nature genetics

An atlas of genetic correlations across human diseases and traits

Brendan Bulik-Sullivan[1–3,9], Hilary K Finucane[4,9], Verneri Anttila[1–3], Alexander Gusev[5,6], Felix R Day[7], Po-Ru Loh[1,5], ReproGen Consortium[8], Psychiatric Genomics Consortium[8], Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3[8], Laramie Duncan[1–3], John R B Perry[7], Nick Patterson[1], Elise B Robinson[1–3], Mark J Daly[1–3], Alkes L Price[1,5,6,10] & Benjamin M Neale[1–3,10]

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

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

The first methods to test for genetic overlap were family studies[3–7]. To estimate the genetic overlap for many pairs of phenotypes, family study designs require the measurement of multiple traits for the same individuals. Consequently, it is challenging to scale these designs to a large number of traits, especially traits that are difficult or costly to measure (for example, low-prevalence diseases). More recently, GWAS have allowed effect size estimates to be obtained for specific genetic variants, so it is possible to test for shared genetics by looking for correlations in effect sizes across traits, which does not require measuring multiple traits per individual.

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

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

To overcome these limitations, we have developed a technique for estimating genetic correlation using only GWAS summary statistics that is not biased by sample overlap. Our method, cross-trait LD Score

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Stanley Center for Psychiatric Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA. [4]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [6]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [7]Medical Research Council (MRC) Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK. [8]A full list of members and affiliations appears in the **Supplementary Note**. [9]These authors contributed equally to this work. [10]These authors jointly supervised this work. Correspondence should be addressed to B.B.-S. (bulik@broadinstitute.org), B.M.N. (bneale@broadinstitute.org), H.K.F. (hilaryf@mit.edu) or A.L.P. (aprice@hsph.harvard.edu).

Received 2 February; accepted 26 August; published online 28 September 2015; doi:10.1038/ng.3406

regression, is a simple extension of single-trait LD Score regression[19] and is computationally very fast. We apply this method to data from 24 GWAS and report genetic correlations for 276 pairs of phenotypes, demonstrating shared genetic bases for many complex traits and diseases.

RESULTS

Overview of the methods

The method presented here for estimating genetic correlation from summary statistics relies on the fact that the GWAS effect size estimate for a given SNP incorporates the effects of all SNPs in linkage disequilibrium (LD) with that SNP[19,20]. For a polygenic trait, SNPs with high LD will have higher χ^2 statistics on average than SNPs with low LD[19]. A similar relationship holds if we replace the χ^2 statistics for a single study with the product of the z scores from two studies of traits with nonzero genetic correlation.

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

$$E[z_{1j}z_2\ell_j] = \frac{\sqrt{N_1N_2}\,\varrho_g}{M}\ell_j + \frac{\varrho N_s}{\sqrt{N_1N_2}} \quad (1)$$

where N_i is the sample size for study i, ϱ_g is the genetic covariance (defined in the Online Methods), ℓ_j is the LD Score[19], N_S is the number of individuals included in both studies and ϱ is the phenotypic correlation among the N_S overlapping samples. We derive this equation in the **Supplementary Note**. If study 1 and study 2 are the same study, then equation (1) reduces to the single-trait result from ref. 19, as the genetic covariance between a trait and itself is heritability, and $\chi^2 = z^2$. As a consequence of equation (1), we can estimate genetic covariance using the slope from the regression of $z_{1j}z_{2j}$ on LD Score, which is computationally very fast (Online Methods).

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

Normalizing genetic covariance by SNP heritabilities yields genetic correlation: $r_g := \varrho_g/\sqrt{h_1^2h_2^2}$ where h_i^2 denotes the SNP heritability[11] from study i. Genetic correlation ranges between −1 and 1. Results similar to equation (1) hold if one or both studies are case-control studies, in which case, genetic covariance is on the observed scale. Details are provided in the **Supplementary Note**. There is no distinction between observed- and liability-scale genetic correlation for case-control traits, so we can define and estimate the genetic correlation between a case-control trait and a quantitative trait and the genetic correlation between pairs of case-control traits without the need to specify a scale (**Supplementary Note**).

Table 1 Simulations with complete sample overlap

Parameter	True value	Estimate	s.d.	s.e.
h^2	0.58	0.58	0.072	0.075
ρ_g	0.29	0.29	0.057	0.058
r_g	0.50	0.49	0.079	0.073

Estimates represent the average cross-trait LD Score regression estimates across 1,000 simulations; s.d. represents the standard deviation of the estimates across 1,000 simulations, and s.e. represents the mean cross-trait LD Score regression standard error across 1,000 simulations. Further details on the simulation setup are given in the Online Methods.

