Estimating Genetic Correlation from GWAS Summary Statistics

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

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

Abstract

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

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

1 Introduction

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

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

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

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

Here, we introduce a computationally trivial method for estimating genetic correlation that requires only GWAS summary statistics and is not biased by sample overlap. We apply this method to data from 25 GWAS and report genetic correlations between 300 pairs of phenotypes. Our results confirm many previously reported genetic overlaps and epidemiological associations, thus validating the utility of genetic correlation as an epidemiological tool. In addition, we report several novel results, including a positive genetic correlation between anorexia nervosa (AN) and schizophrenia and a negative genetic correlation between AN and BMI. These results demonstrate the advantages of a polygenic modeling framework, since there are currently no SNPs associated with AN at genome-wide significance [20].

2 Results

2.1 Overview of Methods

Our method for estimating genetic correlation from summary statistics relies on the fact that the GWAS effect-size estimate for a given SNP incorporates the effects of all SNPs in LD with that SNP[21, 22]. For a polygenic trait, SNPs in regions with strong LD will have higher χ^2-statistics on average than SNPs in regions with little LD [22]. A similar relationship holds if we replace χ^2-statistics for a single study with the product $z_{1j}z_{2j}$, where $z_{i,j}$ denotes the Z-score for study i and SNP j.

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

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

where N_i is the sample size for study i, ρ_g is genetic covariance (defined in methods), ℓ_j is LD Score, N_s is the number of samples shared between study 1 and study 2, and ρ is the phenotypic correlation among the N_s overlapping samples (Supplementary Note). If study one and study two are the same study, then this reduces to the single-phenotype result from [22], because the genetic covariance between a trait and itself is simply h^2, and $\chi^2 = z^2$. A similar relationship holds if one or both of the studies is an ascertained study of a binary phenotype (Supplementary Note), in which case the estimate of genetic covariance will be on the observed scale. As a consequence of equation 2.1, we can estimate genetic covariance using the slope from the regression of $z_{1j}z_{2j}$ on LD Score. This regression does not appear to be sensitive to modeling assumptions about genetic architecture (Methods) This estimate will not be biased even if the two studies share samples, because sample overlap only affects the intercept. Intuitively, bias from sample overlap affects all SNPs equally, similar to cryptic relatedness in the one-trait case [22], and therefore this bias is not correlated with LD Score. Similarly, the estimate of genetic covariance will be only minimally biased by shared population stratification or genotpying artifacts, because these sources of bias are also not correlated with LD Score and will only affect the intercept [22]. If we can rule out shared population stratification, and the intercept term $\rho N_s/\sqrt{N_1 N_2}$ is known ahead of time (*e.g.,* if there is no sample overlap and $N_s = 0$), we can obtain a lower standard error by constraining the intercept to equal this value.

If we normalize genetic covariance to lie in the interval $[-1, 1]$, we obtain genetic correlation: $r_g := \rho_g/\sqrt{h_1^2 h_2^2}$, where h_i^2 denotes the SNP-heritability of trait i. In this paper, we report genetic correlations obtained using genetic covariances estimated with two-phenotype LD Score regression and heritabilities estimated using univariate LD Score regression [22].

2.2 Replication of PGC Cross Disorder Results

For further validation, we replicated the estimates of genetic correlations between psychiatric disorders obtained with individual genotypes and REML in the Psychiatric Genomics Consortium (PGC) Cross-Disorder Group paper [15], using LD Score regression and the summary statistics from [28], downloaded from the PGC website (URLs).

Including an intercept in the LD Score regression protects the results from QC issues such as population stratification (as described in [22]) and sample overlap, but at the cost of larger standard errors. Since the summary statistics from [28] were generated after a careful QC process, and the

3

samples used for each disease were non-overlapping, we also estimated genetic correlation using LD Score regression with constrained intercept.

Results from this analysis are displayed in Figure 1. As expected, the genetic correlation estimates from LD Score regression were similar to the results from REML. LD Score regression without intercept gave standard errors that were only slightly larger than REML, while the standard errors from LD Score regression with intercept were larger, especially for smaller sample sizes (*e.g.,* ADD, ASD).

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

2.3 Application to a Large Set of Publicly Available Summary Statistics

We used LD Score regression to estimate genetic correlations among 25 phenotypes (URLs, Methods). Genetic correlation estimates for all 300 pairwise combinations of the 25 traits are displayed in Figure 2. For clarity of presentation, we pruned to one phenotype from each cluster of highly correlated phenotypes for the main text; highly correlated anthropometric phenotypes are displayed in supplementary table XXX, smoking phenotypes are displayed in supplementary table YYY, and insulin-related traits are displayed in supplementary table ZZZ.

