Estimating Genetic Correlations between Traits from GWAS Summary Statistics

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

September 28, 2014

Contents

List of Figures

List of Tables

1 Abstract

Discovering relationships between phenotypes is a fundamental goal of epidemiology, with implications for drug development, nosology and treatment. The interpretation of phenotypic correlations in observational epidemiological studies can be confounded by environmental factors, so genetic correlations between phenotypes may be more easily interpretable. The largest currently available sources of genotype-phenotype data are genome-wide association studies (GWAS); however, existing methods for estimating genetic correlation from GWAS data require genotype and phenotype data for at least one of the phenotypes, 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. In this paper, we describe a method based on LD Score regression which estimates genetic correlations directly from GWAS summary statistics and is immune to sample overlap. In addition, we relax many common assumptions about genetic architecture, and demonstrate that our method is not confounded when effect size depends on allele frequency or linkage disequilibrium. Since dozens of sets of summary statistics can be freely downloaded from the internet, we can report a much larger number of genetic correlations – 300 in this paper alone – than was previously possible.

2 Introduction

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 $[-1, 1]$. Positive genetic correlation is a stronger condition than pleiotropy. To exhibit positive 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.

Existing methods for estimating genetic correlation from GWAS genotype data, such as restricted maximum likelihood (REML) as implemented in the software package GCTA [32, 33] – hereafter referred to as GCTA-REML – or polgyenic risk scores [10], require individual genotype data, which are often difficult or impossible to obtain due to restrictions on data sharing. Thus, investigations of additive genetic correlations between human traits typically report at most a handful of genetic correlations, usually estimated from samples of at most a few tens of thousands of individuals, and only a few dozen genetic correlations have been estimated using GWAS data to date [20, 31, 6].

Here, we propose a modification of LD Score regression [4] that can estimate the genetic correlation between two traits from GWAS summary statistics, allowing us to estimate hundreds of genetic correlations with computational ease. 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.

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. More 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

$$\mathbb{E}[z_{1,j} z_{2,j}] = \frac{\rho_g \sqrt{N_1 N_2}}{M} \ell_j + \frac{N_s \rho}{\sqrt{N_1 N_2}}, \tag{2.2}$$

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 (2.2) is derived in the supplementary note.

We can 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. We can estimate heritability using LD Score regression (as described in [4]), and use these heritability estimates to transform the estimates of genetic covariance into estimates of genetic correlation.

An equation similar to Equation (2.2) 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 [26, 4], this can introduce model misspecification bias into estimates of heritability and genetic correlation from methods such as LD Score regression and GCTA-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 of this paper and also [11].

In this paper, we describe a series of simulations establishing that the LD Score regression estimates genetic correlation. We then replicate the genetic correlations reported by the PGC Cross-Disorder Group using GCTA-REML in [20], using only the summary statistics from [21]. Finally, we report 300 (mostly novel) genetic correlations between pairs of phenotypes with publicly available GWAS summary data. We find that phenotypes tend to cluster within categories defined by clinical practice and observational epidemiology; nonetheless, we do observe some surprising results. For instance, we estimate genetic correlations close to zero between Alzheimer's and all of the psychiatric traits; although Alzheimer's disease is classified as a psychiatric disorder in ICD-10, it appears to be genetically distinctive. Instead, Alzheimer's disease clusters (weakly, but significantly) with anthopometric and metabolic traits. These results would not have been possible to obtain except with methods that operate on summary statistics, because the consortia in question do not share individual genotype data.

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 $\mathscr{O}(MN)$ time for computing summary statistics and $\mathscr{O}(M)$ time for the regression. For comparison, GCTA-REML takes time $\mathscr{O}(MN^2)$ for computing the genetic relatedness matrix (GRM) and $\mathscr{O}(N^3)$ 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 [4, 11] (URLs).

3 Results

3.1 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.

