Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits

Nicholas Mancuso,[1,*] Huwenbo Shi,[2] Pagé Goddard,[3] Gleb Kichaev,[2] Alexander Gusev,[4,5,6,8] and Bogdan Pasaniuc[1,2,7,8,*]

Although genome-wide association studies (GWASs) have identified thousands of risk loci for many complex traits and diseases, the causal variants and genes at these loci remain largely unknown. Here, we introduce a method for estimating the local genetic correlation between gene expression and a complex trait and utilize it to estimate the genetic correlation due to predicted expression between pairs of traits. We integrated gene expression measurements from 45 expression panels with summary GWAS data to perform 30 multi-tissue transcriptome-wide association studies (TWASs). We identified 1,196 genes whose expression is associated with these traits; of these, 168 reside more than 0.5 Mb away from any previously reported GWAS significant variant. We then used our approach to find 43 pairs of traits with significant genetic correlation at the level of predicted expression; of these, eight were not found through genetic correlation at the SNP level. Finally, we used bi-directional regression to find evidence that BMI causally influences triglyceride levels and that triglyceride levels causally influence low-density lipoprotein. Together, our results provide insight into the role of gene expression in the susceptibility of complex traits and diseases.

Introduction

Although genome-wide association studies (GWASs) have identified tens of thousands of common genetic variants associated with many complex traits,[1] with some notable exceptions,[2,3] the causal variants and genes at these loci remain unknown. Multiple lines of evidence have shown that GWAS risk variants co-localize with genetic variants that regulate expression—i.e., expression quantitative trait loci (eQTLs).[4] This suggests that a substantial proportion of GWAS risk variants influence complex traits by regulating expression levels of their target genes.[4–7] Analyses of genotype, phenotype, and gene expression measurements from multiple tissues in the same set of individuals can directly investigate this plausible chain of causality. However, doing so is challenging because of cost and tissue availability; therefore, GWAS and eQTL datasets remain largely independent (i.e., no overlapping subjects).[8,9] Recent work has shown that one way to integrate GWAS and eQTL data is to predict gene expression levels for GWAS samples and then test for association between the predicted expression and traits.[10–12] This approach, referred to as transcriptome-wide association study (TWAS), can increase power over GWAS when the causal mechanism includes genetic variants that regulate the expression of susceptibility genes. TWAS benefits from a lower multiple-testing burden by probing several thousands of genes, whereas GWAS probes several million SNPs. Although TWAS can also be performed with measured gene expression levels directly, using predicted gene expression has several benefits. First, expression measurements are usually not available in GWAS data. Second, predicted gene expression removes environmental noise by focusing on the genetically regulated component, which can increase statistical power. Third, using the predicted expression to test for association can eliminate potential confounding from reverse causation, where traits affect gene expression levels.[10,11] However, compared with GWAS, TWAS is underpowered when risk is not mediated through expression or when expression data are not available in the right tissue.

In this work, we introduce methods for estimating the genetic correlation between gene expression and a complex trait from summary GWAS and eQTL data. We utilize the local (*cis*) genetic variation near a gene (i.e., ± 0.5 Mb around the transcription start site [TSS]) to estimate the correlation in the genetic effects between gene expression and the trait. We show that under this framework, TWAS can be viewed as a test for non-zero genetic covariance between expression and a trait from summary association data. In addition to identifying susceptibility genes, the predicted expression can also be used for estimating the genome-wide genetic correlation between pairs of complex traits at the level of predicted expression. This is analogous to computing genome-wide genetic correlation between complex traits,[13] whereby correlations are determined over predicted gene expression effects rather than SNP effects, and

[1]Department of Pathology & Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90024, USA; [2]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90024, USA; [3]Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, Los Angeles, CA 90024, USA; [4]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [6]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [7]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90024, USA
[8]These authors contributed equally to this work
*Correspondence: nmancuso@mednet.ucla.edu (N.M.), bpasaniuc@mednet.ucla.edu (B.P.)
http://dx.doi.org/10.1016/j.ajhg.2017.01.031.

can give insights into the component of genetic correlation mediated through expression. We demonstrate through extensive simulations that our approach is approximately unbiased and well calibrated under the null and slightly conservative when true correlation is near the boundaries. Finally, we utilize estimated effects of predicted expression within a bi-directional regression approach[14] to investigate putative causal direction for pairs of complex traits that are genetically correlated.

We analyze summary statistics from 30 GWASs spanning 2.3 million phenotype measurements[15–28] jointly with 45 expression panels[8,29–34] sampled from more than 35 tissues to gain insight into the role of expression in the etiology of complex traits. First, we test each gene-tissue pair across 45 panels to perform a multi-tissue TWAS for each of the 30 traits to identify 1,196 gene associations. For example, at four independent loci, we find 11 genes that do not overlap a genome-wide significant SNP for educational years. Notably, all four loci were replicated in a recent, larger GWAS for educational years.[35] Second, we identify 43 pairs of traits showing a genome-wide-significant genetic correlation at the level of predicted expression. Overall, the predicted-expression correlation was highly concordant with SNP-level genetic correlation from cross-trait linkage disequilibrium (LD) score regression, which suggests that a large component of genetic correlation between complex traits is driven by local regulation of gene expression. Finally, we use our bi-directional analysis to provide evidence of putative causal effects between pairs of these traits. Overall, our results shed light on shared biological mechanisms responsible for susceptibility to disease and complex traits, as well as potential downstream effects between traits.

Material and Methods

Datasets

We used summary association statistics from 30 large-scale (n = 20,000 subjects) GWASs, including various anthropometric[15,27,28] (body mass index [BMI], femoral neck bone mineral density [BMD], forearm BMD, lumbar spine BMD, and height), hematopoietic[23,25,26] (hemoglobin, HbA_{1c}, mean cell hemoglobin [MCH], MCH concentration, mean cell volume, number of platelets, packed cell volume, and red blood cell count), immune-related[17,19] (Crohn disease [OMIM: 266600], inflammatory bowel disease [OMIM: 266600], ulcerative colitis [OMIM: 266600], and rheumatoid arthritis [OMIM: 180300]), metabolic[16,20,22,24] (age of menarche, fasting glucose, fasting insulin, high-density lipoprotein [HDL], HOMA-B, HOMA-IR, low-density lipoprotein [LDL], triglycerides [TG], type 2 diabetes [OMIM: 125853], and total cholesterol [TC] levels), neurological[18] (schizophrenia [OMIM: 181500]), and social[21] (college and educational attainment) phenotypes (see Table S1). We removed SNPs that were strand ambiguous or had a minor allele frequency (MAF) $\leq$ 1% (see Table S1).

Gene expression data from RNA sequencing data were obtained from the CommonMind Consortium[29] (brain, n = 613), the Genotype-Tissue Expression Project[8] (GTEx; 41 tissues; see Table S2 for sample size per tissue), and the Metabolic Syndrome in Men study[31,32] (adipose, n = 563). Expression microarray data were obtained from the Netherlands Twins Registry[34] (NTR; blood, n = 1,247), and the Young Finns Study[30,33] (YFS; blood, n = 1,264).

Performing TWAS with GWAS Summary Statistics

We estimated SNP heritability for observed expression levels partitioned into *cis*-h^2_g (1 Mb region surrounding the TSS) and *trans*-h^2_g (rest of genome) components. We used the AI-REML algorithm implemented in Genome-wide Complex Trait Analysis (GCTA),[36] which allows estimates to fall outside of the (0, 1) boundaries to maintain unbiasedness. To control for confounding, we included batch variables and the top 20 principal components estimated from genome-wide SNPs. Genes with significant *cis*-heritability in expression data were used for prediction (*cis*-h^2_g $p < 0.05$ in a likelihood ratio test between the *cis*-only and joint models). The average number of genes with significant *cis*-h^2_g across expression studies was 816 (min = 70 genes from GTEx small intestine samples; max = 3,704 genes from the YFS).

We performed 45 TWASs for each of the 30 GWASs;[11] for each trait, we used Bonferroni correction for all gene-tissue pairs tested (see Table S2). In brief, we estimated the strength of association between the predicted expression of a gene and a complex trait (z_{TWAS}) as a function of the vector of GWAS summary Z scores at a given *cis*-locus, $\boldsymbol{z}_T'$ (i.e., vector of SNP association Wald statistics), and the LD-adjusted weight vector learned from the gene expression data, $\boldsymbol{w}_{GE}$, as

$$z_{TWAS} = \frac{\boldsymbol{w}_{GE}'\boldsymbol{z}_T}{\sqrt{\mathrm{var}(\boldsymbol{w}_{GE}'\boldsymbol{z}_T)}} = \frac{\boldsymbol{w}_{GE}'\boldsymbol{z}_T}{\sqrt{\boldsymbol{w}_{GE}'\boldsymbol{V}\boldsymbol{w}_{GE}}},$$

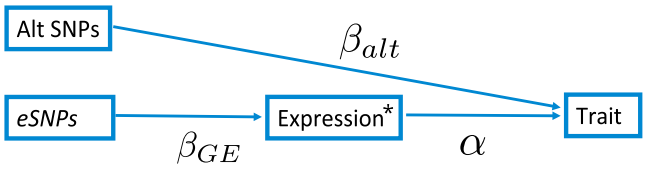
where $\boldsymbol{V}$ is a covariance matrix across SNPs at the locus (i.e., LD). We estimated $\boldsymbol{w}_{GE}$ by using GBLUP[37] from eQTL data and computed z_{TWAS} by using GWAS summary data for all 30 traits and the ~36,000 gene expression measurements across all studies. We removed all loci in the human leukocyte antigen (HLA) region as a result of complex LD patterns.

Estimating the Proportion of Trait Variance Explained by Predicted Expression

We use the LD score regression[38,39] approach described in Guseve et al.[11] to quantify the heritability explained by predicted expression for a complex trait (denoted here as h^2_{GE}). The expected χ^2 statistic under a polygenic trait is $E[\chi^2] = 1 + (N_T\ell/M)h^2_{GE} + N_T a$, where N_T is the number of individuals in the GWAS, M is the number of genes, ℓ is the LD score, and a is the effect of population structure. We estimate ℓ for each gene by predicting expression for 503 European samples in 1000 Genomes[40] by using the GBLUP weights (see above) and then computing sample correlation. For each trait, we perform LD score regression by using z^2_{TWAS} (which follows a χ^2 distribution asymptotically) to infer h^2_{GE}. We estimate heritability for each expression study separately to account for varying sample sizes and repeated gene measurements.

