Estimating Genetic Correlation from GWAS Summary Statistics

Brendan Bulik-Sullivan[*†1,2,3], Hilary K Finucane[4,*], Verneri Antilla[1,2,3], Alexander Gusev[5,6], Felix R. Day[7], the ReproGen Consortium, John R.B. Perry[7], Cecilia Lindgren[1,2,3], Nick Patterson[1], Elise Robinson[1,2,3], Mark J Daly[1,2,3], Alkes L Price[**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, Boston, MA, 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, Cambridge, UK

Abstract

A recent focus in statistical genetics has been leveraging genetic association results from multiple phenotypes simultaneously to assist in the interpretation of results from genome-wide association studies (GWAS). Here we describe a novel method for estimating genetic correlations between traits that requires only GWAS summary statistics as input. We apply our method to 25 sets of publicly available GWAS summary statistics, which include more than 1.5 million phenotyped samples. Our results replicate many known genetic and epidemiological associations, including correlations between psychiatric traits, anthropometric traits, and components of metabolic syndrome. In addition, we report several novel results, including a positive genetic correlation between anorexia and schizophrenia and a negative genetic correlation between anorexia nervosa (AN) and body mass. These results highlight the power of a polygenic modeling framework, since there currently are no genome-wide significant SNPs for AN. We provide open source software implementing these analyses.

---

[*]Co-first authors
[**]Co-last authors
[†]Correspondence should be addressed to BBS (bulik@broadinstitute.org) or BMN (bmneale@broadinstitute.org).

1 Introduction

Discovering relationships between phenotypes is a fundamental goal of epidemiology, with applications to drug development, classification and treatment of disease. The traditional strategy in epidemiology is to search for correlations between phenotypes via large cross-sectional or longitudinal observational studies; however, the interpretation of results from these studies can be confounded by social factors and reverse causation. An alternative strategy that is more robust to confounding is to search instead for pairs of phenotypes with shared genetic etiology [1]. Many such methods have the additional advantage that they do not require measuring the same traits on the same individuals, which is particularly useful for phenotypes that are rare or costly to assay.

The largest currently available sources of genotype-phenotype data are genome-wide association studies (GWAS). There are three main approaches for testing relationships between phenotypes using GWAS data. Mendelian randomization has proved effective for traits for which there exist large-effect genetic variants [2]. However for many complex traits, the heritability is distributed over thousands of variants with small effect sizes [3]. For these traits, Mendelian randomization suffers from low power and weak instrument bias [4], so it is more effective to estimate genetic correlations using techniques such as restricted maximum likelihood (REML) [5, 6, 7] or polygenic scores [8, 9] that model the effects of all SNPs, including those that do not reach genome-wide significance [10, 11, 12].

REML and polygenic scores have only been applied to a small number of traits so far, because these methods requires access to individual genotypes, which are often difficult to obtain due to privacy considerations and informed consent limitations. Here we introduce a method for estimating genetic correlation that requires only GWAS summary statistics as input and is several orders of magnitude less computationally demanding than REML.

We apply this method to all GWAS for which the summary statistics have been made publicly available for download and report genetic correlations between hundreds of pairs of traits. The vast majority of our results are consistent with previously published genetic correlations, observations of overlap at genome-wide significant loci, and epidemiological associations. In addition, we report several novel results, including a positive genetic correlation between anorexia and schizophrenia and a negative genetic correlation between anorexia and BMI. These results demonstrate the advantages of a polygenic modeling framework, since there are currently no SNPs associated with anorexia at genome-wide significance [13].

2 Results

2.1 Overview of Methods

The χ^2-statistic at a given SNP incorporates the effects of all SNPs in LD with that SNP. 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 [14]. A similar relationship holds if we replace χ^2 statistics for a single trait with the product $z_1 z_2$, where z_i denotes the Z-score for trait i.

More precisely, under a polygenic model [5], the expected value of $z_1 z_2$ is

$$\mathbb{E}[z_{1j} z_{2j}] = \frac{\sqrt{N_1 N_2} \rho_g}{M} \ell_j + \frac{\rho N_s}{\sqrt{N_1 N_2}}, \tag{2.1}$$

where N_i is the sample size for study i, ρ_g is genetic covariance, ℓ_j is LD Score, 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. A similar relationship holds if one or both of the studies is an ascertained study of a binary phenotype (Supplementary Note), and in this case, the estimate of genetic covariance will be on the observed scale. As a consequence of equation 2.1, we can estimate genetic covariance using the slope from the regression of $z_{1j}z_{2j}$ on LD Score. This estimate will not be biased even if the two studies share samples, because sample overlap only affects the intercept, not the slope. Intuitively, bias from sample overlap affects all SNPs equally, similar to cryptic relatedness in the one-trait case [14], and therefore this bias is not correlated with LD Score.

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 heritability of trait i. In this paper, we report genetic correlations obtained using genetic covariances estimated with two-phenotype LD Score regression and heritabilities estimated using univariate LD Score regression [14].

2.2 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,

2.3 Misspecified Models of Genetic Architecture

Estimates of heritability and genetic covariance can be biased if the underlying model of genetic architecture is misspecified [15] 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 [16] and Online Methods) similar to that taken by Lee, *et. al.* with REML [17].

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 [18] 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) [19] with MAF above 5%. Results from these simulations are displayed in Supplementary Tables 5.1, 5.2 and 5.3.

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.

3

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

2.4 Replication of PGC Cross Disorder Results

For further validation, we replicated the estimates of genetic correlations between psychiatric disorders obtained with individual genotypes and REML in the Psychiatric Genomics Consortium (PGC) Cross-Disorder Group paper [10], using LD Score regression and the summary statistics from [20], downloaded from the PGC website (URLs).

