An Atlas of Genetic Correlations across Human Diseases and Traits

Brendan Bulik-Sullivan[†*,1,2,3], Hilary K Finucane[*,4], Verneri Anttila[1,2,3], Alexander Gusev[5,6], Felix R. Day[7], ReproGen Consortium[8], Psychiatric Genomics Consortium[8], Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3[8], John R.B. Perry[7], Nick Patterson[1], Elise Robinson[1,2,3], Mark J Daly[1,2,3], Alkes L Price[**,1,5,6], and Benjamin M Neale[†**,1,2,3]

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[2]Stanley Center for Psychiatric Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[3]Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.
[4]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA.
[5]Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA.
[6]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA.
[7]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK
[8]A list of members and affiliations appears in the Supplementary Note.

Abstract

A recent focus in statistical genetics has been combining genetic association results from multiple phenotypes in order understand the relationships among traits. In this paper, we estimate 300 genetic correlations among 25 traits, totaling more than 1.5 million unique phenotype measurements. To enable this analysis, we introduce a statistical method based on LD Score regression for estimating genetic correlation using only GWAS summary statistics. Our results include a positive genetic correlation between anorexia nervosa and schizophrenia and a negative genetic correlation between anorexia nervosa and body mass index, as well as a large number of replications and positive controls. These results highlight the power of a polygenic modeling framework, since there currently are no genome-wide significant SNPs for anorexia nervosa.

[*]Co-first authors
[**]Co-last authors
[†]Address correspondence to BBS (`bulik@broadinstitute.org`) or BMN (`bneale@broadinstitute.org`).

Introduction

Discovering correlations between phenotypes is a fundamental goal of epidemiology, with applications to classification and treatment of disease, as well as to development of pharmaceutical drugs. One classical strategy in epidemiology is to search for correlations between phenotypes via cross-sectional or longitudinal observational studies; however, the interpretation of results from these studies can be confounded and are vulnerable to reverse causation [1, 2]. An alternative strategy that is effective for heritable traits and is more robust to confounding is to search instead for pairs of phenotypes with shared genetic etiology.

The earliest methods for searching for traits with genetic overlap were twin and family studies [3, 4, 5, 6, 7], which have been applied to a wide spectrum of traits. However, family methods have the disadvantage of requiring measurements of different traits on the same individuals. Genome-wide association studies (GWAS) produce effect-size estimates for specific genetic variants, so it is possible to test for shared genetic etiology using these effect-size estimates, circumventing the requirement to measure multiple traits per individual. This can substantially reduce the cost and difficulty of epidemiological studies for uncommon diseases and traits that are expensive to assay.

One widely-used technique for testing relationships between phenotypes using GWAS data is Mendelian randomization [1, 2], which is the specialization to genetics of instrumental variables [8]. Mendelian randomization has proved effective for traits for which large-effect genetic variants have been identified [9, 10]; however for many complex traits, the heritability is distributed over thousands of variants with small effects [11], in which case Mendelian randomization using genome-wide significant SNPs suffers from low power and weak instrument bias [8].

In this paper, our goal is to estimate genetic correlation, a quantity whose definition includes the effects of all SNPs, including those that do not reach genome-wide significance (Methods). In cases where two phenotypes are suspected of having a cause-effect relationship, genetic correlation can be interpreted as a re-scaling of the instrumental variables estimate obtained from an instrument constructed using all SNPs (Methods); however, unlike instrumental variables, genetic correlation is also meaningful for pairs of diseases, and can be interpreted as a genetic analogue of comorbidity. The two main existing techniques for estimating genetic correlation from GWAS data are restricted maximum likelihood (REML) [12, 13, 14, 15, 16, 17] and polygenic scores [18, 19]. These methods have only been applied to a small number of traits so far, because they require individual genotypes, which are often difficult to obtain due to privacy considerations and informed consent limitations.

Here, we introduce a computationally fast method based on LD Score regression [20] for estimating genetic correlation. This method requires only GWAS summary statistics and is not biased by sample overlap. We apply this method to data from 25 GWAS and report genetic correlations between 300 pairs of phenotypes.

Results

Overview of Methods

Our method for estimating genetic correlation from summary statistics relies on the fact that the GWAS effect-size estimate for a given SNP incorporates the effects of all SNPs in LD with that SNP [21, 20]. For a polygenic trait, SNPs in regions with strong LD will have higher χ^2-statistics on average than SNPs in regions with little LD [20]. A similar relationship holds if we replace χ^2-statistics for a single study with the product of Z-scores from two different studies.

More precisely, under a polygenic model [12, 14], the expected value of $z_{1j}z_{2j}$ is

$$\mathbb{E}[z_{1j}z_{2j}] = \frac{\sqrt{N_1 N_2}\rho_g}{M}\ell_j + \frac{\rho N_s}{\sqrt{N_1 N_2}}, \tag{1}$$

where N_i is the sample size for study i, ρ_g is genetic covariance (defined in Methods), ℓ_j is LD Score [20], N_s is the number of samples shared between study 1 and study 2, and ρ is the phenotypic correlation among the N_s overlapping samples (Supplementary Note). If study 1 and study 2 are the same study, then this reduces to the single-phenotype result from [20], because the genetic covariance between a trait and itself is the same as heritability, and $\chi^2 = z^2$. As a consequence of equation 1, we can estimate genetic covariance using the slope from the regression of $z_{1j}z_{2j}$ on LD Score, which is computationally very fast (see Methods). If we normalize genetic covariance to lie in the interval $[-1, 1]$, we obtain genetic correlation: $r_g := \rho_g/\sqrt{h_1^2 h_2^2}$, where h_i^2 denotes the SNP-heritability from study i [12]. Theory and extensive simulations confirm that our method produces robust estimates of genetic correlation, even when summary statistics are affected by population stratification [20] or sample overlap (Supplementary Note).

Simulations

We performed a series of simulations to evaluate the robustness of the model to potential confounders such as sample overlap and misspecified models of genetic architecture, as well as to determine whether the inference procedure produces appropriate type I error,

Estimates of heritability and genetic covariance can be biased if the underlying model of genetic architecture is misspecified [60]; however, estimates of genetic correlation are more robust to model misspecification biases than estimates of heritability or genetic covariance. Since genetic correlation is estimated as a ratio, model misspecification biases that affect the numerator and the denominator in the same direction will tend to cancel. In situations where MAF- or LD-dependent genetic architectures are a particular concern, it is possible to correct for such biases with LD Score regression using a MAF or LD binning approach (see [32] and Online Methods) similar to that taken by Lee, *et al.* with REML [61].

To quantify the bias introduced by MAF- or LD-dependent genetic architectures, we simulated a variety of different LD Scores and genetic architectures. In realistic scenarios, only a subset of causal SNPs are directly genotyped or successfully imputed, so we used a densely imputed panel of 1000 Genomes (1kG) SNPs [57] in order to generate phenotypes and estimate LD Scores, but computed summary statistics only for the 16% of 1kG SNPs that are also in HapMap3 (HM3) [62] with MAF above 5%. Results from these simulations are displayed in Supplementary Tables S2, S3 and S4.

We found that the partitioned LD Scores were not biased by MAF- and LD-dependent genetic architectures when estimating heritability and genetic covariance, but gave substantially higher standard errors than non-partitioned LD Scores. The estimates of genetic correlation from the simpler non-partitioned LD Scores were approximately unbiased in simulations where both heritability and genetic covariance depended on LD, and only minimally biased in simulations where genetic correlation also depended on LD. In addition, the non-partitioned LD Scores gave substantially lower standard errors than the more complex partitioned LD Score models. Thus for the remainder of this paper, we use non-partitioned LD Scores for estimating r_g, since these LD Scores had minimal bias and the lowest variance in simulations.

In all of the simulations described in this section, there was full sample overlap, which confirms that the LD Score regression with unconstrained intercept is not biased by sample overlap.

Replication of Pyschiatric Cross-Disorder Results

For additional validation, we replicated the estimates of genetic correlations between psychiatric disorders obtained with individual genotypes and REML by the Psychiatric Genomics Consortium (PGC) [15], by applying LD Score regression to summary statistics generated from the same data [22]. Since these summary statistics were generated from non-overlapping samples, we applied LD Score regression using both unconstrained and constrained intercepts (Methods). Results from these analyses are displayed in Figure 1. As expected, the genetic correlation estimates from LD Score regression were similar to the results from REML. LD Score regression with constrained intercept gave standard errors that were only slightly larger than REML, while the standard errors from LD Score regression with intercept were larger, especially for traits with small sample sizes (*e.g.,* ADD, ASD).

Application to Summary Statistics From 25 Phenotypes

We used LD Score regression to estimate genetic correlations among 25 phenotypes (URLs, Methods). Genetic correlation estimates for all 300 pairwise combinations of the 25 traits are displayed in Figure 2. For clarity of presentation, the 25 phenotypes were restricted to contain only one phenotype from each cluster of highly correlated phenotypes (Methods). Highly correlated educational, anthropometric, smoking, and insulin-related phenotypes that were excluded from Figure 2 are displayed in Table S6 and Figures S1, S2 and S3, respectively. References and sample sizes are displayed in Table S1.

For the majority of pairs of traits in Figure 2, no GWAS-based genetic correlation estimate has been reported; however, many associations have been described in an *ad-hoc* manner based off the observation of overlap among genome-wide significant loci. Examples of genetic correlations that are consistent with overlap among top loci include the correlations between plasma lipids and cardiovascular disease [10]; age at onset of menarche and obesity [23]; type 2 diabetes, obesity, fasting glucose, plasma lipids and cardiovascular disease [24]; birth weight, adult height and type 2 diabetes [25]; birth length, adult height and infant head circumference [26, 27]; and childhood obesity and adult obesity [26]. For many of these pairs of traits, we can reject the null hypothesis of zero genetic correlation with overwhelming statistical significance (*e.g.,* $p < 10^{-20}$ for age at onset of menarche and obesity).

The first section of table 1 lists genetic correlation results that are consistent with well-known epidemiological associations, but, as far as we are aware, have not previously been reported using genetic data. Our estimates of the genetic correlation between age at onset of menarche and adult height [28], cardiovascular disease [29] and type 2 diabetes [29, 30] are consistent with the epidemiological associations. Our estimate of a negative genetic correlation between AN and obesity (and a similar genetic correlation with BMI) is intriguing given that the most striking feature of anorexia nervosa is the ability to maintain dangerously low BMI [31] and suggests that similar genetic factors may influence normal variation in BMI as well as severely dysregulated BMI in psychiatric illness. This result is consistent with our observation that BMI GWAS findings appear to primarily implicate neuronal, rather than metabolic, cell-types and epigenetic marks [32]. The negative genetic correlation between adult height and coronary artery disease agrees with a
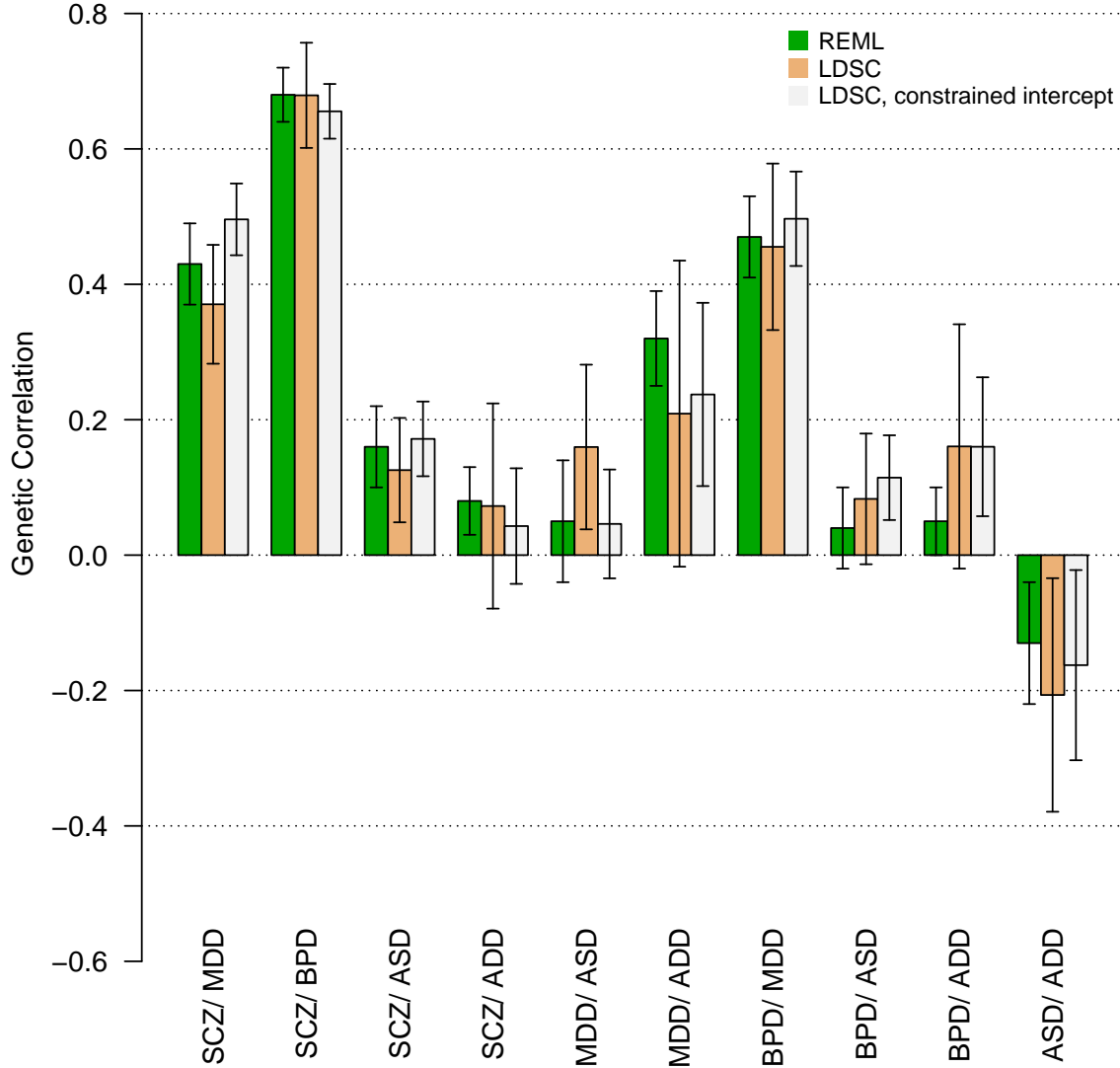
Figure 1: *Replication of PGC Cross-Disorder Results. This plot compares LD Score regression estimates of genetic correlation using the summary statistics from [22] (which were generated from approximately the same data as [15]) to estimates obtained from REML in [15]. The horizontal axis indicates pairs of phenotypes, and the vertical axis indicates genetic correlation. The error bars show standard errors. Colors indicate different estimation procedures. Green is REML, orange is LD Score with intercept and white is LD Score with constrained intercept. The estimates of genetic correlation between psychiatric phenotypes in figure 2 use larger sample sizes; this plot is intended primarily as a technical validation. Abbreviations: ADD = attention deficit hyperactivity disorder; ASD = autism spectrum disorder; BPD = bipolar disorder; MDD = major depressive disorder; SCZ = schizophrenia.*