Estimating Genetic Correlation of Expression and Complex Traits from Summary Data

Let expression and traits be modeled as a linear function of the genotypes in a ~1 Mb locus flanking the gene: $\boldsymbol{\gamma}_{GE} = \boldsymbol{X}\boldsymbol{\beta}_{GE} + \boldsymbol{\epsilon}_{GE}$ and $\boldsymbol{\gamma}_T = \boldsymbol{X}\boldsymbol{\beta}_T + \boldsymbol{\epsilon}_T$, where $\boldsymbol{X}$ is the standardized genotype matrix, $\boldsymbol{\beta}_{GE}$ and $\boldsymbol{\beta}_T$ are the standardized effects for expression and traits,



$$\rho_g = cor([\beta_{GE} \times \alpha; \beta_{alt}]_{T_1}, [\beta_{GE} \times \alpha; \beta_{alt}]_{T_2})$$

$$\rho_{GE} = cor(\alpha_{T_1}, \alpha_{T_2})$$

Figure 1. Causal Diagram Illustrating the Genetic Component of a Trait
The total effect of SNPs on a trait can be partitioned into components that are mediated through *cis*-regulated (i.e., predicted, indicated by an asterisk) gene expression ($\beta_{GE} \times \alpha$) or through alternative pathways (β_{alt}). In contrast to ρ_g, which quantifies the correlation of the total SNP effects between two traits ($\beta_{GE} \times \alpha$; β_{alt}), ρ_{GE} focuses exclusively on the effects of *cis*-regulated gene expression (α).

respectively, and ϵ_{GE} and ϵ_T are the environmental noise for expression and traits, respectively. The local covariance between expression and complex traits is

$$\begin{aligned}\mathrm{cov}(\boldsymbol{y}_{GE}, \boldsymbol{y}_T) &= \mathrm{cov}(\boldsymbol{X}\boldsymbol{\beta}_{GE} + \boldsymbol{\epsilon}_{GE}, \boldsymbol{X}\boldsymbol{\beta}_T + \boldsymbol{\epsilon}_T) \\ &= \boldsymbol{\beta}_{GE}^{'}\mathrm{cov}(\boldsymbol{X}, \boldsymbol{X})\boldsymbol{\beta}_T + \mathrm{cov}(\boldsymbol{\epsilon}_{GE}, \boldsymbol{\epsilon}_T) \\ &= \boldsymbol{\beta}_{GE}^{'}\boldsymbol{V}\boldsymbol{\beta}_T + \mathrm{cov}(\boldsymbol{\epsilon}_{GE}, \boldsymbol{\epsilon}_T),\end{aligned}$$

where $\boldsymbol{V}$ is the LD matrix. If no individuals are shared between studies, then $\mathrm{cov}(\boldsymbol{\epsilon}_{GE}, \boldsymbol{\epsilon}_T) = 0$ (as in eQTL studies and GWASs). The local genetic correlation between expression and traits can be computed as

$$\rho_{g,local} = \frac{\boldsymbol{\beta}_{GE}^{'}\boldsymbol{V}\boldsymbol{\beta}_T}{\sqrt{h^2_{g,local}(GE)}\sqrt{h^2_{g,local}(T)}},$$

where $h^2_{g,local}(GE)$ and $h^2_{g,local}(T)$ are the local SNP heritability[41] for expression and traits, respectively, estimated at the locus. However, this requires knowledge of the true effect sizes. Given association statistics $\boldsymbol{z}_T$, we estimate an LD-adjusted effect size as $\widehat{\boldsymbol{\beta}}_T = \frac{1}{\sqrt{N_T}}\boldsymbol{V}^{-1}\boldsymbol{z}_T$. Hence, an estimate of the local genetic covariance[42] is given by

$$\widehat{\boldsymbol{\beta}}_{GE}^{'}\boldsymbol{V}\widehat{\boldsymbol{\beta}}_T = \frac{1}{\sqrt{N_{GE}}\sqrt{N_T}}(\boldsymbol{z}_{GE}^{'}\boldsymbol{V}^{-1})\boldsymbol{V}(\boldsymbol{V}^{-1}\boldsymbol{z}_T) = \widehat{\boldsymbol{b}}_{GE}^{'}\boldsymbol{V}^{-1}\widehat{\boldsymbol{b}}_T,$$

where $\widehat{\boldsymbol{b}}_{GE}$ and $\widehat{\boldsymbol{b}}_T$ are the marginal (i.e., LD-unadjusted) standardized effect-size estimates.[41,43] It follows that

$$\begin{aligned}\frac{1}{\sqrt{N_T}}z_{TWAS} &= \frac{1}{\sqrt{N_T}}\frac{\widehat{\boldsymbol{\beta}}_{GE}^{'}\boldsymbol{z}_T}{\sqrt{\mathrm{var}\left(\widehat{\boldsymbol{\beta}}_{GE}^{'}\boldsymbol{z}_T\right)}} = \frac{\widehat{\boldsymbol{b}}_{GE}^{'}\boldsymbol{V}^{-1}\widehat{\boldsymbol{b}}_T}{\sqrt{h^2_{g,local}(GE)}} \\ &= \rho_{g,local}\sqrt{h^2_{g,local}(T)}.\end{aligned}$$

We standardize this estimate to obtain our final local genetic correlation estimate as

$$\widehat{\rho}_{g,local} = \frac{z_{TWAS}}{\sqrt{N_T \times h^2_{g,local}(T)}}.$$

In practice, we use the variance explained by the local index SNP (i.e., smallest p value) as a proxy for $h^2_{g,local}(T)$.

Genetic Correlation between Traits at the Level of Predicted Expression

Consider a simple model where the genetic component of a trait can be decomposed into genetic effects that are mediated through *cis*-gene expressions of k genes plus genetic effects not mediated through expression at other loci in the genome:

$$\boldsymbol{y}_T = \sum_{i=1}^{k}(\boldsymbol{X}_i\boldsymbol{\beta}_{\mathbf{GE}_i})\alpha_i + \boldsymbol{X}_{\mathbf{alt}}\boldsymbol{\beta}_{\mathbf{alt}} + \boldsymbol{\epsilon}_T,$$

where $\boldsymbol{X}_i$ is a vector of genotypes at the *cis*-locus of gene i, $\boldsymbol{\beta}_{GE_i}$ is the casual eQTL effect vector for gene i, α_i is the direct effect of gene expression on a trait, and $\boldsymbol{X}_{\mathbf{alt}}$ and $\boldsymbol{\beta}_{\mathbf{alt}}$ refer to the genotype and causal effects, respectively, of variants not mediated through expression. We define the genome-wide genetic correlation at the level of expression between two complex traits as the correlation across the gene effects: $\rho_{GE} = \mathrm{cor}(\alpha_{T_1}, \alpha_{T_2})$. In practice, we do not know α, but we can estimate it as

$$\widehat{\alpha} = \frac{\mathrm{cov}(\boldsymbol{X}\boldsymbol{\beta}_{GE}, \boldsymbol{y}_T)}{\mathrm{var}(\boldsymbol{X}\boldsymbol{\beta}_{GE})} = \frac{\boldsymbol{\beta}_{GE}^{'}\boldsymbol{V}\boldsymbol{\beta}_T}{h^2_{g,local}(GE)} = \widehat{\rho}_{g,local}\Big/\frac{\sqrt{h^2_{g,local}(GE)}}{\sqrt{h^2_{g,local}(\boldsymbol{y}_T)}}$$

to obtain an estimate of expression correlation by using predicted expression ($\widehat{\rho}_{GE}$). In practice, we use the standardized estimates of $\widehat{\alpha}$, which are proportional to $\widehat{\rho}_{g,local}$. Unlike SNP-based genetic correlation (ρ_g), which captures genetic correlation across all common variants in the genome, ρ_{GE} captures only the component of genetic correlation driven by *cis* genetic effects on expression (see Figure 1). For instance, a pair of traits with highly correlated effects in *cis*-regions but weakly correlated effects in *trans*-regions will result in $\rho_{GE} > \rho_g$. In the absence of large *trans*-eQTL effects, we expect $\rho_{GE} \approx \rho_g$. Furthermore, because ρ_{GE} accounts for only the shared effect from predicted expression, any genetic effect on a trait not driven through expression in the measured eQTL data will not be represented in ρ_{GE}. We test for significance by assuming $\widehat{\rho}_{GE}\sqrt{(M-2)/(1-\widehat{\rho}^2_{GE})} \sim t(M-2)$, where M is the number of genes and t is the t distribution with $M-2$ degrees of freedom. This procedure requires the effects of M genes on the trait to be independent, which could be violated in practice; hence, we compute $\widehat{\rho}_{GE}$ by using one gene per 1 Mb locus.

Estimating Putative Casual Relationships between Pairs of Traits

To glean insight into the underlying causal relationship between pairs of traits, we perform a bi-directional regression[14] and estimate two different values of ρ_{GE} by varying gene sets. Before describing the approach, we first review several causal models that explain non-zero ρ_{GE} between two traits (see Figure 2). Models A and B depict causal relationships in which the effects of a gene set are mediated by one trait on the other. We can formally state model A (without loss of generality for B). Let trait 1 (T_1) be defined as $\boldsymbol{y}_{T_1} = \boldsymbol{G}_{T_1}\boldsymbol{\beta}_{T_1} + \boldsymbol{\epsilon}_{T_1}$, where $\boldsymbol{G}_{T_1}$ denotes the matrix of predicted expression at the causal genes, $\boldsymbol{\beta}_{T_1}$ is the effect size, and $\boldsymbol{\epsilon}_{T_1}$ is environmental noise. We define trait 2 (T_2) as

$$\boldsymbol{y}_{T_2} = \boldsymbol{y}_{T_1}\gamma_{T_1} + \boldsymbol{G}_{T_2}\boldsymbol{\beta}_{T_2} + \boldsymbol{\epsilon}_{T_2} = \boldsymbol{G}_{T_1}\boldsymbol{\beta}_{T_1}\gamma_{T_1} + \boldsymbol{G}_{T_2}\boldsymbol{\beta}_{T_2} + \boldsymbol{\epsilon}_{T_2'},$$

where γ_{T_1} is the causal effect of T_1 on T_2, $\boldsymbol{G}_{T_2}$ and $\boldsymbol{\beta}_{T_2}$ are the remaining causal genes and their effects, respectively, for T_2, and $\boldsymbol{\epsilon}_{T_2'}$ is the combined environment component. Under model A, the causal gene set for T_1 will have a non-zero effect on T_2 (i.e.,

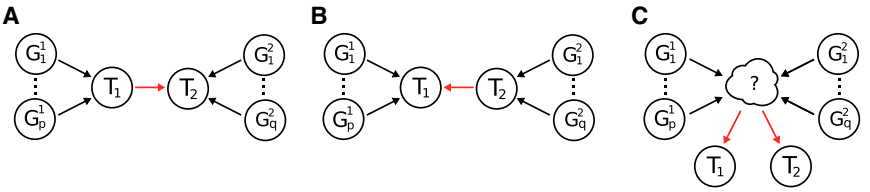


Figure 2. Illustration of Several Causal Models That Explain Expression Correlation for Traits 1 and 2 Given Their Causal Gene Sets

(Model A) Trait 1 directly influences trait 2. In this case, the effect of genes $G_1^1, \ldots, G_p^1$ on trait 2 is mediated by trait 1, which implies $\{G_i^1\}_{i=1}^{p} \subsetneq \{G_i^2\}_{i=1}^{q}$.

(Model B) Trait 2 directly influences trait 1. Similarly, the effect of genes $G_1^2, \ldots, G_q^2$ on trait 1 is mediated by trait 2, which implies $\{G_i^2\}_{i=1}^{q} \subsetneq \{G_i^1\}_{i=1}^{p}$.

(Model C) Traits 1 and 2 are influenced independently through an unobserved trait or traits.

