Estimating Genetic Correlations between Traits from GWAS Summary Statistics

Brendan Bulik-Sullivan*, Hilary Finucane*, ... , *et. al.*

October 14, 2014

Contents

List of Figures

List of Tables

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 has been to search for correlations between phenotypes via large observational epidemiological studies; however the interpretation of results from these studies can be confounded by social factors that are difficult to control for. An alternative strategy that is robust to confounding by social factors is to search instead for pairs of phenotypes with shared genetic etiology.

The largest currently available sources of genotype-phenotype data are genome-wide association studies (GWAS). GWAS have identified tens of thousands of phenotype-SNP associations in total, but these associations are spread across hundreds of phenotypes. At current sample sizes, the number of associated loci per phenotype is usually only in the tens to hundreds, since for most traits, the majority of the genetic contribution to phenotypic variance is accounted for by variants with small effects, which cannot be confidently associated except with very large sample sizes. Therefore, simply scanning for pairs of phenotypes with shared genome-wide significant loci is not a powerful approach for polygenic phenotypes.

A more powerful option is to use information from all genotyped SNPs to estimate the genome-wide (narrow-sense) genetic correlation between phenotypes; however, existing methods for estimating genetic correlation from GWAS data, such as restricted maximum likelihood (REML) as implemented in the software package GCTA [1, 2, 3] require individual genotype and phenotype data, which can be difficult or impossible to obtain due to restrictions on data sharing. For this reason, only a few dozen genetic correlations have been estimated from GWAS data to date [4, 5, 6].

In this paper, we describe a method based on LD Score regression which bypasses data sharing difficulties by estimating genetic correlations using only GWAS summary statistics. These estimates extend the understanding gleaned from the handful of previously estimated genetic correlations, giving us a big-picture view of how phenotypes cluster together as well as allowing us to skim for surprising and interesting genetic correlations.

In addition, we describe methods for relaxing some of the unrealistic assumptions about genetic architecture made by existing methods for estimating genetic correlation, and demonstrate that our method is minimally confounded when effect size depends on allele frequency or linkage disequilibrium. As a sanity check, we replicate the genetic correlations reported by the PGC Cross-Disorder Group using REML in [4], using only the summary statistics from [7].

Since dozens of sets of summary statistics can be freely downloaded from the internet, we are able report a much larger number of genetic correlations – hundreds in this paper alone – than was previously possible. Most genetic correlations tend to cluster within previously defined phenotypic categories; however, we observe several surprising results that would not have been possible to obtain without methods that operate on summary statistics.

The computational demands of our method are very mild. 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. For comparison, REML takes time $\mathcal{O}(MN^2)$ for computing the genetic relatedness matrix (GRM) and $\mathcal{O}(N^3)$ time for maximizing the likelihood. Practically, LD Score regression takes a matter of minutes on a standard laptop. We provide an open-source software package, `ldsc`, written in python, which implements the analyses described in this paper and also the analyses from [8, 9] (URLs).

2 Results

2.1 Overview of Methods

The additive genetic covariance, ρ_g between two phenotypes y_1 and y_2 is the bivariate analogue of heritability, and is defined as the covariance (in the population) between the additive genetic components of y_1 and y_2. The normalized version of genetic covariance is genetic correlation,

$$r_g := \frac{\rho_g}{\sqrt{h_1^2 h_2^2}}, \tag{2.1}$$

where h_i^2 denotes the heritability of trait i; genetic correlation lies in the interval $[-1, 1]$.

The genetic correlation that we estimate in this paper is different from the quantity estimated by family studies. We focus on the genetic correlation with respect to the additive genetic component of phenotype captured by common SNPs; family studies attempt to estimate the total narrow sense genetic correlation.

Unlike the distinction between h_g^2, h^2 and H^2, where we have the inequality $h_g^2 < h^2 < H^2$ (or the same, with $h_{5\text{-}50\%}^2$ in place of h_g^2 [10]), no such inequality holds in general for the narrow-sense genetic correlation among SNPs, the total narrow-sense genetic correlation and the broad-sense genetic correlation.

We can estimate the denominator of equation 2.1 from summary statistics using the methods described in [8]. In this paper, we describe a method for estimating of the numerator based off a simple modification of the regression from [8]. Instead of regressing χ^2-statistics against LD Score, we regress the product of Z-scores from two different GWAS against LD Score, and the slope times a constant gives us an estimator of genetic covariance (Online Methods).

