroskedasticity. We never assume homoskedasticity when computing standard errors or p-values (we use a block jackknife, which is robust to heteroskedasticity), so non-normality of the phenotypes also does not bias our inference. We derive regression weights for ascertained studies of binary phenotypes in section **??**. Note that if the phenotypes are normally distributed, then $\hat{\beta}_j\hat{\gamma}_j$ follows a product-normal distribution, which is not in the exponential family, so this is not a GLM.

3.2 Ascertainment

In this section, we derive the LD Score regression estimators of heritability and genetic covariance for ascertained case/control samples (which was addressed only via simulation in [3]). The fact that this estimator works is *not* a consequence of the equivalence between LD Score regression and HE regression (see section **??**), and the fact that HE regression works in ascertained case/control

[5]For instance, it is sufficient but not necessary to assume that β, γ, δ and ϵ are multivariate normal, and that N_1 and N_2 are sufficiently large that we can invoke the central limit theorem. More generally, the phenotypes will be approximately normal if δ and ϵ are normal and if β and γ are reasonably polygenic. If there are few casual SNPs, then the conditional variance may be larger.

samples [6], because case/control ascertainment induces LD between causal SNPs in the ascertained samples. HE regression accounts for this LD by using a GRM computed from sample genotypes. It is not clear *a priori* that the LD Score regression approach of using population LD as an estimate of sample LD is valid when the sample is ascertained. However, this turns out to be fine, though we do not address this issue directly. Our proof strategy is first to note that GWAS summary statistics can be written in terms of the sample allele frequencies in cases and the sample allele frequencies in controls. Since the sample allele frequency in cases is a consistent estimator of the population allele frequency in cases, and likewise for the sample allele frequency in controls, we can write the large-N limit of our GWAS summary statistics in terms of the population allele frequencies (see section 3.2.1). Since the population allele frequencies depend on population LD rather than ascertained sample LD, LD Score regression with population LD yields a consistent estimators of heritability and genetic covariance.

This section is structured as follows: in 3.2.2 and 3.2.3, we show that using estimates of population LD is a valid, even with ascertained samples, and we derive the usual factors for converting heritabilities and genetic covariances from ascertained samples into estimates of the population heritabilities and genetic covariances.

In 3.2.4, we deal with the special case of the liability threshold model, and derive LD Score regression estimators of the heritability of liability, and genetic covariance between liability and other phenotypes.

3.2.1 Case/Control Test Statistics

Consider a study of size N where the sample prevalence of phenotype y is P. Let $N_{eff} := NP(1-P) = N_0N_1/N$ denote the effective sample size. We compute Z-statistics

$$Z_j := \frac{\sqrt{N_{eff}}(\hat{p}_1 - \hat{p}_0)}{\sqrt{\hat{p}_j(1-\hat{p}_j)}}, \tag{3.3}$$

and χ^2-statistics[6]

$$\chi_j^2 := Z_j^2, \tag{3.4}$$

where $\hat{p}_j$ denotes allele frequency in the entire sample[7], $\hat{p}_1$ denotes allele frequency among cases in the sample and $\hat{p}_0$ denotes allele frequency among controls in the sample. We aim to derive an estimator of heritability from $\mathbb{E}[\chi_j^2 \,|\, \ell_j]$ and an estimator of genetic covariance from $\mathbb{E}[Z_j \,|\, \ell_j]$ in samples where $P \neq K$ under various models of genetic architecture. First, we need a lemma, which allows us to write our χ^2-statistics in terms of population allele frequencies in the large-N limit.

Lemma 3.1. *In an ascertained study with sample size N, sample prevalence P and population prevalence K, the expected Z-statistic of a SNP j conditional on its population allele frequencies in cases and controls is*

$$\mathbb{E}[Z_j \,|\, p_0, p_1] = \tag{3.5}$$

and the expected χ^2 statistic is

$$\mathbb{E}[\chi_j^2 \,|\, p_0, p_1] = +1. \tag{3.6}$$

[6]This is the equal to N times the squared correlation between phenotype and genotype, *i.e.,* the Armitage Trend Test (ATT).

[7]Note that if j has nonzero effect size, the expected value of $\hat{p}_j$ is not equal to p_j unless $P = K$.