$\gamma_{T_1} \neq 0$); however, if T_1 does not cause T_2, this effect will be zero given that unrelated genes have no downstream effect. Bi-directional regression provides a test to distinguish between models A and B by regressing estimated effect sizes for gene sets under model A (i.e., $\beta_{T_1} \sim \beta_{T_1}\gamma_{T_1}$) and comparing to estimates under model B (i.e., $\beta_{T_2} \sim \beta_{T_2}\gamma_{T_2}$). Because the causal gene sets for each trait are unknown, we use their identified susceptibility genes as a proxy. We estimate ρ_{GE} by conditioning on the gene set for trait i and denote its value as $\rho_{j|i}$. We repeat this procedure by ascertaining the gene set for trait j to obtain $\rho_{i|j}$. We perform a Welch's t test[44] to determine whether estimates of $\rho_{i|j}$ and $\rho_{j|i}$ are significantly different, thus providing evidence consistent with a causal direction. To minimize spurious results, we require at least ten genes for estimation in each conditional test. This approach mirrors bi-directional regression analyses of estimated SNP effects on two complex traits.[45,46] We stress that although a bi-directional approach is capable of rejecting model A in favor of model B (or vice versa), it cannot rule out model C, in which a shared pathway (or set of pathways) drives both traits independently (see Figure 2).

Simulation Framework

We simulate gene expression levels by using real genotype data measured in 503 European individuals from the 1000 Genomes Project.[40] Given a gene locus, we generate expression levels under the linear model $\boldsymbol{E} = \boldsymbol{X}\boldsymbol{w} + \epsilon$, where $\boldsymbol{E}$ is a gene expression vector of length N, X is the $N \times 2$ mean-centered and variance-standardized genotype matrix over two randomly selected SNPs in the locus, $\boldsymbol{w}$ is the causal effect, and ϵ is the environmental noise. We sample effect sizes $\boldsymbol{w}_i \sim \mathrm{N}(0, [h_g^2/2])$ for $i = 1$ and 2 and noise from a normal distribution to yield $h_g^2 = 0.1$ (consistent with what we observe in real gene expression data). We consider only SNPs with a MAF ≥ 0.01 and Hardy-Weinberg equilibrium deviation $p \geq 1 \times 10^{-5}$. We simulate a complex trait as a linear function of predicted gene expression for $k = 100$ genes, given by $\boldsymbol{y} = \sum_{i=1}^{k}(\boldsymbol{X}_i\boldsymbol{w}_i)\alpha_i + \epsilon$, where $\boldsymbol{X}_i\boldsymbol{w}_i$ is the predicted expression of the i^{th} gene with effect sizes $\alpha_i \sim \mathrm{N}(0, h_{GE}^2/k)$. For simulations involving ρ_{GE}, we simulate the two traits $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ by using the same process, except effects for the i^{th} gene are drawn from a bivariate normal distribution:

$$\begin{bmatrix} \alpha_{i,1} \\ \alpha_{i,2} \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\alpha,1}^2 & \rho_{GE}\sigma_{\alpha,1}\sigma_{\alpha,2} \\ \rho_{GE}\sigma_{\alpha,1}\sigma_{\alpha,2} & \sigma_{\alpha,2}^2 \end{bmatrix} \right),$$

where $\sigma_{\alpha,*}^2 = (h_{GE,*}^2)/k$. Lastly, we perform an association scan on $\boldsymbol{y}$ by using all SNPs at each gene locus to obtain SNP-level Z scores $\boldsymbol{z}_T$.

Results

Accurate Estimation of Expression-Trait Genetic Correlation in Simulations

To validate our statistical framework for estimating $\rho_{g,local}$, we used real genotype data to perform simulations under various architectures (see Material and Methods). In brief, we simulated gene expression for 100 independent gene loci, which we then used to simulate a complex trait. Using our approach, we performed a GWAS and estimated $\rho_{g,local}$ from TWAS summary statistics (see Material and Methods). We observed unbiased estimates for $\rho_{g,local}$ both when causal variants were typed and when they were masked from the data (see Figure S1). Estimated values of $\rho_{g,local}$ were highly correlated with their true values ($r = 0.73$; $p < 2.2 \times 10^{-16}$), which indicates that using weights inferred from GBLUP maintains moderate power levels. This slight loss in power extended to h_{GE}^2 estimates, which quantify the total effect of predicted expression on a trait ($r = 0.74$; $p < 6.7 \times 10^{-12}$; see Table S3). As eQTL datasets increase in sample size, and predictive models become more accurate, we expect this attenuation bias to decrease.

We next performed extensive simulations to validate our procedure for estimating genetic correlation due to predicted expression (ρ_{GE}) between pairs of traits. We simulated genetically correlated complex traits from predicted expression by sampling effects from a bivariate normal distribution with correlation ρ_{GE} (see Material and Methods). We first estimated $\rho_{g,local}$ for each gene-trait pair, which served as input for estimating ρ_{GE}. Overall, we observed our estimator to be approximately unbiased, with conservative estimates for ρ_{GE} when its underlying value was near the boundaries (see Figure 3). Importantly, estimates were relatively unbiased when causal variants were untyped in the data. Our method appropriately accounted for LD among variants, resulting in a large improvement over the naive SNP correlation approach (which simply correlates the Z scores by ignoring LD). We also assessed our approach for testing for deviations from $\rho_{GE} = 0$ and found estimates consistent with the null distribution with $\lambda_{GC} = 0.97$ (Jack-knife 95% CI = [0.86, 1.08]; see Figure S2). To measure how sensitive our approach is to estimates of $h_{g,local}^2(GE)$ at each gene, we repeated simulations by using variance explained by the top eQTL as a proxy for local heritability. Although estimates were highly similar ($r = 0.99$; $p < 6.6 \times 10^{-7}$), our approach produced estimates closer to the ground truth (see Figure S3).

TWAS Identifies 1,196 Genes Associated with 30 Complex Traits and Diseases

We integrated GWAS summary data of 30 complex traits with gene expression to identify 1,196 susceptibility genes (i.e., genes with at least one significant trait association),

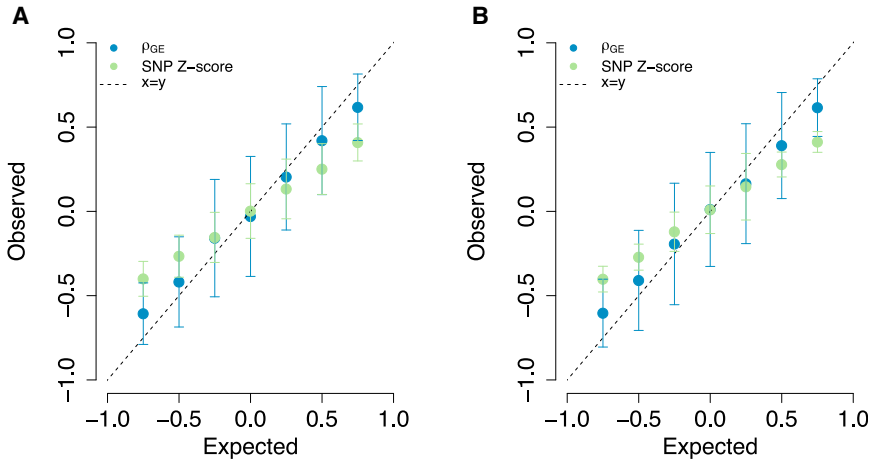


Figure 3. Simulation Results for $\widehat{\rho}_{GE}$ and Correlation of SNP *Z* Scores
Each point represents the mean estimate over 100 simulations. Error bars represent the 95% confidence interval estimated by the mean SE across simulations. The dotted line represents the identity line.
(A) Causal SNPs for gene expression are typed in the data.
(B) Causal SNPs are untyped.

comprising 5,490 total associations (after Bonferroni correction; see Material and Methods). Of these associations, we observed 1,789 distinct gene-trait pairs, of which 783 were found in anthropometric traits, 423 in metabolic traits, 215 in immune-related traits, 213 in hematopoietic traits, 137 in neurological traits (e.g., schizophrenia), and 18 in social traits (see Tables 1, S4, and S5). For example, the 137 susceptibility genes found for schizophrenia included *SNX19* (e.g., GTEx cerebellum; $p < 2.2 \times 10^{-8}$) and *NMRAL1* (e.g., GTEx skeletal muscle; $p < 9.7 \times 10^{-7}$); this is consistent with a previously reported study[12] that used different methods and expression data (see Table S6). We did not find susceptibility genes for forearm BMD, HOMA-B, or MCH concentration, consistent with low GWAS signal for these traits (see Table 1). Indeed, the number of GWAS risk loci strongly correlated with the number of identified susceptibility genes ($r = 0.99$; $p < 2.2 \times 10^{-16}$). Using the PANTHER database,[47] we explored putative molecular function and pathways enriched with identified susceptibility genes but were underpowered to detect molecular function for most individual traits (see Appendix A).

Next, we quantified the overlap of susceptibility genes and GWAS signals. Of the 1,789 identified gene-trait pairs, 168 (9%) were not proximal (more than 0.5 Mb from the TSS) to any genome-wide-significant SNP for that respective trait (see Table 2). This measure was robust to increases in window size, such that 140 (8%) gene-trait pairs did not overlap a genome-wide-significant SNP within 1 Mb of the TSS. We observed increased SNP association statistics at these genes (mean $\chi^2 = 6.5$; see Figure S4), which suggests that GWASs with an increased sample size will discover genome-wide-significant SNPs nearby. We tested this hypothesis by assessing the new TWAS loci for educational years[21] ($n = 126{,}599$) in a recent, much larger GWAS for educational years[35] ($n = 293{,}723$). All four independent loci contained a genome-wide-significant SNP in the larger GWAS (see Table S7). Of the 1,526 GWAS risk loci, 1,405 (92%) overlapped at least one eGene (i.e., a gene with heritable expression levels in at least one of the considered expression panels), and 551 (36%) overlapped at least one susceptibility gene (see Table 1). Focusing on the 1,621 TWAS associations that overlapped a genome-wide-significant SNP, we observed 1,350 (83%) genes that were not the closest, suggesting that the traditional heuristic of prioritizing genes closest to GWAS SNPs is typically not supported by evidence from eQTL data[48] (see Figure S5). This is also supported by the mean χ^2 association statistics for genes closest to index SNPs ($\chi^2 = 43.9$) and the top association ($\chi^2 = 72.9$; see Figure S6). In addition, lead GWAS SNPs typically have a weaker eQTL effect for the proximal gene than for the TWAS-implicated gene in 1,088 of 1,350 TWAS associations. This result, consistent with earlier reports,[11,12] highlights the importance of utilizing the entire locus and estimates of LD to prioritize genes.

Although GWAS SNPs provide the majority of the power in this approach, the flexibility of TWASs to leverage allelic heterogeneity provides a significant gain.[11] We found 219 instances across 19 traits where association signal was stronger (20% higher χ^2 statistics on average) in TWASs than in GWASs. For example, predicted expression in *CCDC88B* (OMIM: 611205; a gene involved in T cell maturation and inflammation[49]) exhibited strong association with Crohn disease ($p_{TWAS} = 6.32 \times 10^{-8}$), whereas the index SNP (i.e., top overlapping GWAS SNP) at site rs11231774 was only suggestive ($p_{GWAS} = 2.47 \times 10^{-6}$). This effect was most dramatic for height, such that 108 susceptibility genes had a stronger signal than GWAS index SNPs. We observed that the χ^2 statistics for predicted expression in *CRELD1* (OMIM: 607170; $p_{TWAS} = 1.55 \times 10^{-10}$) were 2.6× higher than those for the index SNP rs1473183 ($p_{GWAS} = 6.33 \times 10^{-5}$).