Simulations

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

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

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

Replication of psychiatric cross-disorder results

As technical validation, we replicated the estimates of genetic correlation among psychiatric disorders obtained with individual-level genotype data and REML[14] by applying cross-trait LD Score regression to summary statistics from the same data[22]. These summary statistics were generated from non-overlapping samples, so we applied cross-trait LD Score regression using both unconstrained and constrained intercepts (Online Methods). The results from cross-trait LD Score regression were similar to the results from REML (**Fig. 1**). Cross-trait LD Score regression with a constrained intercept gave standard errors that were only slightly larger than those from REML, whereas the standard errors from cross-trait LD Score regression with intercept were substantially larger, especially for traits with small sample sizes

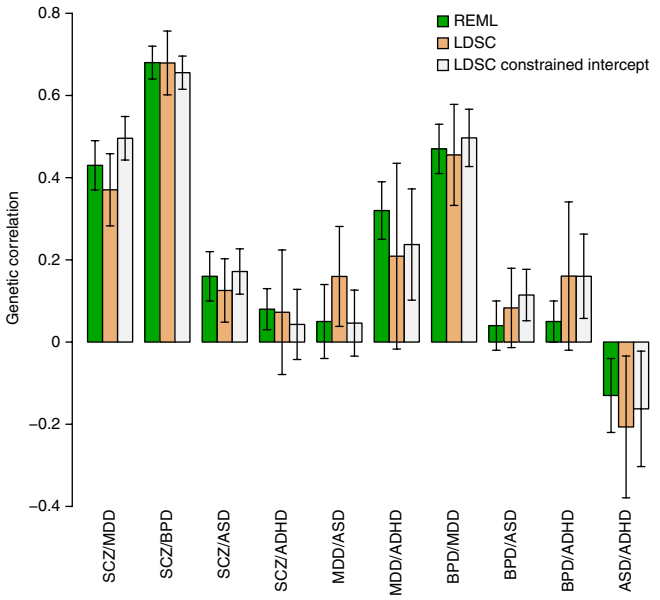

Figure 1 Replication of psychiatric cross-disorder results. This plot compares cross-trait LD Score (LDSC) regression estimates of genetic correlation using summary statistics from ref. 22 to estimates obtained from REML with the same data[14]. The horizontal axis indicates pairs of phenotypes, and the vertical axis indicates genetic correlation. Error bars represent standard errors. The estimates of genetic correlation among psychiatric phenotypes in **Figure 2** use larger sample sizes; the analysis here is intended as a technical validation. ADHD, attention deficit hyperactivity disorder; ASD, autism spectrum disorder; BPD, bipolar disorder; MDD, major depressive disorder; SCZ, schizophrenia.

(for example, attention deficit and hyperactivity disorder (ADHD) and autism spectrum disorder (ASD)).

Application to summary statistics from 24 phenotypes

We used cross-trait LD Score regression to estimate genetic correlations among 24 phenotypes[23–28] (see URLs and the Online Methods). The genetic correlation estimates for all 276 pairwise combinations of the 24 traits are shown in **Figure 2**. For clarity of presentation, the 24 phenotypes were restricted to include only one phenotype from each cluster of closely related phenotypes (Online Methods). The genetic correlations among the educational, anthropometric, smoking and insulin-related phenotypes that were excluded from **Figure 2** are shown in **Supplementary Figures 1–4**, respectively. Corresponding references and sample sizes are shown in **Supplementary Table 3**. A full table of the 1,176 genetic correlations among the 49 traits is provided in **Supplementary Table 4**.

The first section of **Table 2** lists genetic correlation results that are consistent with epidemiological associations but, as far as we are aware, have not previously been reported using genetic data. The estimates of genetic correlation between age at menarche and adult height[29] and between triglyceride levels[30] and type 2 diabetes (T2D)[30,31] are consistent with epidemiological associations. The estimate of a negative genetic correlation between anorexia nervosa and obesity suggests that the same genetic factors influence normal variation in body mass index (BMI) and dysregulated BMI in psychiatric illness. This result is consistent with the observation that GWAS findings for BMI implicate neuronal rather than metabolic cell types and epigenetic marks[32,33]. The negative genetic correlation between adult height and coronary artery disease (CAD) agrees with a replicated epidemiological association[34–36]. We observed several significant associations with the educational attainment phenotypes from Rietveld *et al.*[37]: we estimated a statistically significant negative genetic correlation between college attendance and Alzheimer's disease, which agrees with epidemiological results[38,39]. The positive genetic correlation between college attendance and bipolar disorder is consistent with previous epidemiological reports[40,41]. The estimate of a negative genetic correlation between smoking and college attendance is consistent with observed differences in smoking rate as a function of educational attainment[42].

