Published in final edited form as:

Nat Genet. 2015 November ; 47(11): 1236–1241. doi:10.1038/ng.3406.

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], **Po-Ru Loh**[1,5], **ReproGen Consortium**[8], **Psychiatric Genomics Consortium**[8], **Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3**[8], **Laramie Duncan**[1,2,3], **John R.B. Perry**[7], **Nick Patterson**[1], **Elise B. 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 T.H. Chan School of Public Health, Boston, MA, USA

[6]Department of Biostatistics, Harvard T.H. Chan 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

Abstract

Identifying genetic correlations between complex traits and diseases can provide useful etiological insights and help prioritize likely causal relationships. The major challenges preventing estimation of genetic correlation from genome-wide association study (GWAS) data with current methods are the lack of availability of individual genotype data and widespread sample overlap among meta-analyses. We circumvent these difficulties by introducing a technique – cross-trait LD Score regression – for estimating genetic correlation that requires only GWAS summary statistics and is not biased by sample overlap. We use this method to estimate 276 genetic correlations among 24 traits. The results include genetic correlations between anorexia nervosa and schizophrenia, anorexia and obesity and associations between educational attainment and several diseases. These results highlight the power of genome-wide analyses, since there currently are no significantly associated SNPs for anorexia nervosa and only three for educational attainment.

Address correspondence to BBS bulik@broadinstitute.org, BMN bneale@broadinstitute.org, HKF hilaryf@mit.edu and ALP aprice@hsph.harvard.edu.

[8]A list of members and affiliations appears in the Supplementary Note.

*Co-first authors

**Co-last authors

Introduction

Understanding the complex relationships among human traits and diseases is a fundamental goal of epidemiology. Randomized controlled trials and longitudinal studies are time-consuming and expensive, so many potential risk factors are studied using cross-sectional correlations studies at a single time point. Obtaining causal inferences from such studies can be challenging due to issues such as confounding and reverse causation, which can lead to spurious associations and mask the effects of real risk factors [1, 2]. Genetics can help elucidate cause and effect, since inherited genetic risks cannot be subject to reverse causation and are correlated with a smaller list of confounders.

The first methods for testing for genetic overlap were family studies [3, 4, 5, 6, 7]. In order to estimate genetic overlaps among many pairs of phenotypes, family designs require measuring multiple traits on the same individuals. Consequently, it is challenging to scale family designs to a large number of traits, especially traits that difficult or costly to measure (*e.g.*, low-prevalence diseases). More recently, genome-wide association studies (GWAS) have allowed us to obtain effect-size estimates for specific genetic variants, so it is possible to test for shared genetics by looking for correlations in effect-sizes across traits, which does not require measuring multiple traits per individual.

There exists a large class of methods for interrogating genetic overlap via GWAS that focus only on genome-wide significant SNPs. One of the most influential methods in this class is Mendelian randomization, which uses significantly associated SNPs as instrumental variables to attempt quantify causal relationships between risk factors and disease[1, 2]. Methods that focus on significant SNPs are effective for traits where there are many significant associations that account for a substantial fraction of heritability [8, 9]. For many complex traits, heritability is distributed over thousands of variants with small effects, and the proportion of heritability accounted for by significantly associated variants at current sample sizes is small [10]. In such situations, one can often obtain more accurate results by using genome-wide data, rather than just significantly associated variants [11].

A complementary approach is to estimate genetic correlation, which includes the effects of all SNPs, including those that do not reach genome-wide significance (Methods). The two main existing techniques for estimating genetic correlation from GWAS data are restricted maximum likelihood (REML) [11, 12, 13, 14, 15, 16] and polygenic scores [17, 18]. These methods have only been applied to a few traits, because they require individual genotype data, which are difficult to obtain due to informed consent limitations.

In order to overcome these limitations, we have developed a technique for estimating genetic correlation using only GWAS summary statistics that is not biased by sample overlap. Our method, cross-trait LD Score regression, is a simple extension of single-trait LD Score regression [19] and is computationally very fast. We apply this method to data from 24 GWAS and report genetic correlations for 276 pairs of phenotypes, demonstrating shared genetic bases for many complex diseases and traits.

Results

Overview of Methods

The method presented here 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 linkage disequilibrium (LD) with that SNP [19, 20]. For a polygenic trait, SNPs with high LD will have higher χ^2 statistics on average than SNPs with low LD [19]. A similar relationship holds if we replace χ^2 statistics for a single study with the product of *z*-scores from two studies of traits with non-zero genetic correlation.

More precisely, under a polygenic model [11, 13], the expected value of $z_{1j}z_{2j}$ is

$$E[z_{1j}z_{2j}]=\frac{\sqrt{N_1N_2}\rho_g}{M}\ell_j+\frac{\rho N_s}{\sqrt{N_1N_2}}, \quad (1)$$

where N_i is the sample size for study *i*, ρ_g is genetic covariance (defined in Methods), ℓ_j is LD Score [19], N_s is the number of individuals included in both studies, and ρ is the phenotypic correlation among the N_s overlapping samples. We derive this equation in the Supplementary Note. If study 1 and study 2 are the same study, then Equation 1 reduces to the single-trait result from [19], because genetic covariance between a trait and itself is 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 (Methods).

Sample overlap creates spurious correlation between z_{1j} and z_{2j}, which inflates $z_{1j}z_{2j}$. The expected magnitude of this inflation is uniform across all markers, and in particular does not depend on LD Score. As a result, sample overlap only affects the intercept from this regression (the term $\rho N_s/\sqrt{N_1N_2}$) and not the slope, so the estimates of genetic correlation will not be biased by sample overlap. Similarly, shared population stratification will alter the intercept but have minimal impact on the slope, because the correlation between LD Score and the rate of genetic drift is minimal [19]. If we are willing to assume no shared population stratification, and we know the amount of sample overlap and phenotypic correlation in advance (*i.e.*, the true value of $\rho N_s/\sqrt{N_1N_2}$), we can constrain the intercept to this value. We refer to this approach as constrained intercept LD Score regression. Constrained intercept LD Score regression has lower standard error – often by as much as 30% – than LD Score regression with unconstrained intercept, but will yield biased and misleading estimates if the intercept is misspecified, *e.g.*, if we specify the wrong value of $N_s\rho$ or do not completely control for population stratification.