3.1.1 Sample Overlap

To verify the unbiasedness of our estimation procedure in the presence of sample overlap (which is derived formally in the Supplementary Note), we simulated two GWAS with quantitative phenotypes, using genotypes from the 4,292 individuals in the Wellcome Trust Case/Control Consortium 1 (WTCCC1, [9]) bipolar disorder cohort for the first GWAS and genotypes from the 4,482 individuals in the WTCCC1 coronary artery disease cohort for the second GWAS. These cohorts contain 2,713 overlapping inviduals. Additive genetic effect sizes were drawn from a bivariate point-normal distribution with 10% of SNPs causal and true genetic correlation 0.7. We then estimated genetic correlation using LD Score regression. Results from these simulations are summarized in Supplementary Table 6.1, and confirm that LD Score regression is not confounded by sample overlap.

3.1.2 Case-Control Ascertainment

In the Supplementary Note, we provide a proof that LD Score regression is valid when applied to ascertained samples of binary phenotypes. To verify this result, we performed simulations with case/control ascertainment.

Simulating case/control GWAS under a liability threshold model requires rejection sampling from a large pool of individuals. For instance, in order to simulate 1,000 cases for a phenotype with prevalence of 1%, one would need in expectation to sample 100,000 individuals. This is impractical using real genotypes, so we used simulated genotypes with a simplified LD block LD structure ($r^2 = 0$ or 1). This is the same simulation scheme used in [4].

With these simulated genotypes, we simulated standard case/control ascertainment following a liability threshold model, and estimated the genetic correlation using LD Score regression. Results from these simulations are summarized in supplementary table 6.2, and confirm that LD Score regression recovers the true heritability and genetic correlation, even for low-prevalence diseases. The simplified LD structure should not hinder interpretation of these simulation results, since they are just a confirmation of a mathematically established property of LD Score Regression.

3.1.3 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.* [26] demonstrate that GCTA-REML can be confounded by MAF- or LD-dependent genetic architectures. Estimates of genetic correlation are somewhat more robust. Since genetic correlation is estimated as a ratio $\hat{\rho}_g / \sqrt{\hat{h}_1^2 \hat{h}_2^2}$ (or the weighted

block jackknife estimator of this ratio, see Methods), and the model misspecification bias affects both the numerator and the denominator in the same direction, the bias will tend to approximately cancel, unless genetic correlation (not just heritability and genetic covariance) also depends on MAF or LD.

When genetic correlation depends on MAF or LD, LD Score regression is subject to similar biases as REML; however, it is possible to remove these biases by allowing for MAF- or LD-dependent genetic architectures by using partitioned LD Score regression (see [11] and Methods). We used simulations to explore the behavior of both partitioned and naive (*i.e.,* not partitioned) LD Score regression under three sets of bivariate genetic architectures with MAF- and LD-dependence.

For the first genetic architecture, the MAF- and LD- dependence of ρ_g and h^2 was the same for both phenotypes, and genetic correlation did not vary with MAF or LD. Effect sizes were drawn from a normal distribution so that the magnitude of per-allele effect size was uncorrelated with MAF and variants with LD Score below 100 were 4× enriched for heritability.

For the second genetic architecture, the genetic correlation did not vary with MAF or LD, but the direction of the MAF- and LD-dependence was different for each phenotype. The genetic architecture of the first phenotype matched the first simulation: we drew per-allele effect sizes from a normal distribution such that the variance of per-allele effect sizes were uncorrelated with MAF, and variants with LD Score below 100 were 4× enriched for heritability. Per-allele effect sizes for the second phenotype were drawn from a normal distribution such that the variance of per-allele effect size followed $\sqrt{p(1-p)}$, where p is MAF, and variants with LD Score above 100 were 4× enriched for heritability.

For the third genetic architecture, we allowed not only heritability and genetic covariance to depend on MAF and LD, but also genetic correlation. The parameters of these simulations were the same as the second genetic architecture, except that genetic correlation was 0.2 for variants with LD Score less than 100 and 0.8 for variants with LD Score greater than 100.

We estimated heritability, genetic covariance and genetic correlation using both naive LD Score regression and two partitioned LD Score regression models (one with 30 bins, one with 60 bins) that allow for both MAF- and LD-dependence. Results from these simulations are presented, in the order described, in Supplementary Tables 6.3, 6.4 and 6.5. As expected, the heritability and genetic covariance estimates from naive LD Score regression were badly biased in all cases (similar to the results obtained by Speed, *et. al.*), but the bias in the genetic correlation estimates was much less severe, except in the third set of simulations, where genetic correlation varied with LD. Nevertheless, partitioned LD Score regression was able to remove almost all bias introduced by LD- and MAF-dependence, and the increase in the standard errors was only mild.