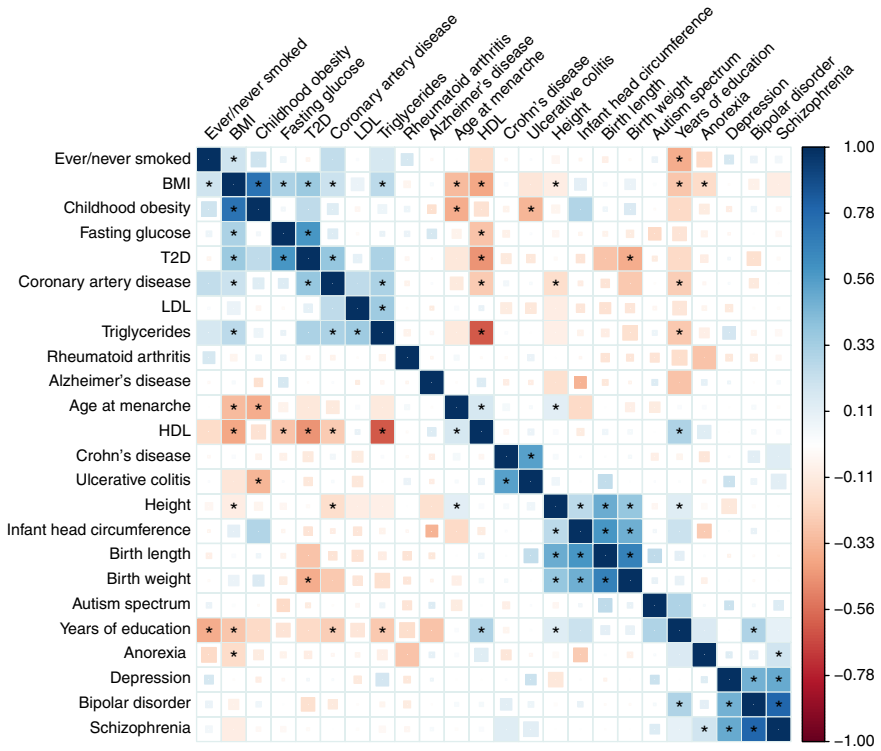

Figure 2 Genetic correlations among 24 traits analyzed by genome-wide association. Blue, positive genetic correlation; red, negative genetic correlation. Larger squares correspond to more significant *P* values. Genetic correlations that are different from zero at a false discovery rate (FDR) of 1% are shown as full-sized squares. Genetic correlations that are significantly different from zero after Bonferroni correction for the 300 tests in this analysis are marked with an asterisk. We show results that do not pass multiple-testing correction as smaller squares, to avoid obscuring positive controls, where the estimates point in the expected direction but do not achieve statistical significance owing to small sample size. The correction for multiple testing is conservative as the tests are not independent. To keep this figure to a reasonable size, we have selected representatives of several clusters of highly correlated traits. Additional genetic correlations are shown in **Supplementary Figures 1–4**. All genetic correlations in this report can be found in tabular form in **Supplementary Table 4**. BMI, body mass index; T2D, type 2 diabetes; LDL, low-density lipoprotein; HDL, high-density lipoprotein.

Table 2 Genetic correlation estimates, standard errors and *P* values for selected pairs of traits

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

Results are grouped into genetic correlations that are new genetic results but are consistent with established epidemiological associations ("Epidemiological"), genetic correlations that are new to both genetics and epidemiology ("New/nonzero") and interesting null results ("New/low"). The *P* values are uncorrected *P* values. Results that passed multiple-testing corrections for the 300 tests in **Figure 2** at an FDR of 1% have a single asterisk; results that passed Bonferroni correction have two asterisks. We present some genetic correlations that agree with epidemiological associations but that did not pass multiple-testing correction in these data.