replicated epidemiological association [33, 34, 35]. We observe several significant associations with the educational attainment phenotypes from Rietveld *et al.* [36]. We estimate a statistically sig-
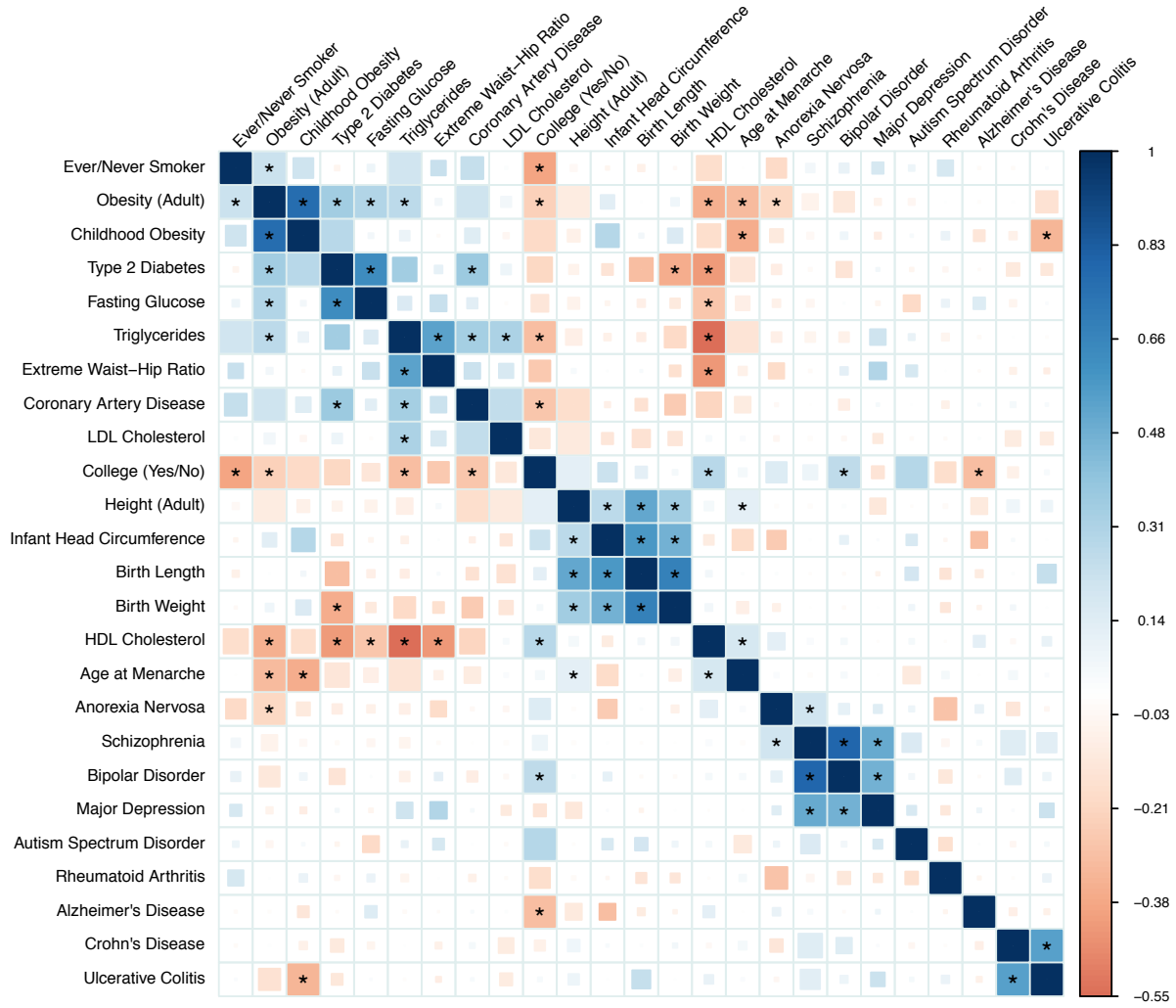
Figure 2: *Genetic Correlations among 25 Published GWAS. Blue corresponds to positive genetic correlations; red corresponds to negative genetic correlation. Larger squares correspond to more significant p-values. Genetic correlations that are different from zero at 1% FDR are displayed as full-sized squares. Genetic correlations that are significantly different from zero at significance level 0.05 after Bonferroni correction for the 300 tests in this figure are given an asterisk. We display results that do not pass multiple testing correction as smaller squares in order to avoid the confusion that would result from whiting our positive controls that do not pass multiple testing correction because of small sample size. This multiple testing correction is conservative, since the tests are not independent.*

nificant negative genetic correlation between college and Alzheimer's disease, consistent with the epidemiological observation that low educational attainment is one of the largest risk factors for

	Phenotype 1	Phenotype 2	r_g (se)	p-value
Epidemiological	Age at Menarche	Height (Adult)	0.11 (0.03)	6×10^{-5} **
	Age at Menarche	Type 2 Diabetes	-0.13 (0.04)	3×10^{-3}
	Age at Menarche	Triglycerides	-0.15 (0.04)	1×10^{-3} *
	Coronary Artery Disease	Age at Menarche	-0.11 (0.05)	4×10^{-2}
	Coronary Artery Disease	College (Yes/No)	-0.278 (0.07)	1×10^{-4} **
	Coronary Artery Disease	Height (Adult)	-0.17 (0.05)	2×10^{-4} *
	Alzheimer's	College (Yes/No)	-0.30 (0.08)	1×10^{-4} **
	Bipolar Disorder	College (Yes/No)	0.026 (0.064)	6×10^{-5} **
	Obesity (Adult)	College (Yes/No)	-0.23 (0.04)	2×10^{-8} **
	Triglycerides	College (Yes/No)	-0.30 (0.04)	5×10^{-12} **
	Anorexia Nervosa	Obesity (Adult)	-0.20 (0.04)	4×10^{-6} **
	Ever/Never Smoker	College (Yes/No)	-0.39 (0.07)	1×10^{-9} **
	Ever/Never Smoker	Obesity (Adult)	0.22 (0.05)	7×10^{-5} **
Novel/Nonzero	Autism Spectrum Disorder	College (Yes/No)	0.28 (0.08)	5×10^{-4} *
	Ulcerative Colitis	Childhood Obesity	-0.33 (0.08)	3.9×10^{-5} **
	Anorexia Nervosa	Schizophrenia	0.19 (0.04)	1.5×10^{-5} **
Novel/Zero	Schizophrenia	Alzheimer's	0.05 (0.05)	0.58
	Schizophrenia	Ever/Never Smoker	0.03 (0.06)	0.26
	Schizophrenia	Triglycerides	-0.05 (0.04)	0.21
	Schizophrenia	LDL Cholesterol	-0.02 (0.03)	0.64
	Schizophrenia	HDL Cholesterol	0.03 (0.04)	0.50
	Schizophrenia	Rheumatoid Arthritis	-0.05 (0.05)	0.38
	Crohn's Disease	Rheumatoid Arthritis	-0.02 (0.09)	0.83
	Ulcerative Colitis	Rheumatoid Arthritis	-0.09 (0.09)	0.33

Table 1: *Genetic correlation estimates, standard errors and p-values for selected pairs of traits. Results are grouped into genetic correlations that are novel genetic results, but are consistent with established epidemiological associations ("Epidemiological"), novel genetic correlations ("Novel/Nonzero") and interesting null results ("Novel/Zero"). The p-values displayed are uncorrected p-values. Results that pass multiple testing correction for the 300 tests in Figure 2 at 1% FDR are given a single asterisk; results that pass Bonferroni correction at significance level 0.05 are given two asterisks. For completeness, we present some genetic correlations that are directionally consistent with epidemiological associations but that do not pass multiple testing correction in our data.*

Alzheimer's [37, 38]. The positive genetic correlation between college and bipolar disorder is consistent with psychiatric literature showing that educational attainment and bipolar disorder status are positively correlated [39, 40], Our estimate of a negative genetic correlation between smoking and college is consistent with the epidemiological result that smoking is more prevalent among less-educated groups [41]. Finally, our estimates of genetic correlations among metabolic traits are consistent with the estimates obtained using REML in Vattikuti *et al.* [16] (Supplementary Table S4), and are directionally consistent with the recent Mendelian randomization results from Wuertz *et al.* [54]. In addition, our estimate of 0.57 (0.074) for the genetic correlation between Crohn's disease and ulcerative colitis is consistent with the estimate of 0.62 (0.042) from Chen *et al.* [17].

The second section of table 1 lists results that are, to the best of our knowledge, new both to genetics and epidemiology. First, we find a positive genetic correlation between anorexia nervosa

and schizophrenia. Comorbidity between eating and psychotic disorders has not been thoroughly investigated in the psychiatric literature [42, 43], and our results raise the intriguing possibility of fundamental similarity between these classes of disease. Second, we estimate a negative genetic correlation between ulcerative colitis and childhood obesity. The relationship between premorbid BMI and ulcerative colitis is at present not well-understood, and our results suggest that exploring this relationship may be a fruitful direction for further investigation. Third, we estimate a positive genetic correlation between autism spectrum disorder (ASD) and educational attainment, which itself has very high genetic correlation with IQ [44, 45, 36]. The ASD summary statistics were generated using a case-pseudocontrol study design, so this result cannot be explained by diagnostic biases, such as the tendency for the parents of children who receive a diagnosis of ASD to be better educated than the general population [46]. The distribution of IQ among individuals with ASD has lower mean than the general population, but with heavy tails [47] (*i.e.,* an excess of individuals with extremely low and extremely high IQ). In addition, there is evidence that the genetic architectures of high IQ and low IQ ASD are dissimilar [48]. We are unable to offer an explanation for this result, but propose that further exploration of this genetic correlation is an interesting direction for future research.

The third section of table 1 lists several instances where the genetic correlation is close to zero with small standard error, in contrast to previous reports of epidemiological correlation or pleiotropy. We estimate genetic correlations close to zero between schizophrenia and rheumatoid arthritis, schizophrenia and smoking, and schizophrenia and plasma lipids. The lack of genetic correlation between schizophrenia and rheumatoid arthritis is interesting because schizophrenia has been observed to be protective for rheumatoid arthritis [49]. The absence of genetic correlation between schizophrenia and smoking is notable because of the high prevalence of smoking among individuals with schizophrenia [50]. Our estimate of zero genetic correlation between schizophrenia and plasma lipid levels or cardiovascular disease contrasts with earlier reports of extensive pleiotropy between schizophrenia and triglycerides [51]. However, the observation from Andreassen, *et al.* [51] could potentially be explained the sensitivity of the method used in the original report to the properties of a few regions with high linkage disequilibrium, rather than trait biology (Table S5). We estimate near-zero genetic correlation between Alzheimer's disease and schizophrenia. The genetic correlations between Alzheimers disease and the other four psychiatric traits (anorexia nervosa, bipolar, major depression, ASD) are also close to zero, but with larger standard errors, due to smaller sample sizes. This suggests that the genetic basis of Alzheimer's disease is distinct from psychiatric conditions. Last, we estimate near zero genetic correlation between rheumatoid arthritis and both Crohn's disease and ulcerative colitis. Although these immune diseases share a large number of associated loci [52, 53], it is as often the case that an allele that is risk-increasing for one immune disease is protective for another as an allele is risk-increasing for two [52], consistent near-zero genetic correlation overall. This is an example of pleiotropy without genetic correlation (Methods).