Normalizing genetic covariance by the SNP-heritabilities yields genetic correlation:

$r_g:=\rho_g/\sqrt{h_1^2h_2^2}$, where h_i^2 denotes the SNP-heritability [11] from study *i*. Genetic correlation ranges between −1 and 1. Results similar to Equation 1 hold if one or both studies is a case/control study, in which case genetic covariance is on the observed scale. Details are provided in the Supplementary Note. There is no distinction between observed

and liability scale genetic correlation for case/control traits, so we can define and estimate genetic correlation between a case/control trait and a quantitative trait and genetic correlation between pairs of case/control traits without the need to specify a scale (Supplementary Note).

Simulations

We performed a series of simulations to evaluate the robustness of the model to potential confounders such as sample overlap and model misspecification, and to verify the accuracy of the standard error estimates (Methods).

Table 1 shows cross-trait LD Score regression estimates and standard errors from 1,000 simulations of quantitative traits. For each simulation replicate, we generated two phenotypes for each of 2,062 individuals in our sample by drawing effect sizes approximately 600,000 SNPs on chromosome 2 from a bivariate normal distribution. We then computed summary statistics for both phenotypes and estimated heritability and genetic correlation with cross-trait LD Score regression. The summary statistics were generated from completely overlapping samples. Results are shown in Table 1. These simulations confirm that cross-trait LD Score regression yields accurate estimates of the true genetic correlation and that the standard errors match the standard deviation across simulations. Thus, cross-trait LD Score regression is not biased by sample overlap, in contrast to estimation of genetic correlation via polygenic risk scores, which is biased in the presence of sample overlap [18]. We also evaluated simulations with one quantitative trait and one case/control study and show that cross-trait LD Score regression can be applied to binary traits and is not biased by oversampling of cases (Supplementary Table 1).

Estimates of heritability and genetic covariance can be biased if the underlying model of genetic architecture is misspecified, *e.g.*, if variance explained is correlated with LD Score or MAF [19, 21]. Because genetic correlation is estimated as a ratio, it is more robust; biases that affect the numerator and the denominator in the same direction tend to cancel. We obtain approximately correct estimates of genetic correlation even in simulations with models of genetic architecture where our estimates of heritability and genetic covariance are biased (Supplementary Table 2).

Replication of Pyschiatric Cross-Disorder Results

As technical validation, we replicated the estimates of genetic correlations among psychiatric disorders obtained with individual genotypes and REML in [14], by applying cross-trait LD Score regression to summary statistics from the same data [22]. These summary statistics were generated from non-overlapping samples, so we applied cross-trait LD Score regression using both unconstrained and constrained intercepts (Methods). Results from these analyses are shown in Figure 1. The results from cross-trait LD Score regression were similar to the results from REML. Cross-trait LD Score regression with constrained intercept gave standard errors that were only slightly larger than those from REML, while the standard errors from cross-trait LD Score regression with intercept were substantially larger, especially for traits with small sample sizes (*e.g.*, ADHD, ASD).

Application to Summary Statistics From 25 Phenotypes

We used cross-trait LD Score regression to estimate genetic correlations among 24 phenotypes (URLs, Methods). Genetic correlation estimates for all 276 pairwise combinations of the 24 traits are shown in Figure 2. For clarity of presentation, the 24 phenotypes were restricted to contain only one phenotype from each cluster of closely related phenotypes (Methods). Genetic correlations among the educational, anthropometric, smoking, and insulin-related phenotypes that were excluded from Figure 2 are shown Supplementary Figures 1, 2, 3 and 4, respectively. A full table of 1176 genetic correlations among 49 traits is provided in Supplementary Table 4. References and sample sizes are shown in Supplementary Table 3.

The first section of Table 2 lists genetic correlation results that are consistent with epidemiological associations, but, as far as we are aware, have not previously been reported using genetic data. The estimates of the genetic correlation between age at menarche and adult height [29], triglycerides [30] and type 2 diabetes [30, 31] are consistent with the epidemiological associations. The estimate of a negative genetic correlation between anorexia nervosa and obesity suggests that the same genetic factors influence normal variation in BMI as well as dysregulated BMI in psychiatric illness. This result is consistent with the observation that BMI GWAS findings implicate neuronal, rather than metabolic, cell-types and epigenetic marks [32, 33]. The negative genetic correlation between adult height and coronary artery disease agrees with a replicated epidemiological association [34, 35, 36]. We observe several significant associations with the educational attainment phenotypes from Rietveld *et al.* [37]: we estimate a statistically significant negative genetic correlation between college and Alzheimer's disease, which agrees with epidemiological results [38, 39]. The positive genetic correlation between college and bipolar disorder is consistent with previous epidemiological reports [40, 41]. The estimate of a negative genetic correlation between smoking and college is consistent with the observed differences in smoking rates as a function of educational attainment [42].

The second section of Table 2 lists three results that are, to the best of our knowledge, new both to genetics and epidemiology. One, 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 [43, 44], and this result raises the possibility of similarity between these classes of disease. Two, we estimate a negative genetic correlation between ulcerative colitis (UC) and childhood obesity. The relationship between premorbid BMI and ulcerative colitis is not well-understood; exploring this relationship may be a fruitful direction for further investigation. Three, we estimate a positive genetic correlation between autism spectrum disorder (ASD) and educational attainment (which has very high genetic correlation with IQ [37, 45, 46]). The ASD summary statistics were generated using a case-pseudocontrol study design, so this result cannot be explained by oversampling of ASD cases from the more highly educated parents, which is observed epidemiologically [47]. The distribution of IQ among individuals with ASD has lower mean than the general population, but with heavy tails [48] (*i.e.*, an excess of individuals with low and high IQ). There is also emerging evidence that the genetic architecture of ASD varies across the IQ distribution [49].

The third section of Table 2 lists interesting examples where the genetic correlation is close to zero with small standard error. The low genetic correlation between schizophrenia and rheumatoid arthritis is interesting because schizophrenia has been observed to be protective for rheumatoid arthritis [50], though the epidemiological effect is weak, so it is possible that there is a real genetic correlation, but it is too small for us to detect. The low genetic correlation between schizophrenia and smoking is notable because of the increased tobacco use (both prevalence and number of cigarettes per day) among individuals with schizophrenia [51]. The low genetic correlation between schizophrenia and plasma lipid levels contrasts with a previous report of pleiotropy between schizophrenia and triglycerides [52]. Pleiotropy (unsigned) is different from genetic correlation (signed; see Methods); however, the pleiotropy reported by Andreassen, *et al.* [52] could be explained by the sensitivity of the method used to the properties of a small number of regions with strong LD, rather than trait biology (Supplementary Figure 5). We estimate near-zero genetic correlation between Alzheimer's disease and schizophrenia. The genetic correlations between Alzheimers disease and the other 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 (RA) and both Crohn's disease (CD) and UC. Although these diseases share many associated loci [53, 54], there appears to be no directional trend: some RA risk alleles are also risk alleles for UC and CD, but many RA risk alleles are protective for UC and CD [53], yielding near-zero genetic correlation. This example highlights the distinction between pleiotropy and genetic correlation (Methods).