One major concern when estimating genetic correlations from GWAS summary statistics is that many pairs of GWAS share large numbers of overlapping samples, so the Z-scores from one study may not be independent of the Z-scores from another study. This turns out not to be a difficulty for LD Score regression. Since sample overlap affects all SNPs equally, sample overlap merely inflates the LD Score regression intercept, and does not affect the slope (Supplementary Note). If the amount of sample overlap and the phenotypic correlation among overlapping individuals is known ahead of time, then it is possible to reduce the standard error of the genetic covariance estimate by constraining the intercept (Online Methods). In addition, because LD Score is minimally correlated with F_{ST} [8], the LD Score regression estimator of genetic covariance is protected from shared population stratification.

2.2 Simulations

In order to check our derivations and verify the robustness of our inference procedure to violations of our modeling assumptions, we performed a variety of simulations.

2.2.1 Shared Population Stratification

Shared population stratification is a potential confounder for estimates of genetic correlation. For example, if a GWAS for phenotype one is confounded by population stratification from North/South European ancestry and the same is true of a GWAS for phenotype two, then the estimate of genetic correlation obtained from REML will be biased upwards

Two-phenotype LD Score regression should be protected from from shared population stratification for exactly the same reasons that single phenotype LD Score regression is protected from shared population stratification [8]: LD Score is essentially uncorrelated with F_{ST}, so shared population stratification will only inflate the intercept of the two-phenotype LD Score regression.

To verify this conclusion, we simulated pairs of phenotypes from real genotype under an additive model, then added an environmental stratification component to aligned with the first principal component of the genotype data to both phenotypes in each pair. The environmental stratification component was scaled to account for 10% of the phenotypic variance, which corresponds to moderate population stratification (the median LD Score regression intercept was 1.02, the median λ_{GC} was 1.32). We then estimated genetic correlation using LD Score regression. Results from these simulations are displayed in SUPPLEMENTARY FIGURE, and confirm that the LD Score regression estimator of genetic correlation is not biased by shared population stratification.

2.2.2 Misspecified Models of Genetic Architecture

Estimates of heritability and genetic covariance can be biased if the underlying model of genetic architecture is misspecified. For example, Speed, *et. al.* [11] demonstrate that REML can be confounded by MAF- or LD-dependent genetic architectures.

Lee, *et. al.* ([12]) showed that it is possible to correct for such biases using MAF-binned REML. One can take a similar approach with LD Score regression to obtain unbiased estimates of heritability and genetic covariance even under MAF or LD-dependent genetic architectures (see [9] and Online Methods). Estimates of genetic correlation are more robust 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.

To quantify the bias introduced by MAF- or LD-dependent genetic architectures, we performed a series of simulations, using a variety of different LD Scores and genetic architectures. In order to simulate the realistic scenario where only a subset of causal SNPs are directly genotyped, we used a densely imputed panel of 1000 Genomes (1kG) SNPs [13] in order to generate phenotypes and estimate LD Scores, but computed summary statistics only for HapMap3 (HM3) SNPs [14] Results from these simulations are displayed in Supplementary Tables 5.1, 5.2 and 5.3.

We found that the binned LD Scores were almost completely immune to MAF- and LD-dependent genetic architectures when estimating heritability and genetic covariance, but gave substantially higher standard errors when estimating genetic correlation. The simplest LD Score model, with LD Scores computed using only "genotyped" SNPs (*i.e.,* HM3 SNPs) – hereafter referred to as HM3 LD Score – was the best performing estimator of genetic correlation. The simple model gave biased heritability and genetic covariance estimates in simulations with MAF- and LD- dependent genetic architectures, but he genetic correlation estimates were approximately unbiased in simulations where heritability and genetic covariance depended on LD, and only minimally biased even in simulations where genetic correlation depended on LD as well.

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

2.3 Real Data

2.3.1 Replication of PGC Cross Disorder Results

For further validation, we replicated the estimates of genetic correlations between psychiatric phenotypes obtained with individual genotypes and REML in the PGC Cross-Disorder Group paper [4], using LD Score regression and the summary statistics from [7], downloaded from the PGC website (URLs). Since the HM3 LD Score was the best-performing estimator of genetic correlation in simulations, we used this LD Score for application to real data.

Including an intercept in the LD Score regression protects the results from QC issues such as population stratification (as described in [8]) and sample overlap, but at the cost of a substantial increase in standard error. Since the summary statistics from [7] 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 intercepts [10].

Results from this analysis are displayed in Figure 1. As expected, the genetic correlation estimates from LD Score regression with HM3 LD Scores were very 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 somewhat larger, especially for the very small studies (*e.g.,* ADD, Autism).

The computational demands of this analysis were trivial: after computing LD Scores and preprocessing the summary statistics, the LD Score regression took about one minute per pair of phenotypes (most of which was spent reading compressed LD Score files into memory) and less than 1GB of RAM.