Discussion

We have described a new method for estimating genetic correlation from GWAS summary statistics. We applied our method to a large dataset of publicly available GWAS summary statistics, spanning 25 traits and more than 1.5 million phenotype-genotype pairs. We replicated many previously-reported GWAS-based genetic correlations, and confirm observations of overlap among genome-

wide significant SNPs, Mendelian randomization results and known epidemiological associations, thus validating the utility of genetic correlation as an epidemiological tool. In addition, we report several novel results that merit further analysis, including a positive genetic correlation between educational attainment and autism spectrum disorder and a positive genetic correlation between anorexia nervosa and schizophrenia.

This method is an advance for several reasons: it does not require individual genotypes, genome-wide significant SNPs or measuring all traits on all individuals; it is not biased by sample overlap; and it is computationally trivial. Although it may be possible to extend existing methods such as LD-pruned polygenic scores to work on pairs of summary statistics, such methods would be biased by sample overlap, and would suffer from the disadvantage that LD-pruning results in loss of information. These advantages allow estimation of genetic correlations between a much large set of pairs of phenotypes than was previously possible.

In general, the difficulties in interpreting genetic correlation are similar to the difficulties in interpretation of Mendelian randomization. Although genetic correlation is immune to environmental confounding, genetic correlation is still subject to genetic confounding, which is similar to the idea of confounding by pleiotropy from the Mendelian randomization literature (Methods). This is best illustrated with an example. The negative genetic correlation between HDL and CAD in Figure 2 could result from a direct causal effect HDL $\rightarrow$ CAD, but could also result from non-causal shared genetic etiology, perhaps mediated by the genetic correlation of -0.55 between HDL and triglycerides (TG), which is the explanation offered in [10, 55]. This scenario can be represented graphically as HDL $\leftarrow$ G $\rightarrow$ TG $\rightarrow$ CAD, where G is the set of genetic variants with effects on both HDL and TG. This diagram would result in a genetic correlation between HDL and CAD, because there is an unblocked path from HDL to CAD [56]. Extending genetic correlation to deal with multiple phenotypes that share causal loci (*e.g.,* genetic correlation between HDL and CAD controlling for TG, analogous to [10, 55]) is an important direction for future work.

We note several limitations of the LD Score regression estimator of genetic correlation. First, LD Score regression requires larger sample sizes than methods such as REML that use individual genotypes in order to give estimates with acceptable standard error. Second, LD Score regression is only currently applicable to studies that sample individuals from non-admixed populations for which a sequenced reference panel (such as 1000 Genomes [57]) is available [20]. Third, LD Score regression performs optimally when applied to traits with highly polygenic genetic architectures, such as psychiatric traits. At very large sample size and for less-polygenic traits, analyzing only the significantly associated SNPs can sometimes be a more powerful strategy. Developing methods that can make optimal use of both confidently associated large-effect SNPs and diffuse polygenic signal from thousands of small-effect SNPs is a direction for future research. Fifth, although LD Score regression is not biased by oversampling of cases in case-control studies, the results may nonetheless be biased by more subtle biases in ascertainment, such as misclassification of cases [15].

Despite these limitations, we believe that the LD Score regression estimator of genetic correlation will be a useful addition to the epidemiological toolbox, since it allows for rapid screening for correlations among a diverse set of traits, without the need for measuring multiple traits on the same individuals or genome-wide significant SNPs.

Methods

Definition of Genetic Covariance and Correlation

Let S denote a set of M SNPs, let X denote the random M-vector of additively (0-1-2) coded genotypes for the SNPs in S, and let y_1 and y_2 denote phenotypes $\beta := \mathrm{argmax}_{\alpha \in \mathbb{R}^M} \mathrm{Cor}\left[y, X\alpha\right]^2$, where the maximization is performed in the population (*i.e.*, in the infinite data limit). This is a projection, so uniqueness of β is guaranteed as we remove SNPs that are linearly dependent (in the population). Then h_S^2, the heritability accounted for by SNPs in S, is defined $h_S^2(y_1) := \sum_j \beta_j^2$, and the genetic covariance among SNPs in S is defined $\rho_S(y_1, y_2) := \sum_{j \in S} \beta_j \gamma_j$. It is more common to instead report genetic correlation, $r_S(y_1, y_2) := \rho_S(y_1, y_2)/\sqrt{h_S^2(y_1) h_S^2(y_2)}$, which lies in the interval [-1,1]. Following [12], we use subscript g (as in h_g^2, ρ_g, r_g) when the set of SNPs is genotyped and imputed SNPs in GWAS.

SNP genetic correlation (*i.e.*, S equals genotyped SNPs) is different from family study genetic correlation. In a family study, the relationship matrix captures information about all genetic variation, not just common SNPs, so family studies estimate the total narrow-sense genetic correlation (*i.e.*, S equals all variants). Unlike the relationship between SNP-heritability [12] and total narrow-sense heritability, for which we have the inequality $h_g^2 \leq h^2$, no similar relationship holds between SNP genetic correlation and total narrow-sense genetic correlation. For example, if β and γ are strongly correlated among common variants, but only weakly correlated among rare variants, then the total narrow-sense genetic correlation will be less than the SNP genetic correlation. Choice of the set of SNPs S is discussed in the Supplementary Note.

Genetic correlation is (asymptotically) proportional to Mendelian randomization estimates. Suppose we wish to estimate the effect b of y_1 on y_2 using two-stage least squares (2SLS) [8] and a genetic instrument $g_i := \sum_{j \in S} X_{ij} \beta_j$, where i indexes individuals and S is a set of SNPs. The 2SLS estimate is $\hat{b}_{2SLS} := g^{\mathsf{T}} y_2 / g^{\mathsf{T}} y_1$, and $\mathbb{E}[g^{\mathsf{T}} y_2] = \rho_S(y_1, y_2)$ and $\mathbb{E}[g^{\mathsf{T}} y_1] = h_S^2(y_1)$. Thus, $\lim_{N \to \infty} \hat{b}_{2SLS} = r_S(y_2, y_1)\sqrt{h_S^2(y_1)/h_S^2(y_2)}$. If S is the set of all SNPs, then this instrument has the (possibly undesirable) property of being symmetric in y_1 and y_2. This is discussed further in the Supplementary Note.

Genetic correlation is different from pleiotropy. Two traits have a pleiotropic relationship if many genetic variants have nonzero effects on both traits. Genetic correlation is a stronger condition than pleiotropy: to exhibit genetic correlation, it is not sufficient for two phenotypes to be influenced by overlapping sets of genetic variants, the directions of effects must also be consistently aligned across the genome. Both quantities are informative: if two phenotypes are influenced by variants at the same loci or in the same pathways, this may indicate important shared biology, even if the direction of the effects do not align.

Cross-Trait LD Score Regression

We estimate genetic covariance (technically $\rho_{5\text{-}50\%}$, see Supplementary Note) by regressing $z_{1j} z_{2j}$ against $\ell_j \sqrt{N_{1j} N_{2j}}$, (where N_{ij} is the sample size for SNP j in study i) then multiplying the resulting slope by M, the number of SNPs in the reference panel with MAF between 5% and 50%.

If we know ahead of time that there is no sample overlap, we can constrain the intercept of this regression to zero (that is, regression through the origin), which gives lower standard error. This can be accomplished with the `--no-intercept` flag in `ldsc`.

Regression Weights

For heritability estimation, we use the LD Score regression weights derived in the Supplementary Note from [20]. If effect sizes for both phenotypes are drawn from a bivariate normal distribution, then the optimal regression weights for genetic covariance estimation are

$$\mathrm{Var}[z_{1j}z_{2j} \,|\, \ell_j] \propto \left(\frac{N_1 h_1^2 \ell_j}{M} + 1 \right) \left(\frac{N_2 h_2^2 \ell_j}{M} + 1 \right) + 2 \left(\frac{\sqrt{N_1 N_2} \rho_g}{M} \ell_j + \frac{\rho N_s}{\sqrt{N_1 N_2}} \right)^2 \qquad (2)$$

(Supplementary Note). This quantity depends on both heritabilities, the genetic covariance and the number of overlapping samples, which are not known a priori, so it is necessary to estimate them from the data. Linear regression with weights estimated from the data is called feasible generalized least squares (FGLS) [8]. We run a first LD Score regression weighted with weights computed using the heritability estimates from the single-phenotype LD Score regressions, N_s set to zero, and ρ_g estimated with the aggregate estimator $\hat{\rho}_{g,agg} := (\bar{\ell}\sqrt{N_1 N_2})^{-1} \sum_j z_{1j}z_{2j}$, where $\bar{\ell}$ denotes the mean LD Score among SNPs included in the regression. We then update the weights using the estimates of ρ_g and ρN_s obtained from the first LD Score regression and use the updated weights for the second and final LD Score regression. In addition, we weight by $1/\ell_j$ (where ℓ_j here denotes LD Score with the sum taken over SNPs included in the regression) in order to downweight SNPs that are overcounted. This is a heuristic: the optimal approach would be to weight by the inverse covariance matrix of χ^2-statistics, but this matrix is difficult to compute.

Assessment of Statistical Significance via Block Jackknife

Under a polygenic model, the χ^2-statistics and Z-score products for SNPs in LD are correlated, so the usual estimator of the regression standard error, which assumes that all data points are independent, will be biased downwards. Therefore, we use a block jackknife over blocks of adjacent SNPs in order to obtain a correlation-robust standard error. The default setting in `ldsc` is 200 blocks per genome, which can be adjusted with the `--num-blocks` flag. This is the same procedure used in [20], and gives accurate standard errors in simulations (Table S5). We also obtain a standard error for the genetic correlation by using a ratio block jackknife over SNPs.

Computational Complexity

If N denotes sample size and M denotes the number of SNPs, then LD Score regression takes $\mathcal{O}(MN)$ time for computing summary statistics and $\mathcal{O}(M)$ time for the regression. Estimating LD Scores takes $\mathcal{O}(MN)$ time, though N for estimating LD Scores does not need to as large as N for GWAS; for instance, we use $N = 378$ Europeans from 1000 Genomes for estimating LD Score. In addition, LD Scores only need to be computed once, and can then be applied to many pairs of phenotypes. For comparison, REML takes time $\mathcal{O}(MN^2)$ for computing the genetic relatedness matrix (GRM) and $\mathcal{O}(N^3)$ time for maximizing the likelihood.

Practically, estimating LD Scores takes about an hour parallelized over chromosomes, and LD Score regression takes about 15 seconds for each pair of phenotypes on a standard 2014 MacBook Air with 1.7 GhZ Intel Core i7 processor.

Summary Statistic Datasets

We selected traits for inclusion in the main text via the following procedure:

1. Begin with all publicly available non-sex-stratified European-only summary statistics that provide signed summary statistics, are imputed to at least HapMap 2 and do not include heritable covariates.

2. Remove all traits with heritability Z-score below 4, based on the observation that genetic correlations for traits with heritability Z-score below 4 are generally too noisy to interpret.

3. Prune sets of highly correlated phenotypes (*e.g.,* obesity classes 1-3, or the several glucose and insulin measures from the MAGIC consortium) by picking the trait from each cluster with the highest heritability heritability Z-score (traits pruned in this step are displayed in the supplementary figures).

We then applied the following filters (encoded in the script `sumstats_to_chisq.py` included with our `ldsc` software package) to the summary statistics:

1. For studies that provide a measure of imputation quality (*e.g.,* INFO score), filter to INFO greater than 0.9.

2. For studies that provide sample MAF, filter to sample MAF above 1%.

3. In order to restrict to well-imputed SNPs in studies that do not provide a measure of imputation quality, filter to HapMap3 [62] SNPs with HapMap3 MAF above 1%, which tend to be well-imputed in most studies. This step should be skipped if INFO scores are available for all studies.

4. If sample size varies from SNP to SNP, remove SNPs with effective sample size less than two thirds of the the 90th percentile of sample size.

5. Remove indels and structural variants.

6. Remove strand-ambiguous SNPs.

7. Remove SNPs whose alleles do not match the alleles reported in 1000 Genomes.

8. Because the presence of outliers can increase the standard error of the LD Score regression estimates, we also removed SNPs with extremely large effect sizes ($\chi^2 > 80$, as in [20]).

If the summary statistics have been GC corrected at any point in the data-generating process, then the heritability and genetic covariance estimates will be biased downwards. The bias in the numerator and denominator of genetic correlation will precisely cancel, so genetic correlation estimates are not affected by GC correction. The majority of the datasets analyzed in this paper were GC corrected at least once, which is why we do not report genetic covariance and heritability.