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

The third section of **Table 2** lists interesting examples where the genetic correlation was close to zero with small standard error. The low genetic correlation between schizophrenia and rheumatoid arthritis is interesting because schizophrenia has been observed to be protective for rheumatoid arthritis[50], although the epidemiological effect is weak; it is thus possible that there is a real genetic correlation but that it was too small for us to detect. The low genetic correlation between schizophrenia and smoking is notable because of the increased tobacco use (in terms of both prevalence and number of cigarettes per day) among individuals with schizophrenia[51]. The low genetic correlation between schizophrenia and plasma lipid levels contrasts with a previous report of pleiotropy between schizophrenia and triglyceride levels[52]. Pleiotropy (unsigned) is different from genetic correlation (signed; Online Methods); however, the pleiotropy reported by Andreassen *et al.*[52] could be explained by the sensitivity of the method used to the properties of a small number of regions with strong LD rather than trait biology (**Supplementary Fig. 5**). We estimated near-zero genetic correlation between Alzheimer's disease and schizophrenia. The genetic correlations between Alzheimer's disease and other psychiatric traits (anorexia nervosa, bipolar disorder, major depression and ASD) were also close to zero but with larger standard errors, due to smaller sample sizes. This suggests that the genetic basis of Alzheimer's disease is distinct from that for psychiatric conditions. Lastly, we estimated near-zero genetic correlation between rheumatoid arthritis and both Crohn's disease and ulcerative colitis. Although these diseases share many associated loci[53,54], there appears to be no directional trend: some rheumatoid arthritis risk alleles are also risk alleles for ulcerative colitis and Crohn's disease, but many rheumatoid arthritis risk alleles are protective for ulcerative colitis and Crohn's disease[53], yielding near-zero genetic correlation. This example highlights the distinction between pleiotropy and genetic correlation (Online Methods).

Finally, the estimates of genetic correlations among metabolic traits are consistent with the estimates obtained using REML in Vattikuti *et al.*[15] (**Supplementary Fig. 6**) and are directionally consistent with the recent Mendelian randomization results from Wurtz *et al.*[55]. The estimate of 0.54 (standard error = 0.07) for the genetic correlation between Crohn's disease and ulcerative colitis is consistent with the estimate of 0.62 (0.04) from Chen *et al.*[16].

DISCUSSION

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

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

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

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

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

Despite these limitations, we believe that the cross-trait LD Score regression estimator of genetic correlation will be a useful addition to the epidemiological toolbox because it allows for rapid screening for correlations among a diverse set of traits, without the need for measuring multiple traits on the same individuals or genome-wide significant SNPs.

URLs. ldsc software, http://www.github.com/bulik/ldsc; PGC (psychiatric) summary statistics, http://www.med.unc.edu/pgc/downloads; GIANT (anthropometric) summary statistics, http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files; EGG (Early Growth Genetics) summary statistics, http://www.egg-consortium.org/; MAGIC (insulin, glucose) summary statistics, http://www.magicinvestigators.org/downloads/; CARDIoGRAM (coronary artery disease) summary statistics, http://www.cardiogramplusc4d.org/; DIAGRAM (type 2 diabetes) summary statistics, http://www.diagram-consortium.org/; rheumatoid arthritis summary statistics, http://www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl_etal_2010NG/; IGAP (Alzheimer's) summary statistics, http://www.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php; IIBDGC (inflammatory bowel disease) summary statistics (we used a newer version of these data with 1000 Genomes Project imputation), http://www.ibdgenetics.org/downloads.html; plasma lipid summary statistics, http://www.broadinstitute.org/mpg/pubs/lipids2010/; SSGAC (educational attainment) summary statistics, http://www.ssgac.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank P. Sullivan, C. Bulik, S. Caldwell, C. Arabica and O. Andreassen for helpful comments. This work was supported by US National Institutes of Health (NIH) grants R01 MH101244 (A.L.P.), R01 HG006399 (N.P.), 1R01 MH101244-02 (B.M.N.), 5U01 MH094432-03 (B.M.N.) and R03 CA173785 (H.K.F.) and by the Fannie and John Hertz Foundation (H.K.F.). Data on anorexia nervosa were obtained by funding from the Wellcome Trust Case Control Consortium 3 project titled "A Genome-Wide Association Study of Anorexia Nervosa" (WT088827/Z/09). Data on glycemic traits were contributed by MAGIC investigators and were downloaded from http://www.magicinvestigators.org/. Data on coronary artery disease and myocardial infarction were contributed by CARDIoGRAMplusC4D investigators and were downloaded from http://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 the analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients and their families. The work with iSelect chips was funded by the French National Foundation on Alzheimer's Disease and Related Disorders. EADI was supported by a LABEX (Laboratory of Excellence Program Investment for the Future) DISTALZ grant, INSERM, Institut Pasteur de Lille, Université de Lille 2 and the Lille University Hospital. GERAD was supported by the Medical Research Council, UK (grant 503480), Alzheimer's Research UK (grant 503176), the Wellcome Trust (grant 082604/2/07/Z) and the German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grants 01GI0102 and 01GI0711, 01GI0420. CHARGE was partly supported by US NIH/National Institute on Aging grants R01 AG033193 and AG081220 and AGES contract N01-AG-12100, National Heart, Lung, and Blood Institute (NHLBI) grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by US NIH/National Institute on Aging grants U01 AG032984, U24 AG021886 and U01 AG016976 and Alzheimer's Association grant ADGC-10-196728.