Recent work[50] applied a similar approach[12] that used summary eQTLs from blood and GWAS data to identify 71 genes for 28 complex traits.[50] Of the investigated traits, 12 overlapped those in our study. Overall, whereas that study reported 63 genes for these traits, we identified 564 genes. Surprisingly, despite using independent methods and expression data, we replicated 40 out of 51 associations for genes assayed in both studies (see Table S8). This increase in power can be attributed to two reasons. First, we integrated many more expression panels sampled from many tissues, leading to many more genes for the assay. Second, we used a method that jointly tests the entire locus rather than the index SNPs. We have shown

Table 1. Summary of GWAS and TWAS Results

Trait	Abbreviation	Number of GWASs				Number of Susceptibility Genes	
		Loci	Loci with an eGene	Loci with a Single Susceptibility Gene	Loci with at Least One Susceptibility Gene	Genes Overlapping GWASs	Genes Not Overlapping GWASs
Age at menarche	AM	70	60	14	19	34	9
Body mass index	BMI	76	60	10	18	44	11
College	COL	5	5	2	2	1	4
Crohn disease	CD	50	48	4	17	65	5
Educational years	EY	7	4	2	2	2	11
Fasting glucose	FG	12	11	2	5	8	1
Fasting insulin	FI	0	0	0	0	0	1
Femoral neck bone mineral density	FN	20	20	2	2	2	1
Forearm bone mineral density	FA	3	3	0	0	0	0
Hemoglobin	HB	22	21	2	5	22	3
HbA_{1c}	–	10	10	0	1	4	0
Height	–	482	454	94	225	669	52
High-density lipoprotein	HDL	100	95	11	29	98	4
HOMA-B	–	4	3	0	0	0	0
HOMA-IR	–	0	0	0	0	0	1
Inflammatory bowel disease	IBD	63	59	12	23	70	11
Low-density lipoprotein	LDL	75	72	8	25	84	3
Lumbar spine	LS	24	23	2	3	4	0
Mean cell hemoglobin concentration	MCHC	5	3	0	0	0	0
Mean cell hemoglobin	MCH	35	31	5	17	46	7
Mean cell volume	MCV	43	40	8	20	49	1
Number of platelets	PLT	35	34	6	13	30	8
Packed cell volume	PCV	14	13	1	3	5	1
Red blood cell count	RBC	25	21	3	10	35	2
Rheumatoid arthritis	RA	44	41	7	13	30	5
Schizophrenia	SCZ	95	74	15	31	113	24
Total cholesterol	TC	88	85	13	40	117	0
Triglycerides	TG	70	67	4	18	59	1
Type 2 diabetes	T2D	12	12	0	1	3	0
Ulcerative colitis	UC	37	36	5	9	27	2
Total		1,526	1,405	232	551	1,621	168

The first four numeric columns summarize GWAS risk loci. The last two numeric columns summarize identified TWAS susceptibility genes. The majority (92%) of GWAS risk loci overlap at least one eGene, of which 40% contain at least one susceptibility gene. We report 168 (9%) identified gene-trait pairs that do not overlap a GWAS variant, providing risk loci for follow up.

that many identified susceptibility genes contain signals of allelic heterogeneity; therefore, using individual SNPs will decrease power.

Genes Associated with Multiple Traits

We investigated the degree of pleiotropic susceptibility genes (i.e., genes associated with more than one trait) in our data and found 380 (32%) genes associated with multiple traits (see Figure S7). For example, *IKZF3* (OMIM: 606221) displayed strong associations with Crohn disease (NTR; $p = 1.6 \times 10^{-9}$), HDL levels (NTR; $p = 6.6 \times 10^{-15}$), inflammatory bowel disease (NTR; $p = 7.9 \times 10^{-16}$), rheumatoid arthritis (NTR; $p = 6.0 \times 10^{-8}$), and ulcerative colitis (NTR; $p = 9.2 \times 10^{-10}$). Indeed, *IKZF3* has been

Table 2. Susceptibility Genes That Do Not Overlap a Genome-wide Significant SNP within 0.5 Mb of the Transcription Start and End Sites for Each Trait

Trait	Genes
AM	*CCDC65*, *COG6*, *INO80E*, *NUCKS1*, *PMS2P5*, *RAB7L1*, *SLC26A9*, *STAG3L2*, and *TMEM180*
BMI	*CDK5RAP3*, *CERCAM*, *DHRS11*, *GGNBP2*, *INO80E*, *RP11-6N17.10*, *RP11-6N17.9*, *SLC27A4*, *STAG3L1*, *TUBA1C*, and *URM1*
CD	*CCDC88B*, *CISD1*, *PPP1R14B*, *RIT1*, and *SMIM19*
COL	*ABCB9*, *AC091729.9*, *AFF3*, and *RNF123*
EY	*ABCB9*, *EIF3CL*, *MIR4721*, *MPHOSPH9*, *NFATC2IP*, *RP11-1348G14.4*, *SDCCAG8*, *SH2B1*, *STK24*, *SULT1A1*, and *TUFM*
FG	*MAPRE3*
FI	*KNOP1*
FN	*FGFRL1*
HB	*CCDC117*, *UBE2Q2*, and *WNT3*
HDL	*HRAS*, *KNOP1*, *RETSAT*, and *TYRO3*
HEIGHT	*ARL17A*, *ATF1*, *ATP5J2*, *C20orf194*, *C9orf156*, *CCDC116*, *CNIH4*, *COX6B1*, *CRELD1*, *CRHR1*, *DAB2IP*, *DESI1*, *DLG5*, *DUS3L*, *ECHDC2*, *FAM35A*, *FUCA2*, *H2AFJ*, *HIBADH*, *INO80E*, *IQGAP1*, *KANSL1*, *LBX2-AS1*, *LRRC37A2*, *MAPT*, *MAT2A*, *MED4*, *MEGF9*, *MGMT*, *MORC2-AS1*, *MSRB2*, *P4HTM*, *PHF19*, *PLEKHA1*, *PSMD5*, *PSMD5-AS1*, *RP11-173M1.8*, *RP11-455F5.3*, *RP11-4O1.2*, *RP11-67A1.2*, *RP13-39P12.3*, *RP4-612B15.3*, *RRN3*, *SFTPD*, *SH3YL1*, *SUSD1*, *TMEM128*, *UBE2L3*, *UTP18*, *WDR60*, *YPEL3*, and *YWHAB*
HOMA-IR	*KNOP1*
IBD	*ADCY3*, *CCDC88B*, *FAM189B*, *GBA*, *GBAP1*, *HCN3*, *PPP1R14B*, *RMI2*, *SATB2*, *TMEM180*, *ZFP90*
LDL	*DHRS13*, *ERAL1*, and *WDR25*
MCH	*AP003419.16*, *GSTP1*, *PABPC4*, *PTPRCAP*, *RP11-69E11.4*, *RP1-18D14.7*, and *RPS6KB2*
MCV	*COX4I2*
PCV	*PLEKHH2*
PLT	*ACTR1A*, *BAZ2A*, *CCDC17*, *IPP*, *MUTYH*, *PRIM1*, *TESK2*, and *TMEM180*
RA	*METTL21B*, *RNF40*, *RPS26*, *SLC26A10*, and *SUOX*
RBC	*COX4I2* and *FBXL20*
SCZ	*ALMS1P*, *ARL14EP*, *CAD*, *CBR3*, *CEBPZ*, *CORO7*, *CPNE7*, *DND1*, *EMB*, *ENDOG*, *EPN2*, *GRAP*, *IK*, *NMRAL1*, *NRBP1*, *PCNX*, *PFDN1*, *PRR12*, *PRRG2*, *RNF112*, *RP11-135L13.4*, *SEPT10*, *SRA1*, and *TMCO6*
TG	*L3MBTL3*
UC	*SATB2* and *TNPO3*

For details on individual genes, expression studies, and association statistics, see Table S4. Genome-wide significance: $p < 5 \times 10^{-8}$.

shown to influence lymphocyte development and differentiation.[51,52] These traits are known to have a strong autoimmune component;[53] hence, association with predicted *IKZF3* expression levels is consistent with a model where *cis*-regulated variation in *IKZF3* product levels contributes to risk. Similarly, we observed three susceptibility genes shared between educational years (EY) and height (see Figure 4): *ABCB9* (OMIM: 605453; GTEx heart left ventricle; $p_{height} = 1.38 \times 10^{-15}$; $p_{EY} = 1.28 \times 10^{-6}$), *BTN2A3P* (OMIM: 613592; GTEx subcutaneous adipose; $p_{height} = 3.82 \times 10^{-12}$; $p_{EY} = 1.90 \times 10^{-7}$), and *MPHOSPH9* (OMIM: 605501; GTEx thyroid; $p_{height} = 5.84 \times 10^{-18}$; $p_{EY} = 1.30 \times 10^{-6}$). Although not direct evidence of co-localization of educational years and height at these loci, this result is consistent with a recent study[13] that reported a non-zero genetic correlation between height and educational years ($\hat{\rho}_g = 0.13$; $p = 3.82 \times 10^{-6}$).

The Effect of *cis* Expression on Traits Is Consistent across Tissues

Having established the importance of individual predicted gene expression levels for these traits, we next estimated the amount of trait variance explained by predicted expression by using all examined genes, including those not significantly associated, and an LD score regression approach (see Material and Methods). We found 108 tissue-trait pairs across 17 traits and 33 tissues where the cumulative effect of all measured genes on the trait was significantly greater ($p < 0.05/45$) than for the significant-only set (see Table S9). For example, in height we estimated $h^2_{GE} = 0.07$ (Jack-knife SE = 0.02; $p = 5.6 \times 10^{-4}$) by using all 3,733 measured genes in YFS and $h^2_{GE} = 0.015$ (Jack-knife SE = 6.9; p = 0.03) by using only the 169 YFS susceptibility genes ($p_{all>sig} = 5.6 \times 10^{-3}$). This suggests that height has additional susceptibility genes, which we are underpowered to detect. Strikingly, the predicted expression from all

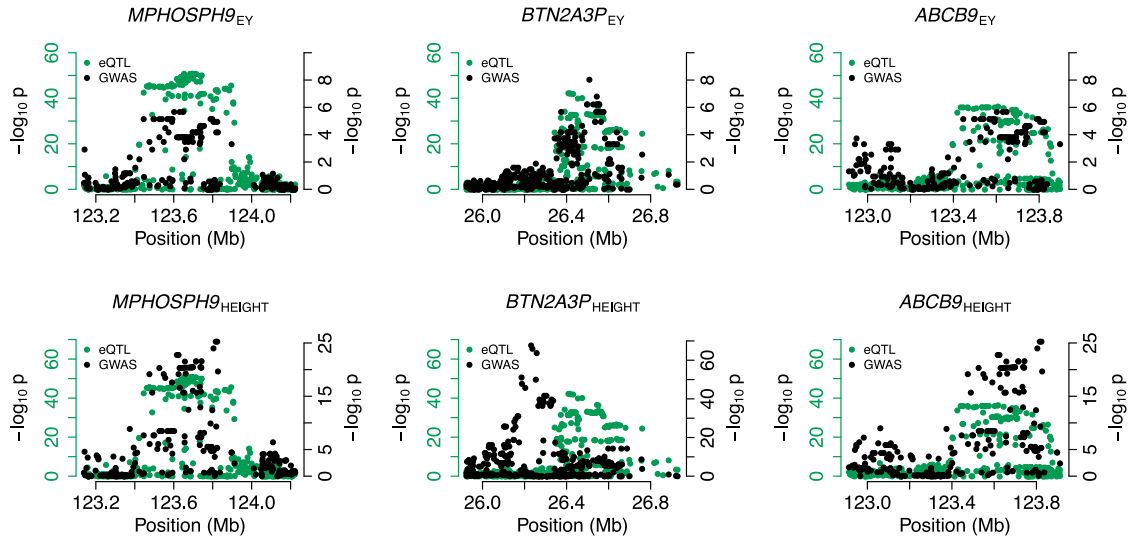


Figure 4. Susceptibility Genes Shared for Educational Years and Height
We indicate $-\log_{10}$ p values for eQTLs in green and trait-specific GWASs in black on separate axes to simplify illustration.