Data on Alzheimer's disease were obtained from the following source:

International Genomics of Alzheimer's Project (IGAP) is a large two-stage study based upon genome-wide association studies (GWAS) on individuals of European ancestry. In stage 1, IGAP used genotyped and imputed data on 7,055,881 single nucleotide polymorphisms (SNPs) to meta-analyze four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The European Alzheimer's Disease Initiative, EADI; the Alzheimer Disease Genetics Consortium, ADGC; The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium, CHARGE; The Genetic and Environmental Risk in AD consortium, GERAD). In stage 2, 11,632 SNPs were genotyped and tested for association in an independent set of 8,572 Alzheimer's disease cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1 and 2.

We only used stage 1 data for LD Score regression.

URLs

1. `ldsc` software:
   `github.com/bulik/ldsc`

2. This paper:
   `github.com/bulik/gencor_tex`

3. PGC (psychiatric) summary statistics:
   `www.med.unc.edu/pgc/downloads`

4. GIANT (anthopometric) summary statistics:
   `www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files`

5. EGG (Early Growth Genetics) summary statistics:
   `www.egg-consortium.org/`

6. MAGIC (insulin, glucose) summary statistics:
   `www.magicinvestigators.org/downloads/`

7. CARDIoGRAM (coronary artery disease) summary statistics:
   `www.cardiogramplusc4d.org`

8. DIAGRAM (T2D) summary statistics:
   `www.diagram-consortium.org`

9. Rheumatoid Arthritis summary statistics:
   `www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl_etal_2010NG/`

10. IGAP (Alzheimers) summary statistics:
    `www.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php`

11. IIBDGC (inflammatory bowel disease) summary statistics:
    `www.ibdgenetics.org/downloads.html`
    We used a newer version of these data with 1000 Genomes imputation.

12. Plasma Lipid summary statistics:
    `www.broadinstitute.org/mpg/pubs/lipids2010/`

13. Educational Attainment summary statistics:
    `www.ssgac.org/`

14. Beans:
    `www.barismo.com`
    `www.bluebottlecoffee.com`

Acknowledgements

Author Contributions

MJD provided reagents BMN and ALP provided reagents CL, ER, VA, JP and FD aided in the interpretation of results. JP and FD provided data. The caffeine molecule is responsible for all that is good about this manuscript. BBS and HKF are responsible for the rest. All authors revised and approved the final manuscript.

Competing Financial Interests

Unfortunately, we have no financial conflicts of interest to declare.

1 References

[1] George Davey Smith and Shah Ebrahim. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 32(1):1–22, 2003.

[2] George Davey Smith and Gibran Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics*, 23(R1):R89–R98, 2014.

[3] SG Vandenberg. Multivariate analysis of twin differences. *Methods and goals in human behavior genetics*, pages 29–43, 1965.

[4] Oscar Kempthorne and Richard H Osborne. The interpretation of twin data. *American journal of human genetics*, 13(3):320, 1961.

[5] John C Loehlin and Steven Gerritjan Vandenberg. *Genetic and environmental components in the covariation of cognitive abilities: An additive model*. Louisville Twin Study, University of Louisville, 1966.

[6] Michael Neale and Lon Cardon. *Methodology for genetic studies of twins and families*. Number 67. Springer, 1992.

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

[8] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

[9] Benjamin F Voight, Gina M Peloso, Marju Orho-Melander, Ruth Frikke-Schmidt, Maja Barbalic, Majken K Jensen, George Hindy, Hilma Hólm, Eric L Ding, Toby Johnson, et al. Plasma hdl cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet*, 380(9841):572–580, 2012.

[10] Ron Do, Cristen J Willer, Ellen M Schmidt, Sebanti Sengupta, Chi Gao, Gina M Peloso, Stefan Gustafsson, Stavroula Kanoni, Andrea Ganna, Jin Chen, et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nature genetics*, 45(11):1345–1352, 2013.

[11] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[12] 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.

[13] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[14] Sang Hong Lee, Jian Yang, Michael E Goddard, Peter M Visscher, and Naomi R Wray. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19):2540–2542, 2012.

[15] Cross-Disorder Group of the Psychiatric Genomics Consortium et al. Genetic relationship between five psychiatric disorders estimated from genome-wide snps. *Nature Genetics*, 2013.

[16] 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.

[17] Guo-Bo Chen, Sang Hong Lee, Marie-Jo A Brion, Grant W Montgomery, Naomi R Wray, Graham L Radford-Smith, Peter M Visscher, et al. Estimation and partitioning of (co) heritability of inflammatory bowel disease from gwas and immunochip data. *Human molecular genetics*, page ddu174, 2014.

[18] Shaun M Purcell, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan, Pamela Sklar, Shaun M Purcell, Jennifer L Stone, Patrick F Sullivan, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, 2009.

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

[20] 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.

[21] Jian Yang, Michael N Weedon, Shaun Purcell, Guillaume Lettre, Karol Estrada, Cristen J Willer, Albert V Smith, Erik Ingelsson, Jeffrey R O'Connell, Massimo Mangino, et al. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19(7):807–812, 2011.

[22] Cross-Disorder Group of the Psychiatric Genomics Consortium et al. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*, 381(9875):1371, 2013.

[23] John RB Perry, Felix Day, Cathy E Elks, Patrick Sulem, Deborah J Thompson, Teresa Ferreira, Chunyan He, Daniel I Chasman, Tõnu Esko, Gudmar Thorleifsson, et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature*, 514(7520):92–97, 2014.

[24] Andrew P Morris, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, Ayellet V Segre, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, 44(9):981, 2012.

[25] Momoko Horikoshi, Hanieh Yaghootkar, Dennis O Mook-Kanamori, Ulla Sovio, H Rob Taal, Branwen J Hennig, Jonathan P Bradfield, Beate St Pourcain, David M Evans, Pimphen

Charoen, et al. New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nature genetics*, 45(1):76–82, 2013.

[26] Early Growth Genetics (EGG) Consortium et al. A genome-wide association meta-analysis identifies new childhood obesity loci. *Nature genetics*, 44(5):526–531, 2012.

[27] H Rob Taal, Beate St Pourcain, Elisabeth Thiering, Shikta Das, Dennis O Mook-Kanamori, Nicole M Warrington, Marika Kaakinen, Eskil Kreiner-Møller, Jonathan P Bradfield, Rachel M Freathy, et al. Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nature genetics*, 44(5):532–538, 2012.

[28] NC Onland-Moret, PHM Peeters, CH Van Gils, F Clavel-Chapelon, T Key, A Tjønneland, A Trichopoulou, R Kaaks, Jonas Manjer, S Panico, et al. Age at menarche in relation to adult height the epic study. *American journal of epidemiology*, 162(7):623–632, 2005.

[29] Felix Day et al. Puberty timing associated with diabetes, cardiovascular disease and also diverse health outcomes in men and women: the uk biobank study. *Submitted*, 2014.

[30] Cathy E Elks, Ken K Ong, Robert A Scott, Yvonne T van der Schouw, Judith S Brand, Petra A Wark, Pilar Amiano, Beverley Balkau, Aurelio Barricarte, Heiner Boeing, et al. Age at menarche and type 2 diabetes risk the epic-interact study. *Diabetes care*, 36(11):3526–3534, 2013.

[31] American Psychiatric Association et al. *DSM 5*. American Psychiatric Association, 2013.

[32] Hilary K. Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, Stephan Ripke, Felix R. Day, The ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R. B. Perry, Yukinori Okada, Brad Bernstein, Soumya Raychaudhuri, Mark Daly, Nick Patterson, Benjamin M. Neale, and Alkes L. Price. Polygenic effects of cell-type-specific functional elements in 17 traits and 1.3 million phenotyped samples. *In preparation*, 2014.

[33] Na Wang, Xianglan Zhang, Yong-Bing Xiang, Gong Yang, Hong-Lan Li, Jing Gao, Hui Cai, Yu-Tang Gao, Wei Zheng, and Xiao-Ou Shu. Associations of adult height and its components with mortality: a report from cohort studies of 135 000 chinese women and men. *International journal of epidemiology*, 40(6):1715–1726, 2011.

[34] Patricia R Hebert, Janet W Rich-Edwards, JE Manson, Paul M Ridker, Nancy R Cook, Gerald T O'Connor, Julie E Buring, and Charles H Hennekens. Height and incidence of cardiovascular disease in male physicians. *Circulation*, 88(4):1437–1443, 1993.

[35] Janet W Rich-Edwards, JoAnn E Manson, Meir J Stampfer, Graham A Colditz, Walter C Willett, Bernard Rosner, Frank E Speizer, and Charles H Hennekens. Height and the risk of cardiovascular disease in women. *American journal of epidemiology*, 142(9):909–917, 1995.

[36] Cornelius A Rietveld, Sarah E Medland, Jaime Derringer, Jian Yang, Tõnu Esko, Nicolas W Martin, Harm-Jan Westra, Konstantin Shakhbazov, Abdel Abdellaoui, Arpana Agrawal, et al. Gwas of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, 340(6139):1467–1471, 2013.

[37] Deborah E Barnes and Kristine Yaffe. The projected effect of risk factor reduction on alzheimer's disease prevalence. *The Lancet Neurology*, 10(9):819–828, 2011.

[38] Sam Norton, Fiona E Matthews, Deborah E Barnes, Kristine Yaffe, and Carol Brayne. Potential for primary prevention of alzheimer's disease: an analysis of population-based data. *The Lancet Neurology*, 13(8):788–794, 2014.

[39] James H MacCabe, Mats P Lambe, Sven Cnattingius, Pak C Sham, Anthony S David, Abraham Reichenberg, Robin M Murray, and Christina M Hultman. Excellent school performance at age 16 and risk of adult bipolar disorder: national cohort study. *The British Journal of Psychiatry*, 196(2):109–115, 2010.

[40] Jari Tiihonen, Jari Haukka, Markus Henriksson, Mary Cannon, Tuula Kieseppä, Ilmo Laaksonen, Juhani Sinivuo, and Jouko Lönnqvist. Premorbid intellectual functioning in bipolar disorder and schizophrenia: results from a cohort study of male conscripts. *American Journal of Psychiatry*, 162(10):1904–1910, 2005.

[41] John P Pierce, Michael C Fiore, Thomas E Novotny, Evridiki J Hatziandreu, and Ronald M Davis. Trends in cigarette smoking in the united states: educational differences are increasing. *Jama*, 261(1):56–60, 1989.

[42] Ruth H Striegel-Moore, Vicki Garvin, Faith-Anne Dohm, and Robert A Rosenheck. Psychiatric comorbidity of eating disorders in men: a national study of hospitalized veterans. *International Journal of Eating Disorders*, 25(4):399–404, 1999.

[43] Barton J Blinder, Edward J Cumella, and Visant A Sanathara. Psychiatric comorbidities of female inpatients with eating disorders. *Psychosomatic Medicine*, 68(3):454–462, 2006.

[44] Ian J Deary, Steve Strand, Pauline Smith, and Cres Fernandes. Intelligence and educational achievement. *Intelligence*, 35(1):13–21, 2007.

[45] Catherine M Calvin, Cres Fernandes, Pauline Smith, Peter M Visscher, and Ian J Deary. Sex, intelligence and educational achievement in a national cohort of over 175,000 11-year-old schoolchildren in england. *Intelligence*, 38(4):424–432, 2010.

[46] Maureen S Durkin, Matthew J Maenner, F John Meaney, Susan E Levy, Carolyn DiGuiseppi, Joyce S Nicholas, Russell S Kirby, Jennifer A Pinto-Martin, and Laura A Schieve. Socioeconomic inequality in the prevalence of autism spectrum disorder: evidence from a us cross-sectional study. *PLoS One*, 5(7):e11551, 2010.

[47] Elise B Robinson, Kaitlin E Samocha, Jack A Kosmicki, Lauren McGrath, Benjamin M Neale, Roy H Perlis, and Mark J Daly. Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *Proceedings of the National Academy of Sciences*, 111(42):15161–15165, 2014.

[48] Kaitlin E Samocha, Elise B Robinson, Stephan J Sanders, Christine Stevens, Aniko Sabo, Lauren M McGrath, Jack A Kosmicki, Karola Rehnström, Swapan Mallick, Andrew Kirby, et al. A framework for the interpretation of de novo mutation in human disease. *Nature genetics*, 46(9):944–950, 2014.

[49] Alan J Silman and Jacqueline E Pearson. Epidemiology and genetics of rheumatoid arthritis. *Arthritis Res*, 4(Suppl 3):S265–S272, 2002.

[50] Jose de Leon and Francisco J Diaz. A meta-analysis of worldwide studies demonstrates an association between schizophrenia and tobacco smoking behaviors. *Schizophrenia research*, 76(2):135–157, 2005.

[51] Ole A Andreassen, Srdjan Djurovic, Wesley K Thompson, Andrew J Schork, Kenneth S Kendler, Michael C O?Donovan, Dan Rujescu, Thomas Werge, Martijn van de Bunt, Andrew P Morris, et al. Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *The American Journal of Human Genetics*, 92(2):197–209, 2013.

[52] Chris Cotsapas, Benjamin F Voight, Elizabeth Rossin, Kasper Lage, Benjamin M Neale, Chris Wallace, Gonçalo R Abecasis, Jeffrey C Barrett, Timothy Behrens, Judy Cho, et al. Pervasive sharing of genetic effects in autoimmune disease. *PLoS genetics*, 7(8):e1002254, 2011.