3.2 Real Data

3.2.1 Replication of PGC Cross Disorder Results

For further validation, we replicated the estimates of genetic correlations between psychiatric phenotypes obtained with individual genotypes and GCTA-REML in the PGC Cross-Disorder Group paper [20], using LD Score regression and the summary statistics from [21], downloaded from the PGC website (URLs).

GCTA-REML with a GRM computed using only ~ 1 million genotyped SNPs estimates a different quantity than LD Score regression with sum r^2 taken over all ∼15 million SNPs in 1kG (1kG LD Scores). GCTA-REML with ∼1 million genotyped SNPs estimates quantity which describe

properties of best linear predictor of phenotype that one could obtain using only the ∼1 million genotyped SNPs. LD Score regression with 1kG LD Scores estimates quantities which are properties of the best linear predictor of phenotype that one could obtain using all 1kG SNPs.

Since LD Score regression with 1kG LD Scores models a much larger proportion of all causal SNPs, we believe that these results are more reliable and biologically meaningful (see section 5.2 in the Methods). As a practical matter, the differences between these quantities tends to be small[1], and will get smaller as the quality of imputation reference panels continues to increase. Thus, to give a fair comparison to GCTA-REML, we include results from LD Score regression results obtained using an LD Score with the sum of r^2 taken only over the 1.2 million autosomal HapMap 3 (HM3) SNPs [8] – which we refer to as HM3 LD Score – as well as results from LD Score regression with a 1kG LD Score partitioned on derived allele frequency (DAF) and recombination rate, to account for potential DAF- and LD-dependence in genetic architecture.

Including an intercept in the LD Score regression protects the results from QC issues such as population stratification (as described in [4]) as well as sample overlap, but at the cost of a substantial increase in standard error, which is non-negligible at the relatively small sample sizes from [21]. Since the summary statistics from [21] were generated after a careful QC process, and the samples used for each disease were non-overlapping, we also fit LD Score regression with a constrained intercept, which – in the non-partitioned case – is equivalent to HE regression (See the section "LD Score Regression is Haseman-Elston Regression" in the Supplementary Note).

Results from this analysis are displayed in Figure 1. The genetic correlation estimates from LD Score regression with HM3 LD Scores were very similar to the results from GCTA-REML. LD Score regression without intercept gave standard errors that were only slightly larger than GCTA-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 results from the DAF- and recombination rate-partitioned LD Score regression with 1kG LD Scores ... WRITE ME

The differences between all four estimators are consistent with noise. ... ?

The computational demands of this analysis were trivial: after computing LD Scores and pre-processing 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.