Finally, the estimates of genetic correlations among metabolic traits are consistent with the estimates obtained using REML in Vattikuti *et al.* [15] (Supplementary Table 6), and are directionally consistent with the recent Mendelian randomization results from Wuertz *et al.* [55]. The estimate of 0.54 (0.07) for the genetic correlation between CD and UC is consistent with the estimate of 0.62 (0.04) from Chen *et al.* [16].

Discussion

We have described a new method for estimating genetic correlation from GWAS summary statistics, which we applied to a dataset of GWAS summary statistics consisting of 24 traits and more than 1.5 million unique phenotype measurements. We reported several new findings that would have been difficult to obtain with existing methods, including a positive genetic correlation between anorexia nervosa and schizophrenia. Our method replicated many previously-reported GWAS-based genetic correlations, and confirmed observations of overlap among genome-wide significant SNPs, MR results and epidemiological associations.

This method is an advance for several reasons: it does not require individual genotypes, genome-wide significant SNPs or LD-pruning (which loses information if causal SNPs are in LD). Our method is not biased by sample overlap and is computationally fast. Furthermore, our approach does not require measuring multiple traits on the same individuals, so it scales easily to studies of thousands of pairs of traits. These advantages

allow us to estimate genetic correlation for many more pairs of phenotypes than was possible with existing methods.

The challenges in interpreting genetic correlation are similar to the challenges in MR. We highlight two difficulties. First, genetic correlation is immune to environmental confounding, but is subject to genetic confounding, analogous to confounding by pleiotropy in MR. For example, the genetic correlation between HDL and CAD in Figure 2 could result from a causal effect *HDL→CAD*, but could also be mediated by triglycerides (TG) [9, 56], represented graphically [57] as *HDL←G→TG→CAD*, where *G* is the set of genetic variants with effects on both HDL and TG. Extending genetic correlation to multiple genetically correlated phenotypes is an important direction for future work [58]. Second, although genetic correlation estimates are not biased by oversampling of cases, they are affected by other forms of biased sampling, such as misclassification [14] and case/control/covariate sampling (*e.g.*, a BMI-matched study of T2D).

We note several limitations of cross-trait LD Score regression as an estimator of genetic correlation. First, cross-trait LD Score regression requires larger sample sizes than methods that use individual genotypes in order to achieve equivalent standard error. Second, cross-trait LD Score regression is not currently applicable to samples from recently-admixed populations. Third, we have not investigated the potential impact of assortative mating on estimates of genetic correlation, which remains as a future direction. Fourth, methods built from polygenic models, such as cross-trait LD Score regression and REML, are most effective when applied to traits with polygenic genetic architectures. For traits where significant SNPs account for a sizable proportion of heritability, analyzing only these SNPs can be more powerful. Developing methods that make optimal use of both large-effect SNPs and diffuse polygenic signal is a direction for future research.

Despite these limitations, we believe that the cross-trait LD Score regression estimator of genetic correlation will be a useful addition to the epidemiological toolbox, because 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

All definitions refer to narrow-sense heritabilities and genetic covariances. Let *S* denote a set of *M* SNPs, let *X* denote a vector of additively (0-1-2) coded genotypes for the SNPs in *S*, and let y_1 and y_2 denote phenotypes. Define $\beta := \mathit{argmax}_{\alpha \in R^M} \mathit{Cor}[y_1, X\alpha]$, where the maximization is performed in the population (*i.e.*, in the infinite data limit). Let γ denote the corresponding vector for y_2. This is a projection, so β is unique modulo SNPs in perfect LD. Define h_S^2, the heritability explained by SNPs in *S*, as $h_S^2(y_1) := \sum_j \beta_j^2$ and $\rho_S(y_1, y_2)$, the genetic covariance among SNPs in *S*, as $\rho_S(y_1, y_2) := \sum_{j \in S} \beta_j \gamma_j$. The genetic correlation among SNPs in *S* is $r_S(y_1, y_2) := \rho_S(y_1, y_2) / \sqrt{h_S^2(y_1) h_S^2(y_2)}$, which lies in [−1,1].

Following [11], we use subscript g (as in h^2_g, ρ_g,r_g) when the set of SNPs is genotyped and imputed SNPs in GWAS.

SNP genetic correlation (r_g) is different from family study genetic correlation. In a family study, the relationship matrix captures information about all genetic variation, not just common SNPs. As a result, family studies estimate the total genetic correlation (S equals all variants). Unlike the relationship between SNP-heritability [11] and total heritability, for which $h^2_g \leq h^2$, no similar relationship holds between SNP genetic correlation and total genetic correlation. If β and γ are more strongly correlated among common variants than rare variants, then the total genetic correlation will be less than the SNP genetic correlation.

Genetic correlation is (asymptotically) proportional to Mendelian randomization estimates. If we use a genetic instrument $g_i := \sum_{j \in S} X_{ij}\beta_j$ to estimate the effect b_{12} of y_1 on y_2, the 2SLS estimate is $\hat{b_{2SLS}} := g^T y_2 / g^T y_1$ [59]. The expectations of the numerator and denominator are $E[g^T y_2] = \rho_S(y_1, y_2)$ and $E[g^T y_1] = h^2_S(y_1)$. Thus, $plim_{N\to\infty} \hat{b}_{2SLS} = r_S(y_2, y_1)\sqrt{h^2_S(y_1)/h^2_S(y_2)}$. If we use the same set S of SNPs to estimate b_{12} and b_{21} (*e.g.*, if S is the set of all common SNPs, as in the genetic correlation analyses in this paper), then this procedure is symmetric in y_1 and y_2.

Genetic correlation is different from pleiotropy. Two traits have a pleiotropic relationship if many variants affect both. Genetic correlation is a stronger condition than pleiotropy: to exhibit genetic correlation, the directions of effect must also be consistently aligned.

Cross-Trait LD Score Regression

Recall from the Overview of Methods that the cross-trait LD Score regression equation is

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

where z_{ij} denotes the z-score for study i and SNPj, N_i is the sample size for study i, ρ_g is genetic covariance, ℓ_j is LD Score [19], N_s is the number of individuals included in both studies, and ρ is the phenotypic correlation among the N_s overlapping samples. We derive this equation in the Supplementary Note. We estimate genetic covariance 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% (technically, this is an estimate of the genetic covariance among SNPs with 5–50% MAF; Supplementary Note).