AUTHOR CONTRIBUTIONS

M.J.D., A.G., P.-R.L., L.D., N.P., B.M.N. and A.L.P. provided reagents. E.B.R., V.A., J.R.B.P. and F.R.D. aided in the interpretation of results. J.R.B.P. and F.R.D. provided data on age at menarche. B.B.-S. and H.K.F. are responsible for the remainder. All authors revised and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Smith, G.D. & Ebrahim, S. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
2. Smith, G.D. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**(R1), R89–R98 (2014).
3. Vandenberg, S.G. in *Methods and Goals in Human Behavior Genetics* 29–43 (Academic Press, 1965).
4. Kempthorne, O. & Osborne, R.H. The interpretation of twin data. *Am. J. Hum. Genet.* **13**, 320–339 (1961).
5. Loehlin, J.C. & Vandenberg, S.G. in *Progress in Human Behavior Genetics* (ed. Vandenberg, S.G.) 261–285 (Johns Hopkins Univ. Press, 1968).
6. Neale, M. & Cardon, L. *Methodology for Genetic Studies of Twins and Families* Number 67 (Springer, 1992).
7. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**, 234–239 (2009).
8. Voight, B.F. *et al.* Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study. *Lancet* **380**, 572–580 (2012).
9. Do, R. *et al.* Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat. Genet.* **45**, 1345–1352 (2013).
10. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

11. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
12. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
13. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. & Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism–derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).
14. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
15. Vattikuti, S., Guo, J. & Chow, C.C. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet.* **8**, e1002637 (2012).
16. Chen, G.-B. *et al.* Estimation and partitioning of (co) heritability of inflammatory bowel disease from GWAS and Immunochip data. *Hum. Mol. Genet.* **23**, 4710–4720 (2014).
17. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
18. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
19. Bulik-Sullivan, B.K. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
20. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
21. Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
22. Cross-Disorder Group of the Psychiatric Genomics Consortium. *et al.* Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371–1379 (2013).
23. Perry, J.R. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).
24. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
25. Horikoshi, M. *et al.* New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nat. Genet.* **45**, 76–82 (2013).
26. Freathy, R.M. *et al.* Type 2 diabetes risk alleles are associated with reduced size at birth. *Diabetes* **58**, 1428–1433 (2009).
27. Early Growth Genetics (EGG) Consortium. A genome-wide association meta-analysis identifies new childhood obesity loci. *Nat. Genet.* **44**, 526–531 (2012).
28. Taal, H.R. *et al.* Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nat. Genet.* **44**, 532–538 (2012).
29. Onland-Moret, N.C. *et al.* Age at menarche in relation to adult height: the EPIC study. *Am. J. Epidemiol.* **162**, 623–632 (2005).
30. Day, F. *et al.* Puberty timing associated with diabetes, cardiovascular disease and also diverse health outcomes in men and women: the UK Biobank study. *Sci. Rep.* **5**, 11208 (2014).
31. Elks, C.E. *et al.* Age at menarche and type 2 diabetes risk: the EPIC-InterAct study. *Diabetes Care* **36**, 3526–3534 (2013).
32. Finucane, H.K. *et al.* Partitioning heritability by functional category using genome-wide association study summary statistics. *Nat. Genet.* doi:10.1038/ng.3404 (28 September 2015).
33. Farooqi, I.S. Defining the neural basis of appetite and obesity: from genes to behaviour. *Clin. Med.* **14**, 286–289 (2014).
34. Wang, N. *et al.* Associations of adult height and its components with mortality: a report from cohort studies of 135,000 Chinese women and men. *Int. J. Epidemiol.* **40**, 1715–1726 (2011).
35. Hebert, P.R. *et al.* Height and incidence of cardiovascular disease in male physicians. *Circulation* **88**, 1437–1443 (1993).
36. Rich-Edwards, J.W. *et al.* Height and the risk of cardiovascular disease in women. *Am. J. Epidemiol.* **142**, 909–917 (1995).
37. Rietveld, C.A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467–1471 (2013).
38. Barnes, D.E. & Yaffe, K. The projected effect of risk factor reduction on Alzheimer's disease prevalence. *Lancet Neurol.* **10**, 819–828 (2011).
39. Norton, S., Matthews, F.E., Barnes, D.E., Yaffe, K. & Brayne, C. Potential for primary prevention of Alzheimer's disease: an analysis of population-based data. *Lancet Neurol.* **13**, 788–794 (2014).
40. MacCabe, J.H. *et al.* Excellent school performance at age 16 and risk of adult bipolar disorder: national cohort study. *Br. J. Psychiatry* **196**, 109–115 (2010).
41. Tiihonen, J. *et al.* Premorbid intellectual functioning in bipolar disorder and schizophrenia: results from a cohort study of male conscripts. *Am. J. Psychiatry* **162**, 1904–1910 (2005).
42. Pierce, J.P., Fiore, M.C., Novotny, T.E., Hatziandreu, E.J. & Davis, R.M. Trends in cigarette smoking in the United States: educational differences are increasing. *J. Am. Med. Assoc.* **261**, 56–60 (1989).
43. Striegel-Moore, R.H., Garvin, V., Dohm, F.-A. & Rosenheck, R.A. Psychiatric comorbidity of eating disorders in men: a national study of hospitalized veterans. *Int. J. Eat. Disord.* **25**, 399–404 (1999).
44. Blinder, B.J., Cumella, E.J. & Sanathara, V.A. Psychiatric comorbidities of female inpatients with eating disorders. *Psychosom. Med.* **68**, 454–462 (2006).
45. Deary, I.J., Strand, S., Smith, P. & Fernandes, C. Intelligence and educational achievement. *Intelligence* **35**, 13–21 (2007).
46. Calvin, C.M., Fernandes, C., Smith, P., Visscher, P.M. & Deary, I.J. Sex, intelligence and educational achievement in a national cohort of over 175,000 11-year-old schoolchildren in England. *Intelligence* **38**, 424–432 (2010).
47. Durkin, M.S. *et al.* Socioeconomic inequality in the prevalence of autism spectrum disorder: evidence from a US cross-sectional study. *PLoS ONE* **5**, e11551 (2010).
48. Robinson, E.B. *et al.* Autism spectrum disorder severity reflects the average contribution of *de novo* and familial influences. *Proc. Natl. Acad. Sci. USA* **111**, 15161–15165 (2014).
49. Samocha, K.E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
50. Silman, A.J. & Pearson, J.E. Epidemiology and genetics of rheumatoid arthritis. *Arthritis Res.* **4** (suppl 3), S265–S272 (2002).
51. de Leon, J. & Diaz, F.J. A meta-analysis of worldwide studies demonstrates an association between schizophrenia and tobacco smoking behaviors. *Schizophr. Res.* **76**, 135–157 (2005).
52. Andreassen, O.A. *et al.* Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am. J. Hum. Genet.* **92**, 197–209 (2013).
53. Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* **7**, e1002254 (2011).
54. Farh, K.K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
55. Wurtz, P. *et al.* Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change. *PLoS Med.* **11**, e1001765 (2014).
56. Burgess, S., Freitag, D.F., Khan, H., Gorman, D.N. & Thompson, S.G. Using multivariable Mendelian randomization to disentangle the causal effects of lipid fractions. *PLoS ONE* **9**, e108891 (2014).
57. Greenland, S., Pearl, J. & Robins, J.M. Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37–48 (1999).
58. Dahl, A., Hore, V., Iotchkova, V. & Marchini, J. Network inference in matrix-variate Gaussian models with non-independent noise. *arXiv* http://arxiv.org/abs/1312.1622 (2013).

ONLINE METHODS

Definitions of genetic covariance and correlation. All definitions refer to narrow-sense heritabilities and genetic covariances. Let S denote a set of M SNPs, let X denote a vector of additively coded genotypes (0, 1 or 2) for the SNPs in S, and let y_1 and y_2 denote phenotypes. Define $\beta := \text{argmax}_{\alpha \in R^M} \text{Cor}[y_1, X\alpha]$, where the maximization is performed in the population (that is, in the infinite data limit). Let γ denote the corresponding vector for y_2. The maximizing value of β is the projection of y onto the span of X, so β is unique modulo SNPs in perfect LD. Define h^2_S, the heritability explained by SNPs in S, as

$$h^2_S(y_1) := \sum_j \beta_j^2$$

and $\varrho_S(y_1, y_2)$, the genetic covariance among SNPs in S, as

$$\varrho_S(y_1, y_2) := \sum_{j \in S} \beta_j \gamma_j$$

The genetic correlation among SNPs in S is $r_S(y_1, y_2) := \varrho_g(y_1, y_2) / \sqrt{h^2_S(y_1) h^2_S(y_2)}$, which lies in $[-1, 1]$. Following ref. 11, we use subscript g (as in h^2_g, ϱ_g, r_g) when the set of SNPs is genotyped and imputed SNPs in GWAS.