2.3.2 Application to a Large Set of Publicly Available Summary Statistics

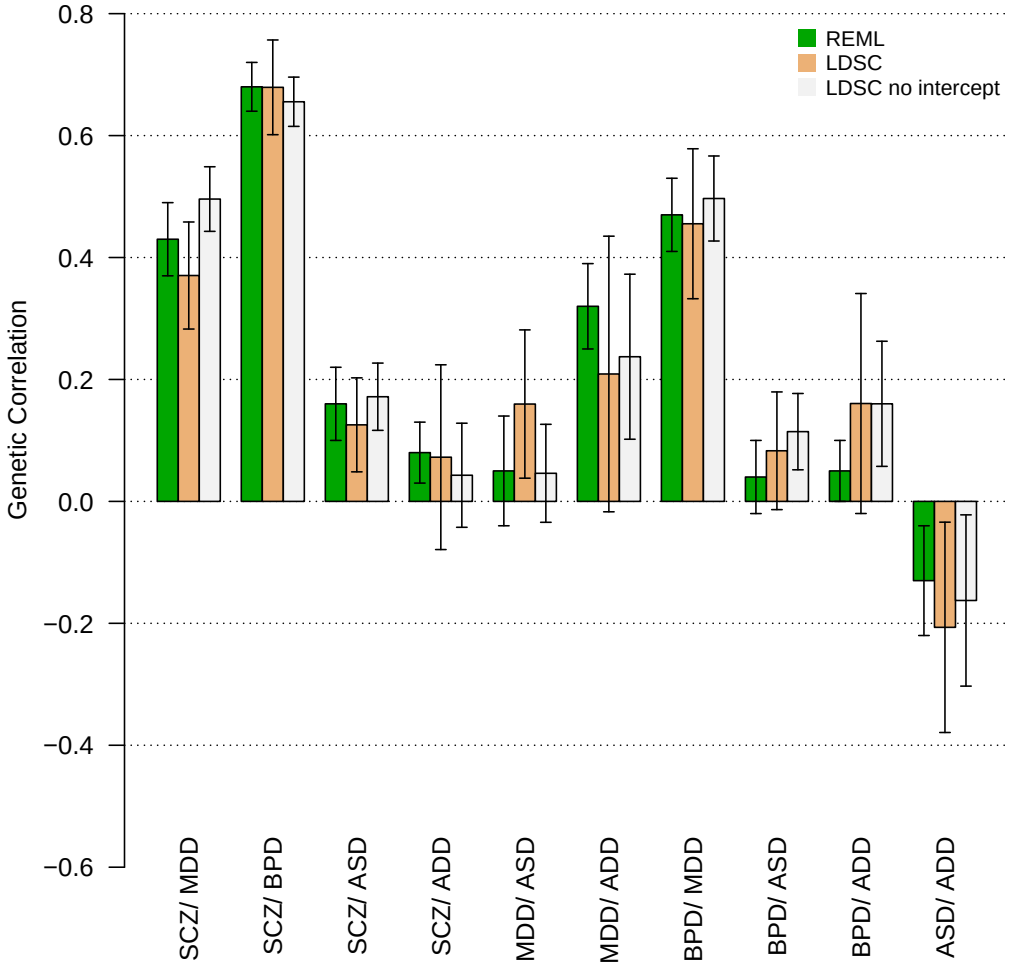
We applied our method to 27 publicly available sets of GWAS summary statistics, including schizophrenia [15], major depression [16], bipolar disorder [17], autism [7], attention-deficit hyperactivity disorder [18], anorexia [19], height [20], body mass index [21], BMI-adjusted waist-hip ratio [22], obesity [23], fasting glucose [24], HOMA-B, HOMA-IR [25], HbA1C [26], cigarettes per day, age of onset of smoking, ever vs never smokers, former vs current smokers [27], coronary artery disease [28], type-2 diabetes [29], rheumatoid arthritis [30], high-density lipoprotein, low-density lipoprotein, triglycerides, total cholesterol [31], ulcerative colitis [32], Crohn's disease [32] and Alzhiemer's disease [33] (see URLs for a full list of links).

Because HM3 LD Score performed best in simulations, we used this LD Score for estimating all pairwise genetic correlations. The majority of these studies share some samples, so we did not constrain any of the LD Score regression intercepts. The subset of genetic correlation estimates that were statistically significantly different from zero after correction for 351 tests at 5% FDR are displayed as a heatmap in Figure 2, and the full set of genetic correlation estimates are provided in tabular (`csv`) format in the Supplementary Data.