Proof. Note that the case where we condition on p_0 and p_1 (*i.e.,* when the only randomness is from sampling and genetic architecture is nonrandom), is the usual case considered in power analyses for GWAS, so we can even obtain the asymptotic distributions of the Z and χ^2 statistics from standard results on Wald statistics. First,

$$Z_j \,|\, p_0, p_1 \sim N()$$

so

$$\mathbb{E}[Z_j \,|\, p_0, p_1] =$$

and χ^2_j follows a noncentral χ^2 distribution with one degree of freedom and non-centrality parameter

$$\text{NCP} =$$

Since the expected value of a noncentral χ^2 distribution with k degrees of freedom and noncentrality parameter λ is $k + \lambda$,

$$\mathbb{E}[\chi^2_j \,|\, p_0, p_1] = +1$$

□

3.2.2 Heritability of the Observed Phenotype

Next, we remove the conditioning on p_0 and p_1 by noting that p_0 and p_1 are fixed conditional on β_{loc}, and can be approximated using 3.10 and 3.11. Thus, the inner term of the numerator from **??** is

$$\begin{aligned}\mathbb{E}[(\hat{p}_1 - \hat{p}_0)^2 \,|\, \beta_{loc}] \approx \frac{p_j^2\phi(\tau)^2\alpha_j^2}{K^2(1-K)^2} &+ \frac{p_j(1-p_j) + \frac{p_j\phi(\tau)\alpha_j}{K}\left(1 - 2p_j - \frac{p_j\phi(\tau)\alpha_j}{K}\right)}{N_1} \\ &+ \frac{p_j(1-p_j) + \frac{p_j\phi(\tau)\alpha_j}{K}\left(2p_j - 1 - \frac{p_j\phi(\tau)\alpha_j}{K}\right)}{N_0}. \\ &\approx \frac{p_j^2\phi(\tau)^2\alpha_j^2}{K^2(1-K)^2} + \mathscr{O}(1/N). \end{aligned} \tag{3.7}$$

Next, we remove the conditioning on p_0 and p_1. Let $C := \phi(\tau)P(1-K)+K(1-P))/(K(1-K))$. Using the approximations from 3.10 and 3.11,

$$\tilde{p}_j \,|\, \beta_{loc} \approx p_j\left(1 + C\alpha_j\right),$$

and

$$\begin{aligned}\mathbb{E}[\tilde{p}_j(1-\tilde{p}_j) \,|\, \beta_{loc}] &= p_j(1 + C\alpha_j)(1 - p_j - p_jC\alpha_j) \\ &= p_j\left(1 - p_j + C\alpha_j(1 - 2p_j - p_jC\alpha_j).\right).\end{aligned}$$

Note that we have already computed $p_0(1-p_0) \,|\, \beta_{loc}$ and $p_1(1-p_1) \,|\, \beta_{loc}$ in 3.7. Thus, the inner

term of the denominator of **??** is

$$\begin{aligned}\mathbb{E}[\hat{p}_j(1-\hat{p}_j)\,|\,\beta_{loc}] &= p_j\left(1-p_j+C\alpha_j(1-2p_j-p_jC\alpha_j)\right)\\ &\quad-\frac{(1-P)^2p_j(1-p_j)+\frac{p_j\phi(\tau)\alpha_j}{K}\left(2p_j-1-\frac{p_j\phi(\tau)\alpha_j}{K}\right)}{N_0}\\ &\quad-\frac{P^2p_j(1-p_j)+\frac{p_j\phi(\tau)\alpha_j}{K}\left(1-2p_j-\frac{p_j\phi(\tau)\alpha_j}{K}\right)}{N_1}\\ &\approx p_j\left(1-p_j+C\alpha_j(1-2p_j-p_jC\alpha_j)\right)+\mathscr{O}(1/N). \end{aligned} \tag{3.8}$$

We now introduce randomness into β. For this section, model the entries of β as *i.i.d.* draws from a distribution with expectation zero and variance h^2/M. The heritability of liability (in the population) under this model is h^2; $\mathbb{E}[\alpha_j]=0$; $\mathbb{E}[Z_j]=0$, and

$$\mathbb{E}[\alpha_j^2]=\frac{(1-p_j)h^2}{p_jM}\ell_j.$$

The expected numerator of the χ^2-statistic is therefore

$$\begin{aligned}N_{eff}\mathbb{E}[(\hat{p}_1-\hat{p}_0)^2] &= p_j(1-p_j)\left(\frac{N_{eff}\phi(\tau)^2h^2\ell_j}{K^2(1-K)^2M}\left(1+\frac{N(1-K)^2}{N_0N_1K^2}\right)+\frac{NN_{eff}}{N_0N_1}\right)\\ &\approx p_j(1-p_j)\left(\frac{N_{eff}\phi(\tau)^2h^2\ell_j}{MK^2(1-K)^2}+1\right),\end{aligned}$$

where the approximation is justified by the fact that $N/N_0N_1 \ll 1$, which is a reasonable approximation when P is not so far from $1/2$ (a balanced study). For brevity, let c denote

$$c := \frac{P(1-P)\phi(\tau)^2}{K^2(1-K)^2},$$

which is the factor used to convert between liability scale heritability, h^2 and observed scale heritability, $h^2_{obs} := ch^2$.