Our estimates of genetic correlations among metabolic traits are generally consistent with the estimates obtained using REML in Vattikuti *et. al.* [16], though our standard errors are much lower, since operating on summary statistics allows us access to larger sample sizes (Supplementary Table 4.1). In addition, our estimate of 0.57 (0.074) for the genetic correlation between Crohn's disease and ulcerative colitis is consistent with the estimate of 0.62 (0.042) from Chen *et. al.* [17].

For the majority of pairs of traits in Figure 2, no GWAS-based genetic correlation estimate has been reported; however, many associations have been described in an *ad-hoc* manner based off the observation of overlap among genome-wide significant loci, or, more formally, via Mendelian randomization analyses. Examples of genetic correlations that are consistent with overlap among top loci include the correlations between plasma lipids and cardiovascular disease [10]; age at onset of menarche and obesity [29]; type 2 diabetes, obesity, fasting glucose, plasma lipids and cardiovascular disease [30]; birth weight, adult height and type 2 diabetes [31]; birth length, adult height and infant head circumference [32, 33]; and childhood obesity and adult obesity [32]. For many of these pairs of traits, we can reject the null hypothesis of zero genetic correlation with overwhelming statistical significance (*e.g.,* $p = 6 \times 10^{-24}$ for age at onset of menarche and obesity; $p = 6 \times 10^{-52}$ for obesity and childhood obesity). Our results on metabolic traits are directionally consistent with the recent Mendelian randomization results from [34].

The first row of table 1 lists genetic correlation results that are consistent with well-known epidemiological association, but, to the best of our knowledge, have never been reported using genetic data. For example, our estimates of the genetic correlation between age at onset of menarche and adult height [35], cardiovascular disease [36] and type 2 diabetes [36, 37] are consistent with the epidemiological associations. We estimate a statistically significant negative genetic correlation between educational attainment and Alzheimer's disease, consistent with the epidemiological observation that low educational attainment is one of the largest risk factors for Alzheimer's [38, 39]. The positive genetic correlation between educational attainment and bipolar disorder is consistent with psychiatric literature showing that education and bipolar disorder status are positively correlated [40, 41], Our estimate of a negative genetic correlation between AN and obesity (and a similar
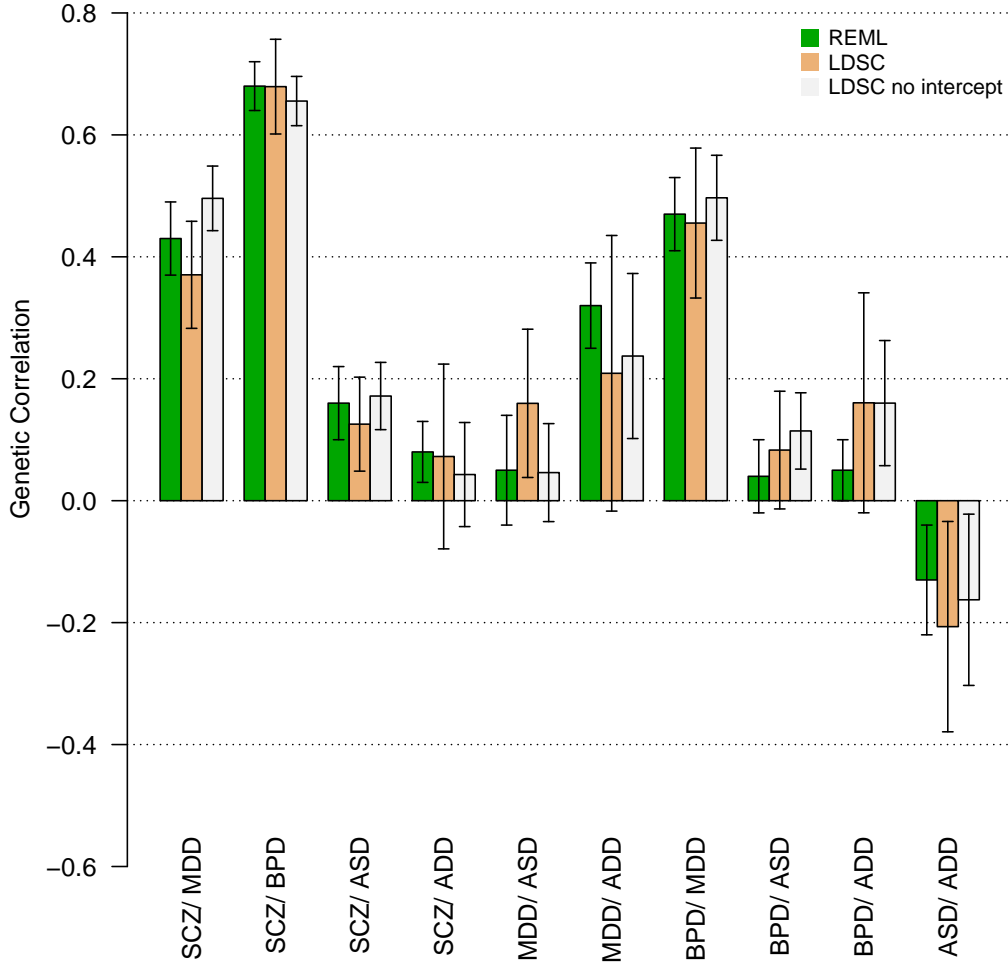
Figure 1: *Replication of PGC Cross-Disorder Results. This plot compares LD Score regression estimates of genetic correlation using the summary statistics from [28] (which were generated from approximately the same data as [15]) to estimates obtained from REML in [15]. The horizontal axis indicates pairs of phenotypes, and the vertical axis indicates genetic correlation. The error bars show standard errors. Colors indicate different estimation procedures. Green is REML, orange is LD Score with intercept and white is LD Score without intercept. The estimates of genetic correlation between psychiatric phenotypes in figure 2 use larger sample sizes; this plot is intended primarily as a technical validation. Abbreviations: ADD = attention deficit hyperactivity disorder; ASD = autism spectrum disorder; BPD = bipolar disorder; MDD = major depressive disorder; SCZ = schizophrenia.*