YFS genes accounts for 12% of SNP heritability measured in height.[54] However, for most trait-tissue pairs, we did not observe a significant difference at our given sample sizes. Indeed, we measured a significant association between expression-study sample size and number of eGenes ($r = 0.73$; SE = 0.10; $p = 1.3 \times 10^{-8}$), which indicates that smaller studies lack power to find eGenes and thus underestimate the total h^2_{GE}.

We next asked whether any tissues are burdened with increased levels of risk for a given trait. To test this hypothesis, we examined the difference between estimated trait variance explained per gene and the average. Our results did not suggest tissue-specific enrichment at the current sample sizes (see Table S10). We observed a significant correlation between gene expression sample size and tissue enrichment estimates ($p = 62.4 \times 10^{-6}$). One explanation for this relationship is that the number of eGenes identified per study increases with sample size, which increases h^2_{GE} estimates. Given no observable difference in tissue-specific risk, we expect local estimates of genetic correlation to be highly similar across tissues. When estimating $\rho_{g,local}$, we observed consistent effect-size estimates in both sign and magnitude estimates across tissues (mean tissue-tissue $r = 0.82$; see Figure 5). These results are compatible with earlier work that found that *cis* effects on expression are largely consistent across tissues.[55] To obtain a meta-estimate of local genetic correlation for gene-trait pairs with measurements in multiple tissues, we used the mean genetic correlation across all expression panels in all of the following analyses.

Genetic Correlation between Traits at the Level of Predicted Expression

To evaluate the shared contribution of predicted expression on pairs of traits, we used nominally significant ($p < 0.05$) genes to compute the genome-wide genetic correlation at levels of predicted expression (see Material and Methods). For 435 distinct pairs, we discovered 43 significant expression correlations, 22 of which had previously reported non-zero genetic correlations[13] (see Figure 6 and Table 3). For example, age of menarche and BMI had $\hat{\rho}_{GE} = -0.32$ (95% CI = [−0.32, −0.21]; $p = 7.97 \times 10^{-8}$). This negative correlation is consistent with estimates published in

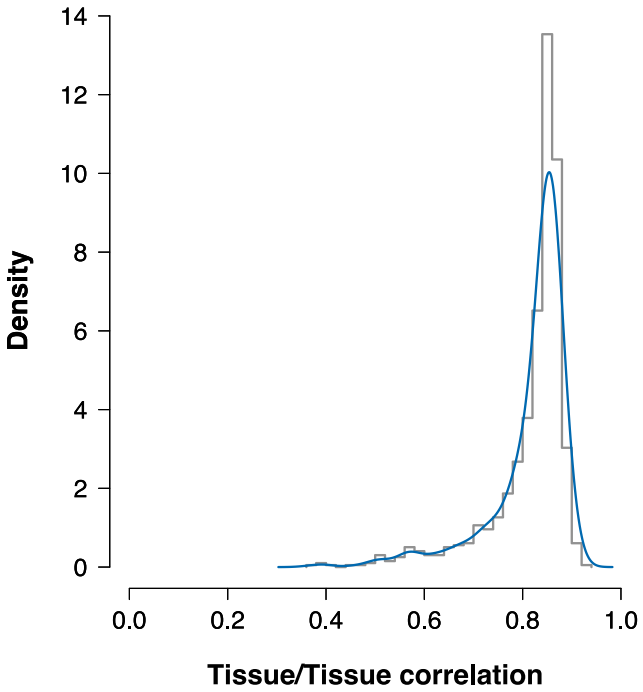


Figure 5. Histogram and Density Estimate for Correlation of $\rho_{g,local}$ across Tissues
We computed the correlation across pairs of different tissues by using local estimates of genetic correlation between expression and traits. Most tissues exhibited a high correlation over the underlying gene effects on traits with an estimated mean of $r = 0.82$.

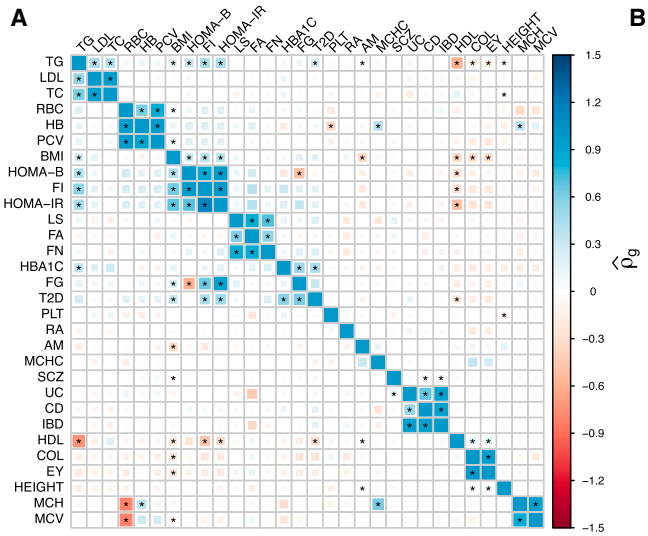


Figure 6. Estimates of Genetic Correlation $\hat{\rho}_g$ Obtained from LD Scores versus Estimates of Expression Correlation $\hat{\rho}_{GE}$ from Nominally Significant TWAS Results

(A) Correlation matrix for 30 traits. The lower triangle contains $\hat{\rho}_{GE}$, and the upper triangle contains $\hat{\rho}_g$ estimates. Correlation estimates that are significantly non-zero ($p < 0.05/435$) are marked with an asterisk (*). The strength and direction of correlation are indicated by size and color. We found 43 significantly correlated traits by using predicted expression and 62 by using genome-wide SNPs.

(B) Linear relationship between estimates of $\hat{\rho}_{GE}$ and $\hat{\rho}_g$. We indicate whether individual estimates were significant in either approach by color. Non-significant trait pairs are reduced in size for visibility.

epidemiological studies,[56] in addition to studies probing genetic correlation across complex traits.[13] To determine whether estimates were sensitive to changes in scale, we recomputed $\hat{\rho}_{GE}$ by using the top eQTL as a proxy for local heritability of gene expression and observed similar results ($r = 0.99$; $p = 2.2 \times 10^{-16}$; see Figure S8). Results were also robust to increasing window size for gene pruning, such that there was no significant difference in estimates between 2 and 4 Mb windows ($r_{2Mb} = 0.99$; $r_{4Mb} = 0.98$). Using estimates of $\hat{\rho}_{GE}$, we clustered traits and observed groups forming naturally in the trait-trait matrix (see Figure 6). Interestingly, BMI clustered with insulin-related traits (HOMA-B, HOMA-IR, and fasting insulin). Our estimates were highly consistent with the results of LD score regression (see Figure 6 and Table S11). Out of 435 pairs of traits, 35 demonstrated significance for $\hat{\rho}_{GE}$ and $\hat{\rho}_g$, whereas 8 and 27 were exclusive to $\hat{\rho}_{GE}$ and $\hat{\rho}_g$, respectively. Given the high degree of concordance between estimates, we tested for significant differences and found four insulin-related pairs of traits and three blood-related pairs with more extreme values for $\hat{\rho}_{GE}$ (see Table S11). Differences for these pairs of traits can be partially explained by overconfident standard errors for $\hat{\rho}_{GE}$ (see Table S12). Overall, we found $\hat{\rho}_{GE}$ to explain most of the variation in $\hat{\rho}_g$ ($r^2 = 0.72$). We compared this to the naive approach of computing the correlation of SNP Z scores across susceptibility gene loci and observed a much smaller proportion of variance explained in $\hat{\rho}_g$ ($r^2 = 0.46$). This reinforces that, compared to the naive approach, our method incorporates LD to aggregate signal.

Bi-directional Regression Suggests Putative Causal Relationships

Given pairs of traits with significant estimates of ρ_{GE}, we aimed to distinguish among possible causal explanations by performing bi-directional regression analyses (see Material and Methods). To empirically validate our approach, we regressed HDL, LDL, and TG with TC. TC is the direct consequence of summing over TG, HDL, and LDL levels, so we expected to observe higher signal for $\rho_{TC\,|\,lipid}$ than for $\rho_{lipid\,|\,TC}$. Of these three, we found evidence that TG influences TC ($p = 2.34 \times 10^{-3}$). We observed consistent, but not significant, evidence for the effects of LDL on TC ($p = 0.07$) and HDL on TC ($p = 0.55$; see Figure 7). These results suggest that point estimates from the bi-directional approach favor the correct model but might not have adequate power required for significance.