The expected denominator of the χ^2-statistic is

$$\begin{aligned}\mathbb{E}[\hat{p}_j(1-\hat{p}_j)] &= p_j(1-p_j)-\mathscr{O}(\ell_j/M)+\mathscr{O}(1/N)\\ &\approx p_j(1-p_j).\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}[\chi_j^2] &\approx c\frac{Nh^2}{M}\ell_j+1\\ &= \frac{Nh^2_{obs}}{M}\ell_j+1.\end{aligned} \tag{3.9}$$

3.2.3 Genetic Covariance with the Observed Phenotype

In this section, we derive a genetic covariance estimator that works when both studies are ascertained to oversample cases and may include overlapping samples. This derivation does not cover more complicated ascertainment schemes (*e.g.,* attempting to estimate genetic covariance between T2D and BMI from a T2D GWAS consisting of low-BMI cases and high-BMI controls).

3.2.4 Heritability and Genetic Covariance of Liability

For phenotypes generated according to the liability threshold model, we can estimate not only the heritability of the observed phenotype (genetic covariance between the observed phenotype and other phenotypes), but also the heritability of the unobserved liability (genetic covariance between unobserved liability and other phenotypes).

The derivation is the same as in 3.2.2 and 3.2.3, except with one extra step: we need to write $\mathbb{P}[y_i = 1 \,|\, G_{ij} = 1]$ in terms of the heritability of liability.

Let $\alpha_j := \mathbb{E}[\psi \,|\, G_{ij} = 1] = p_j^{-\frac{1}{2}}(1-p_j)^{\frac{1}{2}} \sum_{\{k \,|\, r_{jk} \neq 0\}} r_{jk}\beta_k$ denote the marginal per-normalized genotype effect size of SNP j on liability. This is the per-normalized-genotype effect size that one would obtain from regressing liability against the genotype at SNP j at infinite sample size.

Then if $\phi(x, \mu, \sigma^2)$ denotes the density of a normal distribution with expectation μ and variance σ^2 evaluated at x,

$$\begin{aligned}
\mathbb{P}[y_i = 1 \,|\, G_{ij} = 1] &= \int_{\tau}^{\infty} \phi(x, \alpha_j, 1-\alpha_j^2)dx \\
&= \int_{\tau-\alpha_j}^{\infty} \phi(x(1-\alpha_j^2)^{\frac{1}{2}}; 0, 1)dx \\
&= (1-\alpha_j^2)^{-\frac{1}{2}} \int_{(\tau-\alpha_j)(1-\alpha_j^2)^{\frac{1}{2}}}^{\infty} \phi(u; 0, 1)du \\
&= (1-\alpha_j^2)^{-\frac{1}{2}}(1-\Phi(\tau')),
\end{aligned}$$

where $\tau' := (\tau - \alpha_j)(1-\alpha_j^2)^{\frac{1}{2}}$. Similarly,

$$\begin{aligned}
\mathbb{P}[y_i = 0 \,|\, G_{ij} = 1] &= 1 - \mathbb{P}[y_i = 1 \,|\, G_{ij} = 1] \\
&= 1 - (1-\alpha_j^2)^{-\frac{1}{2}}(1-\Phi(\tau')).
\end{aligned}$$

We approximate $\Phi(\tau')$ with a first-order Taylor expansion[8] around τ:

$$\begin{aligned}
\Phi(\tau') &\approx \Phi(\tau) + \phi(\tau)(\tau' - \tau) \\
&= 1 - K + \phi(\tau)(\tau' - \tau).
\end{aligned}$$

We also make the approximation that

$$(1-\alpha_j^2)^{\frac{1}{2}} \approx 1.$$

Thus we have

$$\begin{aligned}
p_1 \,|\, \beta_{loc} &= \frac{p_j}{K}(1-\alpha_j^2)^{-\frac{1}{2}}(1-\Phi(\tau')) \\
&\approx p_j \left(1 + \frac{\phi(\tau)\alpha_j}{K}\right), \qquad (3.10)
\end{aligned}$$

and

$$\begin{aligned}
p_0 \,|\, \beta_{loc} &= \frac{p_j}{1-K}\left(1 - (1-\alpha_j^2)^{-\frac{1}{2}}(1-\Phi(\tau'))\right) \\
&\approx p_j \left(1 - \frac{\phi(\tau)\alpha_j}{1-K}\right). \qquad (3.11)
\end{aligned}$$

[8]This is a reasonable approximation for small α_j, *e.g.,* for polygenic traits and away from loci with huge effects.

We can plug these results into the expressions from 3.2.2 in order to obtain an estimator of the heritability of liability:

WRITE MEEEE

3.2.5 Regression Weights

Lorem Ipsum Dolor Sic Amet

3.2.6 Comparison with Polygenic Scoring

In this section, we compare the power of summary-statistic based approaches with polygenic risk scoring as tests of the null hypothesis $H_0 : r_g = 0$. LD Score regression with unconstrained intercept is robust to several common confounders, such as unknown sample overlap and shared population stratification, but this robustness comes at the cost of a substantial increase in standard error. As a result, LD Score regression with unconstrained intercept is underpowered as a test of H_0, except at very large sample sizes. For purposes of testing H_0 given only summary statistics, we recommend eliminating confounders via other methods (*e.g.,* by quantifying sample overlap and correcting for population stratification with PCA) then using LD Score regression with constrained intercept, which (modulo weighting) is equivalent to testing for association between $\hat{\beta}$ and $\hat{\gamma}$.