genetic correlation with BMI) is intriguing given that the most striking feature of anorexia nervosa is a biologically perplexing ability to achieve and maintain dangerously low BMI [42] and suggests that similar genetic factors may influence normal variation in BMI as well as severely dysregulated BMI in psychiatric illness. This result is consistent with our observation that BMI GWAS findings appear to primarily implicate neuronal, rather than metabolic, cell-types and epigenetic marks
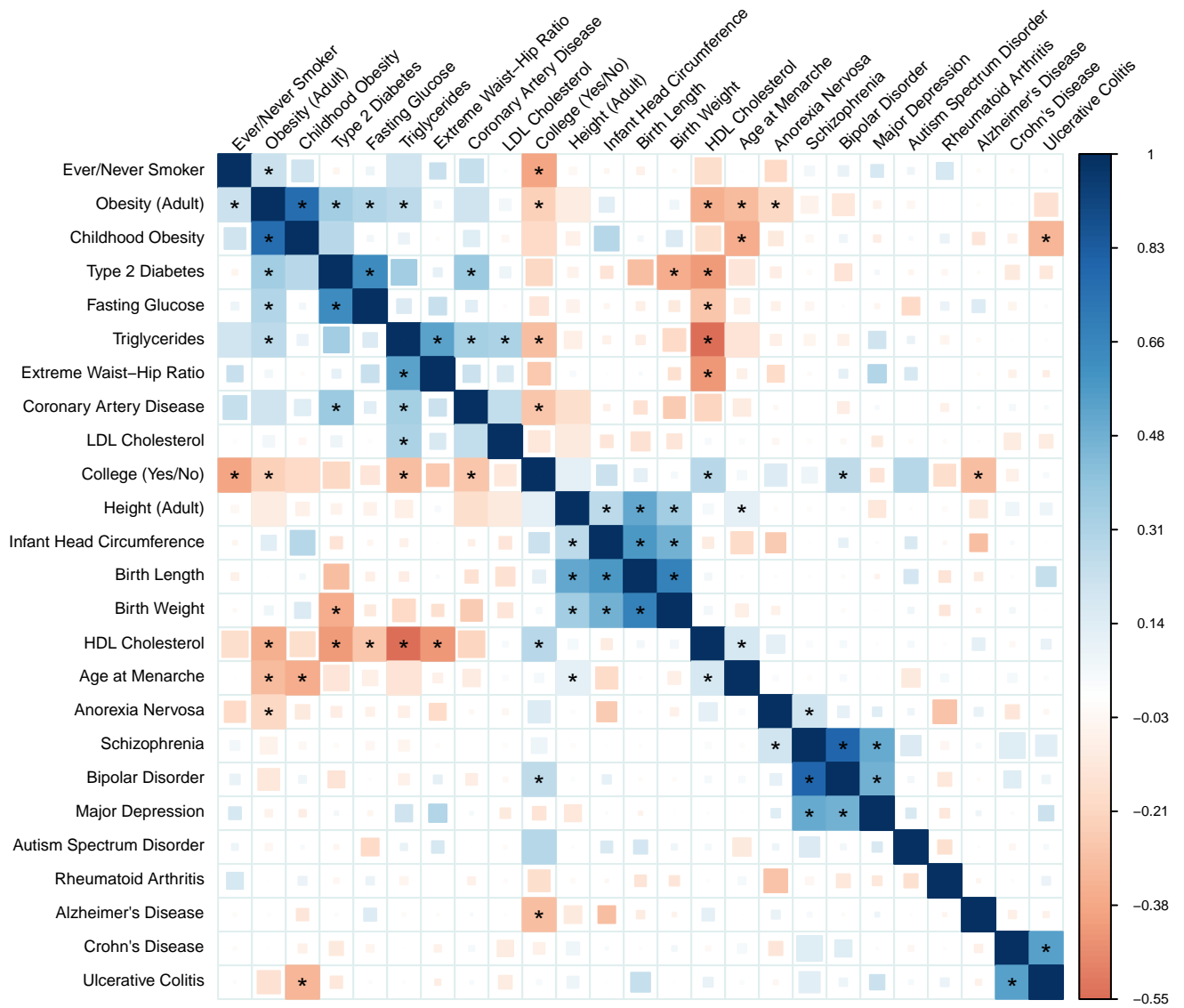
Figure 2: *Genetic Correlations among 25 Published GWAS. The color of the square indicates the value of the genetic correlation. Blue corresponds to positive genetic correlations; red corresponds to negative genetic correlation. The size of the colored squares indicates the magnitude of the p-value. Larger squares correspond to lower p-values, and all genetic correlations that are different from zero at 1% FDR are displayed as full-sized squares. Genetic correlations that are significantly different from zero at significance level 0.05 after Bonferroni correction for 300 tests are given an asterisk. Note that this multiple testing correction is conservative, since many of the tests are not independent.*