Including an intercept in the LD Score regression protects the results from QC issues such as population stratification (as described in [14]) and sample overlap, but at the cost of larger standard errors. Since the summary statistics from [20] were generated after a careful QC process, and the samples used for each disease were non-overlapping, we also fit LD Score regression with constrained intercept.

Results from this analysis 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 without 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 smaller sample sizes (*e.g.,* ADD, ASD).

The computational demands of this analysis were trivial: after computing LD Scores (which takes an hour parallelized over chromosomes and only needs to be done once), the LD Score regression took a few seconds per pair of phenotypes and less than 1GB of RAM.

2.5 Application to a Large Set of Publicly Available Summary Statistics

We used LD Score regression to estimate genetic correlations between 25 phenotypes (Table 1, URLs), including all phenotypes with publicly available summary statistics and sufficiently large sample size and heritability (see Methods). For clarity of presentation, we pruned to one phenotype from each cluster of highly correlated phenotypes for the main text; highly correlated anthopometric phenotypes are displayed in supplementary table XXX, smoking phenotypes are displayed in supplementary table YYY, and insulin-related traits are displayed in supplementary table ZZZ. Genetic correlation estimates for the 25 traits are displayed in Figure 2.

Our estimates of genetic correlation between metabolic traits are generally consistent with the estimates obtained using REML in [11], though our standard errors are much lower, since operating on summary statistics allows us access to larger sample sizes (Supplementary Table 4.1). 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 [12].

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, or, more formally, via Mendelian randomization analyses. Examples of genetic correlations that are consistent with overlap among top loci include the correlations between plasma lipids and cardiovascular disease [21]; age at onset of menarche and obesity [22]; type 2 diabetes and obesity, fasting glucose, plasma lipids, and cardiovascular disease [23]; birth weight, adult height and type 2 diabetes [24]; birth length, adult height and infant head circumference [25, 26]; and childhood obesity and adult obesity [25]. For many of these pairs of traits, we can reject the null hypothesis of zero genetic correlation with
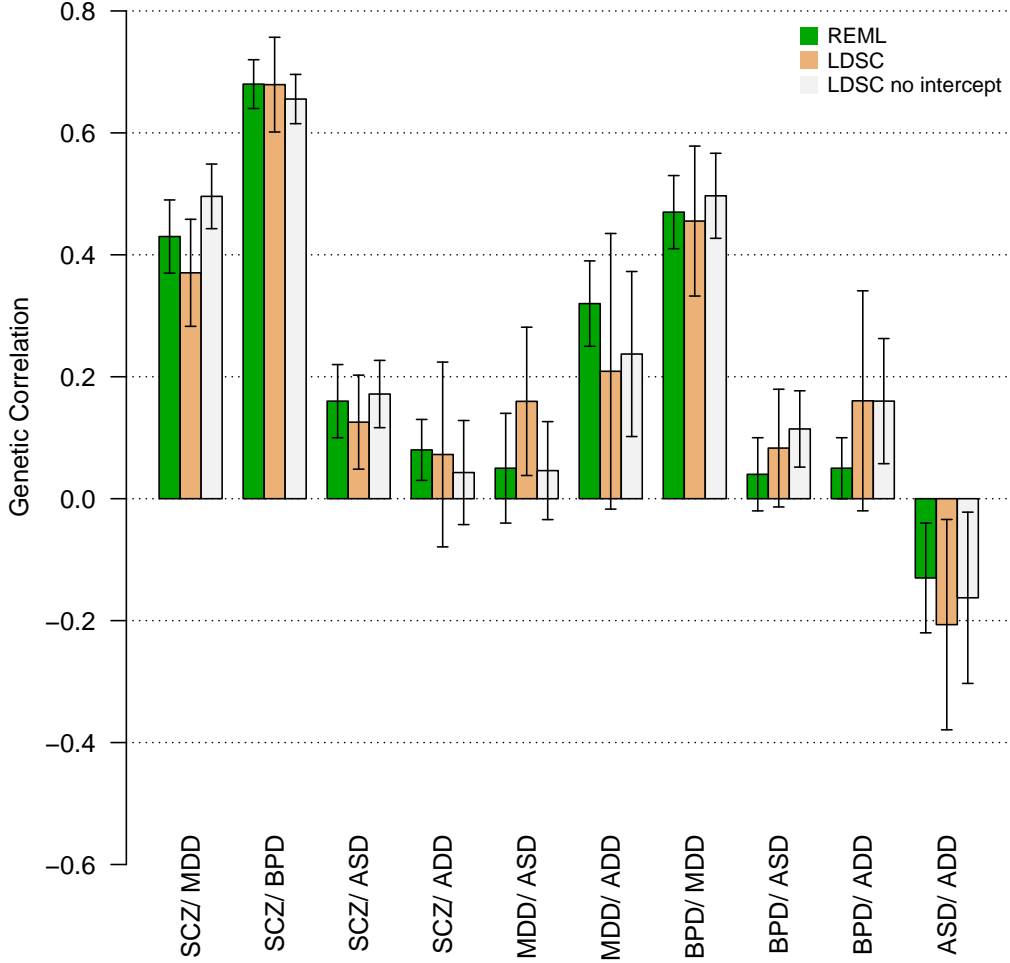
4

Figure 1: *Replication of PGC Cross Disorder Results. This plot compares LD Score regression estimates of genetic correlation using the summary statistics from [20] (which were generated from approximately the same data as [10]) to estimates obtained from REML in [10]. 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 without 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.*

overwhelming statistical significance (*e.g.*, $p = 6 \times 10^{-24}$ for age at onset of menarche and obesity; $p = 6 \times 10^{-52}$ for obesity and childhood obesity).