For simplicity, we consider only the case of LD pruned markers and independent samples for the comparison. One can of course increase power by not LD pruning (both for prediction and with summary statistics), and the samples need not be independent.

The polygenic scoring method is described in detail by Dudbridge in [4]. Briefly, given a vector $\hat{\beta}$ of effect size estimates for trait 1, and a genotype matrix X for individuals phenotyped for trait 2, denoted Y, we can test the null hypothesis H_0 by testing whether $\mathrm{Cor}[X\hat{\beta}, Y]$ is significantly different from zero.

As derived by Dudbridge [4], the NCP for the χ^2 test of association between $X\hat{\beta}$ and Y is

$$\lambda_{PRS} = \frac{N_2 R^2_{PRS}}{1 - R^2_{PRS}}, \tag{3.12}$$

where R^2_{PRS} is the expected squared Pearson correlation between the predicted and true phenotypes, which is given by

$$R^2_{PRS} = \frac{N_1 \rho_g^2}{M(N_1 h_1^2/M + 1)}, \tag{3.13}$$

(assuming Y has been pre-normalized to variance one).

The squared correlation between $\hat{\beta}$ and $\hat{\gamma}$ (estimated in independent samples) is

$$\begin{aligned} R^2_S :&= \mathrm{Cor}[\hat{\beta}, \hat{\gamma}] \\ &= \frac{\mathrm{Cov}[\hat{\beta}, \hat{\gamma}]^2}{\mathrm{Var}[\hat{\beta}]\mathrm{Var}[\hat{\gamma}]} \\ &= \frac{\mathrm{Cov}[\beta + \delta, \gamma + \epsilon]^2}{\mathrm{Var}[\beta + \delta]\mathrm{Var}[\gamma + \epsilon]} \\ &\approx \frac{N_1 N_2 \rho_g^2}{M^2(N_1 h_1^2/M + 1)(N_2 h_2^2/M + 1)} \end{aligned}$$

where the approximation is $\text{Var}[\hat{\beta} + \epsilon] \approx h_1^2/M + N_1^{-1}$, which is the same approximation from [4], and holds when genetic effects are small. Thus, the NCP of the χ^2 test for association between $\hat{\beta}$ and $\hat{\gamma}$ is

$$\lambda_S = \frac{MR_S^2}{1 - R_S^2}. \tag{3.14}$$

Since MR_S is equal to $N_2 R_{PRS}/(N_2 h_2^2/M + 1)$, and the term in the denominator is greater than or equal to one, we have the inequality $\lambda_S \leq \lambda_{PRS}$, with equality if and only if $h_2^2 = 0$. The magnitude of the difference will increase with increasing $h_2^2 N_2$. Thus, with no LD, no p-value thresholding, PRS is a more powerful test for genetic correlation than LD Score regression with constrained intercept (which under these conditions is equivalent to computing $\text{Cor}[\hat{\beta}, \hat{\gamma}]$).