	Phenotype 1	Phenotype 2	r_g (se)	p-value
	Age at Menarche	Height (Adult)	0.11 (0.03)	6×10^{-5} **
	Age at Menarche	Type 2 Diabetes	-0.13 (0.04)	3×10^{-3}
	Age at Menarche	Triglycerides	-0.15 (0.04)	1×10^{-3} *
	Coronary Artery Disease	Age at Menarche	-0.11 (0.05)	4×10^{-2}
	Coronary Artery Disease	College (Yes/No)	-0.278 (0.07)	1×10^{-4} **
	Coronary Artery Disease	Height (Adult)	-0.17 (0.05)	2×10^{-4} *
	Alzheimer's	College (Yes/No)	-0.30 (0.08)	1×10^{-4} **
Epidemiological	Bipolar Disorder	College (Yes/No)	0.026 (0.064)	6×10^{-5} **
	Obesity (Adult)	College (Yes/No)	-0.23 (0.04)	2×10^{-8} **
	Triglycerides	College (Yes/No)	-0.30 (0.04)	5×10^{-12} **
	Anorexia Nervosa	Obesity (Adult)	-0.20 (0.04)	4×10^{-6} **
	Schizophrenia	Crohn's Disease	0.013 (0.04)	3×10^{-3}
	Schizophrenia	Ulcerative Colitis	0.012 (0.05)	1×10^{-2}
	Ever/Never Smoker	College (Yes/No)	-0.39 (0.07)	1×10^{-9} **
	Ever/Never Smoker	Obesity (Adult)	0.22 (0.05)	7×10^{-5} **
	Autism Spectrum Disorder	College (Yes/No)	0.28 (0.08)	5×10^{-4} *
Novel	Ulcerative Colitis	Childhood Obesity	-0.33 (0.08)	3.9×10^{-5} **
	Anorexia Nervosa	Schizophrenia	0.19 (0.04)	1.5×10^{-5} **
	Anorexia Nervosa	Alzheimer's	0.09 (0.09)	0.32
	Major Depression	Alzheimer's	0.08 (0.15)	0.60
	Bipolar Disorder	Alzheimer's	-0.021 (0.09)	0.82
	Autism Spectrum Disorder	Alzheimer's	0.12 (0.12)	0.30
	Schizophrenia	Alzheimer's	0.05 (0.05)	0.58
Null	Schizophrenia	Ever/Never Smoker	0.03 (0.06)	0.26
	Schizophrenia	Triglycerides	-0.05 (0.04)	0.21
	Schizophrenia	LDL Cholesterol	-0.02 (0.03)	0.64
	Schizophrenia	HDL Cholesterol	0.03 (0.04)	0.50
	Schizophrenia	Rheumatoid Arthritis	-0.05 (0.05)	0.38
	Crohn's Disease	Rheumatoid Arthritis	-0.02 (0.09)	0.83
	Ulcerative Colitis	Rheumatoid Arthritis	-0.09 (0.09)	0.33