Our results on metabolic traits are directionally consistent with the recent Mendelian randomization results from [27], except that we identify a statistically significant positive genetic correlation between fasting glucose and obesity, and they identify a positive effect of BMI on LDL that is not

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

Table 1: *Datasets used in the main analyses.*

apparent in our data. Some of these results have also been reported in the twin study literature; for example, our estimate of 0.12 (0.03) ($p = 5 \times 10^{-4}$) for the genetic correlation between adult height and educational attainment is consistent with the genetic correlation estimates of 0.08 in males and 0.17 in females between adult height and IQ (for which educational attainment is a proxy phenotype) from [28]. This result has also been reported in the epidemiological literature [29], though we stress that this genetic correlation should not be given a causal interpretation. Both height and educational attainment have genetic correlations with many other phenotypes, and the observed genetic correlation between height and educational attainment could be mediated by any of these (see Discussion). The negative genetic correlation between birth weight and type 2 diabetes that we report is also consistent with overlap among top loci observed in [30].

Many of our results are consistent with well-known epidemiological association, but, to the best of our knowledge, have never been reported using genetic data. For example, our estimates of the genetic correlation between age at onset of menarche and adult height [31], cardiovascular disease [32] and type 2 diabetes [32, 33] are consistent with the epidemiological associations. We estimate a genetic correlation of -0.30 (0.08), $p = 1 \times 10^{-4}$, between educational attainment and Alzheimer's disease, consistent with the epidemiological observation that low educational attainment is one
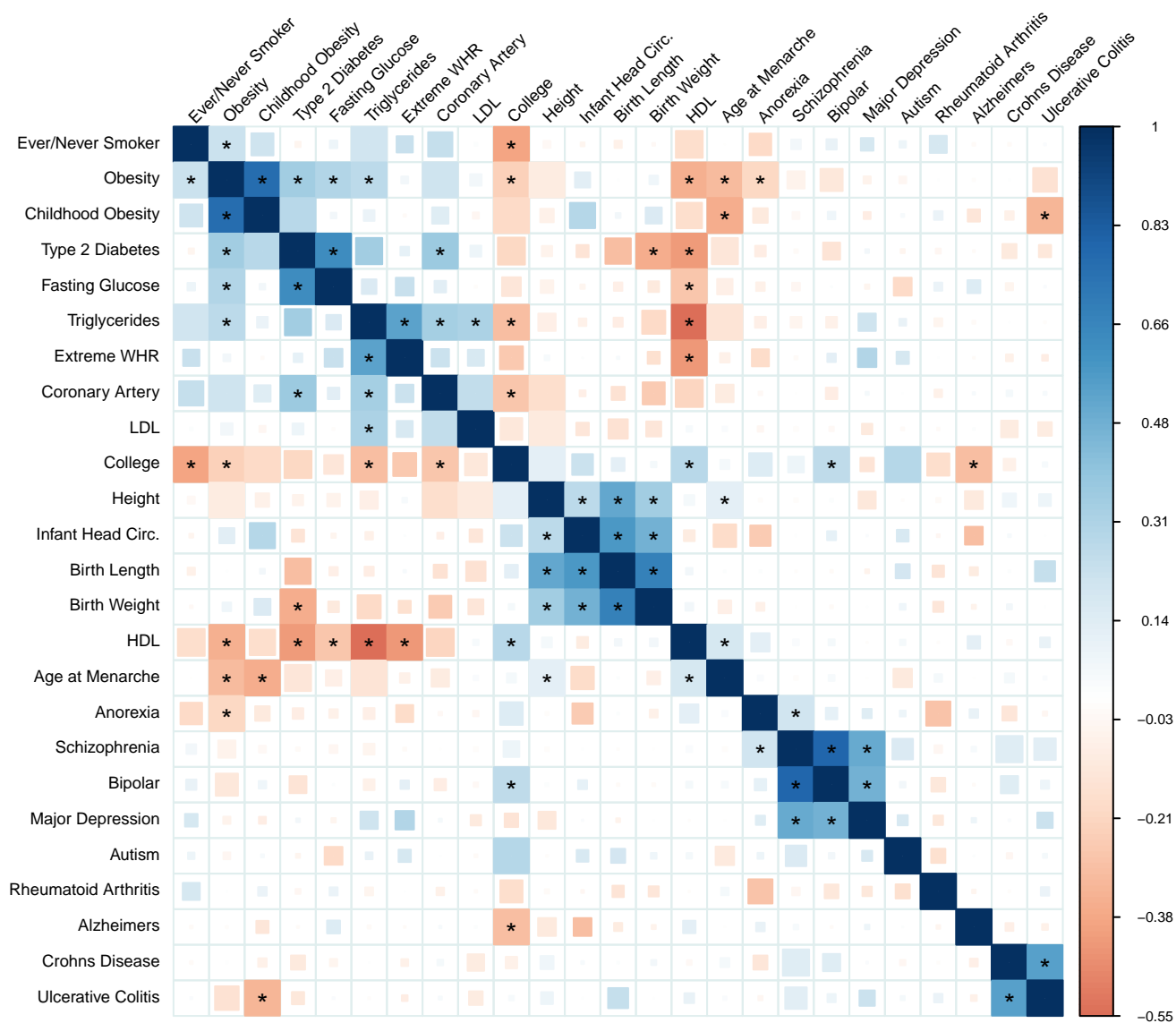
Figure 2: *Genetic Correlations between 25 Published GWAS. The color of the square indicates the value of the genetic correlation. Blue corresponds to positive genetic correlations; red corresponds to negative genetic correlation. The size of the colored squares indicates the magnitude of the p-value. Larger squares correspond to lower p-values, and all 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 $\alpha = 0.05$ after Bonferroni correction for 300 tests are given an asterisk. Note that this multiple testing correction is conservative, since many of the tests are not independent.*