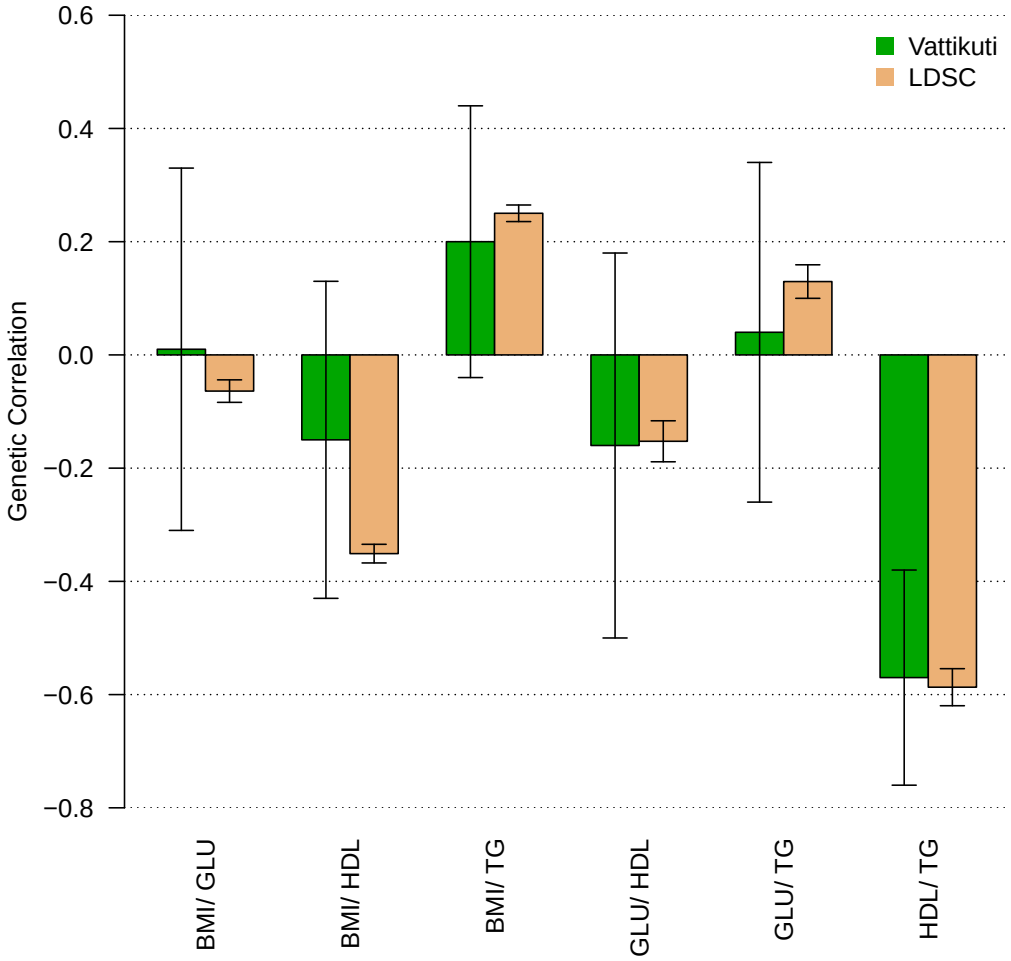
In the more realistic scenario where there is LD, PRS loses power as a result of LD pruning, which always removes some signal of heritability (though we note that LD pruning need not be a necessary step in risk prediction). In addition, LD Score regression gains power via efficient regression weights and by the ease with which one can incorporate functional priors into the regression (simply by replacing $\sum r_{jk}^2$ with $\sum w_k r_{jk}^2$, where w_k is an estimate of the per-SNP heritability accounted for by a SNP having the same annotations as k.

3.2.7 p-Value Thresholding

If (as is likely) some SNPs are not causal, then we can improve power for both PRS and methods based on summary statistics by removing SNPs that have no effect on the phenotypes in question. We can accomplish this either with functional priors or via p-value thresholding. The utility of p-value thresholding increases with sample size. If we have genotypes for trait 1, and this study has much larger sample size than the study of trait 2, then our only option for polygenic risk scoring is to attempt to predict trait 1 via a risk score obtained from the study of trait 2. However, in this case we can only impose a p-value threshold based on the p-values from study 2, which will be much less useful than a p-value threshold based on p-values from the much larger study 2. In situations like this, summary statistic based methods may be more powerful than PRS, even under the simple no-LD model considered here.