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

[24]. Our estimate of the genetic correlations between schizophrenia and inflammatory bowel disease (IBD – Crohn's and UC) are not statistically significantly different from zero after correction for multiple tests, but the direction of the genetic correlation (positive) is consistent with recent epidemiological work showing an increased risk for schizophrenia after IBD diagnosis [43] and an increased risk for IBD after schizophrenia diagnosis [44], although we emphasize that the evidence

for genetic correlation is not yet conclusive. The positive genetic correlation between smoking and obesity is in the opposite direction of the epidemiological association (smoking is associated with lower BMI, although smoking cessation often leads to weight gain) and Mendelian randomization results using rs1051730 [46]. We propose that the resolution of this apparent discrepancy is an interesting direction for future research. Our estimate of a negative genetic correlation between smoking and educational attainment is consistent with the epidemiological result that smoking is more prevalent among less-educated groups [47]. Last, the negative genetic correlation between height and coronary artery disease (possibly related to the negative genetic correlation between height and LDL, a known risk factor for coronary artery disease [10, 48]) is consistent with a replicated epidemiological association [49, 50, 51].

The second row of table 1 lists results that are, to the best of our knowledge, novel. First, we find a positive genetic correlation between between anorexia nervosa and schizophrenia. Suggestive evidence for comorbidity between anorexia nervosa and schizophrenia has been reported previously [52, 53]; however, our result is more statistically significant. Second, we estimate a negative genetic correlation between ulcerative colitis and childhood obesity [**waiting on advice from Mark Daly about what to say here**]. Third, we estimate a positive genetic correlation between autism spectrum disorder (ASD) and educational attainment (a proxy for cognitive function [54, 55, 56], at least after controlling for socioeconomic status). The ASD summary statistics were generated using a case/pseudocontrol study design, so this result cannot be explained by diagnostic biases, such as the tendency for the parents of children who receive a diagnosis of ASD to be better educated than the general population [57]. The distribution of IQ among individuals with ASD has lower mean than the general population, but with heavy tails [58] (*i.e.,* an excess of individuals with extremely low and extremely high IQ). In addition, there is evidence that the genetic architectures of high IQ and low IQ ASD are dissimilar [59]. We are unable to offer an explanation for this result, but propose that further exploration of this genetic correlation is an interesting direction for future research.

Finally, the third row of table 1 lists several instances where the genetic correlation is close to zero, in contrast to previous reports of epidemiological correlation or genetic overlap. We estimate genetic correlations close to zero between schizophrenia and rheumatoid arthritis, schizophrenia and smoking, and schizophrenia and plasma lipids. The lack of genetic correlation between schizophrenia and rheumatoid arthritis is interesting because schizophrenia has been observed to be protective for rheumatoid arthritis [60]. The absence of genetic correlation between schizophrenia and smoking is notable because of the high prevalence of smoking among individuals with schizophrenia [61]. Our estimate of zero genetic correlation between schizophrenia and plasma lipids or cardiovascular disease contrasts with earlier reports of extensive pleiotropy between schizophrenia and triglycerides [62]. However, the observation from Andreassen, *et. al.* [62] could potentially be explained the sensitivity of the method used to the properties of a few regions with high linkage disequilibrium, rather than trait biology (Supplementary Table XXX). We estimate near-zero genetic correlations between Alzheimer's disease and five psychiatric disorders. This is notable given the high genetic correlations among the five psychiatric disorders, and suggests that the genetic basis of Alzheimer?s disease is distinct from these other conditions.