If we know the correct value of the intercept term $\rho N_s\sqrt{N_1 N_2}$ ahead of time, we can reduce the standard error by constraining the intercept to this value using the -constrain-intercept flag in ldsc (for pairs of binary traits, we give a corresponding expression in terms

of the number of overlapping cases and controls in the Supplementary Note). Note that this works even when there is known nonzero sample overlap

We recommend using the in-sample estimate of ρ (denoted $\hat{\rho}$), rather than the population value of ρ. Under unbiased sampling $\hat{\rho}$ is consistent for ρ with $O(1/N)$ variance, so in this case, the distinction between ρ and $\hat{\rho}$ is not of great importance. Under biased sampling (as discussed in the previous section), the expected LD Score regression intercept depends on the expected sample correlation $E[y_{i1}y_{i2}|s=1]$ (which is estimated consistently by $\hat{\rho}$), not population ρ. Thus, we advise to use $\hat{\rho}$ rather than ρ when constraining the intercept.

Regression Weights

For heritability estimation, we use the regression weights from [19]. If effect sizes for both phenotypes are drawn from a bivariate normal distribution, then the optimal regression weights for genetic covariance estimation are

$$Var[z_{1j}z_{2j}|\ell_j] = \left(\frac{N_1\ell_j}{M}+1\right)\left(\frac{N_2\ell_j}{M}+1\right)+\left(\frac{\sqrt{N_1N_2}\rho_g}{M}\ell_j+\frac{\rho N_s}{\sqrt{N_1N_2}}\right)^2 \quad (3)$$

(Supplementary Note). This quantity depends on several parameters ($h_1^2, h_2^2, \rho_g, \rho, N_s$) which are not known a priori, so it is necessary to estimate them from the data. We compute the weights in two steps:

1. The first regression is weighted using heritabilities from the single-trait LD Score regressions, $\rho N_s=0$, and ρ_g estimated as $\hat{\rho}_g := (\sqrt{N_1N_2})^{-1}\sum_j z_{1j}z_{2j}$.
2. The second regression is weighted using the estimates of ρN_s and ρ_g from step 1. The genetic covariance estimate that we report is the estimate from the second regression.

Linear regression with weights estimated from the data is called feasible generalized least squares (FGLS). FGLS has the same limiting distribution as WLS with optimal weights, so WLS *p*-values are valid for FGLS [59]. We multiply the heteroskedasticity weights by $1/\ell_j$ (where ℓ_j is LD Score with sum over regression SNPs) in order to downweight SNPs that are overcounted. This is a heuristic: the optimal approach is to rotate the data so that it is de-correlated, but this rotation matrix is difficult to compute.

Two-Step Estimator

As noted in [19], SNPs with very large effect sizes can result in large LD Score regression standard errors for single-trait LD Score regression with unconstrained intercept; cross-trait LD Score regression with unconstrained intercept behaves similarly. This is due to the well-known fact that linear regression deals poorly with outliers in the response variable (LD Score regression with constrained intercept is not nearly as adversely affected by large-effect SNPs). The solution proposed in [19] was to remove SNPs with $\chi^2>80$ from the LD Score regression. This is a satisfactory solution when the goal is to estimate the LD Score regression intercept. If the goal is to distinguish polygenicity from population stratification,

and we are willing to assume that the population stratification is subtle, such that SNPs with χ^2>80 are much more likely to be real causal SNPs rather than artifacts, then we can make the task much easier by removing those SNPs. However, this is unsatisfactory if the goal is to estimate h^2: ignoring large-effect SNPs with χ^2>80 would bias estimates of h^2 and ρ_g towards zero. Therefore, for estimating h^2 or ρ_g, we take a two step approach. The first step is to estimate the LD Score regression intercept with all SNPs with χ^2>30 removed (*i.e.*, all genome-wide significant SNPs; the threshold can be adjusted with the -two-step flag in ldsc). The second step is to estimate h^2 or ρ_g using all SNPs and constrained intercept LD Score regression with the intercept constrained to the value from the first step (note that we account for uncertainty in the intercept when computing a standard error; see the next section).

Assessment of Statistical Significance via Block Jackknife

Summary statistics for SNPs in LD are correlated, so the OLS standard error will be biased downwards. We estimate a heteroskedasticity-and-correlation-robust standard error with a block jackknife over blocks of adjacent SNPs. This is the same procedure used in [19], and gives accurate standard errors in simulations (Table 1). We obtain a standard error for the genetic correlation by using a ratio block jackknife over SNPs. The default setting in ldsc is 200 blocks per genome, which can be adjusted with the -num-blocks flag.

For the two-step estimator, if we were to estimate the intercept in the first step, then obtain a jackknife standard error for the second step treating the intercept as fixed, the standard error would be biased downwards, because it would not take into account the uncertainty in the intercept. Instead, we jackknife both steps of the procedure, which appropriately accounts for uncertainty in the intercept and yields a valid standard error.

Reverse Causation

Consider a scenario where a risk factor E_1 causes a disease D, but incidence of disease D changes postmorbid levels of E_1 (this could occur *e.g.*, incidence of disease persuades affected individuals to change their behavior in ways that lower E_1). If D is sufficiently common in our GWAS sample, then the genetic correlation may be affected by reverse causation. LD Score regression (or any genetic correlation estimator) will yield a consistent estimate of the cross-sectional genetic correlation between E_1 and D at the given timepoint; however, the cross-sectional genetic correlation between E_1 and D will be attenuated relative to the genetic correlation between disease and *pre-morbid* levels of E_1. The genetic correlation between disease and pre-morbid levels of the risk factor will typically be the more interesting quantity to estimate, because it is more closely related to the causal effect of E_1 on D. We can estimate this quantity by excluding all post-morbid measurements of the risk factor from the risk factor GWAS. This allows us to circumvent reverse causation, at the cost of a small decrease in sample size. If D is uncommon, then modification of behavior after onset of D will account for only a small fraction of the population variance in E_1, so the effect of reverse causation on the genetic correlation will be small. Thus, reverse causation is primarily a concern for high-prevalence diseases.

Non-Random Ascertainment

We show in the Supplementary Note that LD Score regression is robust to oversampling of cases in case/control studies, modulo transformation observed and liability scale heritability and genetic covariance. Oversampling of cases is the most common form of biased sampling, but there are many other forms of biased sampling. For example, consider case/control/covariate ascertainment, where the sampling of cases and controls takes into account a covariate. As as concrete example, we know that high BMI is a major risk factor for T2D. If we wish to discover genetic variants that influence risk for T2D via mechanisms other than BMI, we may wish to perform a case/control study for T2D where we compare BMI-matched cases and controls. If we were to use such a T2D study and a random population study of BMI to compute the genetic correlation between BMI and T2D, the result would be substantially attenuated relative to the population genetic correlation between T2D and BMI. (Note that this example holds irrespective of whether there is sample overlap and applies to all genetic correlation estimators, not just LD Score).