[53] Kyle Kai-How Farh, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J Housley, Samantha Beik, Noam Shoresh, Holly Whitton, Russell JH Ryan, Alexander A Shishkin, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 2014.

[54] Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change, author=Wurtz, Peter and others, journal=PLoS Medicine, year=2014, publisher=Public Library of Science.

[55] Stephen Burgess, Daniel F Freitag, Hassan Khan, Donal N Gorman, and Simon G Thompson. Using multivariable mendelian randomization to disentangle the causal effects of lipid fractions. *PLoS one*, 9(10):e108891, 2014.

[56] Sander Greenland, Judea Pearl, and James M Robins. Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48, 1999.

[57] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[58] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

[59] Andy Dahl, Victoria Hore, Valentina Iotchkova, and Jonathan Marchini. Network inference in matrix-variate gaussian models with non-independent noise. *arXiv preprint arXiv:1312.1622*, 2013.

[60] Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics*, 91(6):1011–1021, 2012.

[61] S Hong Lee, Jian Yang, Guo-Bo Chen, Stephan Ripke, Eli A Stahl, Christina M Hultman, Pamela Sklar, Peter M Visscher, Patrick F Sullivan, Michael E Goddard, et al. Estimation of snp heritability from dense genotype data. *American journal of human genetics*, 93(6):1151, 2013.

[62] International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

[63] Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014.

[64] Pamela Sklar, Stephan Ripke, Laura J Scott, Ole A Andreassen, Sven Cichon, Nick Craddock, Howard J Edenberg, John I Nurnberger, Marcella Rietschel, Douglas Blackwood, et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near odz4. *Nature genetics*, 43(10):977, 2011.

[65] Stephan Ripke, Naomi R Wray, Cathryn M Lewis, Steven P Hamilton, Myrna M Weissman, Gerome Breen, Enda M Byrne, Douglas HR Blackwood, Dorret I Boomsma, Sven Cichon, et al. A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular psychiatry*, 18(4):497–511, 2012.

[66] Vesna Boraska, Christopher S Franklin, James AB Floyd, Laura M Thornton, Laura M Huckins, Lorraine Southam, N William Rayner, Ioanna Tachmazidou, Kelly L Klump, Janet Treasure, et al. A genome-wide association study of anorexia nervosa. *Molecular psychiatry*, 2014.

[67] Tobacco, Genetics Consortium, et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature genetics*, 42(5):441–447, 2010.

[68] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 2013.

[69] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I Berndt, Michael N Weedon, Fernando Rivadeneira, Cristen J Willer, Anne U Jackson, Sailaja Vedantam, Soumya Raychaudhuri, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

[70] Sonja I Berndt, Stefan Gustafsson, Reedik Mägi, Andrea Ganna, Eleanor Wheeler, Mary F Feitosa, Anne E Justice, Keri L Monda, Damien C Croteau-Chonka, Felix R Day, et al. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nature genetics*, 45(5):501–512, 2013.

[71] Heribert Schunkert, Inke R König, Sekar Kathiresan, Muredach P Reilly, Themistocles L Assimes, Hilma Holm, Michael Preuss, Alexandre FR Stewart, Maja Barbalic, Christian Gieger, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*, 43(4):333–338, 2011.

[72] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, Cristen J Willer, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713, 2010.

[73] Alisa K Manning, Marie-France Hivert, Robert A Scott, Jonna L Grimsby, Nabila Bouatia-Naji, Han Chen, Denis Rybin, Ching-Ti Liu, Lawrence F Bielak, Inga Prokopenko, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature genetics*, 44(6):659–669, 2012.

[74] Ralf JP van der Valk, Eskil Kreiner-Møller, Marjolein N Kooijman, Mònica Guxens, Evangelia Stergiakouli, Annika Sääf, Jonathan P Bradfield, Frank Geller, M Geoffrey Hayes, Diana L Cousminer, et al. A novel common variant in dcst2 is associated with length in early life and height in adulthood. *Human molecular genetics*, page ddu510, 2014.

[75] Luke Jostins, Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, L Philip Schumm, Yashoda Sharma, Carl A Anderson, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, 2012.

[76] Eli A Stahl, Soumya Raychaudhuri, Elaine F Remmers, Gang Xie, Stephen Eyre, Brian P Thomson, Yonghong Li, Fina AS Kurreeman, Alexandra Zhernakova, Anne Hinks, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature genetics*, 42(6):508–514, 2010.

[77] 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.

Supplementary Note

Flavors of Genetic Correlation

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 $y_1 = Y\beta + \delta$, and $y_2 = Z\gamma + \epsilon$, where Y and Z are matrices of genotypes normalized to mean zero and variance one[1], 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, including SNPs that are not genotyped.

We model all of $Y, Z, \beta, \gamma, \delta$ and ϵ as random variables. Suppose that (β, γ) has mean zero and 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 (δ, ϵ) has mean zero and covariance matrix

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

Finally, suppose that there are N_s individuals who appear in both Y and Z, and that the genotypes from all unique individuals are *i.i.d.* draw from a distribution with covariance matrix R (*i.e.*, r is an LD matrix with $r_{jk} = \mathbb{E}[Y_{ij}Y_{ik}]$. 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 hides an implicit assumption that rare SNPs have larger per-allele effect sizes than common SNPs. As discussed in the simulations section of the main text and in our earlier work [20], LD Score regression is robust to moderate violations of this assumption, though it may break down in extreme cases, for instance if all causal variants are rare. In situations where a different model for the variance of β is more appropriate, all proofs in this note go through with LD Score replaced by weighted LD Scores, $\ell_j = \sum_k \mathrm{Var}[\beta_j]r_{jk}^2$.

Proposition 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 $\sum_j X_{ij}\beta_j$, and the additive genetic component of y_2 for this individual is $\sum_j X_{ij}\gamma_j$. Thus, the genetic covariance

---

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

between y_1 and y_2 is

$$\mathrm{Cov}\left[\sum_j X_j\beta_j, \sum_j X_j\gamma_j\right] = \mathbb{E}\left[\left(\sum_j X_j\beta_j\right)\left(\sum_j X_j\gamma_j\right)\right]$$

$$= \sum_j\sum_k \mathbb{E}[X_jX_k\beta_j\gamma_k]$$

$$= \sum_j \mathbb{E}[X_j^2\beta_j\gamma_j]$$

$$= \sum_j \mathbb{E}[X_j^2]\mathbb{E}[\beta_j\gamma_j]$$

$$= \rho_g.$$

$\square$

Suppose we directly genotype SNP j. We compute Z-scores $z_{1j} := Y_j^\mathsf{T}y_1/\sqrt{N_1}$ and $z_{2j} := Y_j^\mathsf{T}y_1/\sqrt{N_2}$ from linear regression.

Proposition 2. *Let j denote a genotyped SNP. Under the model described above,*

$$\mathbb{E}[z_{1j}z_{2j}] = \frac{\sqrt{N_1N_2}\rho_g}{M}\ell_j + \frac{N_s\rho}{\sqrt{N_1N_2}}. \tag{3}$$

Proof. By the law of total expectation,

$$\mathbb{E}[z_{1j}z_{2j}] = \mathbb{E}[\mathbb{E}[z_{1j}z_{2j}\,|\,Y,Z]]$$

First,

$$\mathbb{E}[z_{1j}z_{2j}\,|\,Y,Z] = \frac{1}{\sqrt{N_1N_2}}\mathbb{E}[Y_j^\mathsf{T}y_1y_2^\mathsf{T}Z_j]$$

$$= \frac{1}{\sqrt{N_1N_2}}Y_j^\mathsf{T}\mathbb{E}[(Y\beta+\delta)(Z\gamma+\epsilon)]Z_j$$

$$= \frac{1}{\sqrt{N_1N_2}}Y_j^\mathsf{T}\left(Y\mathbb{E}[\beta\gamma^\mathsf{T}]Z + \mathbb{E}[\delta^\mathsf{T}\epsilon]\right)Z_j$$

$$= \frac{1}{\sqrt{N_1N_2}}\left(\frac{\rho_g}{M}Y_j^\mathsf{T}YZ_j^\mathsf{T}Z + \rho_eY_j^\mathsf{T}Z_j\right).$$

To remove the conditioning on Y and Z, we compute

$$\frac{1}{\sqrt{N_1N_2}}\mathbb{E}[Y_j^\mathsf{T}Z_j] = \frac{N_s}{\sqrt{N_1N_2}},$$

and

$$\frac{1}{\sqrt{N_1N_2}}\mathbb{E}[Y_j^\mathsf{T}YZ_j^\mathsf{T}Z] = \ell_j + \frac{MN_s}{\sqrt{N_1N_2}}.$$

Thus,

$$\mathbb{E}[z_{1j}z_{2j}] = \frac{\sqrt{N_1 N_2}\rho_g}{M}\ell_j + \frac{N_s(\rho_g + \rho_e)}{\sqrt{N_1 N_2}}$$
$$= \frac{\sqrt{N_1 N_2}\rho_g}{M}\ell_j + \frac{N_s\rho}{\sqrt{N_1 N_2}}.$$

Note that if study 1 and study 2 are the same study, then $N_1 = N_2 = N_s$, $\rho_g = h_g^2$, and this reduces to the LD Score regression equation for a single trait from [20]. □

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 an additional distributional assumption. Therefore we derive the conditional variance function for the case where Z-scores follow a multivariate normal distribution[2].

If phenotypes are normally distributed, then by the central limit theorem, $\hat{\beta}$ and $\hat{\gamma}$ are approximately distributed as bivariate normal with expectation zero. Then, from a standard formula for double second moments of the bivariate normal distribution

$$\mathrm{Var}[\hat{\beta}_j\hat{\gamma}_j \,|\, \ell_j] = \mathbb{E}[\hat{\beta}^2\hat{\gamma}^2]$$
$$= \mathrm{Var}[\hat{\beta}]\mathrm{Var}[\hat{\gamma}] + 2\mathbb{E}[\hat{\beta}\hat{\gamma}]^2$$
$$= \left(\frac{h_1^2\ell_j}{M} + \frac{1}{N_1}\right)\left(\frac{h_2^2\ell_j}{M} + \frac{1}{N_2}\right) + 2\left(\frac{\rho_g\ell_j}{M} + \frac{\rho N_s}{N_1 N_2}\right)^2. \tag{4}$$

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; however, in this case LD Score regression may be inefficient, because the regression weights will be suboptimal.

Liability Threshold Model

In the liability threshold model, binary traits are determined by an unobserved continuous liability ψ generated following the usual model for quantitative traits: $\psi_i = \sum_j X_{ij}\beta_j + \epsilon_i$. The observed binary phenotype is then has population prevalence K, and is generated from the unobserved liability via the liability threshold model: $y_i := \mathbf{1}[\psi_i > \tau]$, where τ denotes the liability threshold. If ψ is normally distributed, then setting $\tau := \Phi^{-1}(1 - K)$ (where Φ denotes the cdf of the standard normal distribution) yields a population prevalence of K.

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).

---

[2]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.

Note that the liability threshold construction is surprisingly general. If y is a binary phenotype, then we can define a normally distributed liability for y by defining ψ_i to be a random draw from a normal distribution truncated at the left by τ if $y_i = 1$ and truncated at the right by τ if $y_i = 0$. Therefore, we prefer to think of liability scale heritability as a useful way to compare the heritabilities of binary phenotypes with different prevalences, rather than a quantity that is only meaningful if one is willing to make strong assumptions about the data generating process.

In this section, we derive population case and control allele frequencies $p_{cas,j} = \mathbb{P}[G_{ij} = 1 \mid y_i = 1]$ and $p_{con,j} = \mathbb{P}[G_{ij} = 1 \mid y_i = 0]$ in terms of the heritability of liability generated by a polygenic model, where G_{ij} denotes an un-normalized genotype (note that since we are only modeling additive effects and we are willing to assume Hardy-Weinberg equilibrium, we lose nothing and simplify notation by dealing with haploid genotypes).

Proposition 3. *Suppose unobserved liabilities ψ, φ for traits y_1, y_2 with thresholds τ_1, τ_2 corresponding to prevalences K_1, K_2 are generated according to the usual polygenic model for quantitative traits, i.e., $\psi_i = \sum_j X_{ij}\beta_j + \delta$, $\varphi_i = \sum_j X_{ij}\gamma_j + \epsilon$, with X random with a given LD 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

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

Let ζ_j and ξ_j denote the marginal per-allele effect sizes of j on ψ and φ, and let $p_{cas,kj} := \mathbb{P}[G_{ij} = 1 \mid y_{ik} = 1]$ for $k = 1, 2$. Then

$$\mathbb{E}[p_{cas,1j} - p_{con,1j}] = 0,$$
$$\mathbb{E}[p_{cas,2j} - p_{con,2j}] = 0,$$
$$\mathrm{Var}[p_{cas,1j} - p_{con,1j}] = \frac{p_j(1 - p_j)\phi(\tau_1)^2 h_1^2}{MK_1^2(1 - K_1)^2)}\ell_j,$$
$$\mathrm{Var}[p_{cas,2j} - p_{con,2j}] = \frac{p_j(1 - p_j)\phi(\tau_2)^2 h_2^2}{MK_2^2(1 - K_2)^2}\ell_j,$$
$$\mathrm{Cov}[p_{cas,1j} - p_{con,1j}, p_{cas,2j} - p_{con,2j}] = \frac{p_j(1 - p_j)\phi(\tau_1)\phi(\tau_2)\rho_g}{MK_1(1 - K_1)K_2(1 - K_2)}\ell_j.$$

Note that these are different from the usual expressions for conversion from observed to liability scale, because these expressions only account for the effect of transforming from a continuous liability to a binary observed phenotype and not for ascertainment. We account for ascertainment in the next section.