We tested the 43 pairs of traits identified above (see Table 3) while ascertaining susceptibility genes and observed asymmetric effects at $p < 0.05$ for BMI-TG and LDL-TG (see Figure 8 and Table 4). For example, in the bi-directional analysis on BMI and TG, we observed a significant effect for $\rho_{TG\,|\,BMI} = 0.62$ (95% CI = [0.27, 0.83]; $p = 2.06 \times 10^{-3}$). By contrast, the reverse analysis estimate overlapped 0 at $\rho_{BMI\,|\,TG} = -0.04$ (95% CI = [−0.49, 0.42]; $p = 0.86$). Individual estimates for $\rho_{TG\,|\,BMI}$ and $\rho_{BMI\,|\,TG}$ were significantly different ($p = 0.01$, Welch's t test), which is consistent with a model where BMI directly influences TG levels. In practice, we used susceptibility genes found through a TWAS ($p \sim 1 \times 10^{-6}$), but this could be too strict an inclusion threshold for genes for which we lack power to detect. We conducted analyses with weaker thresholds and observed similar results (see Tables S13 and S14). Our results reinforce previous estimates of putative causal effects where BMI influences TG levels.[45,57]

Discussion

In this work, we described an approach to estimate the local genetic covariance and correlation between gene expression and complex traits by using GWAS summary data. We also introduced a method of estimating genome-wide genetic correlation between complex traits at the level of predicted expression. Using simulations, we demonstrated that both approaches are relatively unbiased under realistic

Table 3. Pairs of Traits with Significant Estimates of ρ_{GE}

		All Nominally Significant Genes			
Trait 1	**Trait 2**	$\widehat{\rho}_{GE}$	**95% CI**		**M**
AM	BMI	−0.33	−0.43	−0.21	257
BMI	COL	−0.31	−0.44	−0.18	190
BMI	EY	−0.31	−0.43	−0.18	210
BMI	FI	0.39	0.25	0.51	164
BMI	HDL	−0.34	−0.45	−0.23	256
BMI	HOMA-B	0.31	0.17	0.44	168
BMI	HOMA-IR	0.36	0.22	0.49	162
BMI	TG	0.29	0.17	0.41	233
CD	IBD	0.93	0.91	0.94	366
CD	UC	0.51	0.41	0.60	218
COL	EY	0.95	0.94	0.96	363
FA	FN	0.57	0.44	0.67	149
FA	LS	0.60	0.49	0.69	170
FG	FI	0.65	0.53	0.74	133
FG	HOMA-B	−0.60	−0.70	−0.47	125
FG	HOMA-IR	0.92	0.89	0.94	136
FI	HDL	−0.31	−0.44	−0.17	168
FI	HOMA-B	0.97	0.96	0.98	243
FI	HOMA-IR	0.99	0.99	0.99	383
FI	TG	0.57	0.45	0.66	152
FN	LS	0.86	0.83	0.89	264
HB	MCH	0.37	0.23	0.50	156
HB	MCHC	0.40	0.23	0.55	105
HB	PCV	0.97	0.96	0.97	338
HB	PLT	−0.36	−0.49	−0.20	141
HB	RBC	0.95	0.94	0.96	260
HbA_{1c}	T2D	0.46	0.30	0.59	110
HbA_{1c}	TG	0.37	0.21	0.50	137
HDL	HOMA-IR	−0.32	−0.46	−0.18	159
HDL	T2D	−0.32	−0.45	−0.19	186
HDL	TG	−0.74	−0.79	−0.69	274
HOMA-B	HOMA-IR	0.97	0.96	0.98	227
HOMA-B	TG	0.43	0.27	0.56	127
HOMA-IR	TG	0.48	0.34	0.60	138
IBD	UC	0.96	0.95	0.96	415
LDL	TC	0.97	0.96	0.97	452
LDL	TG	0.54	0.44	0.63	231
MCH	MCHC	0.63	0.51	0.72	127
MCH	MCV	0.96	0.95	0.97	320
MCH	RBC	−0.81	−0.85	−0.76	207
MCV	RBC	−0.80	−0.85	−0.75	208
PCV	RBC	0.96	0.95	0.97	278
TC	TG	0.61	0.53	0.68	248

Estimates were computed with M pruned genes that were nominally significant ($p < 0.05$) in both traits.

scenarios. We used GWAS summary statistics from 30 complex traits and diseases jointly with expression data collected across 45 expression panels to identify 1,196 susceptibility genes for complex traits. Interestingly, susceptibility genes that were identified for educational years and not proximal to a genome-wide significant SNP were validated in a much larger GWAS.[35] We leveraged estimates of local genetic correlation between gene expression and traits to compute ρ_{GE} for 435 trait pairs. This quantified the shared effect of predicted expression levels between two complex traits. To provide evidence of possible causal direction, we adapted a recently proposed causality test[45] to operate at the level of predicted gene expression. Our results suggest that TG influences LDL and that BMI influences TG. As more GWAS and eQTL summary results become publicly available, we expect additional studies to integrate cross-trait information to make inferences about mechanistic bases for complex traits. Indeed, recent work has combined chromatin phenotypes with alternatively spliced introns and total gene expression (the latter of which overlaps expression used in this study) to identify regulatory mechanisms for schizophrenia.[58]

Under the assumption that gene expression mediates the effect of genetics on complex traits, testing for association between predicted gene expression and traits is equivalent to a two-sample Mendelian randomization test for a causal effect of expression on a trait.[59,60] This test for causality is valid if SNPs do not exhibit pleiotropic effects, which is difficult to prove; therefore, TWAS associations do not provide direct evidence of causal relationships between gene expression and complex traits but rather reflect associations between expression levels and traits. This set of assumptions extends to our bi-directional approach to inferring causal direction. A bi-directional regression is capable of distinguishing between directions of effect but cannot rule out pleiotropy. Therefore, our results show consistency with a putative causal mechanism and should not be interpreted as direct proof of causality.

We conclude with several caveats. First, we note that using estimates of genetic correlation to find susceptibility genes could still be biased as a result of confounding. The expression weights used for TWASs could tag variants that are causal through other genes or non-genic mechanisms. In principle, this can be partially remedied by joint testing of multiple genes and a trait. In this work, we combined

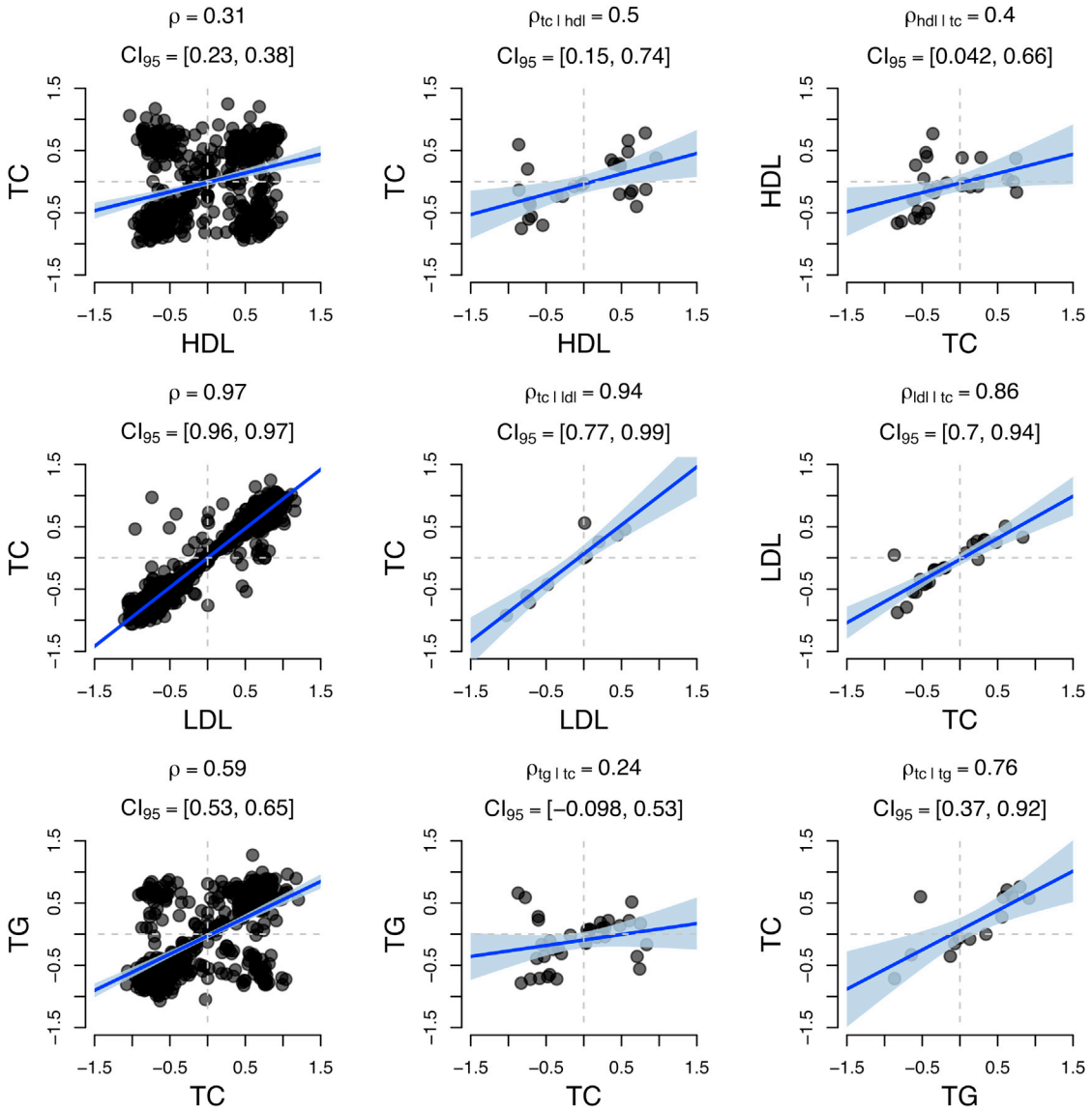


Figure 7. Estimates of Expression Correlation ρ_{GE} between TC and HDL, LDL, and TG
(Left column) Estimates of ρ_{GE} with the use of nominally significant genes ($p < 0.05$).
(Middle column) We repeated the analysis by using only susceptibility genes found in the x axis trait but not found in the y axis trait.
(Right right) Same analysis as in the middle column but with the other trait's susceptibility genes.
All three analyses resulted in stronger point estimates for $\rho_{TC\ |lipid}$ when conditioning on HDL, LDL, and TG genes than for $\rho_{lipid\ |TC}$; however, significance was observed only for $\rho_{TC|TG}$ ($p = 2.34 \times 10^{-3}$). Shaded regions indicate the estimated 95% confidence interval for the regression line.

estimates across tissues by taking the mean effect to compute the genetic correlation between traits and expression. This approach is unbiased but could be inefficient. Recent work[61] has described a random-effect model that combines estimates across tissues to increase power. Finally, our method of estimating correlation between traits by using the genetically predicted component of gene expression makes several simplifying assumptions. First, we remedied the non-independence of genes by sampling single genes within a 1 Mb region, an approach that has been used previously.[46] However, a more powerful approach could take correlations across genes into account. Second, we limited predictive models to the local (or *cis*) effects on gene expression, which ignores distal (or *trans*) effects that regulate gene expression. Although the predictive accuracy of models for gene expression used in this study can account for most of the variation due to genetics,[11] we believe that incorporating additional sources of genomic information (e.g., functional priors on SNP effects[39,62,63]) could make additional refinement possible.