of the largest risk factors for Alzheimer's [34, 35]. The positive genetic correlation between educational attainment and bipolar disorder is consistent with psychiatric literature showing that education and bipolar disorder status are positively correlated [36, 37], Our estimate of a genetic correlation of -0.20 (0.04) ($p = 4 \times 10^{-6}$) between anorexia and obesity is sensible considering that low BMI is one of the diagnostic criteria for anorexia, though it is interesting that the genetic factors that influence normal variation in BMI are similar to the genetic factors that influence unhealthily low BMI in the context of 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 [16]. Our estimate of the genetic correlations between schizophrenia and inflammatory bowel disease (IBD – Crohn's and UC) are not statistically significantly different from zero after correction for multiple tests, but the direction of the genetic correlation (positive) is consistent with recent epidemiological work showing an increased risk for schizophrenia after IBD diagnosis [38] and an increased risk for IBD after schizophrenia diagnosis [39]. These observations are particularly interesting in light of recent results showing that immune cell-type-specific epigenetic marks are enriched for genetic variants that influence risk for schizophrenia [16, 40], although we emphasize that the evidence for genetic correlation is not yet conclusive. The positive genetic correlation between smoking and obesity is in the opposite direction of the epidemiological association (smoking is associated with lower body mass, although smoking cessation often leads to weight gain) and Mendelian randomization results using rs1051730 [41]. We propose that the resolution of this apparent discrepancy is an interesting direction for future research. Our estimate of a negative genetic correlation between smoking and educational attainment is consistent with the epidemiological result that smoking is more prevalent among less-educated groups [43]. Last, the genetic correlation of -0.17 (0.04) ($p = 2 \times 10^{-4}$) between height and coronary artery disease (possibly related to the genetic correlation of -0.11 (0.03) ($p = 0.001$) between height and LDL, a known risk factor for coronary artery disease [21, 44]) is consistent with a replicated epidemiological association [45, 46, 47].

We report three results that are, to the best of our knowledge, novel. First, we found a genetic correlation of 0.19 (0.04) between between anorexia nervosa and schizophrenia ($p = 1.5 \times 10^{-5}$). Suggestive evidence for comorbidity between anorexia nervosa and schizophrenia has been reported previously [48, 49]. This raises the intriguing possibility of a fundamental similarity between these disorders, despite many differences in clinical course, sex ratios, and response to antipsychotics.

Second, we estimate a genetic correlation of -0.33 (0.08) between ulcerative colitis and childhood obesity ($p = 3.9 \times 10^{-5}$). We suggest that this should be interpreted with caution, given the unusually high LD Score regression for UC (intercept 1.07 compared to mean χ^2 of 1.19). This either indicates substantial confounding due to population stratification or (more likely) that the LD Score regression model is a poor fit for the genetic architecture of UC, which features many low frequency variants with large odds ratios. Third, we estimate a genetic correlation of 0.28 (0.08) ($p = 5 \times 10^{-4}$) between autism spectrum disorder (ASD) and educational attainment (a proxy for intelligence [50]). 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 [51]. The distribution of IQ among individuals with ASD has lower mean than the general population, but with heavy tails [52] (*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 [53]. We are unable to offer an explanation for this result, but propose that further

8

exploration of this genetic correlation is an interesting direction for future research.

Finally, we report several instances where the genetic correlation is close to zero, in contrast to previous reports of epidemiological correlation or genetic overlap. 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 [54]. The absence of genetic correlation between schizophrenia and smoking is notable because of the high prevalence of smoking among individuals with schizophrenia [55]. Our estimate of zero genetic correlation between schizophrenia and plasma lipids or cardiovascular disease contrasts with earlier reports of extensive pleiotropy between schizophrenia and triglycerides [56], which may have been artifacts of inadequate control for LD (see Discussion). 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, it appears that the directions of effect are in the opposite direction as often as they are concordant, yielding close to zero genetic correlation overall.

3   Discussion

We have described a new method for estimating genetic correlation from GWAS summary statistics. This method is an advance for five main reasons: it does not require individual genotypes; it does not require genome-wide significant SNPs; it is not biased by sample overlap; it does not require measuring all phenotypes on all individuals, and it is is computationally trivial. These advantages allow estimation of genetic correlations between a much large set of pairs of phenotypes than was previously possible.

We applied our method to a large dataset of publicly available GWAS summary statistics, spanning 25 traits and more than 1.5 million phenotyped individuals. We replicated many previously-reported GWAS-based genetic correlations, observations of overlap among genome-wide significant SNPs, and known epidemiological associations, thus validating the utility of genetic correlation as an epidemiological tool. In addition, we report several novel results, including a positive genetic correlation between educational attainment and ASD, which may be related to emerging differences between the genetic architectures of high IQ and low IQ ASD.

We note some limitations on the interpretation of genetic correlation. 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. This is best illustrated with an example. Consider the strong negative genetic correlation between HDL and CAD in Figure 2. This genetic correlation could result from a direct causal effect HDL $\rightarrow$ CAD, but could also result from non-causal shared genetic etiology, perhaps mediated by the strong genetic correlation of -0.55 between HDL and triglycerides (TG). 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. Indeed, the second mechanism appears to account for a large proportion of the genetic correlation between HDL and CAD, since the association between HDL and CAD is attenuated after controlling for other plasma lipids [21, 44]. Extending genetic correlation to deal with multiple phenotypes (*e.g.,* genetic correlation between HDL and CAD controlling for TG, analogous to [21, 44], possibly with extension to nonlinear models) is an important direction for future work.