Last, we estimate near zero genetic correlation between rheumatoid arthritis and both Crohn's disease and ulcerative colitis. Although these immune diseases share a large number of associated loci [63, 64], it is as often the case that an allele that is risk-increasing for one immune disease is protective for another as an allele is risk-increasing for two [63], consistent near-zero genetic

correlation overall. This is an example of pleiotropy without genetic correlation (see Methods for a discussion of the distinction).

3 Discussion

We have described a new method for estimating genetic correlation from GWAS summary statistics. We applied our method to a large dataset of publicly available GWAS summary statistics, spanning 25 traits and more than 1.5 million phenotype-genotype pairs. We replicated many previously-reported GWAS-based genetic correlations, and confirm observations of overlap among genome-wide significant SNPs, Mendelian randomization results and known epidemiological associations, thus validating the utility of genetic correlation as an epidemiological tool. In addition, we report several novel results that merit further analysis, including a positive genetic correlation between educational attainment and ASD and a positive genetic correlation between anorexia nervosa and schizophrenia.

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

We note some limitations on the interpretation of genetic correlation. In general, the difficulties in interpreting genetic correlation are similar to the difficulties in interpretation of Mendelian randomization. Although genetic correlation is immune to environmental confounding, genetic correlation is still subject to genetic confounding, which is normally called confounding via pleiotropy in the Mendelian randomization literature. This is best illustrated with an example. Consider the strong negative genetic correlation between HDL and CAD in Figure 2. This genetic correlation could result from a direct causal effect HDL $\rightarrow$ CAD, but could also result from non-causal shared genetic etiology, perhaps mediated by the genetic correlation of -0.55 between HDL and triglycerides (TG). This scenario can be represented graphically as HDL $\leftarrow$ G $\rightarrow$ TG $\rightarrow$ CAD, where G is the set of genetic variants with effects on both HDL and TG. This diagram would result in a genetic correlation between HDL and CAD, because there is an unblocked path from HDL to CAD [65]. Indeed, the second mechanism appears to account for a large proportion of the genetic correlation between HDL and CAD, since the association between HDL and CAD is attenuated after controlling for other plasma lipids [10, 48]. Extending genetic correlation to deal with multiple phenotypes (*e.g.,* genetic correlation between HDL and CAD controlling for TG, analogous to [10, 48], possibly with extension to nonlinear models) is an important direction for future work.

We note several limitations of the LD Score regression method. First, LD Score regression requires large sample sizes in order to give estimates with acceptable standard error (as a rule of thumb, Nh_g^2 should be greater than 3000, where h_g^2 is on the observed scale for case/control studies). Constrained intercept LD Score regression gives lower standard errors, but this estimator requires knowing the number of overlapping samples, which can often be difficult to determine for pairs of large meta-analyses. Methods such as REML that use individual genotypes are more statistically efficient. Second, LD Score regression assumes that all individuals in the GWAS were sampled from

populations with similar LD Scores, and that these LD Scores can be estimated from sequenced reference data such as 1000 Genomes [26]. Differences between LD patterns in the reference panel and the GWAS sample may result in bias [22]. Third, there is currently no procedure for applying LD Score regression to samples from admixed populations. Fourth, LD Score regression performs optimally when applied to traits with highly polygenic genetic architectures, such as psychiatric traits. At large sample size and for less-polygenic traits (*e.g.,* traits where a substantial fraction of heritability is concentrated among a handful of SNPs), it is sometimes possible to obtain better power by analyzing only the large-effect SNPs. Developing methods that can make optimal use of both confidently associated large-effect SNPs and diffuse polygenic signal from thousands of small-effect SNPs is a direction for future research. Fifth, while LD Score regression is a consistent estimator of genetic correlation in ascertained case/control studies, the results may nonetheless be skewed by more subtle biases in ascertainment, such as misclassification of cases.

We have developed a method that allows us to rapidly screen hundreds of pairs of phenotypes for associations without having to measure all phenotypes on all individuals. Our results show that the matrix of genetic correlations – especially among metabolic traits – is dense, and that genome-wide genetic pleiotropy is the rule rather than the exception. These findings highlight the importance of methods that account naturally for pleiotropy [10, 48], or even use pleiotropy to gain power [66, 63].