More generally, let s_i=1 denote the event that individual i is selected into our study, and let C_i denote a vector of covariates describing individual i (which may include the phenotype of individual i). Then we can represent an arbitrary biased sampling scheme by specifying the selection probabilities $f(C_i):=P[s_i=1|C_i]$ (note that case/control ascertainment is the special case where $C_i=y_i$). Suppose that phenotypes are generated following the model from Section 1.1 of the Supplementary Note, but that our sample is selected following the biased sampling scheme f. Let a_{ij} denote the additive genetic component for phenotype j in inidividual i. If there is no direct ascertainment on genotype (*i.e.*, if C_i does not include genotypes), then the proof of Proposition 1 in the Supplementary Note goes through, except that ρ is replaced with $E[y_{i1}y_{i2}|s_i=1]$ and ρ_g is replaced with $E[a_{i1}a_{i2}|s_i=1]$.

This has two practical implications: first, in studies with biased sampling schemes and sample overlap, if one wishes to constrain the intercept, one should use the sample correlation between phenotypes $\hat{\rho}$ rather than the population correlation ρ. Under biased sampling, $plim_{N\to\infty}\hat{\rho}=E[y_{i1}y_{i2}|s_i=1]$, which is typically not equal to ρ. Second, even if there is no sample verlap, biased sampling can affect the genetic correlation estimate. If the biased sampling mechanism (*i.e.*, the function $f(C_i):=P[s_i=1|C_i]$) is known, then it may be possible to explicitly model the biased sampling and derive a function for converting genetic correlation estimates from the biased sample to population genetic correlations (similar to the derivations in sections 1.3 and 1.4 of the Supplementary Note). If the biased sampling mechanism can only be described qualitatively, then it should at least be possible to guess the magnitude and direction of the bias by reasoning about $E[y_{i1}y_{i2}|s_i=1]$ and $E[a_{i1}a_{i2}|s_i=1]$.

Computational Complexity

Let N denote sample size and M the number of SNPs. The computational complexity of the steps involved in LD Score regression are as follows:

1. Computing summary statistics takes $O(MN)$ time.
2. Computing LD Scores takes $O(MN)$ time, though the N for computing LD Scores need not be large. We use the N=378 Europeans from 1000 Genomes.

3. LD Score regression takes $O(M)$ time and space.

For a user who has already computed summary statistics and downloads LD Scores from our website (URLs), the computational cost of LD Score regression is $O(M)$ time and space. For comparison, REML takes time $O(MN^2)$ for computing the GRM and $O(N^3)$ time for maximizing the likelihood.

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

Simulations

We simulated quantitative traits under an infinitesimal model in 2062 controls from a Swedish study. To simulate the standard scenario where many causal SNPs are not genotyped, we simulated phenotypes by drawing causal SNPs from 622,146 best-guess imputed 1000 Genomes SNPs on chromosome 2, then retained only the 90,980 HM3 SNPs with MAF above 5% for LD Score regression.

We note that the simulations in [19] show that single-trait LD Score regression is only minimally biased by uncorrected population stratification and moderate ancestry mismatch between the reference panel used for estimating LD Scores and the population sampled in GWAS. In particular, LD Scores estimated from the 1000 Genomes reference panel are suitable for use with European-ancestry meta-analyses. Put another way, LD Score is only minimally correlated with F_{ST}, and the differences in LD Score among European populations are not so large as to bias LD Score regression. Since we use the same LD Scores for cross-trait LD Score regression as for single-trait LD Score regression, these results extend to cross-trait LD Score regression.

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.
2. Remove studies that do not provide signed summary statistics.
3. Remove studies not imputed to at least HapMap 2.
4. Remove studies that adjust for heritable covariates [60].
5. Remove all traits with heritability z-score below 4. Genetic correlation estimates for traits with heritability z-score below 4 are generally too noisy to report.
6. Prune clusters of correlated phenotypes (*e.g.*, obesity classes 1–3) by picking the trait from each cluster with the highest heritability heritability z-score.

We then applied the following filters (implemented in the script munge_sumstats.py included with ldsc):

1. For studies that provide a measure of imputation quality, filter to INFO above 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 [61] SNPs with 1000 Genomes EUR MAF above 5%, 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 0.67 times the 90th percentile of sample size.
5. For specialty chip (*e.g.*, metabochip) meta-analyses, remove SNPs with *N* above the maximum GWAS *N*.
6. Remove indels and structural variants.
7. Remove strand-ambiguous SNPs.
8. Remove SNPs whose alleles do not match the alleles in 1000 Genomes.

Genomic control (GC) correction at any stage biases the heritability and genetic covariance estimates downwards (see the Supplementary Note of [19]. The biases in the numerator and denominator of genetic correlation cancel exactly, so genetic correlation is not biased by GC correction. A majority of the studies analyzed in this paper used GC correction, so 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank P. Sullivan, C. Bulik, S. Caldwell, C. Arabica, O. Andreassen for helpful comments. This work was supported by NIH grants R01 MH101244 (ALP), R01 HG006399 (NP), R03 CA173785 (HKF) and by the Fannie and John Hertz Foundation (HKF).

Data on anorexia nervosa were obtained by funding from the WTCCC3 WT088827/Z/09 titled “A genome-wide association study of anorexia nervosa”.

Data on glycaemic traits have been contributed by MAGIC investigators and have been downloaded from www.magicinvestigators.org.

Data on coronary artery disease/myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from www.CARDIOGRAMPLUSC4D.ORG

We thank the International Genomics of Alzheimer’s Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. The i-Select chips was funded by the French National Foundation on Alzheimer’s disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Université de Lille 2 and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant 503480), Alzheimer’s Research UK (Grant 503176), the Wellcome Trust (Grant 082604/2/07/Z) and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01-AG-12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer’s Association grant ADGC-10-196728.