We note several limitations of the LD Score regression method. First, LD Score regression requires large sample sizes in order to give estimates with reasonable standard error (as a rule of thumb, Nh_g^2 should be greater than 3000, where h_g^2 is on the observed scale for case/control studies). Constrained intercept LD Score regression gives lower standard errors, but this estimator requires knowing the number of overlapping samples, which can often be difficult to determine for pairs of large meta-analyses. REML and genetic risk scores are more efficient estimators of genetic correlation, though both require access to individual genotypes. Second, LD Score regression assumes that all individuals in the GWAS were sampled from populations with similar LD Scores, and that these LD Scores can be estimated from sequenced reference data such as 1kG. Differences between LD patterns in the reference panel and the GWAS sample may result in bias [14]. Third, there is currently no procedure for applying LD Score regression to samples from admixed populations. Fourth, LD Score regression performs optimally when applied to traits with highly polygenic genetic architectures, such as psychiatric traits. At large sample size and for less-polygenic traits (*e.g.,* traits where a substantial fraction of heritability is concentrated among a handful of SNPs), it is sometimes possible to obtain better power by analyzing only the large-effect SNPs. 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, while LD Score regression is a consistent estimator of genetic correlation in ascertained case/control studies, the results may nonetheless be skewed by more subtle biases in ascertainment, such as misclassification of cases.

Finally, we have developed a method that allows us to rapidly screen hundreds of pairs of phenotypes for associations without having to measure all phenotypes on all individuals. Our results show that the matrix of genetic correlations – especially among metabolic traits – is dense, and that genome-wide genetic pleiotropy is the rule rather than the exception. These findings highlight the importance of methods that account naturally for pleiotropy [21, 44], or even use pleiotropy to gain power [57, 58].

4 Online Methods

4.1 Statistical Framework

A discussion of modeling assumptions and a derivation of the two-phenotype LD Score regression equation is provided in the Supplementary Note. This section focuses on definitions of quantitates estimated and details of the estimation procedure.

4.2 Heritability

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 denote a phenotype. Define

$$\beta := \mathrm{argmax}_{\alpha \in \mathbb{R}^M} \mathrm{Cor}\left[y, X\alpha\right]^2, \tag{4.1}$$

where the maximization is performed in the population (*i.e.,* at infinite sample size), rather than in a finite sample. 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

$$h_S^2 := \sum_{j=1}^M \beta_j^2. \tag{4.2}$$

We obtain the Yang/Visscher parameter h_g^2 by taking S to be the set of genotyped SNPs. If we let S denote the set of all SNPs in 1000 Genomes Europeans [18], and let S' denote the set of SNPs with MAF$> 5\%$, then

$$h_{\text{5-50\%}}^2 := \sum_{j \in S'} \beta_j^2. \tag{4.3}$$

We choose 5% as the lower bound, because we can estimate LD Scores for 5% SNPs reasonably well from the $N = 387$ samples in 1000 Genomes. Technically, we should write $h_{\text{5-50\%},1kG}^2$ to indicate that we are only accounting for SNPs in 1000 Genomes, but 1000 Genomes has sufficiently good power to observe 5% and higher SNPs that we feel justified in omitting 1kG from the subscript. With larger sample sizes in future sequenced reference panels, this lower bound can be pushed lower.

There are two main distinctions between $h_{\text{5-50\%}}^2$ and h_g^2, First, h_g^2 does not include the effects of common SNPs that are not tagged by the set of genotyped SNPs g. Second, the effects of causal 4% SNPs are not counted towards $h_{\text{5-50\%}}^2$. In practice, neither of these distinctions makes a large difference, since most GWAS arrays focus on common variation and manage to assay or tag almost all common variants.

Estimating $h_{\text{5-50\%}}^2$ involves only as simple modification of LD Score regression: the raw slope from the regression divided by sample size yields an estimate of the average value of β_j^2. If we multiply this number by M, the number of SNPs in 1000 Genomes, then technically we obtain an estimate of the heritability explained by all 1000 Genomes SNPs; however, this interpretation amounts to assuming that our estimate of heritability per SNP obtained from common SNP GWAS data also applies to rare variants. This is unreasonable, since GWAS data contains very little information about rare variants, Therefore, we instead multiply the slope by $M_{\text{5-50\%}}$, the number of 1000 Genomes SNPs with MAF between 5% and 50%, in order to obtain an estimate of $h_{\text{5-50\%}}^2$. The default option in `ldsc` is to estimate $h_{\text{5-50\%}}^2$; this can be overridden with the `--not-M-5-50` flag.

The value of $h^2_{5\text{-}50\%}$ will always be less than the total narrow-sense heritability, h^2, since h^2 takes into account all forms of genetic variation – variants with MAF under 5%, microsatellites, indels, copy number variants – not just common SNPs. In addition, estimates of h^2 from family studies may be biased upwards by non-additive genetic architectures [59].

4.3 Genetic Covariance and Correlation

For this section, we keep the same notation from the previous section, except with two phenotypes y_1 and y_2. Define

$$\beta := \text{argmax}_{\alpha \in \mathbb{R}^M} \text{Cor}[y_1, X\alpha]^2, \tag{4.4}$$

and

$$\gamma := \text{argmax}_{\alpha \in \mathbb{R}^M} \text{Cor}[y_2, X\alpha]^2, \tag{4.5}$$

Then the genetic covariance among SNPs in S is defined

$$\rho_S := \sum_{j \in S} \beta_j \gamma_j, \tag{4.6}$$

Next, let S denote the set of all SNPs in 1000 Genomes Europeans [18], and let S' denote the set of SNPs with MAF$> 5\%$. Then

$$\rho_{5\text{-}50\%} := \sum_{j \in S'} \beta_j \gamma_j. \tag{4.7}$$

The distinctions between these quantities are the same as the distinctions between h^2_g and $h^2_{5\text{-}50\%}$. Rescaling genetic covariance to lie in the range $[-1, 1]$ makes it easier to interpret, so it is more common to instead report genetic correlation,

$$r_S := \frac{\rho_S}{\sqrt{h^2_{S,1} h^2_{S,2}}}. \tag{4.8}$$

The quantity $r_{5\text{-}50\%}$ is defined by replacing S with 5-50% in equation 4.8. As a practical matter, the difference between r_g (the genetic correlation among genotyped SNPs) and $r_{5\text{-}50\%}$ will be quite small when g contains a large proportion of all common SNPs. This is supported by the simulations described in Section 2.3 of the main text. Technically, LD Score regression with HM3 LD Score is an estimator of $r_{HM3(5\text{-}50\%)}$, the genetic correlation among SNPs in HM3 with MAF between 5 and 50%, but the resulting estimates of genetic correlation were almost identical to the estimates of $r_{1kG(5\text{-}50\%)}$ obtained with 1kG LD Scores (Supplementary Tables 5.1, 5.2 and 5.3), which is why we do not emphasize this distinction in the main text.