Appendix A: Pathway Analysis

We used the PANTHER database[47] to explore putative molecular function and pathways enriched with identified

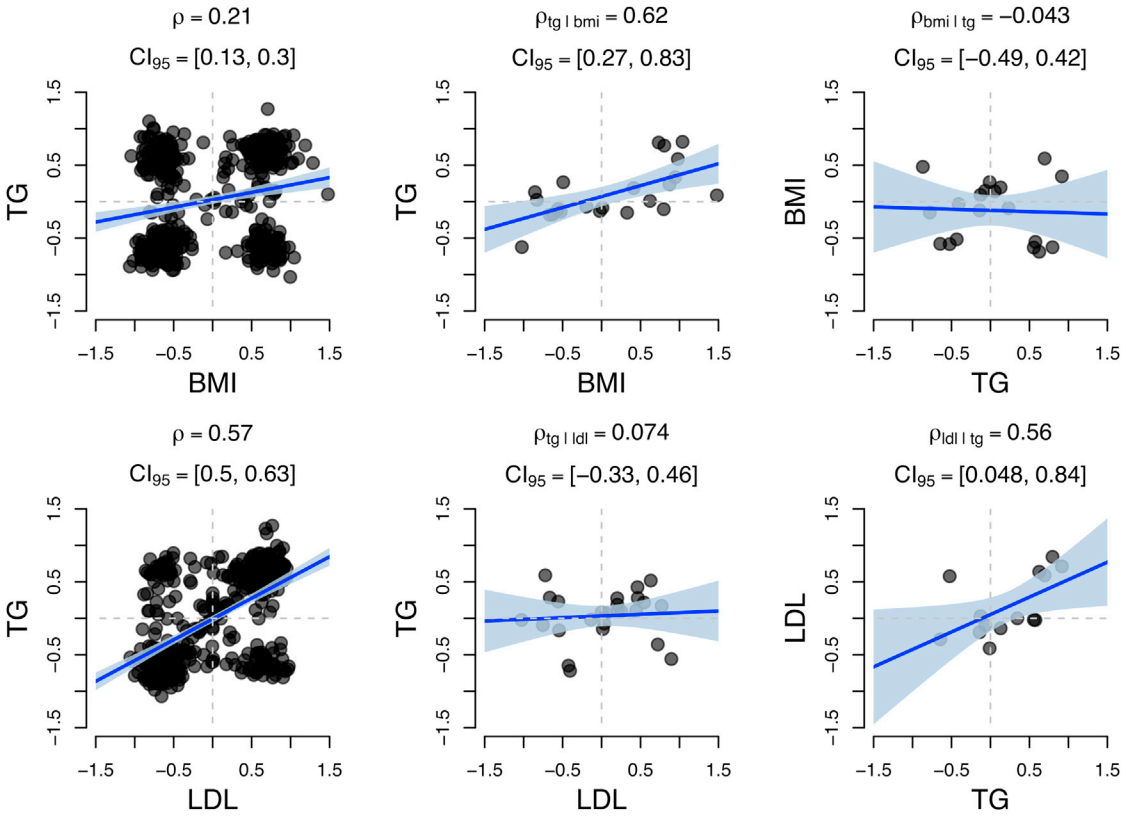


Figure 8. Estimates of $\hat{\rho}_{GE}$ for TG with BMI and for TG with LDL
We present results for pairs of traits that displayed a significant difference ($p < 0.05$, Welch's t test) in their conditional estimates. These results are consistent with a causal model where BMI influences TG and TG influences LDL. Shaded regions indicate the estimated 95% confidence interval for the regression line.

susceptibility genes. Using all susceptibility genes across all traits, we found 13 significantly enriched categories, of which seven were related to binding functions. Catalytic activity exhibited the strongest enrichment at 1.3× (GO: 0003824; $p = 5.17 \times 10^{-9}$; see Figure S9). We next focused on individual traits (see Figure S10); however, most individually tested gene sets did not indicate significant enrichment, except for height, LDL, and TC. For example, height had a significant enrichment of genes with catalytic activity (1.31×; $p = 4.77 \times 10^{-4}$). We next looked at biological processes and found TWAS genes enriched at 1.2× for metabolic processes (GO: 0008152; $p = 7.29 \times 10^{-11}$) and 1.57× cellular catabolic processes (GO: 0044248; $p = 2.51 \times 10^{-2}$; see Figures S11 and S12). Enrichment was most pronounced in susceptibility genes specific to height (1.3×; $p = 1.03 \times 10^{-6}$).

Supplemental Data

Supplemental Data include 12 figures and 14 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.01.031.

Acknowledgments

We would like to thank Valerie Arboleda, Robert Brown, Kathy Burch, and Malika Kumar for helpful discussions and feedback. We also thank Dr. Nicole Soranzo for sharing summary data for the

Table 4. Bi-directional Estimates of Genome-wide Genetic Correlation at the Level of Predicted Expression

		Results when Ascertaining for Trait 1				Results when Ascertaining for Trait 2				Test for Difference		
Trait 1	**Trait 2**	$\hat{\rho}_{GE}$	**SE**	**p**	***M***	$\hat{\rho}_{GE}$	**SE**	**p**	***M***	***t***	**p**	**~*M***
BMI	TG	0.62	0.10	2.06×10^{-3}	22	−0.04	0.22	8.62×10^{-1}	19	2.74	1.12×10^{-2}	25
LDL	TG	0.07	0.19	7.25×10^{-1}	25	0.56	0.13	3.55×10^{-2}	14	−2.17	3.69×10^{-2}	36
TC	TG	0.24	0.14	1.63×10^{-1}	36	0.76	0.08	1.79×10^{-3}	14	−3.22	2.34×10^{-3}	47

We denote the number of ascertained genes used in the test as *M*. We tested for a difference as a *t* statistic, where $t = \frac{|\rho_{GE,1} - \rho_{GE,2}|}{\sqrt{SE_1^2 + SE_2^2}} \sim t(df)$ and df is the approximate degrees of freedom determine by the Welch-Satterthwaite equation.

platelet traits. This research was funded in part by NIH awards GM105857, GM053275, and HG009120. G.K. is supported by the Biomedical Big Data Training Program (NIH-NCI T32CA201160). CMC data were generated as part of the CommonMind Consortium, supported by funding from Takeda Pharmaceuticals, F. Hoffman-La Roche, and NIH grants R01MH085542, R01MH093725, P50MH066392, P50MH080405, R01MH097276, RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881, R37MH057881S1, HHSN271201300031C, AG02219, AG05138, and MH06692. Brain tissue for the study was obtained from the following brain bank collections: the Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania Alzheimer Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain and Tissue Repositories, and the National Institute of Mental Health (NIMH) Human Brain Collection Core. CommonMind Consortium leadership includes Pamela Sklar, Joseph Buxbaum (Icahn School of Medicine at Mount Sinai), Bernie Devlin, David Lewis (University of Pittsburgh), Raquel Gur, Chang-Gyu Hahn (University of Pennsylvania), Keisuke Hirai, Hiroyoshi Toyoshiba (Takeda Pharmaceuticals), Enrico Domenici, Laurent Essioux (F. Hoffman-La Roche), Lara Mangravite, Mette Peters (Sage Bionetworks), Thomas Lehner, Barbara Lipska (NIMH).

Received: August 31, 2016
Accepted: January 23, 2017
Published: February 23, 2017

Web Resources

CommonMind Consortium, https://www.synapse.org
FUSION software package, http://gusevlab.org/projects/fusion/
GCTA, http://cnsgenomics.com/software/gcta/
Gene Ontology, http://www.geneontology.org/
GTEx Portal, http://www.gtexportal.org/home/
OMIM, http://www.omim.org
PLINK, https://www.cog-genomics.org/plink2/
RhoGE software, https://github.com/bogdanlab/RHOGE

References

1. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.
2. Claussnitzer, M., Dankel, S.N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N. Engl. J. Med. *373*, 895–907.
3. Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C., Patsopoulos, N.A., Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S.E., et al.; International Multiple Sclerosis Genetics Consortium; and Wellcome Trust Case Control Consortium 2 (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature *476*, 214–219.
4. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. *6*, e1000888.
5. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423–428.
6. Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. *6*, e1000895.
7. Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. *16*, 197–212.
8. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al.; GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.
9. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.
10. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., Im, H.K.; and GTEx Consortium (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.
11. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.
12. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., and Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*, 481–487, advance online publication.
13. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., Duncan, L., Perry, J.R., Patterson, N., Robinson, E.B., et al.; ReproGen Consortium; Psychiatric Genomics Consortium; and Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3 (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.
14. Davey Smith, G., and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum. Mol. Genet. *23* (R1), R89–R98.
15. Zheng, H.F., Forgetta, V., Hsu, Y.H., Estrada, K., Rosello-Diez, A., Leo, P.J., Dahia, C.L., Park-Min, K.H., Tobias, J.H., Kooperberg, C., et al.; AOGC Consortium; and UK10K Consortium (2015). Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature *526*, 112–117.
16. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al.; Wellcome Trust Case Control Consortium; Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. *44*, 981–990.
17. Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al.;

International Multiple Sclerosis Genetics Consortium; and International IBD Genetics Consortium (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. *47*, 979–986.
18. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
19. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al.; RACI consortium; and GARNET consortium (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature *506*, 376–381.
20. Perry, J.R.B., Day, F., Elks, C.E., Sulem, P., Thompson, D.J., Ferreira, T., He, C., Chasman, D.I., Esko, T., Thorleifsson, G., et al.; Australian Ovarian Cancer Study; GENICA Network; kConFab; LifeLines Cohort Study; InterAct Consortium; and Early Growth Genetics (EGG) Consortium (2014). Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature *514*, 92–97.
21. Rietveld, C.A., Medland, S.E., Derringer, J., Yang, J., Esko, T., Martin, N.W., Westra, H.-J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al.; LifeLines Cohort Study (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science *340*, 1467–1471.
22. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al.; Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.
23. Soranzo, N., Sanna, S., Wheeler, E., Gieger, C., Radke, D., Dupuis, J., Bouatia-Naji, N., Langenberg, C., Prokopenko, I., Stolerman, E., et al.; WTCCC (2010). Common variants at 10 genomic loci influence hemoglobin A_1(C) levels via glycemic and nonglycemic pathways. Diabetes *59*, 3229–3239.
24. Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., Gloyn, A.L., et al.; DIAGRAM Consortium; GIANT Consortium; Global BPgen Consortium; Anders Hamsten on behalf of Procardis Consortium; and MAGIC investigators (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat. Genet. *42*, 105–116.
25. Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y., et al. (2011). New gene functions in megakaryopoiesis and platelet formation. Nature *480*, 201–208.
26. van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature *492*, 369–375.
27. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al.; LifeLines Cohort Study; ADIPOGen Consortium; AGEN-BMI Working Group; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GLGC; ICBP; MAGIC Investigators; MuTHER Consortium; MIGen Consortium; PAGE Consortium; ReproGen Consortium; GENIE Consortium; and International Endogene Consortium (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197–206.
28. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; and LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.
29. Fromer, M., Roussos, P., Sieberts, S.K., Johnson, J.S., Kavanagh, D.H., Perumal, T.M., Ruderfer, D.M., Oh, E.C., Topol, A., Shah, H.R., et al. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat. Neurosci. *19*, 1442–1453.
30. Raitakari, O.T., Juonala, M., Rönnemaa, T., Keltikangas-Järvinen, L., Räsänen, L., Pietikäinen, M., Hutri-Kähönen, N., Taittonen, L., Jokinen, E., Marniemi, J., et al. (2008). Cohort profile: the cardiovascular risk in Young Finns Study. Int. J. Epidemiol. *37*, 1220–1226.
31. Stancáková, A., Civelek, M., Saleem, N.K., Soininen, P., Kangas, A.J., Cederberg, H., Paananen, J., Pihlajamäki, J., Bonnycastle, L.L., Morken, M.A., et al. (2012). Hyperglycemia and a common variant of GCKR are associated with the levels of eight amino acids in 9,369 Finnish men. Diabetes *61*, 1895–1902.
32. Stancáková, A., Javorský, M., Kuulasmaa, T., Haffner, S.M., Kuusisto, J., and Laakso, M. (2009). Changes in insulin sensitivity and insulin release in relation to glycemia and glucose tolerance in 6,414 Finnish men. Diabetes *58*, 1212–1221.
33. Nuotio, J., Oikonen, M., Magnussen, C.G., Jokinen, E., Laitinen, T., Hutri-Kähönen, N., Kähönen, M., Lehtimäki, T., Taittonen, L., Tossavainen, P., et al. (2014). Cardiovascular risk factors in 2011 and secular trends since 2007: the Cardiovascular Risk in Young Finns Study. Scand. J. Public Health *42*, 563–571.
34. Wright, F.A., Sullivan, P.F., Brooks, A.I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y.-H., et al. (2014). Heritability and genomics of gene expression in peripheral blood. Nat. Genet. *46*, 430–437.
35. Okbay, A., Beauchamp, J.P., Fontana, M.A., Lee, J.J., Pers, T.H., Rietveld, C.A., Turley, P., Chen, G.-B., Emilsson, V., Meddens, S.F.W., et al.; LifeLines Cohort Study (2016). Genome-wide association study identifies 74 loci associated with educational attainment. Nature *533*, 539–542.
36. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
37. de Los Campos, G., Vazquez, A.I., Fernando, R., Klimentidis, Y.C., and Sorensen, D. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. *9*, e1003608.
38. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M.; and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
39. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al.; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; and RACI Consortium (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
40. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean,