We find that phenotypes tend to cluster into categories defined by clinical practice and observational epidemiology; for instance, we observe high genetic correlations between anthopometric traits, between psychiatric traits, between metabolic traits and between autoimmune traits. Reassuringly, most genetic correlations across pre-defined phenotypic categories were within not significantly different from zero.

Our results on genetic correlation between metabolic traits are generally consistent with the results from [5], though our standard errors are lower by an order of magnitude, since requiring

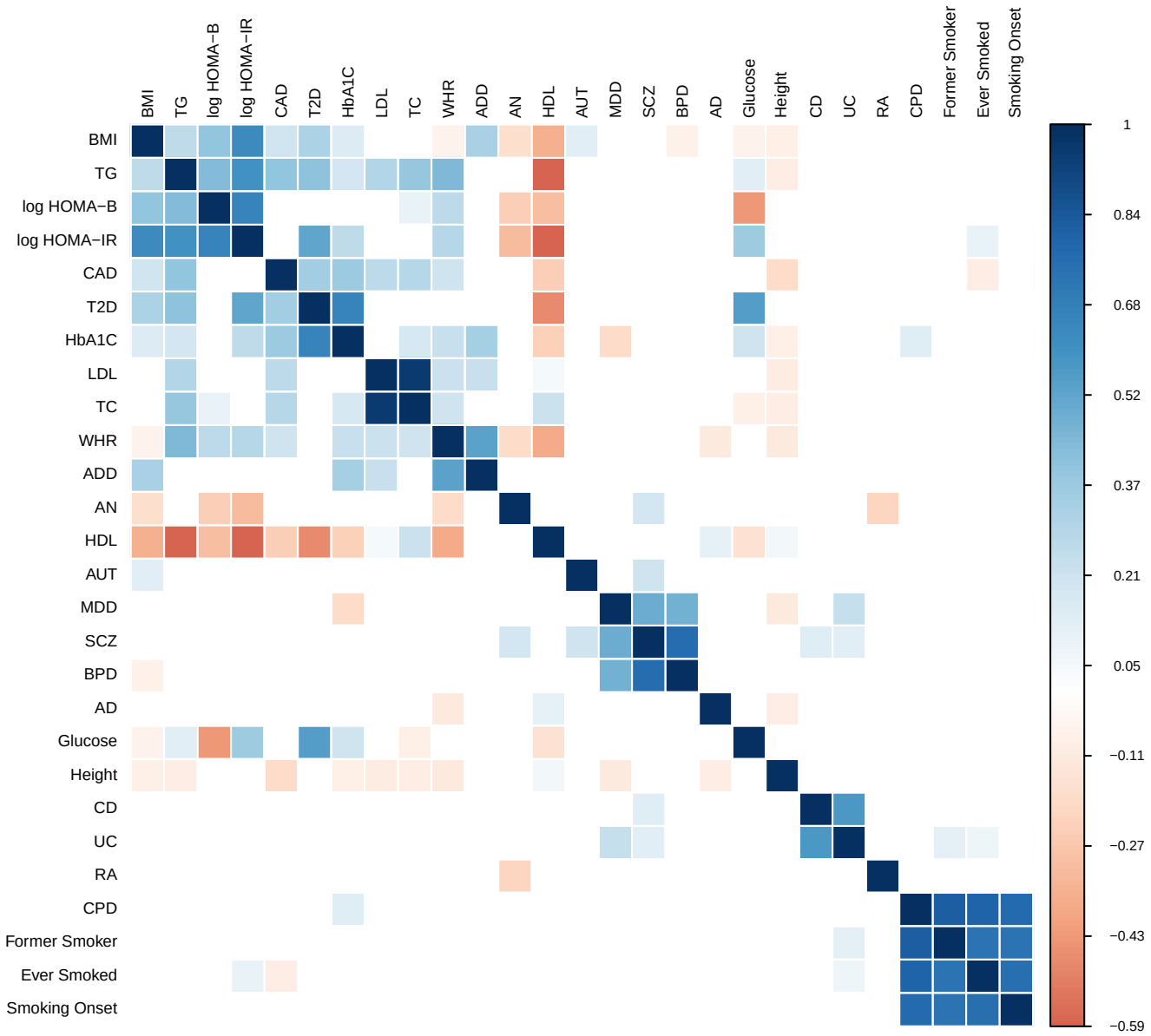
Figure 1: Replication of PGC Cross Disorder Results