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

Genetic correlation is (asymptotically) proportional to Mendelian randomization estimates. If we use a genetic instrument

$$g_i := \sum_{j \in S} X_{ij} \beta_j$$

to estimate the effect b_{12} of y_1 on y_2, the 2SLS estimate is $\hat{b}_{2\text{SLS}} := g^T y_2 / g^T y_1$ (ref. 59). The expectations of the numerator and denominator are $E[g^T y_2] = \varrho_S(y_1, y_2)$ and $E[g^T y_1] = h^2_S(y_1)$. Thus, $\text{plim}_{N \to \infty} \hat{b}_{2\text{SLS}} = r_S(y_2, y_1) \sqrt{h^2_S(y_1) / h^2_S(y_2)}$. If we use the same set S of SNPs to estimate b_{12} and b_{21} (for example, if S is the set of all common SNPs, as in the genetic correlation analyses in this report), then this procedure is symmetric in y_1 and y_2.

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

Cross-trait LD score regression. The cross-trait LD Score regression equation is

$$E[z_{1j} z_2 \mid \ell_j] = \frac{\sqrt{N_1 N_2}\, \varrho_g}{M} \ell_j + \frac{\varrho N_S}{\sqrt{N_1 N_2}} \qquad (2)$$

where z_{ij} denotes the z score for study i and SNPj, N_i is the sample size for study i, e_g is the genetic covariance, ℓ_j is the LD Score[19], N_S is the number of individuals included in both studies and ϱ is the phenotypic correlation among the N_S overlapping samples. We derive this equation in the **Supplementary Note**. We estimate genetic covariance by regressing $z_{1j} z_{2j}$ against $\ell_j \sqrt{N_{1j} N_{2j}}$, (where N_{ij} is the sample size for SNP j in study i) and then multiplying the resulting slope by M, the number of SNPs in the reference panel with MAF between 5% and 50% (technically, this is an estimate of the genetic covariance among SNPs with 5–50% MAF; **Supplementary Note**).

If we know the correct value of the intercept term $\varrho N_S / \sqrt{N_1 N_2}$ ahead of time, we can reduce the standard error by constraining the intercept to this value, using the –constrain-intercept flag in ldsc (for pairs of binary traits, we give corresponding expression in terms of the number of overlapping cases and controls in the **Supplementary Note**). Note that this works even when there is known nonzero sample overlap.

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

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

$$\text{var}[z_{1j} z_{2j} \mid \ell_j] = \left(\frac{N_1 h_1^2}{M} \ell_j + 1 \right) \left(\frac{N_2 h_2^2}{M} \ell_j + 1 \right) + \left(\frac{\sqrt{N_1 N_2}\, \varrho_g}{M} \ell_j + \frac{\varrho N_s}{\sqrt{N_1 N_2}} \right)^2 \qquad (3)$$

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

1. The first regression is weighted using heritabilities from the single-trait LD Score regressions, $\varrho N_S = 0$, and ϱ_g is estimated as:
$$\hat{\varrho}_g := (\sqrt{N_1 N_2})^{-1} \sum_j z_{1j} z_{2j}$$
2. The second regression is weighted using the estimates of ϱN_S and ϱ_g from step 1. The genetic covariance estimate that we report is the estimate from the second regression.

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

Two-step estimator. As noted in ref. 19, SNPs with very large effect sizes can result in large LD Score regression standard errors for single-trait LD Score regression with an unconstrained intercept; cross-trait LD Score regression with an unconstrained intercept behaves similarly. This is because of the well-known fact that linear regression deals poorly with outliers in the response variable (LD Score regression with constrained intercept is not nearly as adversely affected by large-effect SNPs). The solution proposed in ref. 19 was to remove SNPs with $\chi^2 > 80$ from the LD Score regression. This is a satisfactory solution when the goal is to estimate the LD Score regression intercept. If the goal is to distinguish polygenicity from population stratification and we are willing to assume that the population stratification is subtle, such that SNPs with $\chi^2 > 80$ are much more likely to be real causal SNPs than artifacts, then we can make the task much easier by removing these SNPs. However, this is unsatisfactory if the goal is to estimate h^2: ignoring large-effect SNPs with $\chi^2 > 80$ would bias estimates of h^2 and ϱ_g toward zero. Therefore, for estimating h^2 or ϱ_g, we take a two-step approach. The first step is to estimate the LD Score regression intercept with all SNPs with $\chi^2 > 30$ removed (all genome-wide significant SNPs; the threshold can be adjusted with the –two-step flag in ldsc). The second step is to estimate h^2 or ϱ_g using all SNPs and constrained-intercept LD Score regression with the intercept constrained to the value from the first step (note that we account for uncertainty in the intercept when computing standard error).