Proof. This proof is accomplished in two steps. First, we compute allele frequencies conditional on the marginal effects on liability. To do this, we reverse the conditional probability using Bayes' theorem, which reduces the problem to a series of [Taylor approximations to] gaussian integrals. Second, we remove the conditioning on the marginal effects on liability and use the results of the previous section (on quantitative traits) to express the expectations of the allele frequencies in terms of h_1^2, h_2^2, ρ_g and ℓ_j.

By Bayes' theorem,

$$\mathbb{P}[G_{ij} = 1 \,|\, y_{i1} = 1, \zeta_j] = \frac{\mathbb{P}[y_{i1} = 1 \,|\, G_{ij} = 1, \zeta_j]\mathbb{P}[G_{ij} = 1]}{\mathbb{P}[y_{i1} = 1]}$$

$$= \frac{p_j}{K_1}\mathbb{P}[y_{i1} = 1 \,|\, G_{ij} = 1, \zeta_j]$$

$$= \frac{p_j}{K_1}\mathbb{P}[\psi_i > \tau_1 \,|\, G_{ij} = 1, \zeta_j].$$

The distribution of ψ is $\psi \,|\, G_{ij} = 1, \zeta_j = N(\zeta_j, 1)$ (where the approximation that the variance equals one holds when the marginal heritability explained by j is small, which is the typical case in GWAS). Thus $\mathbb{P}[\psi_i > \tau_1 \,|\, G_{ij} = 1]$ is simply a Gaussian integral. Because we wish to obtain a closed-form expression for the variance of $\mathbb{P}[\psi_i > \tau_1 \,|\, G_{ij} = 1, \zeta_j]$ over random ζ_j, we approximate this probability with a first-order Taylor expansion around τ_1:

$$\mathbb{P}[\psi_i > \tau_1 \,|\, G_{ij} = 1, \zeta_j] = 1 - \Phi(\tau_1 - \zeta_j)$$

$$\approx K_1 + \phi(\tau_1)\zeta_j,$$

where Φ and ϕ denote the cdf and density of the standard normal distribution. Therefore,

$$\mathbb{P}[G_{ij} = 1 \,|\, y_{i1} = 1, \zeta_j] = \frac{p_j}{K}\left(K + \phi(\tau_1)\zeta_j\right). \tag{5}$$

A similar argument shows that

$$\mathbb{P}[G_{ij} = 1 \,|\, y_{i1} = 0, \zeta_j] = \frac{p_j}{1 - K_1}\left(1 - K - \phi(\tau_1)\zeta_j\right), \tag{6}$$

so

$$\mathbb{P}[G_{ij} = 1 \,|\, y_{i1} = 1, \zeta_j] - \mathbb{P}[G_{ij} = 1 \,|\, y_{i1} = 0, \zeta_j] = p_j\frac{\phi(\tau_1)\zeta_j}{K_1(1 - K_1)}. \tag{7}$$

Similar results hold for trait 2, replacing ζ with ξ and subscript 1 with subscript 2.

We have written the probabilities in question in terms of constants and marginal effects on liability. Since liability is simply a quantitative trait, the means, variances, and covariances of the marginal effects on liability are described by the LD Score regression equation from the section on quantitative traits. Precisely, $\mathbb{E}[\xi_j] = \mathbb{E}[\zeta_j] = 0$, $\mathrm{Var}[\xi_j] = (1 - p_j)h_1^2\ell_j/p_kM$, $\mathrm{Var}[\zeta_j] = (1 - p_j)h_2^2\ell_j/p_jM$ and $\mathrm{Cov}[\zeta_j, \xi_j] = (1 - p_j)\rho_g\ell_j/p_jM$ (note that $\zeta_j = \mathbb{E}[\psi \,|\, G_{ij} = 1]$. Since $X_i = (G_{ij} - p_j)/\sqrt{p_j(1 - p_j)})$, we have $X_j\beta_j = \xi_j$, which explains the factors of $(1 - p_j)/p_j$. Thus it follows from equation 7 that $\mathbb{E}[p_{cas,1j} - p_{con,1j}] = 0$;

$$\mathrm{Var}[p_{cas,1j} - p_{con,1j}] = \mathrm{Var}\left[\frac{p_j\phi(\tau_1)\zeta_j}{K_1(1 - K_1)}\right]$$

$$= \frac{p_j(1 - p_j)\phi(\tau_1)^2h_1^2}{MK_1^2(1 - K_1)^2}\ell_j, \tag{8}$$

(and similarly for trait two); and last

$$\mathrm{Cov}[p_{cas,1j} - p_{con,1j}, p_{cas,2j} - p_{con,2j}] = \mathrm{Cov}\left[\frac{p_j\phi(\tau_1)\zeta_j}{K_1(1 - K_1)}, \frac{p_j\phi(\tau_2)\xi_j}{K_2(1 - K_2)}\right]$$

$$= \frac{p_j(1 - p_j)\phi(\tau_1)\phi(\tau_2)\rho_g}{MK_1(1 - K_1)K_2(1 - K_2)}\ell_j. \tag{9}$$

$\square$

Ascertained Studies of Liability Threshold Traits

In this section, we derive an analogous LD Score regression equation for ascertained case/control studies. We keep the same notation from the previous section, except here y_1 and y_2 are binary phenotypes. In addition, let P_i denote the sample prevalence of y_i in study i for $i = 1, 2$, and K_i denote the population prevalence.

We compute z-scores $z_j := \frac{\sqrt{NP(1-P)}(\hat{p}_{cas}-\hat{p}_{con})}{\sqrt{\hat{p}_j(1-\hat{p}_j)}}$, where $\hat{p}_j$ denotes allele frequency in the entire sample[3], $\hat{p}_{cas}$ denotes allele frequency among cases in the sample and $\hat{p}_{con}$ denotes allele frequency among controls in the sample.

Proposition 4. *Under a polygenic model,*

$$\mathbb{E}[z_{1j}z_{2j}] \approx \frac{\sqrt{N_1 N_2}\rho_{g,obs}}{M}\ell_j + c\sqrt{N_1 N_2}\left(\sum_{a,b\in\{cas,con\}} \frac{N_{a,b}(-1)^{1+\mathbf{1}[\mathbf{a}=\mathbf{b}]}}{N_{a,1}N_{b,2}}\right) \tag{10}$$

where

$$\rho_{obs} := \frac{\rho_g \phi(\tau_1)\phi(\tau_2)\sqrt{P_1(1-P_1)P_2(1-P_2)}}{K_1(1-K_1)K_2(1-K_2)}$$

denotes observed scale genetic covariance, and $N_{a,b}$ for $a,b \in \{cas, con\}$ denote the number of individuals with phenotype a in study 1 and b in study two (e.g., $N_{cas,con}$ is the number of individuals who are cases in study 1 and controls in study 2).

Proof. The full form of $z_{1j}z_{2j}$ is

$$z_{1j}z_{2j} = \frac{c\sqrt{N_1 N_2}(\hat{p}_{cas,1j} - \hat{p}_{con,1j})(\hat{p}_{cas,2j} - \hat{p}_{con,2j})}{\sqrt{\hat{p}_{1j}(1-\hat{p}_{1j})\hat{p}_{2j}(1-\hat{p}_{2j})}}, \tag{11}$$

where $c := \sqrt{P_1(1-P_1)P_2(1-P_2)}$. Our strategy for obtaining the expectation is

$$\mathbb{E}[z_{1j}z_{2j}] \approx c\sqrt{N_1 N_2}\frac{\mathbb{E}[(\hat{p}_{cas,1j} - \hat{p}_{con,1j})(\hat{p}_{cas,2j} - \hat{p}_{con,2j})]}{\mathbb{E}[\sqrt{\hat{p}_{1j}(1-\hat{p}_{1j})\hat{p}_{2j}(1-\hat{p}_{2j})}]} \tag{12}$$

$$\approx c\sqrt{N_1 N_2}\frac{\mathbb{E}[(\hat{p}_{cas,1j} - \hat{p}_{con,1j})(\hat{p}_{cas,2j} - \hat{p}_{con,2j})]}{\sqrt{\mathbb{E}[\hat{p}_{1j}(1-\hat{p}_{1j})\hat{p}_{2j}(1-\hat{p}_{2j})]}} \tag{13}$$

$$= c\sqrt{N_1 N_2}\frac{\mathbb{E}\left[\mathbb{E}[(\hat{p}_{cas,1j} - \hat{p}_{con,1j})(\hat{p}_{cas,2j} - \hat{p}_{con,2j})\,|\,\zeta_j,\xi_j]\right]}{\sqrt{\mathbb{E}\left[\mathbb{E}[\hat{p}_{1j}(1-\hat{p}_{1j})\hat{p}_{2j}(1-\hat{p}_{2j})\,|\,\zeta_j,\xi_j]\right]}}, \tag{14}$$

where ζ_j and ξ_j denote the marginal effects of j. The first approximation follows from the fact that we incur only $\mathscr{O}(1/N)$ error by moving from the expectation of a ratio to a ratio of expectations. The second approximation follows from the fact that we incur only $\mathscr{O}(1/N)$ error by moving from the expectation of a square root to a square root of expectations. The equality in the third line follows from application of the law of total expectation to both the numerator and denominator.

This differs from the proof strategy used to derive the LD Score regression equation for non-ascertained studies in that here we first condition on ζ_j and ξ_j instead of conditioning on the

---

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

genotype matrices X and Y. This is because X and Y depend on β and γ in an ascertained study: oversampling cases increases the sample allele frequencies of risk alleles.

The technical essence of the proof is deriving $\mathbb{E}[\hat{p}_{cas,1j} - \hat{p}_{con,1j}]$ and showing that $\text{Var}[\hat{p}_{cas,1j} - \hat{p}_{con,1j} \,|\, \zeta_j, \xi_j] \approx \text{Var}[\hat{p}_{cas,1j} - \hat{p}_{con,1j}]$.

First, we compute the numerator. We begin by taking ζ_j and ξ_j as fixed, so the allele frequencies $\mathbb{P}[G_{ij} = 1 \,|\, y_{i1} = 1, y_{i2} = 1]$, etc are fixed. In this case, the only variance in the sample allele frequencies $\hat{p}_{cas,1}, \hat{p}_{con,1}, \hat{p}_{cas,2}, \hat{p}_{con,2}$ comes from finite sampling error. However, if study 1 and study 2 share samples, then the sampling error in study 1 may be correlated to the sampling error in study 2. By linearity of expectation, we have

$$
\mathbb{E}[(\hat{p}_{cas,1j} - \hat{p}_{con,1j})(\hat{p}_{cas,2j} - \hat{p}_{con,2j})] \,|\, \zeta_j, \xi_j] = \mathbb{E}[\hat{p}_{cas,1j}\hat{p}_{cas,2j}] + \mathbb{E}[\hat{p}_{cas,1j}\hat{p}_{con,2j}] \\
+ \mathbb{E}[\hat{p}_{con,1j}\hat{p}_{cas,2j}] + \mathbb{E}[\hat{p}_{con,1j}\hat{p}_{con,2j}]
$$

We can write $\hat{p}_{cas,1j}\hat{p}_{cas,2j} = (p_{cas,1j} + \eta)(p_{cas,2j} + \nu)$, where η and ν denote sampling error. By the [bivariate] central limit theorem,

$$
\mathbb{E}[\hat{p}_{cas,1j}\hat{p}_{cas,2j} \,|\, \zeta_j, \xi_j] = p_{cas,1j}p_{cas,2j} + \mathbb{E}[\eta\nu]
$$
$$
= p_{cas,1j}p_{cas,2j} + \frac{N_{cas,cas}\sqrt{p_{cas,1j}(1 - p_{cas,1j})p_{cas,2j}(1 - p_{cas,2j})}}{N_{cas,1}N_{cas,2}}
$$
$$
\approx p_{cas,1j}p_{cas,2j}\left(1 + \frac{N_{cas,cas}}{N_{cas,1}N_{cas,2}}\right),
$$

where the approximation comes from ignoring the difference between $\sqrt{p_{cas,1j}(1 - p_{cas,1j})p_{cas,2j}(1 - p_{cas,2j})}$ and $p_j(1-p_j)$. This step is justified in the derivation of the denominator, below. A similar relationship holds for other combinations of phenotypes, except we need to flip the sign when the phenotype in study 1 and the phenotype in study 2 disagree.