This plot compares LD Score regression estimates of genetic correlation using the summary statistics from [7] (which were generated from approximately the same data as [4]) to estimates obtained from REML in [4]. The horizontal axis indicates pairs of phenotypes, and the vertical axis indicates genetic correlation. Colors indicate different estimation procedures. Grey is REML, Orange is HM3 LD Score with intercept; blue is HM3 LD Score without intercept, The estimates of genetic correlation between psychiatric phenotypes presented under the header "Application to a Large Set of Publicly Available Summary Statistics" use larger sample sizes, and so are more reliable; this plot is intended primarily as a technical sanity check. Abbreviations: ADD = Attention Deficit Hyperactivity Disorder (1947 trio cases, 1947 trio pseudocontrols, 840 cases, 688 controls); ASD = Autism Spectrum Disorder (4788 trio cases, 4788 trio pseudocontrols, 161 cases, 526 controls); BPD = Bipolar Disorder (6990 cases, 4820 controls); MDD = Major Depressive Disorder (9227 cases, 7383 controls); SCZ = Schizophrenia (9379 cases, 7736 controls).

only summary statistics gives us access to sample sizes that are larger by an order of magnitude (Supplementary Table 4.1. The only discrepancies came from waist-hip ratio (WHR). The data that

Figure 2: Genetic Correlations Between 27 Published GWAS



This figure displays genetic correlations between the 27 phenotypes analyzed estimated with HM3 LD Score. The genetic correlations are color coded according to sign and magnitude. Only the genetic correlations that are statistically significantly different from zero at an FDR of 0.05 are displayed; the rest have been whited out.

we used were from a GWAS for BMI-adjusted WHR [22]; whereas Vattikuti *et. al.* use unadjusted WHR. These are very different phenotypes, so it is unsurprising that their genetic correlation profiles differ (Supplementary Table 5.5). In addition, our estimate of the genetic correlation between Crohn's disease and ulcerative colitis (0.57 SE 0.063) is consistent with the estimate from [6] (0.62 SE 0.042).

SUGGESTED COOL EXAMPLES

1. smoking traits all have near total rg despite no shared gwsig SNPs (goes against the interpretation offered by TAG)
2. first ever SNP rg's for AN (BMI/SCZ have best p-values, replicate in PGC-AN)
3. zero rg between alzheimers and psychiatric stuff
4. alzheimers and height (Z = 9)???
5. dense metabolic cluster at the top left
6. CD/UC and SCZ [34, 35]
7. 0 rg for SCZ and RA (though a small effect may be meaningful, and we are not well-powered to detect anything between 0 and -0.1)

2.4 Heritability

The majority of the summary statistics that we analyzed were "corrected" via genomic control correction. LD Score regression estimates of heritability and genetic covariance from GC corrected summary statistics are biased downwards [8] (the genetic correlation estimates are fine, because the bias in the numerator and denominator cancels exactly). Thus, we can only present heritability estimates for the subset of GWAS that did not use GC correction.

We had information on sample MAF and imputation quality for schizophrenia, CD and UC, so we used all 1kG SNPs with sample MAF above 3% and INFO above 0.9 for these LD Score regressions, since using a larger set of SNPs or the regression decreases the standard error. We estimated heritability using the same partitioned LD Scores as in figure 2, both with and without constraining the LD Score regression intercept to equal one. Heritability estimates for UC, CD, SCZ and BPD are displayed in Table 1.

We also display estimates of h_g^2 from earlier publications, though we caution that h_g^2 and $h_{5\text{-}50\%}^2$ are not the same quantity, so any comparison between them is technically apples-to-oranges (see the Section 4.2 in the Methods). The estimates of h_g^2 in Table 1 were obtained using REML on ascertained samples, and so are biased downwards [36, **?**]. The heritability estimates with constrained intercept differed from the estimates with unconstrained intercept by less than one standard error in all cases except for UC, where the difference was substantial. This could be explained either by a small amount of residual population stratification in the summary statistics (even after careful QC), reference/target LD mismatch or model misspecification bias not fully accounted for by the partitioning.

Table 1: Heritability Estimates

	Crohn's	UC	Schizophrenia	Bipolar
$h^2_{5\text{-}50\%}$, no intercept	0.31 (0.022)	0.26 (0.016)	0.31 (0.01)	0.33 (0.028)
$h^2_{5\text{-}50\%}$, intercept	0.3 (0.05)	0.16 (0.034)	0.29 (0.021)	0.31 (0.072)
REML h^2_g	0.26 (0.011)	0.19 (0.008)	0.23 (0.008)	0.25 (0.012)
Prevalence	0.002	0.003	0.010	0.010

Liability scale heritability estimates for CD, UC, SCZ and BPD. The first row contains estimates of $h^2_{5\text{-}50\%}$ from LD Score regression with intercept constrained to one. The second row contains estimates of $h^2_{5\text{-}50\%}$ from LD Score regression with unconstrained intercept. The LD Score regressions for CD, UC and SCZ used approximately 6 million well-imputed (INFO > 0.9) 1kG SNPs. The LD Score regression for BPD used 1.2 million HM3 SNPs since this GWAS did not use 1kG imputation. All LD Score regressions in this table used a DAF × LD Score partitioned LD Score regression model with 35 bins. The estimates of h^2_g for CD and UC are from [6]; the estimates of h^2_g for SCZ and BPD are from [4]. The prevalence row contains the population prevalences used for transformation to the liability scale. Estimates of liability scale heritability should be interpreted cautiously, since for many diseases, estimates of prevalence are fuzzy.