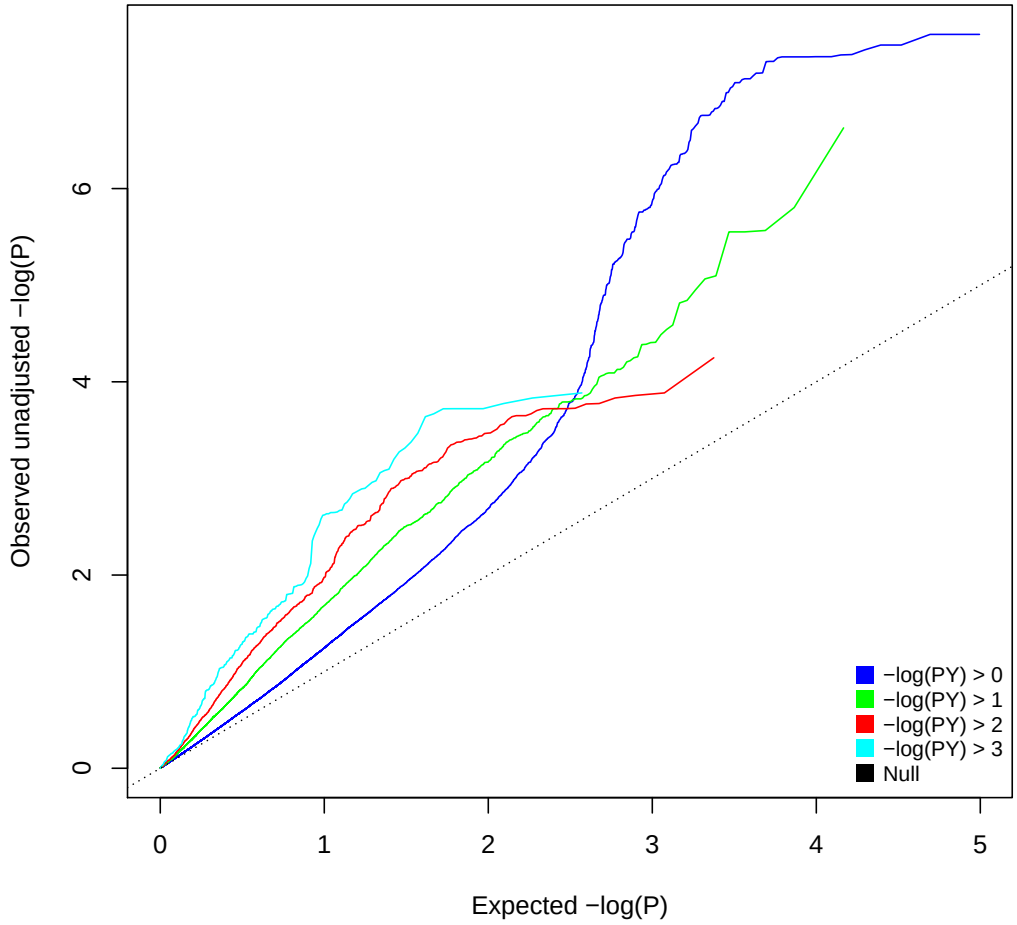
4 Supplementary Figures

4.1 Comparison of Metabolic Genetic Correlations from LDSC to Results from Vattikuti, et al



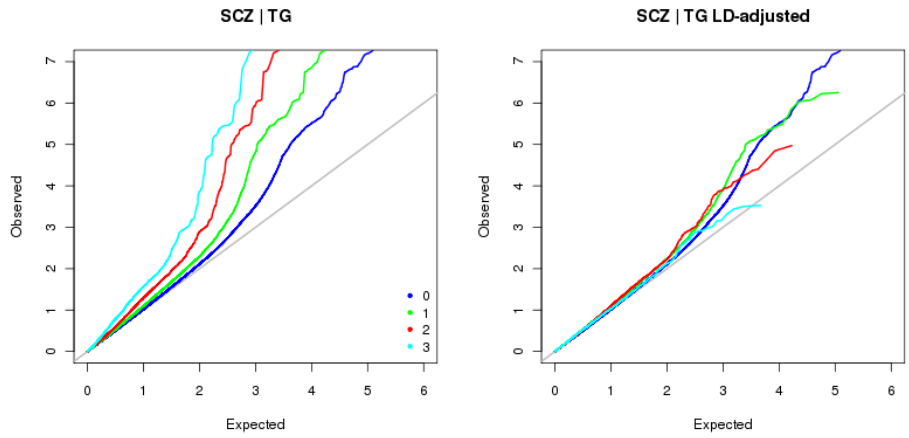
This figure compares the estimates of genetic correlations between metabolic traits from table 3 of [8] to the results obtained using roughly 10x larger sample sizes and LD Score regression in this paper.

4.2 Linkage Disequilibrium May Create False Positive Pleiotropy



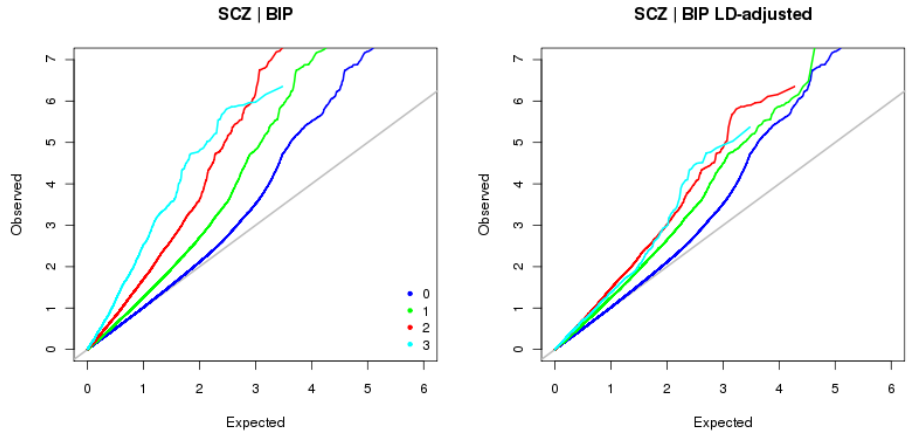
Conditional QQ plot from two completely independent simulated phenotypes with phenotypes generated according to the infinitesimal model ($y = X\beta + \epsilon$ with β and ϵ drawn from normal distributions). The two phenotypes were independent with zero genetic correlation, but the conditional QQ plots show substantial enrichment (*i.e.,* the light blue line is well above the dark blue line), and the correlation between log p-values was 0.14. This effect can be explained by the tendency for p-values to be lower at SNPs with high LD Score noted in [3]. The set of SNPs with p-values for trait 1 below some threshold is enriched for SNPs with high LD Scores, which will tend to have lower p-values for trait 2 as well.