4 Online Methods

4.1 Statistical Framework

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

4.2 Heritability

Let S denote a set of M SNPs, let X denote the random M-vector of additively (0-1-2) coded genotypes for the SNPs in S, and let y denote a phenotype. Define

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

where the maximization is performed in the population (*i.e.,* in the infinite data limit), rather than in a finite sample. This is a projection, so uniqueness of β is guaranteed as we remove SNPs that are linearly dependent (in the population). Then h_S^2, the heritability accounted for by SNPs in S, is defined

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

We obtain the Yang/Visscher parameter h_g^2 by taking S to be the set of genotyped SNPs. If we let S denote the set of all SNPs in 1000 Genomes Europeans [26], and let S' denote the set of SNPs with MAF$> 5\%$, then

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

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

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

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

The value of $h^2_{5\text{-}50\%}$ will always be less than the total narrow-sense heritability, h^2, since h^2 takes into account all forms of genetic variation – variants with MAF under 5%, microsatellites, indels, copy number variants – not just common SNPs. In addition, estimates of h^2 from family studies may be biased upwards by non-additive genetic architectures [67].

4.3 Genetic Covariance and Correlation

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

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

and

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

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 [26], and let S' denote the set of SNPs with MAF$> 5\%$. Then

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

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

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

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

It is however important to note that all of the flavors of GWAS genetic covariance and correlation (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 r_g and $r_{5\text{-}50\%}$ to the total narrow-sense genetic correlation. For example, if β and γ are strongly correlated among common variants, but only weakly correlated among rare variants, then the total narrow-sense genetic correlation will be less than r_g and $r_{5\text{-}50\%}$.

4.4   Genetic Correlation is Different from Pleiotropy

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

Nevertheless, both quantities are informative. If two phenotypes are influenced by variants at the same loci or in the same pathways, this may well indicate interesting shared biology, even if the direction of the effects do not align. However, quantifying genome-wide pleiotropy from GWAS summary statistics remains an open challenge [68]. For example, since p-values tend to decrease with LD Score [22] and near coding and regulatory regions [69], the observation that low p-values for one phenotype tend to predict low p-values for another (as in [70, 71, 62]) may merely reflect properties shared by almost all pairs of heritable phenotypes, rather than a special property of a specific pair of phenotypes (Supplementary Figures 4.2, 4.3, 4.4). Genetic correlation estimates from LD Score regression are not affected by these concerns, and so may be more easily interpretable than currently-available estimates of genome-wide pleiotropy.

4.5   Two-Phenotype LD Score Regression

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

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

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

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

We can estimate heritability using LD Score regression (as described in [22]), and use these heritability estimates to transform the estimates of genetic covariance into estimates of genetic correlation (see Section 4.7 in the Methods).

An equation similar to Equation (4.9) holds if one or both of the studies is an ascertained study of a binary phenotype, and so the same method can be used regardless of whether the Z-scores are from studies of quantitative or case-control traits (Supplementary Note). If the variance of effect sizes depends on minor allele frequency (MAF) or linkage disequilibrium (LD), as discussed in [23, 22], 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 [24].

4.6   Genetic Covariance Regression Weights

For heritability estimation, we use the LD Score regression weights derived in the Supplementary Note from [22]. If effect sizes for both phenotypes are drawn from a bivariate normal distribution, then the optimal regression weights for genetic covariance estimation are

$$\text{Var}[z_{1,j}z_{2,j} \,|\, \ell_j] \propto \left(\frac{h_1^2 \ell_j}{M} + \frac{1}{N_1} \right) \left(\frac{h_2^2 \ell_j}{M} + \frac{1}{N_2} \right) + 2 \left(\frac{\rho_g \ell_j}{M} + \frac{\rho N_s}{N_1 N_2} \right)^2 \tag{4.10}$$

(Supplementary Note). This quantity depends on both heritabilities, the genetic covariance and the number of overlapping samples, which are not known a priori, so it is necessary to estimate them from the data. Linear regression with weights estimated from the data is feasible generalized least squares (FGLS), which gives valid inference.

In order to estimate genetic covariance with efficient regression weights, we run a first LD Score regression weighted with weights computed using the heritability estimates from the single-phenotype LD Score regressions, N_s set to zero, and ρ_g estimated with the aggregate estimator

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

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

We then update the weights using the estimates of ρ_g and ρN_s obtained from the first LD Score regression.