3 Discussion

3.1 Genetic Correlation is Different from Pleiotropy

Genetic correlation is defined as the correlation in the population between the additive genetic components of phenotypes. Pleiotropy is defined as the tendency for two phenotypes to be influenced by the same genetic loci. Thus, 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 loci: the directions of effect of the variants that influence the phenotypes 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. For example, since P values tend to decrease with LD Score [8] and near coding and regulatory regions ([37]), the observation that low P-values for one phenotype tend to predict low P-values for another (as in [38, 39, 40]) may merely reflect properties shared by all pairs of heritable phenotypes, rather than a special property of a specific pair of phenotypes (Supplementary Figures NNN). 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.

3.2 Limitations and Future Directions

We note some limitations of LD Score regression and interpretation of estimates of genetic correlation from GWAS data in general. First, LD Score regression requires large sample sizes (at least several thousand) in order to give estimates with reasonable standard error. For smaller sample sizes, REML is a more statistically efficient estimator of genetic correlation when its assumptions are met (in particular, for non-ascertained studies) and individual genotypes are available. In our opinion, figuring out how to obtain an efficient estimate of genetic correlation from a small and

highly ascertained sample (*e.g.,* a study of a rare polygenic disease, where finding even a few thousand cases to genotype may be challenging) is an open question.

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 sequence data. For multi-continental GWAS, the solution is to run LD Score regression on each continental subcohort separately, but nevertheless, LD Score regression can only be applied to samples from populations for which there exists a large sequenced reference panel. Currently, LD Score regression cannot be applied to samples from admixed populations.

Third, while genetic correlations are less confounded than overall phenotypic correlations from observational epidemiology, genetic correlations cannot be interpreted as causal effects, even genetic correlations between intermediate phenotypes and disease. For example, consider the strong negative genetic correlation between HDL and CAD in Figure 2. This genetic correlation could result from a causal effect, *e.g.,* HDL $\rightarrow$ CAD, but could also result from non-causal shared genetic etiology *e.g.,* HDL $\leftarrow$ G $\rightarrow$ CAD, where G is some set of unknown genetic factors. Indeed, the second explanation seems most likely in this case, because the genetic correlation between HDL and CAD has been observed to vanish after accounting for other lipid fractions [41].

Fifth, while LD Score regression is a consistent estimator of genetic correlation given samples where cases have been oversampled, it may be biased by misclassification of phenotypes or by more subtle forms of ascertainment. Thus, it is best to interpret estimation of genetic correlation from GWAS summary statistics as a hypothesis-generating exercise. This method allows us to screen hundreds of pairs of phenotypes with computational ease, but any findings should be followed-up with confirmatory analyses, including a detailed review of the ascertainment procedures in each GWAS.

Finally, it is impossible to obtain any estimates of genetic correlation at all unless researchers are willing to share full sets of GWAS summary statistics! Since summary statistics contain such a large amount of information about genetic architecture and the relationships between phenotypes, it is surely the case that the benefits of data sharing outweigh the costs. To facilitate data sharing, we have written a short guide describing the necessary information under the header "Minimum Viable Summary Statistics" in the Methods.

3.3 Conclusion

We have described a method for estimating common variant heritability and genetic correlation that requires only GWAS summary statistics as input data. Our method is immune to case/control ascertainment, sample overlap, shared population stratification [8], and can be easily modified to be protected from bias from MAF- and LD-dependent genetic architectures. The computational demands of our method are mild: a few minutes on a standard laptop, irrespective of sample size. Although LD Score regression does not use individual-level data, the increase in standard error relative to methods that do use individual genotype and phenotype data is only moderate. In fact, LD Score regression without intercept is equivalent to HE regression [10].

...etc

4 Online Methods

4.1 Statistical Framework

See the supplementary note for a thorough derivation of the models behind LD Score regression.

4.2 Definitions

4.2.1 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

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

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^2_S, the heritability accounted for by SNPs in S, is defined

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

We obtain the Yang/Visscher parameter h^2_g by taking S to be the set of genotyped SNPs. Next, let$S$ denote the set of all SNPs in 1000 Genomes Europeans [13], and let S' denote the set of SNPs with MAF$>$ 5%. Then

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

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 $h^2_{\text{5-50\%},1kG}$ 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. It is perhaps more important to also note that we are only accounting for autosomal variation. Most GWAS do not report summary statistics for SNPs on the sex chromosomes or in mitochondrial DNA.