It is however important to note that all of the flavors of GWAS genetic covariance and correlation ($\rho_g, \rho_{5\text{-}50\%}, r_g$ and $r_{5\text{-}50\%}$) are different from the quantities estimated from family studies. In a family study, the relationship matrix captures information about all genetic variation, not just common SNPs, so family studies attempt to estimate the total narrow-sense genetic covariance and the total narrow-sense genetic correlation. Unlike the relationship between h^2_g or $h^2_{5\text{-}50\%}$ and the total narrow-sense heritability h^2, there is no simple inequality relating r_g and $r_{5\text{-}50\%}$ to the 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 r_g and $r_{5\text{-}50\%}$.

4.4 Genetic Correlation is Different from Pleiotropy

Genetic correlation is defined as the correlation across SNPs of additive per-normalized genotype effect sizes for two traits. Pleiotropy is defined as the tendency for the same variants to have nonzero effects on two or more phenotypes. Genetic correlation is a strictly stronger condition than pleiotropy: to exhibit genetic correlation, it is not sufficient for two phenotypes to be influenced by the same genetic variants, the directions of effects must also be consistently aligned across the genome.

Nevertheless, both quantities are informative. If two phenotypes are influenced by variants at the same loci or in the same pathways, this may well indicate interesting shared biology, even if the direction of the effects do not align. However, quantifying genome-wide pleiotropy from GWAS summary statistics remains an open challenge [60]. For example, since p-values tend to decrease with LD Score [14] and near coding and regulatory regions [61], the observation that low p-values for one phenotype tend to predict low p-values for another (as in [62, 63, 56]) may merely reflect properties shared by almost all pairs of heritable phenotypes, rather than a special property of a specific pair of phenotypes (Supplementary Figures 4.2, 4.3, 4.4). Genetic correlation estimates from LD Score regression are not affected by these concerns, and so may be more easily interpretable than currently-available estimates of genome-wide pleiotropy.

4.5 Two-Phenotype LD Score Regression

The method described in this paper is based on a simple regression equation relating the product of Z-scores of a given SNP from two GWAS's to the LD Score of the SNP, the genetic correlation and the phenotypic correlation. Precisely, let $z_{1,j}$ and $z_{2,j}$ be the Z-scores for a SNP j from two GWAS's, and let ℓ_j be the LD Score of SNP j; *i.e.,* $\ell_j = \sum_k r^2(j,k)$. Then, assuming a simple model of genetic architecture where for each phenotype, SNP effect sizes are drawn in an uncorrelated fashion from distributions with mean zero and a fixed variance and covariance, we have

$$\mathbb{E}[z_{1,j}z_{2,j}] = \frac{\rho_g\sqrt{N_1 N_2}}{M}\ell_j + \frac{\rho N_s}{\sqrt{N_1 N_2}}, \tag{4.9}$$

where N_1 and N_2 are the sample sizes of the two studies, N_s is the number of shared samples, ρ is the overall phenotypic correlation and ρ_g is the genetic covariance. Since sample overlap affects the term $z_{1,j}z_{2,j}$ equally for all SNPs, and the quantity N_s appears only in the intercept term. Equation (4.9) is derived in the Supplementary Note.

We can therefore estimate the genetic covariance, ρ_g, by regressing $z_{1,j}z_{2,j}$ against ℓ_j, and multiplying the resulting slope by $M/\sqrt{N_1 N_2}$. Because sample overlap only affects the intercept of this regression, and the LD Score regression estimator of genetic covariance is based on the slope, the LD Score regression estimator of genetic covariance is not biased by sample overlap. Indeed, if ρ is known (*e.g.,* if both studies assay the same phenotype and $\rho = 1$), the intercept from this regression times a constant can be used as an estimator of the number of shared samples. Alternatively, if both N_s and ρ are known ahead of time, constraining the intercept can substantially reduce the standard error, though this also has the side effect of removing the robustness to shared population stratification. Constraining the intercept is accomplished via the `--no-intercept` or `--constrain-intercept` flags in `ldsc`.

We can estimate heritability using LD Score regression (as described in [14]), and use these heritability estimates to transform the estimates of genetic covariance into estimates of genetic correlation (see Section 4.7 in the Methods).

An equation similar to Equation (4.9) holds if one or both of the studies is an ascertained study of a binary phenotype, and so the same method can be used regardless of whether the Z-scores are from studies of quantitative or case-control traits (Supplementary Note). If the variance of effect sizes depends on minor allele frequency (MAF) or linkage disequilibrium (LD), as discussed in [15, 14], this can introduce model misspecification bias into estimates of heritability and genetic correlation from methods such as LD Score regression and REML. However, we can easily accommodate MAF- and LD-dependent genetic architectures using partitioned LD Score regression, as described in the results and methods sections as well as [16].