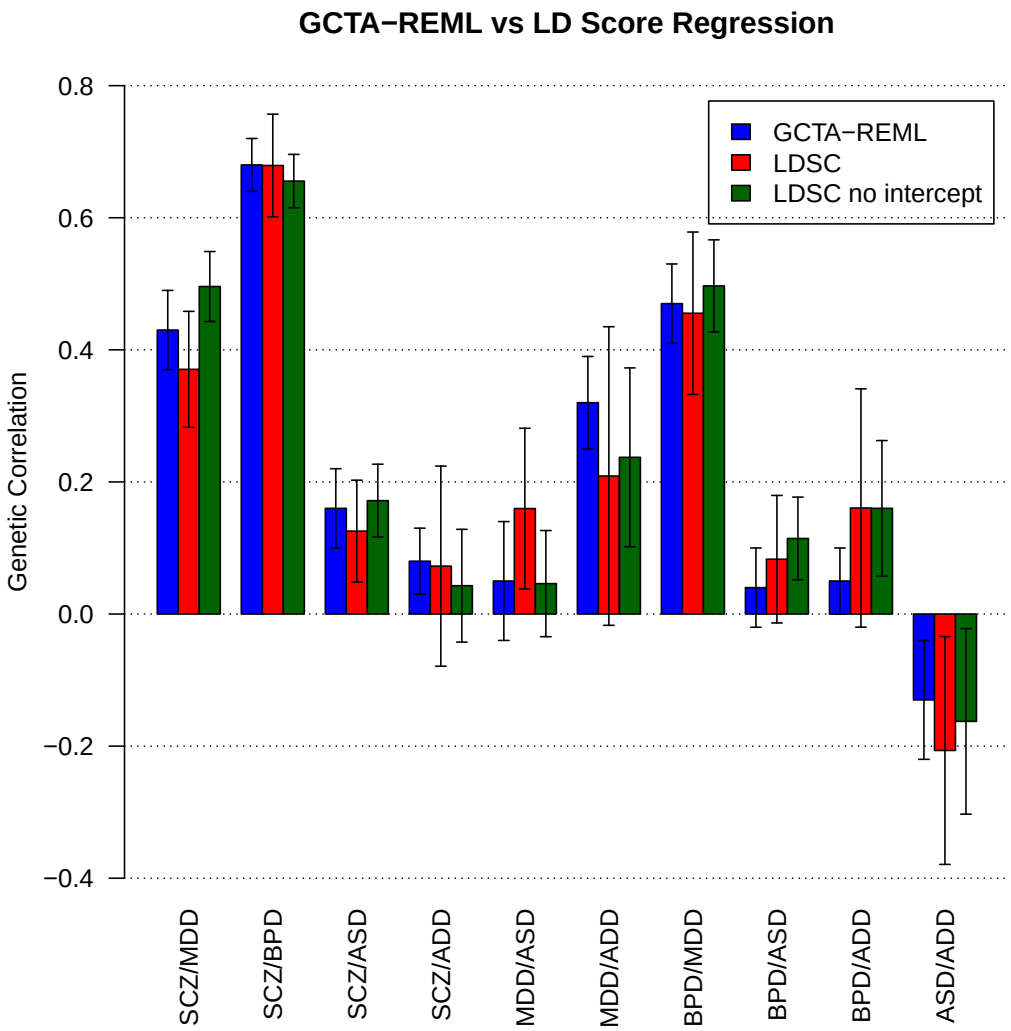
3.2.2 Application to a Large Set of Publicly Available Summary Statistics

We applied our method to 23 [[I think this number is wrong]] publicly available sets of GWAS summary statistics, including schizophrenia [22], major depression [23], bipolar disorder [25], autism [21], attention-deficit hyperactivity disorder [19], anorexia [3], height [1], body mass index [27], waist-hip ratio [13], obesity [2], various insulin- and glucose- related traits [16], cigarettes per day, age of onset of smoking, ever vs never smokers, former vs current smokers [30], coronary artery disease [24], type-2 diabetes [18], rheumatoid arthritis [28], high-density lipoprotein, low-density lipoprotein, triglycerides, total cholesterol [29], ulcerative colitis [14], Crohn's disease [14] and Alzhiemer's disease [15].

We estimated all pairwise genetic correlations between these phenotypes using partitioned LD Score regression. Where information on sample overlap and phenotypic correlation was available, we seeded the regression weights with this information in order to reduce the standard error (Methods,

[1]Except for partitioned heritability, where results from a million genotyped SNPs are not biologically meaningful.

Figure 1: Replication of PGC Cross Disorder Results



This plot compares LD Score regression estimates of genetic correlation using the summary statistics from [21] (which were generated from approximately the same data as [20]) to estimates obtained from GCTA-REML in [20]. The horizontal axis indicates pairs of phenotypes, and the vertical axis indicates genetic correlation. Colors indicate different estimation procedures, as described in the main text. 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).

Supplementary Table **NNN**). LD Score regression heritability estimates are biased downwards by genomic control correction (GC) [4], so we report heritability estimates only for those GWAS that did not use GC correction (the GWAS for psychiatric diseases and inflammatory bowel disease). Note that we strongly recommend against using GC correction in all future meta-analyses, for reasons described in [4].

Heritability estimates are displayed in table **NNN**. These are technically estimates of the heritability accounted for by SNPs with 5-50% MAF that appear in 1000 Genomes, denoted $h^2_{5\text{-}50\%}$ (see Methods). This quantity is not in general the same as the quantity estimated by GCTA-REML described as the "variance explained by genotyped SNPs", and denoted either h^2_g or h^2_{SNP} Indeed, estimates of h^2_g from different GWAS of the same trait are strictly speaking not directly comparable to one another, because the definition of the parameter h^2_g depends on the set g of SNPs used for computing the kinship matrix.