Next, we remove the conditioning on ζ_j and ξ_j using results on the liability threshold model from the previous section. It follows directly from equation 9 that

$$
\mathbb{E}[(p_{cas,1j} - p_{con,1j})(p_{cas,2j} - p_{con,2j})] = \frac{p_j(1 - p_j)\phi(\tau_1)\phi(\tau_2)\rho_g}{MK_1(1 - K_1)K_2(1 - K_2)}\ell_j.
$$

Therefore, we approximate the expectation of the numerator as

$$
\mathbb{E}[(\hat{p}_{cas,1j} - \hat{p}_{con,1j})(\hat{p}_{cas,2j} - \hat{p}_{con,2j})] \approx p_j(1 - p_j)\left(\frac{\phi(\tau_1)\phi(\tau_2)\rho_g}{c'M}\ell_j + \sum_{a,b\in\{cas,con\}}\frac{N_{a,b}(-1)^{1 + \mathbf{1}[\mathbf{a=b}]}}{N_{a,1}N_{b,2}}\right),
$$
$$(15)$$

where $c' := K_1(1 - K_1)K_2(1 - K_2)$.

Next, we derive the expectation of the denominator. Conditional on β, $\hat{p}_{1j}(1 - \hat{p}_{1j})$ is $P_1p_{cas,1j} + (1 - P_1)p_{con,1j}$ plus $\mathscr{O}(1/N)$ sampling variance. If studies 1 and 2 share samples, the $\mathscr{O}(1/N)$ sampling variance in $\hat{p}_{1j}(1 - \hat{p}_{1j})$ and $\hat{p}_{2j}(1 - \hat{p}_{2j})$ will be correlated, but this still only amounts to $\mathscr{O}(N_s/N_1N_s)$ error. If we now view β as random, then as derived above, $P_1p_{cas,1j} + (1 - P_1)p_{con,1j}$ is equal to $p_j(1 - p_j)$ plus $\mathscr{O}(h_{1,obs}^2\ell_j/M)$ error from uncertainty in β. The correlation between uncertainty in β and uncertainty in γ is governed by $\rho_{g,obs}$, so the expectation of the denominator is $\mathbb{E}\left[\sqrt{\hat{p}_{1j}(1 - \hat{p}_{1j})\hat{p}_{2j}(1 - \hat{p}_{2j})}\right] = p_j(1 - p_j) + \mathscr{O}(N_s/N_1N_2) + \mathscr{O}(\rho_{g,obs}\ell_j/M)$. Since we aim to

apply LD Score regression to common variants, a lower bound on $p_j(1 - p_j)$ is around 10^{-2} if $p_j = 0.01$, which is roughly the lowest MAF for which we can estimate LD Score with any degree of precision. For $\ell_j = 100$ (roughly the median 1kG LD Score), $M = 10^7$ and $\rho_{g,obs} = 1$, we get $\rho_{g,obs}\ell_j/M = 10^{-5}$. A worst-case value for $N_s/N_1 N_2$ might be $N_s = N_1 = N_2 = 10^4$, in which case $N_s/N_1 N_2 = 10^{-4}$. Thus, in typical scenarios, both $\rho_{g,obs}\ell_j/M$ and $N_s/N_1 N_2$ will be at least 2-3 orders of magnitude smaller than $p_j(1 - p_j)$. Therefore, we make the approximation that

$$\mathbb{E}\left[\sqrt{\hat{p}_{1j}(1 - \hat{p}_{1j})\hat{p}_{2j}(1 - \hat{p}_{2j})}\right] \approx p_j(1 - p_j). \tag{16}$$

We obtain the desired result by dividing equation $c\sqrt{N_1 N_2}$ times equation 15 by equation 16. $\quad\square$

Corollary 1. *For a single liability threshold model trait, we have the LD Score regression equation*

$$\mathbb{E}[\chi_j^2] = \frac{Nh_{obs}^2}{M}\ell_j + 1, \tag{17}$$

where $h_{obs}^2 = h^2\phi(\tau)^2 P(1 - P)/K^2(1 - K)^2$.

Proof. This follows from the previous proposition with study 1 equal to study 2 and noting that the observed scale genetic covariance between a trait and itself is simply observed scale heritability. To show that the intercept is one, observe that if study 1 and study 2 are the same, then

$$
\begin{aligned}
c\sqrt{N_1 N_2}\left(\sum_{a,b\in\{cas,con\}} \frac{N_{a,b}(-1)^{1+\mathbf{1}[\mathbf{a}=\mathbf{b}]}}{N_{a,1}N_{b,2}}\right) &= NP(1 - P)\left(\frac{1}{N_{cas}} + \frac{1}{N_{con}}\right)\\
&= \frac{NP(1 - P)(N_{cas} + N_{con}))}{N_{cas}N_{con}}\\
&= \frac{N^2 P(1 - P)}{N_{cas}N_{con}}. \tag{18}
\end{aligned}
$$

But $NP = N_{cas}$ and $N(1 - P) = N_{con}$, so equation 18 simplifies to 1. $\quad\square$

Supplementary Tables

Sample Sizes

Trait	Reference	Sample Size
Schizophrenia	PGC Schizophrenia Working Group, *Nature,* 2014 [63]	70,100
Bipolar disorder	PGC Bipolar Working Group, *Nat Genet,* 2011 [64]	16,731
Major depression	PGC MDD Working Group, *Mol Psych,* 2013 [65]	18,759
Anorexia Nervosa	Boraska, *et al., Mol Psych,* 2014 [66]	17,767
Autism Spectrum Disorder	PGC Cross-Disorder Group, *Lancet,* 2013 [22]	10,263
Ever/Never Smoked	TAG Consortium, 2010 *Nat Genet,* [67]	74,035
Alzheimer's	Lambert, *et al., Nat Genet,* 2013 [68]	54,162
College	Rietveld, *et al., Science,* 2013 [36]	101,069
Height	Lango Allen, *et al., Nature* 2010 [69]	133,858
Obesity Class 1	Berndt, *et al., Nat Genet,* 2013 [70]	98,000
Extreme Waist-Hip Ratio	Berndt, *et al., Nat Genet,* 2013 [70]	10,000
Coronary Artery Disease	Schunkert, *et al., Nat Genet,* 2011 [71]	86,995
Triglycerides	Teslovich, *et al., Nature,* 2010 [72]	96,598
LDL Cholesterol	Teslovich, *et al., Nature,* 2010 [72]	95,454
HDL Cholesterol	Teslovich, *et al., Nature,* 2010 [72]	99,900
Type-2 Diabetes	Morris, *et al., Nat Genet,* 2012 [24]	69,033
Fasting Glucose	Manning, *et al., Nat Genet,* 2012 [73]	46,186
Childhood Obesity	EGG Consortium, *Nat Genet,* 2012 [26]	13,848
Birth Length	van der Valk, *et al., HMG,* 2014 [74]	22,263
Birth Weight	Horikoshi, *et al., Nat Genet,* 2013 [25]	26,836
Infant Head Circumference	Taal, *et al., Nat Genet,* 2012 [27]	10,767
Age at Menarche	Perry, *et al., Nature,* 2014 [23]	132,989
Crohn's Disease	Jostins, *et al., Nature,* 2012 [75]	20,883
Ulcerative Colitis	Jostins, *et al., Nature,* 2012 [75]	27,432
Rheumatoid Arthritis	Stahl, *et al., Nat Genet,* 2010 [76]	25,708

Table S1: *Sample sizes and references for traits analyzed in the main text.*

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)

Table S2: *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 [20]. 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.*

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)

Table S3: *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 S2 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.*

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)

Table S4: *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.*

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

Table S5: *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 S2 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.*

Educational Attainment

Phenotype 1	Phenotype 2	r_g	se
College (Yes/No)	Years of Education	1.00	0.014

Table S6: *Genetic correlation between the two educational attainment phenotypes studied by Rietveld, et al. [36].*

Supplementary Figures
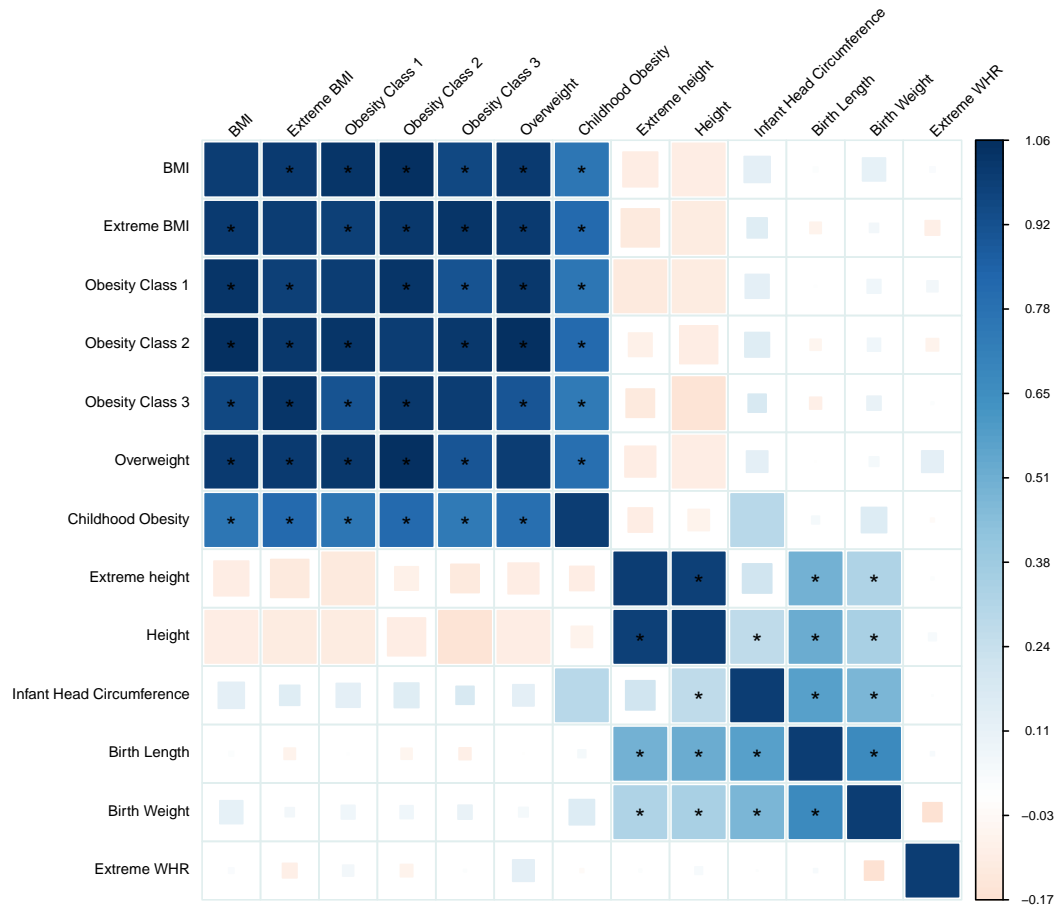
Genetic Correlations among Anthropometric Traits

Figure S1: *Genetic correlations among highly correlated anthropometric traits from studies by the GIANT and EGG consortia. The structure of the figure is the same as Figure 2 in the main text: blue corresponds to positive genetic correlations; red corresponds to negative genetic correlation. Larger squares correspond to more significant p-values. Genetic correlations that are different from zero at 1% FDR are displayed as full-sized squares. Genetic correlations that are significantly different from zero at significance level 0.05 after Bonferroni correction are given an asterisk.*

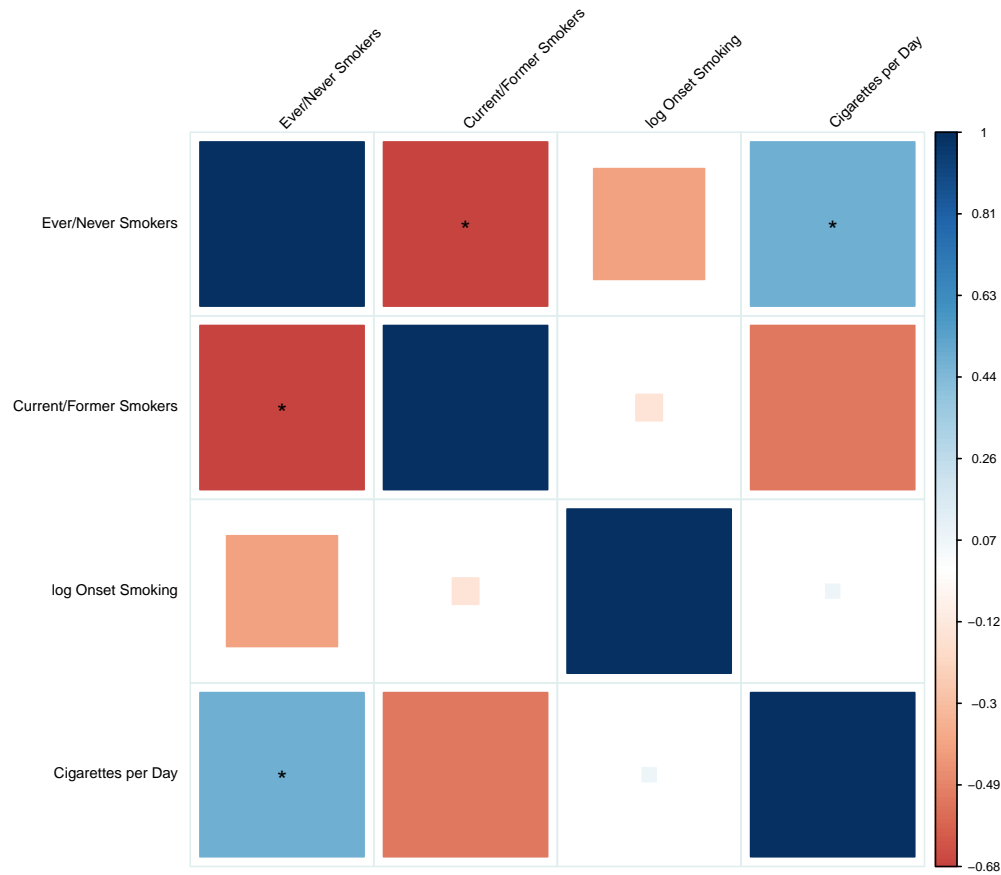# Genetic Correlations among Smoking Traits