References

1. Smith, George Davey; Ebrahim, Shah. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? International journal of epidemiology. 2003; 32(1):1–22. [PubMed: 12689998]
2. Smith, George Davey; Hemani, Gibran. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human molecular genetics. 2014; 23(R1):R89–R98. [PubMed: 25064373]
3. Vandenberg SG. Multivariate analysis of twin differences. Methods and goals in human behavior genetics. 1965:29–43.
4. Kempthorne, Oscar; Osborne, Richard H. The interpretation of twin data. American journal of human genetics. 1961; 13(3):320. [PubMed: 13752449]
5. Loehlin, John C.; Vandenberg, Steven Gerritjan. Genetic and environmental components in the covariation of cognitive abilities: An additive model. Louisville Twin Study, University of Louisville; 1966.
6. Neale, Michael; Cardon, Lon. Methodology for genetic studies of twins and families. Springer; 1992.
7. Lichtenstein, Paul, et al. Common genetic determinants of schizophrenia and bipolar disorder in swedish families: a population-based study. The Lancet. 2009; 373(9659):234–239.
8. Voight, Benjamin F., et al. Plasma hdl cholesterol and risk of myocardial infarction: a mendelian randomisation study. The Lancet. 2012; 380(9841):572–580.
9. Ron, Do, et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. Nature genetics. 2013; 45(11):1345–1352. [PubMed: 24097064]
10. Visscher, Peter M.; Brown, Matthew A.; McCarthy, Mark I.; Yang, Jian. Five years of gwas discovery. The American Journal of Human Genetics. 2012; 90(1):7–24. [PubMed: 22243964]
11. Yang, Jian, et al. Common snps explain a large proportion of the heritability for human height. Nature Genetics. 2010; 42(7):565–569. [PubMed: 20562875]
12. Yang, Jian; Hong Lee, S.; Goddard, Michael E.; Visscher, Peter M. Gcta: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics. 2011; 88(1):76–82. [PubMed: 21167468]
13. Lee, Sang Hong; Yang, Jian; Goddard, Michael E.; Visscher, Peter M.; Wray, Naomi R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012; 28(19):2540–2542. [PubMed: 22843982]

14. Cross-Disorder Group of the Psychiatric Genomics Consortium et al. Genetic relationship between five psychiatric disorders estimated from genome-wide snps. Nature Genetics. 2013
15. Vattikuti, Shashaank; Guo, Juen; Chow, Carson C. Heritability and genetic correlations explained by common snps for metabolic syndrome traits. PLoS genetics. 2012; 8(3):e1002637. [PubMed: 22479213]
16. Chen, Guo-Bo, et al. Estimation and partitioning of (co) heritability of inflammatory bowel disease from gwas and immunochip data. Human molecular genetics. 2014:ddu174.
17. Purcell, Shaun M., et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460(7256):748–752. [PubMed: 19571811]
18. Dudbridge, Frank. Power and predictive accuracy of polygenic risk scores. PLoS genetics. 2013; 9(3):e1003348. [PubMed: 23555274]
19. Bulik-Sullivan, Brendan, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics. 2015
20. Yang, Jian, et al. Genomic inflation factors under polygenic inheritance. European Journal of Human Genetics. 2011; 19(7):807–812. [PubMed: 21407268]
21. Speed, Doug; Hemani, Gibran; Johnson, Michael R.; Balding, David J. Improved heritability estimation from genome-wide snps. The American Journal of Human Genetics. 2012; 91(6):1011–1021. [PubMed: 23217325]
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. 2013; 381(9875):1371. [PubMed: 23453885]
23. Perry, John RB., et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature. 2014; 514(7520):92–97. [PubMed: 25231870]
24. Morris, Andrew P., et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics. 2012; 44(9):981. [PubMed: 22885922]
25. Horikoshi, Momoko, et al. New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. Nature genetics. 2013; 45(1):76–82. [PubMed: 23202124]
26. Freathy, Rachel M., et al. Type 2 diabetes risk alleles are associated with reduced size at birth. Diabetes. 2009; 58(6):1428–1433. [PubMed: 19228808]
27. Early Growth Genetics (EGG) Consortium et al. A genome-wide association meta-analysis identifies new childhood obesity loci. Nature genetics. 2012; 44(5):526–531. [PubMed: 22484627]
28. Rob Taal H, et al. Common variants at 12q15 and 12q24 are associated with infant head circumference. Nature genetics. 2012; 44(5):532–538. [PubMed: 22504419]
29. Onland-Moret NC, et al. Age at menarche in relation to adult height the epic study. American journal of epidemiology. 2005; 162(7):623–632. [PubMed: 16107566]
30. Day, Felix, et al. Puberty timing associated with diabetes, cardiovascular disease and also diverse health outcomes in men and women: the uk biobank study. Scientific Reports. 2014
31. Elks, Cathy E., et al. Age at menarche and type 2 diabetes risk the epic-interact study. Diabetes care. 2013; 36(11):3526–3534. [PubMed: 24159179]
32. Finucane, Hilary K., et al. Partitioning heritability by functional category using GWAS summary statistics. In Press at Nature Genetics. 2015
33. Sadaf Farooqi I. Defining the neural basis of appetite and obesity: from genes to behaviour. Clinical Medicine. 2014; 14(3):286–289. [PubMed: 24889574]
34. Wang, Na, et al. 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. 2011; 40(6):1715–1726. [PubMed: 22268239]
35. Hebert, Patricia R., et al. Height and incidence of cardiovascular disease in male physicians. Circulation. 1993; 88(4):1437–1443. [PubMed: 8403290]
36. Rich-Edwards, Janet W., et al. Height and the risk of cardiovascular disease in women. American journal of epidemiology. 1995; 142(9):909–917. [PubMed: 7572971]