The full list of 300 genetic correlation estimates are provided in tabular (csv) format in the Supplementary Data, and are displayed as a heatmap in Figure 2. 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 ($\hat{h}^2/\text{se}(\hat{h}^2)$), the lower the standard error for genetic correlation, even for GC-corrected data. This is a general feature of ratio estimators: it is difficult to produce an accurate estimate of $1/x$ when the random variable x is close to zero.

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 categories were non-significantly different from zero.

We observed some interesting individual results (Table **NNN**). A few examples: We estimate genetic correlations close to zero between Alzheimer's and all of the psychiatric traits. Even though Alzheimer's disease is classified as a psychiatric disorder in ICD-10, it appears to be genetically distinctive. Instead, Alzheimer's disease clusters (weakly, but significantly) with anthopometric traits. We estimate genetic correlation close to zero between rheumatoid arthritis (RA) and schizophrenia, and between smoking traits and schizophrenia, despite reduced risk of RA in patients with schizophrenia and high rate of smoking in patients with schizophrenia

... etc more examples

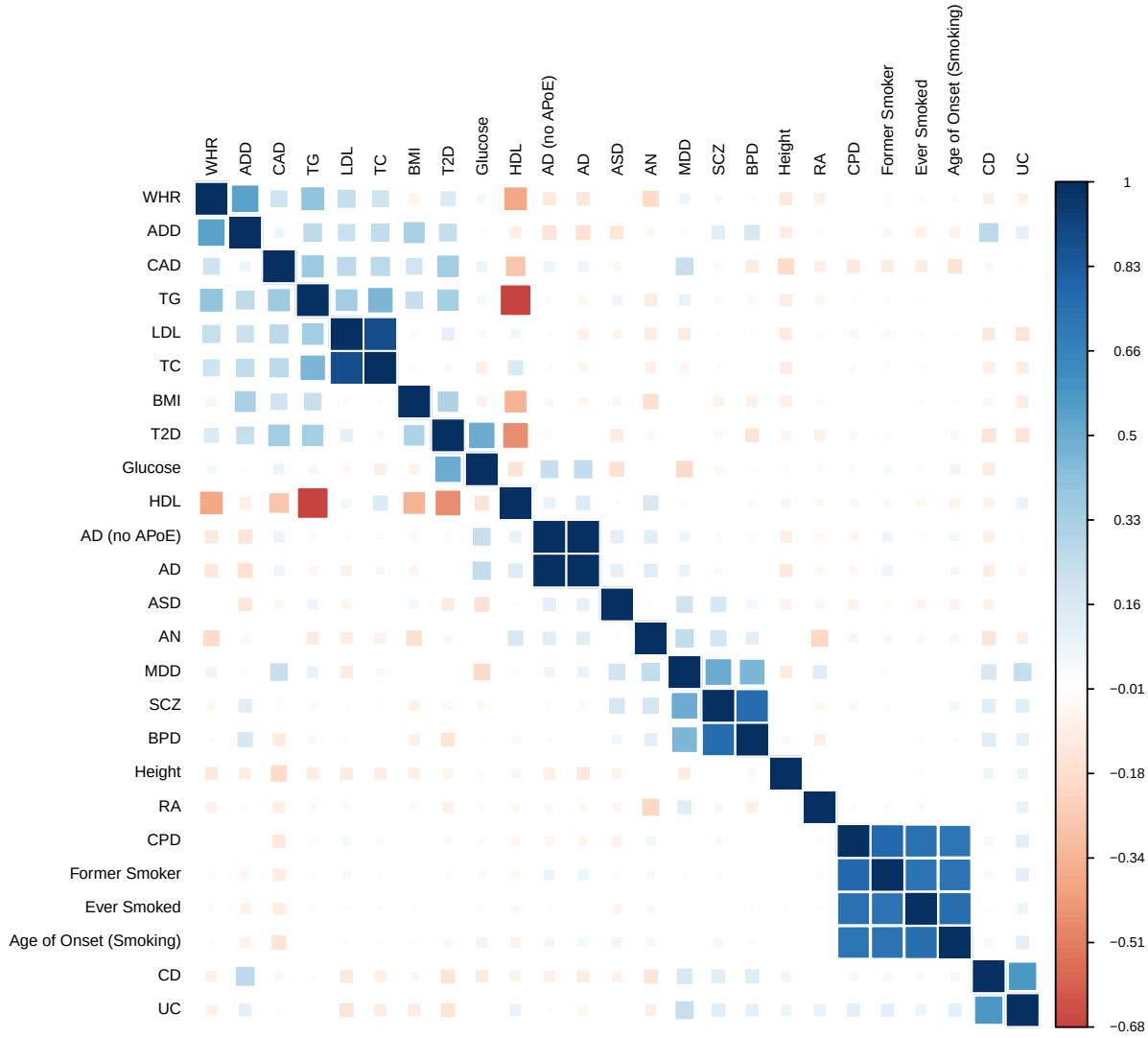
4 Discussion

4.1 Limitations and Future Directions

We note some limitations of LD Score regression. First, LD Score regression requires large sample sizes (at least a few thousand) in order to give estimates with reasonable standard error. For smaller sample sizes, GCTA-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 is 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

Figure 2: Genetic Correlations Between 25 Published GWAS



Placeholder figure (HM3 LD Score). Caption goes here.

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. We provide a flowchart in Supplementary Figure **NNN** describing the situations in which LD Score regression is likely to be useful for estimating genetic correlation.

Finally, 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 positive genetic correlation between LDL and CAD in Figure 2. This genetic correlation could result from a causal effect, *e.g.,* LD $\rightarrow$ CAD, but could also result from shared genetic etiology *e.g.,* LDL $\leftarrow$ G $\rightarrow$ CAD, where G is some set of genetic factors.

Developing methods to distinguish between these models based on genetic data is a particularly exciting direction for future research.

4.2 Data Sharing

Under the header "Minimum Viable Summary Statistics" in the Methods.