4.6 Genetic Covariance Regression Weights

For heritability estimation, we use the LD Score regression weights derived in the Supplementary Note from [14]. If effect sizes for both phenotypes are drawn from a bivariate normal distribution, then the optimal regression weights for genetic covariance estimation are

$$\text{Var}[\hat{\beta}_j \hat{\gamma}_j \,|\, \ell_j] = \left(\frac{h_1^2 \ell_j}{M} + \frac{1}{N_1} \right) \left(\frac{h_2^2 \ell_j}{M} + \frac{1}{N_2} \right) + 2 \left(\frac{\rho_g \ell_j}{M} + \frac{\rho N_s}{N_1 N_2} \right)^2; \tag{4.10}$$

(Supplementary Note) however, this quantity depends on both heritabilities, the genetic covariance and the number of overlapping samples, which are not known a priori, so an approximation is required. In order to obtain approximate regression weights, we use heritability estimates from the single-phenotype LD Score regressions, then we assume that N_s is close enough to zero that the term $\rho N_s / N_1 N_2$ is negligible (this default can be adjusted using the `--overlap` and `--rho` flags in `ldsc`), and estimate a rough genetic covariance (used only for the regression weights) using the aggregate estimator

$$\hat{\rho}_{g,agg} := \frac{1}{\bar{\ell} \sqrt{N_1 N_2}} \sum_{j=1}^{M} z_{1,j} z_{2,j},$$

where $\bar{\ell}$ denotes the mean LD Score among SNPs included in the regression.

Users of our `ldsc` software package should note that when attempting to compute the genetic correlation between a trait and itself using the same GWAS data twice, the result will generally be different from one unless the weights are set appropriately. With the default weights (which are set for zero sample overlap), `ldsc` is simply computing the ratio between the slope of and LD Score regression with efficient weights and the slope of an LD Score regression with inefficient regression weights, which is a noisy estimate of 1.

4.7 Weighted Block Jackknife Genetic Correlation

This section describes the implementation of the `--rg` flag in `ldsc`. Genetic correlation is defined as a ratio of quantities:

$$r_g := \frac{\rho_g}{\sqrt{h_1^2 h_2^2}}.$$

Instead of the naive estimator of this ratio,

$$\hat{r}_g := \frac{\hat{\rho}_g}{\sqrt{\hat{h}_1^2 \hat{h}_2^2}},$$

we use the weighted block jackknife estimator [64] of the ratio, with the jackknife taken over blocks of adjacent SNPs

$$\hat{r}_{g,jack} := n_b \hat{r}_g - \sum_{i=1}^{n_b} \left(1 - \frac{m_i}{M_g}\right) \hat{r}_{g,i} \tag{4.11}$$

where n_b is the number of blocks, and $\hat{r}_{g,i}$ is the naive estimate of genetic correlation obtained by deleting the ith block of SNPs, m_i is the number of SNPs in block i, and M_g is the number of SNPs included in the regression. The weighted block jackknife ratio estimator is less biased than the naive estimate (though this is not so important at the sample sizes that we consider), and comes with a convenient nonparametric heteroskedasticity- and correlation-robust variance estimator [64],

$$\widehat{\mathrm{Var}}\left[\hat{r}_{g,jack}\right] := \frac{1}{n_b} \sum_{i=1}^{n_b} \frac{1}{h_i - 1} \left((h_i - n_b)\hat{r}_g - (h_i - 1)\hat{r}_{g,i} + \sum_{j=1}^{n_b} \left(1 - \frac{m_i}{M_g}\hat{r}_{g,j}\right)\right), \tag{4.12}$$

where $h_i := M_g/m_i$. Weighted block jackknife standard errors (over blocks of adjacent SNPs) are robust to the correlated error structure of GWAS χ^2-statistics, so long as the block size exceeds the typical range of LD. See references [14, 16, 65] for examples of papers in the statistical and population genetics literature that use this technique. We checked the reliability of our standard errors via simulations with real genotypes (Supplementary Table 5.4), and found that the `ldsc` default setting of 2000 blocks genome-wide (which can be adjusted with the `--num-blocks` flag) gives standard error estimates that agree well with the empirical standard deviation across simulation replicates, even though many of the SNPs included in the regression are in LD.

The standard error of the genetic correlation estimate depends not only on sample size but also heritability. As a rule of thumb, the higher the heritability Z-score ($\hat{h}^2/\mathrm{se}(\hat{h}^2)$), the lower the standard error for genetic correlation. This is a general feature of ratio estimators and is not specific to LD Score regression; it is always difficult to obtain an accurate estimate of $1/x$ when the random variable x is close to zero.

In another set of simulations with much lower power (not shown), we observed that the LD Score regression genetic correlation estimates became unstable when either sample size or heritability was so low that at least one of the two heritability estimates was not significantly different from zero. This is a general difficult with attempting to estimate a ratio where the denominator is close to zero, and is not specific to LD Score regression. As a rule of thumb, we recommend discarding any genetic correlation estimates where of the block jackknife SE for the genetic correlation estimate is greater than 0.25. If this occurs, `ldsc` will print an error message by default.

4.8 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.

Practically, estimating LD Scores takes about an hour parallelized over chromosomes, and LD Score regression takes about one minute for each pair of phenotypes on a standard laptop, most of which is spent reading files into memory.

For comparison, REML takes time $\mathscr{O}(MN^2)$ for computing the genetic relatedness matrix (GRM) and $\mathscr{O}(N^3)$ time for maximizing the likelihood.

4.9 Selection of 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, in order to remove traits where the genetic correlation estimates will be 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).

4.10 Summary Statistics

The minimum summary data required for estimating genetic correlation with LD Score regression are the following:

1. Genome-wide summary statistics from cohorts with similar ancestry

2. The summary statistics must be *signed* (allele and direction of effect)

3. The summary statistics should *never* be "corrected" via genomic control (GC) correction. Using GC'ed summary statistics will result in downward bias in the LD Score regression estimates of heritability and genetic covariance, and deflated LD Score regression intercepts. Estimates of genetic correlation will not be affected.