4.3 Pleiotropy Between Triglycerides and Schizophrenia may be Confounded by LD



We reproduced the conditional QQ plot comparing schizophrenia (SCZ) and triglycerides (TG) from [1] (left). We then residualized the TG p-values on LD Score and plotted a new conditional QQ-plot (right). Residualizing on LD Score removed almost all of the enrichment, which is consistent with the near-zero genetic correlation between schizophrenia and TG estimated via LD Score regression.

4.4 Pleiotropy Between Bipolar and Schizophrenia Remains after Correction for LD



As a positive control, we performed the same experiment from the previous figure with schizophrenia (SCZ) and bipolar disorder (BIP), which have a strong positive genetic correlation. In this case, the conditional QQ plot continued to show signal of pleiotropy after residualizing the BIP p-values on LD Score.

5 Supplementary Tables

5.1 Simulations with Parallel LD- and MAF-Dependence

LD Score	h^2(5-50%)	ρ_g(5-50%)	r_g(5-50%)
Truth	0.83	0.42	0.5
HM3	0.53 (0.08)	0.28 (0.07)	0.52 (0.1)
PNG	0.36 (0.08)	0.18 (0.06)	0.5 (0.13)
30 Bins	0.81 (0.12)	0.41 (0.08)	0.51 (0.09)
60 Bins	0.81 (0.12)	0.41 (0.09)	0.51 (0.09)

This table displays simulations with MAF- and LD-dependent genetic architecture where the MAF- and LD- dependence was the same for both phenotypes and genetic correlation did not vary with MAF or LD. Precisely, effect sizes were drawn from a normal distribution so that the magnitude of per-allele effect sizes were uncorrelated with MAF and variants with LD Score below 100 were 4× enriched for heritability.

In all simulations, the sample size was 2062 individuals with full sample overlap between studies; the causal SNPs were best-guess imputed 1000 Genomes SNPs on chromosome 2, and the SNPs retained for the LD Score regression were HapMap 3 SNPs.

Estimates are averages across 100 simulations. Standard deviations (in parentheses) are calculated as the empirical standard deviation across 100 simulations.

LD Scores were estimated using in-sample LD and a 1cM window. HM3 LD Score is $\sum r^2$ with the sum taken over SNPs in HapMap 3. The PNG LD Score is $\sum r^2$ with the sum taken over all SNPs in 1kG as in [3]. The 30 bins LD Score is a per-allele LD Score binned on a MAF by LD Score grid with MAF breaks at 0.05, 0.1, 0.2, 0.3 and 0.4 and LD Score breaks at 35, 75, 150 and 400. The 60 bins LD Score is a per-allele LD Score binned on a MAF by LD Score grid with MAF breaks at 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4 and 0.45 and LD Score breaks at 30, 60, 120, 200 and 300

These simulations demonstrate that naive LD Score regression can give accurate genetic correlation estimates even in situations where the heritability and genetic covariance estimates are badly biased, so long as genetic correlation does not depend on MAF or LD. In addition, these simulations demonstrate that MAF- and LD-binned LD Score regression can give accurate estimates of heritability and genetic covariance even for genetic architectures with MAF- and LD-dependence.

5.2 Simulations with Antiparallel LD- and MAF-Dependence

LD Score	h_1^2(5-50%)	h_2^2(5-50%)	ρ_g(5-50%)	r_g(5-50%)
Truth	0.83	0.89	0.33	0.38
HM3	0.54 (0.09)	1.38 (0.1)	0.4 (0.08)	0.47 (0.08)
PNG	0.36 (0.08)	1.13 (0.08)	0.32 (0.07)	0.5 (0.09)
30 Bins	0.8 (0.14)	0.93 (0.12)	0.34 (0.11)	0.39 (0.1)
60 Bins	0.79 (0.14)	0.93 (0.12)	0.33 (0.11)	0.39 (0.1)

This table displays simulations with MAF- and LD-dependent genetic architecture where the MAF- and LD- dependence was in opposite directions for each phenotype, and genetic correlation did not vary with MAF or LD. Precisely, per-allele effect sizes for the first phenotype were drawn from a normal distribution so that the variance of per-allele effect sizes were uncorrelated with MAF, and variants with LD Score below 100 were 4× enriched for heritability. Per-allele effect sizes for the second phenotype were drawn from a normal distribution so that the variance of per-allele effect size followed $\sqrt{p(1-p)}$, where p is MAF, and variants with LD Score above 100 were 4× enriched for heritability. Otherwise, the parameters of these simulations were the same as in 5.1 These simulations demonstrate that naive LD Score regression can give approximately accurate genetic correlation estimates even in situations where the heritability and genetic covariance estimates are badly biased, so long as genetic correlation does not depend on MAF or LD. In addition, these simulations demonstrate that MAF- and LD-binned LD Score regression can give accurate estimates of heritability and genetic covariance even for genetic architectures with MAF- and LD-dependence.