37. Rietveld, Cornelius A., et al. Gwas of 126,559 individuals identifies genetic variants associated with educational attainment. Science. 2013; 340(6139):1467–1471. [PubMed: 23722424]
38. Barnes, Deborah E.; Yaffe, Kristine. The projected effect of risk factor reduction on alzheimer's disease prevalence. The Lancet Neurology. 2011; 10(9):819–828. [PubMed: 21775213]
39. Norton, Sam; Matthews, Fiona E.; Barnes, Deborah E.; Yaffe, Kristine; Brayne, Carol. Potential for primary prevention of alzheimer's disease: an analysis of population-based data. The Lancet Neurology. 2014; 13(8):788–794. [PubMed: 25030513]
40. MacCabe, James H., et al. Excellent school performance at age 16 and risk of adult bipolar disorder: national cohort study. The British Journal of Psychiatry. 2010; 196(2):109–115. [PubMed: 20118454]
41. Tiihonen, Jari, et al. Premorbid intellectual functioning in bipolar disorder and schizophrenia: results from a cohort study of male conscripts. American Journal of Psychiatry. 2005; 162(10): 1904–1910. [PubMed: 16199837]
42. Pierce, John P.; Fiore, Michael C.; Novotny, Thomas E.; Hatziandreu, Evridiki J.; Davis, Ronald M. Trends in cigarette smoking in the united states: educational differences are increasing. Jama. 1989; 261(1):56–60. [PubMed: 2908995]
43. Striegel-Moore, Ruth H.; Garvin, Vicki; Dohm, Faith-Anne; Rosenheck, Robert A. Psychiatric comorbidity of eating disorders in men: a national study of hospitalized veterans. International Journal of Eating Disorders. 1999; 25(4):399–404. [PubMed: 10202650]
44. Blinder, Barton J.; Cumella, Edward J.; Sanathara, Visant A. Psychiatric comorbidities of female inpatients with eating disorders. Psychosomatic Medicine. 2006; 68(3):454–462. [PubMed: 16738079]
45. Deary, Ian J.; Strand, Steve; Smith, Pauline; Fernandes, Cres. Intelligence and educational achievement. Intelligence. 2007; 35(1):13–21.
46. Calvin, Catherine M.; Fernandes, Cres; Smith, Pauline; Visscher, Peter M.; Deary, Ian J. Sex, intelligence and educational achievement in a national cohort of over 175,000 11-year-old schoolchildren in england. Intelligence. 2010; 38(4):424–432.
47. Durkin, Maureen S., et al. Socioeconomic inequality in the prevalence of autism spectrum disorder: evidence from a us cross-sectional study. PLoS One. 2010; 5(7):e11551. [PubMed: 20634960]
48. Robinson, Elise B., et al. Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. Proceedings of the National Academy of Sciences. 2014; 111(42): 15161–15165.
49. Samocha, Kaitlin E., et al. A framework for the interpretation of de novo mutation in human disease. Nature genetics. 2014; 46(9):944–950. [PubMed: 25086666]
50. Silman, Alan J.; Pearson, Jacqueline E. Epidemiology and genetics of rheumatoid arthritis. Arthritis Res. 2002; 4(Suppl 3):S265–S272. [PubMed: 12110146]
51. de Leon, Jose; Diaz, Francisco J. A meta-analysis of worldwide studies demonstrates an association between schizophrenia and tobacco smoking behaviors. Schizophrenia research. 2005; 76(2):135–157. [PubMed: 15949648]
52. Andreassen, Ole A., et al. Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. The American Journal of Human Genetics. 2013; 92(2):197–209. [PubMed: 23375658]
53. Cotsapas, Chris, et al. Pervasive sharing of genetic effects in autoimmune disease. PLoS genetics. 2011; 7(8):e1002254. [PubMed: 21852963]
54. Farh, Kyle Kai-How, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2014
55. Wurtz, Peter, et al. Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change. PLoS Medicine. 2014
56. Burgess, Stephen; Freitag, Daniel F.; Khan, Hassan; Gorman, Donal N.; Thompson, Simon G. Using multivariable mendelian randomization to disentangle the causal effects of lipid fractions. PloS one. 2014; 9(10):e108891. [PubMed: 25302496]
57. Greenland, Sander; Pearl, Judea; Robins, James M. Causal diagrams for epidemiologic research. Epidemiology. 1999:37–48. [PubMed: 9888278]

58. Dahl, Andy; Hore, Victoria; Iotchkova, Valentina; Marchini, Jonathan. Network inference in matrix-variate gaussian models with non-independent noise. 2013 arXiv preprint arXiv:1312.1622.

59. Angrist, Joshua D.; Pischke, Jörn-Steffen. Mostly harmless econometrics: An empiricist's companion. Princeton university press; 2008.

60. Aschard, Hugues; Vilhjálmsson, Bjarni J.; Joshi, Amit D.; Price, Alkes L.; Kraft, Peter. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. The American Journal of Human Genetics. 2015

61. International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467(7311):52–58. [PubMed: 20811451]

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. SSGAC (educational attainment) summary statistics:

    www.ssgac.org/

14. Beans:

    www.barismo.com

    www.bluebottlecoffee.com

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 on age at menarche. 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

The authors declare no competing financial interests.

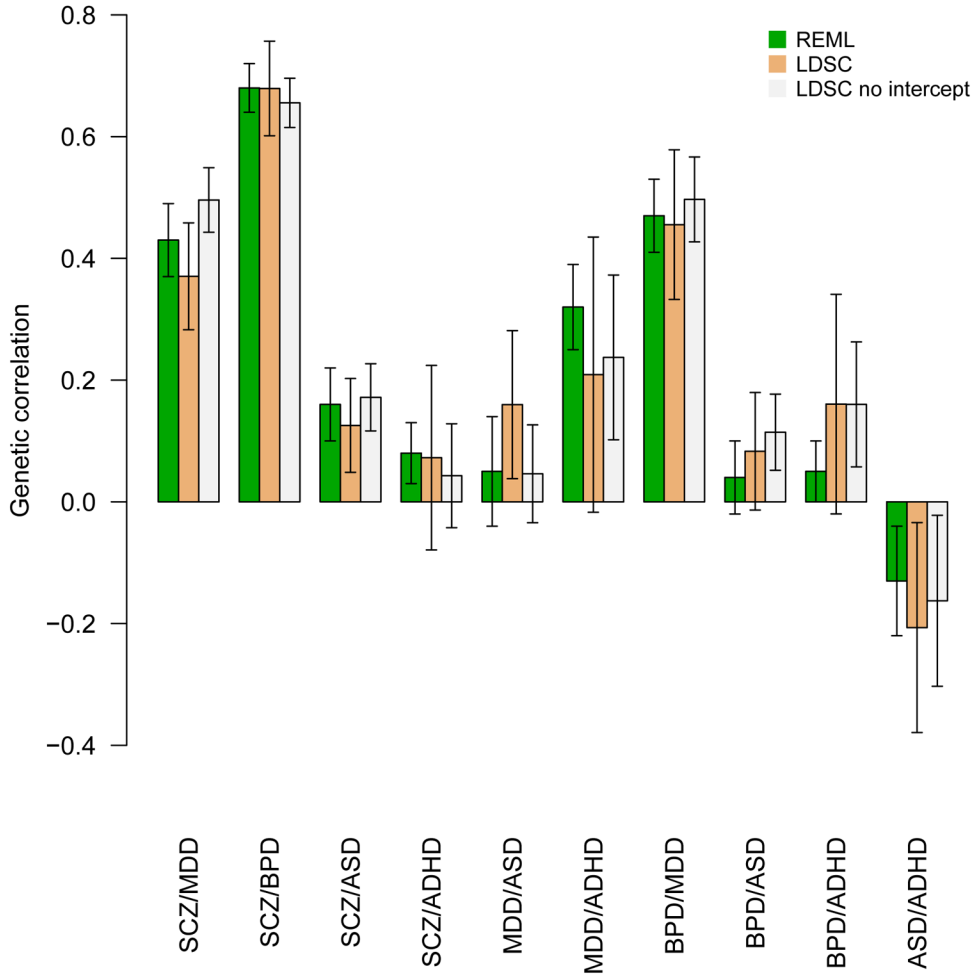

Figure 1.