Assessment of statistical significance via block jackknife. Summary statistics for SNPs in LD are correlated, so the ordinary least squares (OLS) standard error will be biased downward. We estimate a heteroskedasticity- and

correlation-robust standard error with a block jackknife over blocks of adjacent SNPs. This is the same procedure used in ref. 19 and gives accurate standard errors in simulations (**Table 1**). We obtain the standard error for a genetic correlation by using a ratio block jackknife over SNPs. The default setting in ldsc is 200 blocks per genome, which can be adjusted with the –num-blocks flag.

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

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

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

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

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

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

1. Computing summary statistics takes $O(MN)$ time.
2. Computing LD Scores takes $O(MN)$ time, although the N for computing LD Scores need not be large. We use $N = 378$ Europeans from the 1000 Genomes Project.
3. LD Score regression takes $O(M)$ time and space.

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

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

Simulations. We simulated quantitative traits under an infinitesimal model in 2,062 controls from a Swedish study. To simulate the standard scenario where many causal SNPs are not genotyped, we simulated phenotypes by drawing causal SNPs from 622,146 best-guess imputed 1000 Genomes Project SNPs on chromosome 2 and then retained only the 90,980 HapMap 3 SNPs with MAF above 5% for LD Score regression.

We note that the simulations in ref. 19 showed that single-trait LD Score regression is only minimally biased by uncorrected population stratification and moderate ancestry mismatch between the reference panel used to estimate LD Scores and the population sampled in the GWAS. In particular, LD Scores estimated from the 1000 Genomes Project reference panel are suitable for use with European-ancestry meta-analyses. Put another way, LD Score is only minimally correlated with the per-SNP Wright's fixation index (F_{ST}) and the differences in LD Score among European populations are not so large as to bias LD Score regression. Because we use the same LD Scores for cross-trait LD Score regression as for single-trait LD Score regression, these results extend to cross-trait LD Score regression.

Summary statistic data sets. We selected traits for inclusion in the analysis for the main text via the following procedure:

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

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

1. For studies that provide a measure of imputation quality, filter to INFO above 0.9.
2. For studies that provide sample MAF, filter to sample MAF above 1%.
3. To restrict to well-imputed SNPs in studies that do not provide a measure of imputation quality, filter to SNPs in the HapMap 3 panel[61] with a 1000

doi:10.1038/ng.3406

Genomes Project EUR (European) MAF above 5%, which tend to be well imputed in most studies. This step should be skipped if INFO scores are available for all studies.
4. If the sample size varies from SNP to SNP, remove SNPs with an effective sample size less than 0.67 times the 90th percentile of sample size.
5. For meta-analyses with specialty chips (for example, the Metabochip), remove SNPs with a sample size above the maximum GWAS sample size.
6. Remove indels and structural variants.
7. Remove strand-ambiguous SNPs.
8. Remove SNPs whose alleles do not match the alleles in the 1000 Genomes Project.

Genomic control (GC) correction at any stage biases heritability and genetic covariance estimates downward (see the Supplementary Note of ref. 19). Biases in the numerator and denominator of genetic correlation cancel exactly so that genetic correlation is not biased by GC correction. A majority of the studies analyzed in this report used GC correction, so we do not report genetic covariance and heritability.

Data on Alzheimer's disease were obtained from the International Genomics of Alzheimer's Project (IGAP). IGAP is a large two-stage study based on GWAS of individuals of European ancestry. In stage 1, IGAP used genotyped and imputed data for 7,055,881 SNPs to perform meta-analysis of 4 previously published GWAS data sets including 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 Alzheimer's Disease 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 the results from stages 1 and 2. We only used stage 1 data for LD Score regression.

59. Angrist, J.D. & Pischke, J-S. *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton Univ. Press, 2008).
60. Aschard, H., Vilhjálmsson, B.J., Joshi, A.D., Price, A.L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am. J. Hum. Genet.* **96**, 329–339 (2015).
61. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).