4.7   Weighted Block Jackknife Genetic Correlation

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

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

Instead of the naive estimator of this ratio,

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

we use the weighted block jackknife estimator [72] of the ratio, with the jackknife taken over blocks of adjacent SNPs

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

14

where n_b is the number of blocks, and $\hat{r}_{g,i}$ is the naive estimate of genetic correlation obtained by deleting the ith block of SNPs, m_i is the number of SNPs in block i, and M_g is the number of SNPs included in the regression. The weighted block jackknife ratio estimator is less biased than the naive estimate (though this is not so important at the sample sizes that we consider), and comes with a convenient nonparametric heteroskedasticity- and correlation-robust variance estimator [72],

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

where $h_i := M_g/m_i$. Weighted block jackknife standard errors (over blocks of adjacent SNPs) are robust to the correlated error structure of GWAS χ^2-statistics, so long as the block size exceeds the typical range of LD. See references [22, 24, 73] for examples of papers in the statistical and population genetics literature that use this technique. We checked the reliability of our standard errors via simulations with real genotypes (Supplementary Table 5.4), and found that the `ldsc` default setting of 2000 blocks genome-wide (which can be adjusted with the `--num-blocks` flag) gives standard error estimates that agree well with the empirical standard deviation across simulation replicates, even though many of the SNPs included in the regression are in LD.

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

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

4.8 Simulations

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

Estimates of heritability and genetic covariance can be biased if the underlying model of genetic architecture is misspecified [23] Estimates of genetic correlation are more robust to model misspecification biases than estimates of heritability or genetic covariance. Since genetic correlation is estimated as a ratio, model misspecification biases that affect the numerator and the denominator in the same direction will tend to cancel. In situations where MAF- or LD-dependent genetic architectures are a particular concern, it is possible to correct for such biases with LD Score regression using a MAF or LD binning approach (see [24] and Online Methods) similar to that taken by Lee, *et. al.* with REML [25].

To quantify the bias introduced by MAF- or LD-dependent genetic architectures, we simulated a variety of different LD Scores and genetic architectures. In realistic scenarios, only a subset of causal SNPs are directly genotyped or successfully imputed, so we used a densely imputed panel

of 1000 Genomes (1kG) SNPs [26] in order to generate phenotypes and estimate LD Scores, but computed summary statistics only for the 16% of 1kG SNPs that are also in HapMap3 (HM3) [27] with MAF above 5%. Results from these simulations are displayed in Supplementary Tables 5.1, 5.2 and 5.3.

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

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

4.9   Complexity

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

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

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

4.10   Selection of Datasets

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

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

2. Remove all traits with heritability Z-score below 4, in order to remove traits where the genetic correlation estimates will be too noisy to interpret.

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

4.11 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 not be "corrected" via genomic control (GC) correction. Using GC'ed summary statistics will result in downward bias in the LD Score regression estimates of heritability and genetic covariance, and deflated LD Score regression intercepts. Estimates of genetic correlation will not be affected.

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

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

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

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

3. Sample MAF.

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

If these data are not available, we recommend retaining only HM3 SNPs with reference panel MAF above 5% for the LD Score regression as a workaround (note: for *regression*, not for estimation of the LD Scores), since HM3 SNPs seem to be well-imputed in most studies. For newer studies with dense imputation, restricting to HM3 SNPs is an inefficient use of data. Using a larger set of SNPs for the regression will lower the standard error (*e.g.,* using all common 1kG SNPs instead of all common HM3 SNPs for the regression reduces the standard error by about 10% in simulations).

4.12 Large Effect Loci

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

4.13 IGAP

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

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

We only used stage 1 data for LD Score regression.

5   URLs

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6 Acknowledgements

7 Author Contributions

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

8 Competing Financial Interests

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

9 References

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

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

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

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

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

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

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

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

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

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

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

[12] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, 2010.

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

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

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

[16] Shashaank Vattikuti, Juen Guo, and Carson C Chow. Heritability and genetic correlations explained by common snps for metabolic syndrome traits. *PLoS genetics*, 8(3):e1002637, 2012.

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

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

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

[20] 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.

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

[22] 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.

[23] 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.

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

[25] 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.

[26] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

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

[28] 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.

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

[30] 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.

[31] Momoko Horikoshi, Hanieh Yaghootkar, Dennis O Mook-Kanamori, Ulla Sovio, H Rob Taal, Branwen J Hennig, Jonathan P Bradfield, Beate St Pourcain, David M Evans, Pimphen Charoen, et al. New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nature genetics*, 45(1):76–82, 2013.

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

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

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

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

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

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

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

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

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

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

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

[43] 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.

[44] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[62] 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.

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

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

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

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

[67] 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.

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

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

[70] 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.

[71] 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.

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

[73] 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.