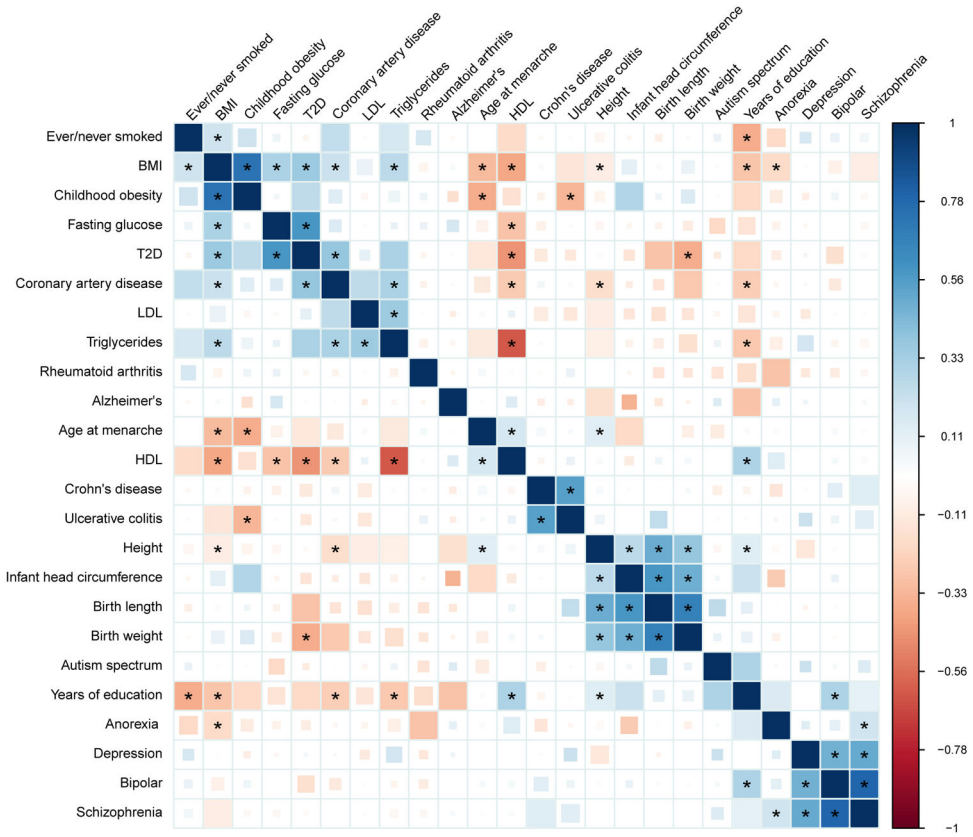

Figure 2.

Table 1

Simulations with complete sample overlap. Truth shows the true parameter values. Estimate shows the average cross-trait LD Score regression estimate across 1000 simulations. SD shows the standard deviation of the estimates across 1000 simulations, and SE shows the mean cross-trait LD Score regression SE across 1000 simulations. Further details of the simulation setup are given in the Methods.

Parameter	Truth	Estimate	SD	SE
h^2	0.58	0.58	0.072	0.075
ρ_g	0.29	0.29	0.057	0.058
r_g	0.50	0.49	0.079	0.073

Table 2

Genetic correlation estimates, standard errors and p-values for selected pairs of traits. Results are grouped into genetic correlations that are new genetic results, but are consistent with established epidemiological associations ("Epidemiological"), genetic correlations that are new both to genetics and epidemiology ("New/Nonzero") and interesting null results ("New/Low"). The p-values are uncorrected p-values. Results that pass multiple testing correctionw for the 300 tests in Figure 2 at 1% FDR have a single asterisk; results that pass Bonferroni correction have two asterisks. We present some genetic correlations that agree with epidemiological associations but that do not pass multiple testing correction in these data.

	Phenotype 1	Phenotype 2	rg(se)	p-value
Epidemiological	Age at menarche	Adult height	0.13 (0.03)	2×10^{-6}**
	Age at menarche	Type 2 diabetes	−0.13 (0.04)	2×10^{-3}*
	Age at menarche	Triglycerides	−0.12 (0.04)	1×10^{-3}*
	Coronary artery disease	Age at menarche	−0.12 (0.05)	3×10^{-2}
	Coronary artery disease	Years of education	−0.25 (0.06)	1×10^{-4}**
	Coronary artery disease	Adult height	−0.17 (0.04)	1×10^{-5}**
	Alzheimer's	Years of education	−0.29 (0.1)	5×10^{-3}*
	Bipolar disorder	Years of education	0.30 (0.06)	9×10^{-7}**
	BMI	Years of education	−0.28 (0.03)	6×10^{-16}**
	Triglycerides	Years of education	−0.26 (0.06)	2×10^{-8}**
	Anorexia nervosa	BMI	−0.18 (0.04)	3×10^{-7}**
	Ever/never smoker	Years of education	−0.36 (0.06)	2×10^{-8}**
	Ever/never smoker	BMI	0.20 (0.04)	8×10^{-7}**
New/Nonzero	Autism spectrum disorder	Years of education	0.30 (0.08)	2×10^{-4}*
	Ulcerative colitis	Childhood obesity	−0.34 (0.08)	3.1×10^{-5}**
	Anorexia nervosa	Schizophrenia	0.19 (0.04)	2×10^{-5}**
New/Low	Schizophrenia	Alzheimer's	0.04 (0.06)	>0.1
	Schizophrenia	Ever/never smoker	0.04 (0.06)	>0.1
	Schizophrenia	Triglycerides	−0.04 (0.04)	>0.1
	Schizophrenia	LDL cholesterol	−0.04 (0.04)	>0.1
	Schizophrenia	HDL cholesterol	0.03 (0.04)	>0.1
	Schizophrenia	Rheumatoid arthritis	−0.04 (0.05)	>0.1
	Crohn's disease	Rheumatoid arthritis	−0.03 (0.08)	>0.1
	Ulcerative colitis	Rheumatoid arthritis	0.09 (0.08)	>0.1