4.3 Recap

5 Online Methods

5.1 Statistical Framework

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

5.2 $h^2_{5\text{-}50\%}$

Let S denote the set of all SNPs in 1000 Genomes (or whatever much larger sequenced reference panel future readers are more familiar with); let X_j denote the random variable whose value is the 0-1-2 coded genotype at SNP j, and let y denote a phenotype. Let

$$\beta := \text{argmax}_\alpha \left(\text{Cor} \left[y, \sum_{j \in S} X_j \alpha_j \right] ^2 \right) \tag{5.1}$$

(note that uniqueness of β is guaranteed because this is a projection). Let S' denote the set of SNPs with MAF$>5\%$. Then

$$h^2_{5\text{-}50\%} := \sum_{j \in S'} \beta_j^2. \tag{5.2}$$

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. With larger sample sizes in future sequenced reference panels, this lower bound can be pushed lower.

The main distinction between $h^2_{5\text{-}50\%}$ and h^2_g is that the effects of causal 4% SNPs are not counted towards $h^2_{5\text{-}50\%}$ This differs from the definition of the quantity h^2_g estimated by GCTA-REML, in that GCTA-REML considers a set of SNPs g, and then projects the phenotype onto those SNPs, without accounting for SNPs that are not in the set g. Thus if there is a 5% SNP in set g that is in high LD with a 4% SNP that happens to be causal for the phenotype of interest but is not in g, then the effect of the 4% SNP is counted towards h^2_g (or at least the component of the effect of the 4% SNP that is tagged by the 5% SNP that is in g). There tends not to be very much LD between SNPs with different MAFs, so in the specific case of a MAF cutoff, this distinction likely makes only a small difference.

Technically, we should write $h^2_{5\text{-}50\%,1kG}$ 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 for notational simplicity. 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.

Note that estimates of $h^2_{5\text{-}50\%}$ should generally be less than pedigree-based estimates of heritability (modulo standard error), since pedigree estimators take into account all forms of genetic variation rare variants, microsatellites, indels, copy number variants, non-autosomal variation, etc).

The genetic covariance and genetic correlation quantitates that we estimate are direct bivariate analogues of $h^2_{5\text{-}50\%}$. It is possible that the genetic covariance between two phenotype may be different among MAF$>5\%$ variants than among rare variants; however, we could not detect such a phenomenon with GWAS data, since current GWAS only reliably assay common variation.