The definition of the Yang-Visscher parameter h^2_g [1] is obtained by replacing 1kG with the set g of genotyped SNPs in equation 4.2. There are two main distinctions between $h^2_{\text{5-50\%}}$ and h^2_g, First, h^2_g 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 $h^2_{\text{5-50\%}}$. 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 $h^2_{\text{5-50\%}}$ 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 extrapolating our estimate of heritability per SNP among common SNPs to rare variants. This is unreasonable, since GWAS data contains very little information about rare variants. Therefore, we

instead multiply the slope by $M_{5\text{-}50\%}$, the number of 1000 Genomes SNPs with MAF between 5% and 50%, in order to obtain an estimate of $h^2_{5\text{-}50\%}$. The default option in `ldsc` is to estimate $h^2_{5\text{-}50\%}$; 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 – rare variants, 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 [42].

4.2.2 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 [13], 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, which is defined as

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

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 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% MAF, 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 (ρ_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 ρ_g and $\rho_{5\text{-}50\%}$ to the total narrow-sense

genetic covariance, or 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 $r_{5\text{-}50\%}$.

4.3 Two-Phenotype LD Score Regression

Our method is based on a simple 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 the product $z_{1,j} z_{2,j}$ of Z-scores from two GWAS against ℓ_j, the LD Score of SNP j, and dividing the resulting slope by $\frac{\sqrt{N_1 N_2}}{M}$. Because sample overlap only affects the intercept, 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 [8]), and use these heritability estimates to transform the estimates of genetic covariance into estimates of genetic correlation (see section 4.6 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 [11, 8], 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 [9].

4.4 Partitioned LD Score Regression

In partitioned LD Score regression, we cut the set of SNPs in our reference panel into bins, for example, we might use five MAF bins, corresponding to MAF 0-10%, 10-20%, ... , 40-50% (as in supplementary table 4 of [4]). This allows the variance explained per SNP to vary from bin to bin, but assume that variance explained per SNP is (roughly) equal within each bin. Concretely, if we

wish to run MAF-binned LD Score regression with five MAF bins $B_1, \ldots, B_5$, we compute five LD Scores, where the i^{th} LD Score is defined

$$\ell_j(i) := \sum_{k \in B_i} r_{jk}^2, \tag{4.10}$$

then run multivariate LD Score regression $\chi_j^2 \sim \ell_j(1) + \ldots \ell_j(5)$ (or $z_{1,j} z_{2,j} \sim \ell_j(1) + \ldots \ell_j(5)$ in the two-phenotype case). This amounts to approximating the unknown function that maps MAF to variance explained per SNP with a locally constant approximation.

Partitoned LD Score regression presents a bias/variance tradeoff: if the mesh of our locally constant approximation is too coarse (*e.g.,* if we were to use two MAF bins instead of five), our locally constant approximation would be poor, and this would result in bias if the underlying genetic architecture were MAF-dependent. On the other hand, if we use too many bins, the standard error of our estimates will increase. In this paper, we use non-partitioned LD Scores for the estimates of genetic correlation, since genetic correlation is less subject to MAF- and LD-biases than heritability, and LD Scores with 30 bins for heritability estimation.

MAF- and LD- partitioned LD Scores can be estimated using the `--cts-bin` and `--cts-breaks` flags from our `ldsc` software.

4.5 Genetic Covariance Regression Weights

For heritability estimation, we use the LD Score regression weights derived in the supplementary note from [8]. The optimal regression weights for genetic covariance estimation are

$$\mathrm{Var}[\hat{\beta}_j \hat{\gamma}_j \,|\, \ell_j] = \left(\frac{h_1^2 \ell_j}{M} + \frac{1}{N_1}\right)\left(\frac{h_2^2 \ell_j}{M} + \frac{1}{N_2}\right) + 2\left(\frac{\rho_g \ell_j}{M} + \frac{\rho N_s}{N_1 N_2}\right)^2; \tag{4.11}$$

(Supplementary Note) however, this quantity depends on both heritabilities, the genetic covariance and the number of overlapping samples, which are often not known a priori, so some 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 (which we only use 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. These regression weights are only an approximation to the optimal weights, but this will not introduce bias into the regression; it will only increase the standard error. The standard errors for LD Score regressions with summary statistics from GWAS with sample size below 10,000 are low enough to be interpretable, so non-optimality of the regression weights does not seem to be a major hindrance.