5.3 Simulations with LD- and MAF-Dependent Genetic Correlation

LD Score	h_1^2(5-50%)	h_2^2(5-50%)	ρ_g(5-50%)	r_g(5-50%)
Truth	0.78	0.72	0.48	0.53
HM3	0.91 (0.1)	0.93 (0.09)	0.46 (0.08)	0.5 (0.06)
PNG	0.77 (0.08)	0.83 (0.08)	0.4 (0.07)	0.5 (0.06)
30 Bins	0.8 (0.14)	0.69 (0.12)	0.4 (0.11)	0.54 (0.09)
60 Bins	0.79 (0.14)	0.68 (0.13)	0.4 (0.11)	0.54 (0.09)

In these simulations, effect sizes for the first phenotype were drawn from a normal distribution with mean zero and variance $(pq)^{0.6}(1 + \log(\ell_j)/50)^2$, effect sizes for the second phenotype were drawn form a normal distribution with mean zero and variance $(pq)^{0.3}(1 - \log(\ell_j)/50)^2$, then noise following a normal distribution with mean zero and variance $1 + (7 - 1)\ell_j/700$, was added to the effect sizes for the second phenotype, so that genetic correlation was roughly 0.35 for low LD SNPs and 0.65 for high LD SNPs.

5.4 Comparison of Standard Error Estimates to Empirical Standard Deviation across Simulations

LD Score	$\widehat{se}(\hat{h}^2)$	$sd(\hat{h}^2)$	$\widehat{se}(\hat{\rho}_g)$	$sd(\hat{\rho}_g)$	$\widehat{se}(\hat{r}_g)$	$sd(\hat{r}_g)$
HM3	0.1 (0.01)	0.08	0.08 (0)	0.07	0.09 (0.02)	0.10
PNG	0.1 (0.01)	0.08	0.08 (0)	0.06	0.09 (0.03)	0.13
30 Bins	0.1 (0.01)	0.12	0.08 (0.01)	0.08	0.09 (0.01)	0.09
60 Bins	0.1 (0.01)	0.12	0.08 (0.01)	0.09	0.09 (0.01)	0.09

This table compares the block jackknife standard errors from ldsc (denoted $\widehat{se}$, which represents the mean standard error estimate across 100 simulation replicates) in the simulations from 5.1 to the empirical standard deviations of the parameter estimates (denoted sd) across 100 simulation replicates. The block jackknife standard errors closely match the empirical standard deviations. This confirms that block jackknife standard error estimates are approximately unbiased even with locally correlated error terms, so long as the block size is sufficiently large.

5.5 Comparison of BMI-Adjusted WHR Genetic Correlations from LDSC to Unadjusted WHR from Vattikuti, et al

Trait 1	Trait 2	Vattikuti	LDSC
BMI	WHR	0.91 (0.18)	-0.06 (0.02)
GLU	WHR	0.09 (0.32)	0.06 (0.03)
HDL	WHR	-0.06 (0.3)	-0.38 (0.03)
TG	WHR	0.32 (0.23)	0.44 (0.02)

This table contrasts the genetic correlation profiles of waist-hip ratio (WHR) and BMI-adjusted WHR. In this paper, we estimated genetic correlations with BMI-adjusted WHR using the summary statistics from [7]. Vattikuti *et. al.* [8] estimated genetic correlations between unadjusted WHR and other metabolic traits using REML. The genetic correlation profiles of these two phenotypes are quite dissimilar, especially with regards to BMI.

References

[1] Ole A Andreassen, Wesley K Thompson, Andrew J Schork, Stephan Ripke, Morten Mattingsdal, John R Kelsoe, Kenneth S Kendler, Michael C O'Donovan, Dan Rujescu, Thomas Werge, et al. Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genetics*, 9(4):e1003455, 2013.

[2] Brendan Bulik-Sullivan. Inferring Genetic Architecture with LD Score Kernel Regression. *In Preparation*, 2014.

[3] Brendan Bulik-Sullivan, Po-Ru Loh, Hilary Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *bioRxiv*, 2014.

[4] Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, 9(3):e1003348, 2013.

[5] Hilary K Finucane and Brendan Bulik-Sullivan. Partitioning heritability with ld score regression. *In preparation*, 2014.

[6] David Golan and Saharon Rosset. Narrowing the gap on heritability of common disease by direct estimation in case-control gwas. *arXiv preprint arXiv:1305.5363*, 2013.

[7] Iris M Heid, Anne U Jackson, Joshua C Randall, Thomas W Winkler, Lu Qi, Valgerdur Steinthorsdottir, Gudmar Thorleifsson, M Carola Zillikens, Elizabeth K Speliotes, Reedik Mägi, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature genetics*, 42(11):949–960, 2010.

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

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