4. The summary statistics should not be meta-analyzed with targeted genotyping at significant loci (*e.g.,* specialty genotyping arrays like immunochip, exome chip, psychchip, metabochip, or replication cohorts). LD Score regression is currently not applicable to data generated using custom genotyping arrays

The next details are convenient, but are only used for filtering SNPs, and so are not strictly necessary:

1. A measure of imputation quality (*e.g.,* INFO) for each SNP

2. Sample size at each SNP (for binary traits, number of cases and number of controls)

3. Sample MAF

Alternatively, one can simply share a list of SNPs passing some MAF and INFO cutoffs (*e.g.,* sample MAF above 3% and INFO above 0.9).

If these data are not available, we recommend retaining only HM3 SNPs with reference panel MAF above 5% for the LD Score regression as a workaround (note: for *regression*, not for estimation

of the LD Scores), since HM3 SNPs seem to be well-imputed in most studies. For newer studies with dense imputation, restricting to HM3 SNPs is an inefficient use of data. Using a larger set of SNPs for the regression will lower the standard error (*e.g.,* using all common 1kG SNPs instead of all common HM3 SNPs for the regression reduces the standard error by about 10% in simulations).

4.11 Huge Effect Loci

Though the derivation of LD Score regression makes no distributional assumptions about effect sizes, the LD Score regression standard error can become very large if effect sizes are drawn from a heavy-tailed distribution, *i.e.,* if there are huge-effect loci. The `ldsc` default is to remove SNPs with $\chi^2 > \max[0.01N, 80]$ (this can be disabled with the `--no-filter-chisq` flag or modified with the `--max-chisq` flag). The resulting genetic correlation estimate can be interpreted as the genetic correlation among SNPs outside of the huge effect loci.

4.12 IGAP

IGAP (which provided the summary statistics for Alzheimer's disease) requests that we include the following text in our methods section:

> *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.

5   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) summay 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`

6   Acknowledgements

7   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.

8   Competing Financial Interests

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

9   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] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Matthew C Keller, Christine E Garver-Apgar, Margaret J Wright, Nicholas G Martin, Robin P Corley, Michael C Stallings, John K Hewitt, and Brendan P Zietsch. The genetic correlation between height and iq: shared genes or assortative mating? *PLoS genetics*, 9(4):e1003451, 2013.

[29] Patrik KE Magnusson, Finn Rasmussen, and Ulf B Gyllensten. Height at age 18 years is a strong predictor of attained education later in life: cohort study of over 950 000 swedish men. *International Journal of Epidemiology*, 35(3):658–663, 2006.

[30] Stefan Johansson, Anastasia Iliadou, Niklas Bergvall, Ulf dé Fairé, Michael S Kramer, Yudi Pawitan, Nancy L Pedersen, Mikael Norman, Paul Lichtenstein, and Sven Cnattingius. The association between low birth weight and type 2 diabetes: contribution of genetic factors. *Epidemiology*, 19(5):659–665, 2008.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] Michael E Benros, Philip R Nielsen, Merete Nordentoft, William W Eaton, Susanne O Dalton, and Preben B Mortensen. Autoimmune diseases and severe infections as risk factors for schizophrenia: a 30-year population-based register study. *American Journal of Psychiatry*, 168(12):1303–1310, 2011.

[39] Michael E Benros, Marianne G Pedersen, Helle Rasmussen, William W Eaton, Merete Nordentoft, and Preben B Mortensen. A nationwide study on the risk of autoimmune diseases in individuals with a personal or a family history of schizophrenia and related psychosis. *American Journal of Psychiatry*, 171(2):218–226, 2014.

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

[41] Bjørn O Åsvold, Johan H Bjørngaard, David Carslake, Maiken E Gabrielsen, Frank Skorpen, George Davey Smith, and Pål R Romundstad. Causal associations of tobacco smoking with cardiovascular risk factors: a mendelian randomization analysis of the hunt study in norway. *International journal of epidemiology*, page dyu113, 2014.

[42] Amy E Taylor, Richard W Morris, Meg E Fluharty, Johan H Bjorngaard, Bjørn Olav Åsvold, Maiken E Gabrielsen, Archie Campbell, Riccardo Marioni, Meena Kumari, Jenni Hällfors, et al. Stratification by smoking status reveals an association of chrna5-a3-b4 genotype with body mass index in never smokers. *PLoS genetics*, 10(12):e1004799, 2014.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

[56] 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.

[57] Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4):407–409, 2014.

[58] 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.

[59] 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.

[60] Nadia Solovieff, Chris Cotsapas, Phil H Lee, Shaun M Purcell, and Jordan W Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495, 2013.

[61] Alexander Gusev, S Hong Lee, Benjamin M Neale, et al. Regulatory variants explain much more heritability than coding.

[62] Andrew J Schork, Wesley K Thompson, Phillip Pham, Ali Torkamani, J Cooper Roddey, Patrick F Sullivan, John R Kelsoe, Michael C O'Donovan, Helena Furberg, Nicholas J Schork, et al. All snps are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated snps. *PLoS Genetics*, 9(4):e1003449, 2013.

[63] 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.

[64] Frank MTA Busing, Erik Meijer, and Rien Van Der Leeden. Delete-m jackknife for unequal m. *Statistics and Computing*, 9(1):3–8, 1999.

[65] Priya Moorjani, Nick Patterson, Joel N Hirschhorn, Alon Keinan, Li Hao, Gil Atzmon, Edward Burns, Harry Ostrer, Alkes L Price, and David Reich. The history of african gene flow into southern europeans, levantines, and jews. *PLoS Genetics*, 7(4):e1001373, 2011.