Figure S2: *Genetic correlations among highly correlated smoking-related traits from the Tobacco and Genetics (TAG) consortium. The structure of the figure is the same as Figure 2 in the main text: blue corresponds to positive genetic correlations; red corresponds to negative genetic correlation. Larger squares correspond to more significant p-values. Genetic correlations that are different from zero at 1% FDR are displayed as full-sized squares. Genetic correlations that are significantly different from zero at significance level 0.05 after Bonferroni correction are given an asterisk.*

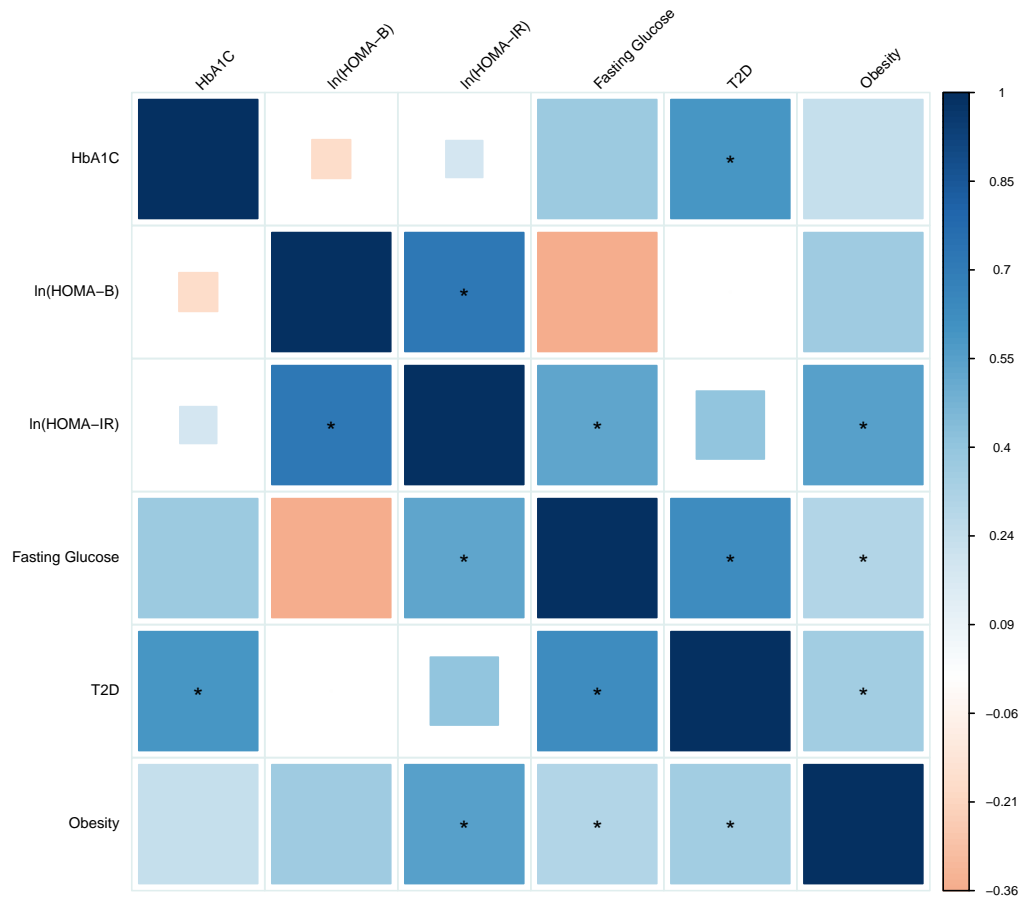## Genetic Correlations among Insulin-Related Traits

Figure S3: *Genetic correlations among highly correlated insulin-related traits from studies by the MAGIC consortium. The structure of the figure is the same as Figure 2 in the main text: blue corresponds to positive genetic correlations; red corresponds to negative genetic correlation. Larger squares correspond to more significant p-values. Genetic correlations that are different from zero at 1% FDR are displayed as full-sized squares. Genetic correlations that are significantly different from zero at significance level 0.05 after Bonferroni correction are given an asterisk.*

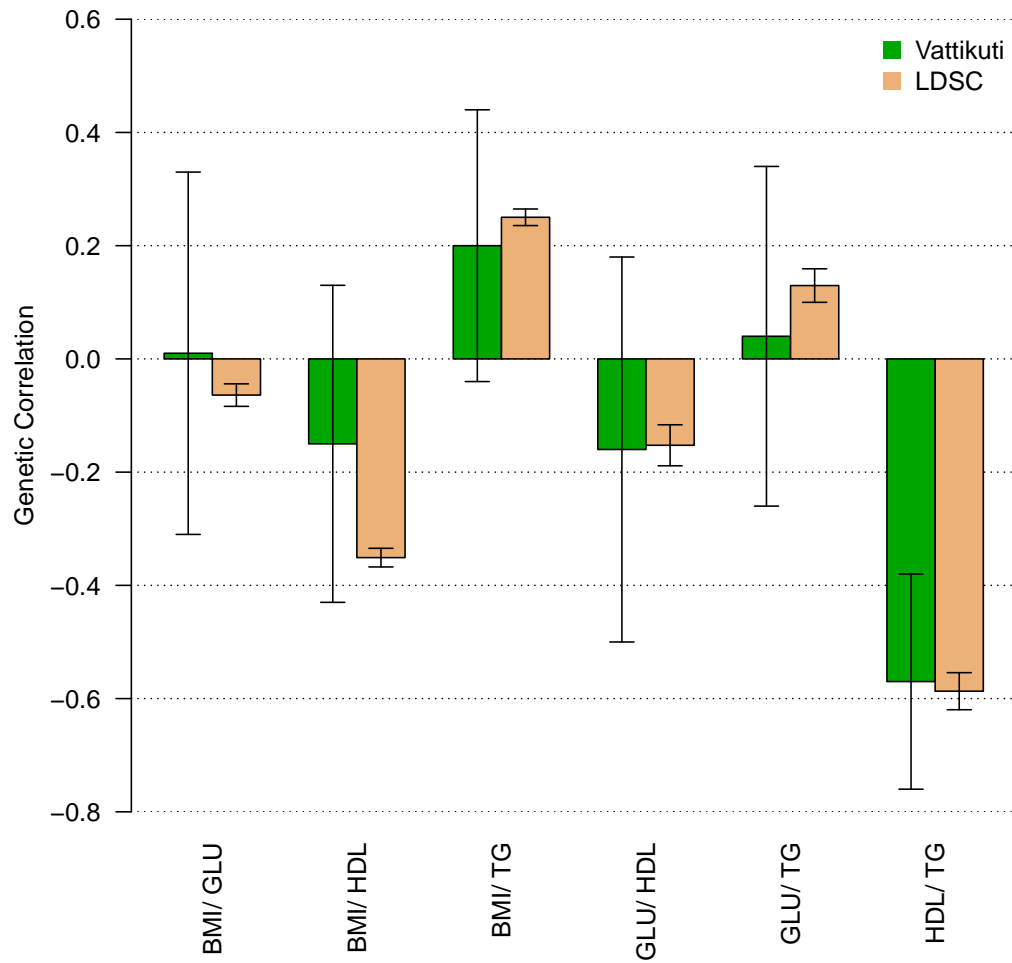# Comparison of Metabolic Genetic Correlations from LDSC to Results from Vattikuti, et al

Figure S4: *This figure compares the estimates of genetic correlations between metabolic traits from table 3 of Vattikuti, et al. [16] to the results obtained using LD Score regression (with much larger sample sizes) in this paper.*

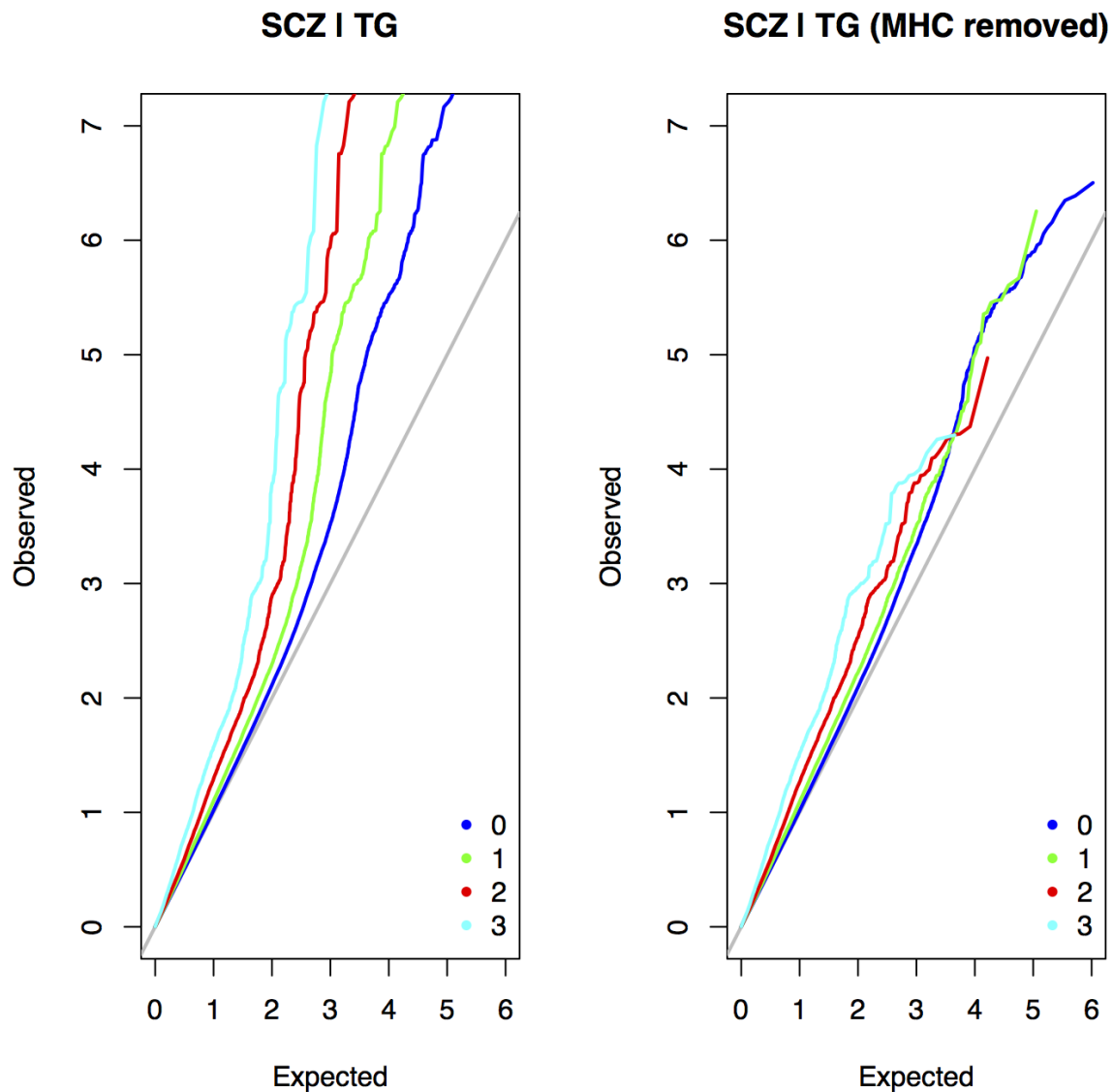# Schizophrenia / TG Conditional QQ Plot with and without the MHC

Figure S5: *We reproduced the conditional QQ plot comparing schizophrenia (SCZ) and triglycerides (TG) from Andreassen et al. [77] (left). The major histocompatibility complex (MHC, approximately chr6, 25-35 MB) is a genomic region featuring SNPs with exceptionally high LD Scores and the strongest GWAS association to schizophrenia. Removing the MHC removes almost all signal from the conditional QQ plot (right), which is consistent with the near-zero genetic correlation between schizophrenia and TG estimated via LD Score regression.*

Collaborators

Supplementary References