5.3 Estimation of LD Scores

We estimated LD Scores from the European samples in the 1000 Genomes Project [7] reference panel using the `--l2` flag in the `ldsc` software package by the authors (URLs) as in [4]. We estimated per-allele LD Scores using the `--per-allele` flag in `ldsc`, and we estimated MAF-binned LD Scores using the `--cts-bin` and `--cts-breaks` flags in `ldsc`. Following [4], we estimated LD Scores using a 1 centiMorgan (cM) window (with the `ldsc` flag `--ld-wind-cm 1`). Unlike [4], we used a MAF cutoff of 1% when estimating LD Scores, in order to reduce the impact of LD measurement error on our regressions. Since we only include variants with MAF $> 5\%$ in LD Score regressions for estimating genetic correlation, and because there is very little LD between variants with MAF $> 5\%$ and variants with MAF $< 5\%$, this is unlikely to impact our results.

For the analyses with HapMap 3 [8] LD Scores, we took the sum of r^2's over the same subset of HapMap 3 SNPs retained for LD Score regression in [4], (that is, HapMap 3 SNPs with MAF $> 1\%$, excluding centromeres and regions with long-range LD) using the `--keep` flag in `ldsc`.

5.4 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 [20]). We allow the variance explained per SNP to vary from bin to bin, but assume that variance explained per SNP is (roughly) equal within each bin. This amounts to approximating the unknown function that maps MAF to variance explained per SNP with a locally constant approximation. This 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. On the other hand, if we use too many bins, the standard error of our estimates will increase. However, we show in the simulations under the header "Misspecified Models of Genetic Architecture" that we can remove almost all MAF- and LD-bias under realistic parameter settings using only a few tens of bins, which increases the standard error only modestly. MAF- and LD- partitioned LD Scores can be estimated using the `--cts-bin` and `--cts-breaks` flags from our `ldsc` software.

5.5 Genetic Covariance Regression Weights

For heritability estimation, we use the LD Score regression weights derived in the supplementary note from [4]. The optimal regression weights for genetic covariance estimation are

$$\mathrm{Var}[\hat{\beta}_j\hat{\gamma}_j \,|\, \ell_j] = \left(\frac{h_1^2\ell_j}{M} + \frac{1-h_1^2}{N_1}\right)\left(\frac{h_2^2\ell_j}{M} + \frac{1-h_2^2}{N_2}\right) + 2\left(\frac{\rho_g\ell_j}{M} + \frac{\rho N_s}{N_1N_2}\right); \tag{5.3}$$

(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 $\rho N_s/N_1N_2$ is negligible (though 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.

5.6 Weighted Block Jackknife Genetic Correlation

This section describes the implementation of the `--sumstats-gencor` 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 [5] of the ratio, with the jackknife taken over blocks of adjacent SNPs

$$\hat{r}_{g,jack} := n_b \hat{r}_g - \sum_{i=1}^{n_b} \left(1 - \frac{m_i}{M_g}\right) \hat{r}_{g,i} \tag{5.4}$$

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 [5],

$$\widehat{\mathrm{Var}}\left[\hat{r}_{g,jack}\right] := \frac{1}{n_b} \sum_{i=1}^{n_b} \frac{1}{h_i - 1} \left((h_i - n_b)\hat{r}_g - (h_i - 1)\hat{r}_{g,i} + \sum_{j=1}^{n_b} \left(1 - \frac{m_i}{M_g}\hat{r}_{g,j}\right), \right) \tag{5.5}$$

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 [4, 11, 17] 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 6.6), 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.

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 (or at least being very cautious with) any genetic correlation estimates where either of the following conditions is met:

1. Either of the heritability estimates is less than 2 SE's from zero, or
2. The block jackknife SE for the genetic correlation estimate is greater than 0.2.

5.7 GWAS Data

5.7.1 Minimum Viable 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)

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).

5.7.2 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).

5.7.3 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.

Note that we only used stage 1 data for LD Score regression.

6 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`
    Note that 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`

7 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.

8 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.

9 Competing Financial Interests

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

10 References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

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

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

[9] International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219, 2011.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.