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 equal to one in expectation, but with noise.

4.6 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 [45] of the ratio, with the jackknife taken over blocks of adjacent SNPs

$$\hat{r}_{g,jack} := n_b \hat{r}_g - \sum_{i=1}^{n_b} \left(1 - \frac{m_i}{M_g}\right) \hat{r}_{g,i} \tag{4.12}$$

where n_b is the number of blocks, and $\hat{r}_{g,i}$ is the naive estimate of genetic correlation obtained by deleting the i^{th} 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 our sample sizes), and comes with a convenient nonparametric variance estimator [45],

$$\widehat{\mathrm{Var}}\left[\hat{r}_{g,jack}\right] := \frac{1}{n_b} \sum_{i=1}^{n_b} \frac{1}{h_i - 1} \left((h_i - n_b)\hat{r}_g - (h_i - 1)\hat{r}_{g,i} + \sum_{j=1}^{n_b} \left(1 - \frac{m_i}{M_g}\hat{r}_{g,j}\right), \right) \tag{4.13}$$

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 [8, 9, 46] 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.

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, even for GC-corrected data. This is a general feature of ratio estimators and is not specific to LD Score regression: it is difficult to produce 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.

4.7 GWAS Data

4.7.1 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, though the genetic correlation estimates will be fine.
4. The summary statistics must 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 Although REML has been applied to such datasets, we suspect these results may have been biased due to the non-random selection of genotyped SNPs.

The next details are nice to have, but are only used for filtering SNPs:

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

If these data are not available, we recommend retaining only HapMap 3 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 HapMap3 SNPs seem to be well-imputed in most studies. For newer studies with dense imputation, restricting to HapMap 3 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).

Since INFO scores are only used for filtering, researchers unwilling to share INFO scores should consider sharing a binary variable indicating the SNPs for which the INFO score was above 0.9.

4.7.2 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 highly kurtotic distribution, *i.e.,* if there are huge-effect loci. The `ldsc` default is to remove a window around SNPs with $\chi^2 > 0.01N$ (this can be disabled with the `--no-filter-chisq` flag). The resulting genetic correlation estimate can be interpreted as the genetic correlation among SNPs outside of the huge effect loci.

4.7.3 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. LD block genotype simulation code:
   `github.com/bulik/ldsc-sim`
3. This paper:
   `github.com/bulik/gencor_tex`
4. PGC (psychiatric) summary statistics:
   `www.med.unc.edu/pgc/downloads`
5. GIANT (anthopometric) summary statistics:
   `www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files`
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. Beans:
    `www.barismo.com`
    `www.bluebottlecoffee.com`

6 Acknowledgements

We would like to thank P. Sullivan and S. Caldwell for helpful discussion. This work was supported by NIH grants R01 HG006399 (ALP), R03 CA173785 (HKF) and by the Fannie and John Hertz Foundation (HKF). The coffee that Brendan drank while writing this paper was roasted by Barismo in Arlington, MA and Blue Bottle Coffee in Oakland, CA.

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, Universit 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.

7 Author Contributions

The caffeine molecule is responsible for everything that is good about this manuscript. BBS and HKF are probably responsible for the other bits. 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] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] Brendan Bulik-Sullivan. Rrelationship between Haseman-Elston Regression and LD Score Regression. *In Preparation*, 2014.

[11] 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.

[12] 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.

[13] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

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

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

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

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

[18] Benjamin M Neale, Sarah E Medland, Stephan Ripke, Philip Asherson, Barbara Franke, Klaus-Peter Lesch, Stephen V Faraone, Thuy Trang Nguyen, Helmut Schäfer, Peter Holmans, et al. Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 49(9):884–897, 2010.

[19] 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.

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

[21] Elizabeth K Speliotes, Cristen J Willer, Sonja I Berndt, Keri L Monda, Gudmar Thorleifsson, Anne U Jackson, Hana Lango Allen, Cecilia M Lindgren, Jian'an Luan, Reedik Mägi, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics*, 42(11):937–948, 2010.

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

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

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

[25] Josée Dupuis, Claudia Langenberg, Inga Prokopenko, Richa Saxena, Nicole Soranzo, Anne U Jackson, Eleanor Wheeler, Nicole L Glazer, Nabila Bouatia-Naji, Anna L Gloyn, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics*, 42(2):105–116, 2010.

[26] Nicole Soranzo. Genetic determinants of variability in glycated hemoglobin (hba1c) in humans: Review of recent progress and prospects for use in diabetes care. *Current diabetes reports*, 11(6):562–569, 2011.

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

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

[29] 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.

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

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

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

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

[34] 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.

[35] 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.

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

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

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

[46] 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.