G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

41. Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am. J. Hum. Genet. *99*, 139–153.
42. Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2016). Local genetic correlation gives insights into the shared genetic architecture of complex traits. bioRxiv. http://dx.doi.org/10.1101/092668.
43. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A.F., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J., et al.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. *44*, 369–375, S1–S3.
44. Welch, B.L. (1947). The generalisation of student's problems when several different population variances are involved. Biometrika *34*, 28–35.
45. Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. Nat. Genet. *48*, 709–717.
46. Do, R., Willer, C.J., Schmidt, E.M., Sengupta, S., Gao, C., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., et al. (2013). Common variants associated with plasma triglycerides and risk for coronary artery disease. Nat. Genet. *45*, 1345–1352.
47. Mi, H., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2013). Large-scale gene function analysis with the PANTHER classification system. Nat. Protoc. *8*, 1551–1566.
48. Won, H., de la Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F., Lu, D., et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature *538*, 523–527.
49. Kennedy, J.M., Fodil, N., Torre, S., Bongfen, S.E., Olivier, J.-F., Leung, V., Langlais, D., Meunier, C., Berghout, J., Langat, P., et al. (2014). CCDC88B is a novel regulator of maturation and effector functions of T cells during pathological inflammation. J. Exp. Med. *211*, 2519–2535.
50. Pavlides, J.M.W., Zhu, Z., Gratten, J., McRae, A.F., Wray, N.R., and Yang, J. (2016). Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. Genome Med. *8*, 84.
51. Hosokawa, Y., Maeda, Y., Takahashi, E.-i., Suzuki, M., and Seto, M. (1999). Human aiolos, an ikaros-related zinc finger DNA binding protein: cDNA cloning, tissue expression pattern, and chromosomal mapping. Genomics *61*, 326–329.
52. Quintana, F.J., Jin, H., Burns, E.J., Nadeau, M., Yeste, A., Kumar, D., Rangachari, M., Zhu, C., Xiao, S., Seavitt, J., et al. (2012). Aiolos promotes TH17 differentiation by directly silencing Il2 expression. Nat. Immunol. *13*, 770–777.
53. Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.
54. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., Robinson, M.R., Perry, J.R.B., Nolte, I.M., van Vliet-Ostaptchouk, J.V., et al.; LifeLines Cohort Study (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. *47*, 1114–1120.
55. Gutierrez-Arcelus, M., Ongen, H., Lappalainen, T., Montgomery, S.B., Buil, A., Yurovsky, A., Bryois, J., Padioleau, I., Romano, L., Planchon, A., et al. (2015). Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. PLoS Genet. *11*, e1004958.
56. Parsons, T.J., Power, C., Logan, S., and Summerbell, C.D. (1999). Childhood predictors of adult obesity: a systematic review. Int. J. Obes. Relat. Metab. Disord. *23* (*Suppl 8*), S1–S107.
57. Fall, T., Hägg, S., Mägi, R., Ploner, A., Fischer, K., Horikoshi, M., Sarin, A.-P., Thorleifsson, G., Ladenvall, C., Kals, M., et al.; European Network for Genetic and Genomic Epidemiology (ENGAGE) consortium (2013). The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. PLoS Med. *10*, e1001474.
58. Gusev, A., Mancuso, N., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Oh, E., McCaroll, S., Neale, B., Ophoff, R., et al. (2016). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. bioRxiv. http://dx.doi.org/10.1101/067355.
59. Pickrell, J. (2015). Fulfilling the promise of Mendelian randomization. bioRxiv. http://dx.doi.org/10.1101/018150.
60. Smith, G.D., and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int. J. Epidemiol. *32*, 1–22.
61. Wang, J., Gamazon, E.R., Pierce, B.L., Stranger, B.E., Im, H.K., Gibbons, R.D., Cox, N.J., Nicolae, D.L., and Chen, L.S. (2016). Imputing Gene Expression in Uncollected Tissues Within and Beyond GTEx. Am. J. Hum. Genet. *98*, 697–708.
62. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.
63. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. *10*, e1004722.

The American Journal of Human Genetics, Volume *100*

Supplemental Data

Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits

Nicholas Mancuso, Huwenbo Shi, Pagé Goddard, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc

Figure S1. Estimates of $\rho_{g,local}$ between gene expression and trait are unbiased in simulations. Starting from real genotype data, we simulated gene expression at independent loci. We then simulated complex trait as a linear function of predicted expression at these loci. We performed a GWAS using complex trait and subsequent TWAS at each gene (using GBLUP weights) which was used as input to estimate $\rho_{g,local}$. A) Results for 2,500 simulations where the causal SNPs driving gene expression were typed in the data. B) Results for 2,500 simulates where causal SNPs driving gene expression were untyped.

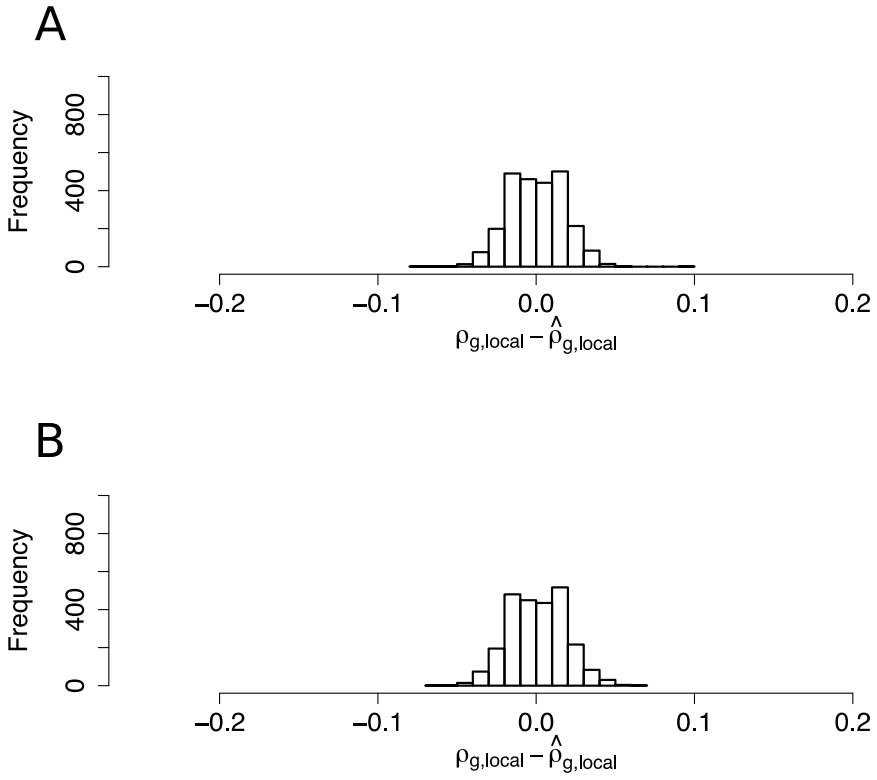


Figure S2. QQ-plot of null distribution in simulations measuring ρ_{GE}. The red line represents the identity line and the gray area is the 95% confidence interval of the null.

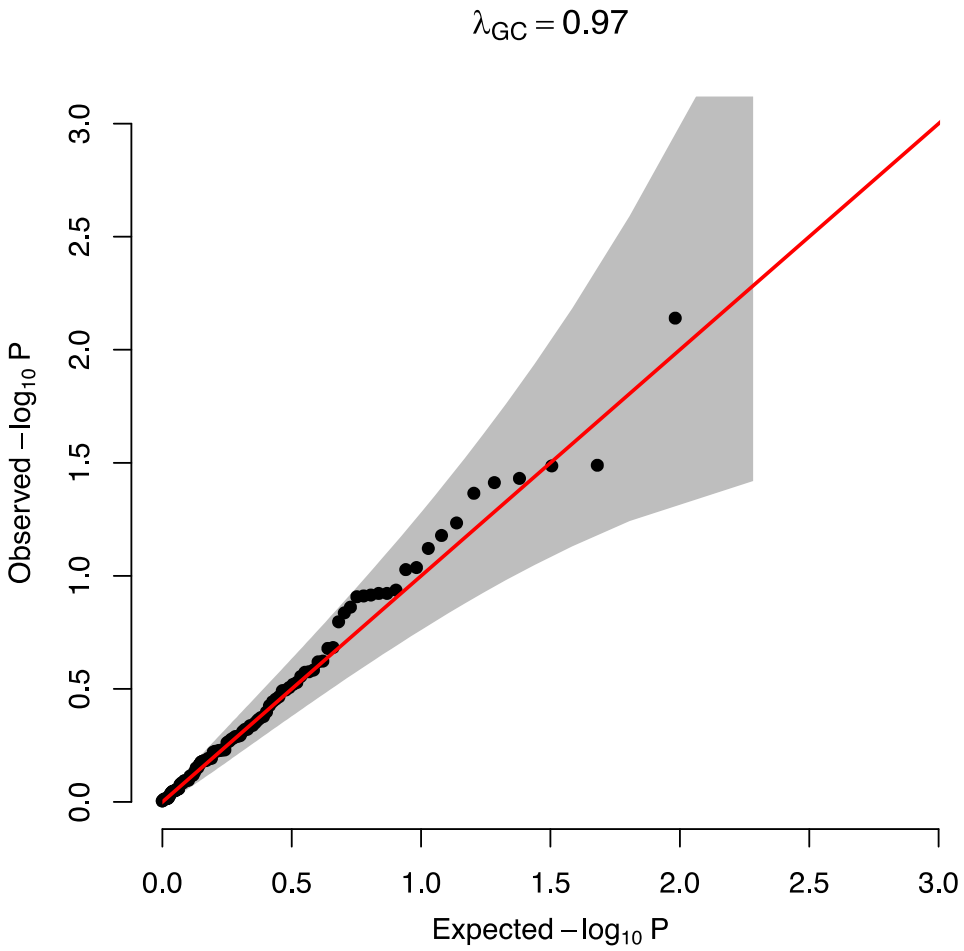


Figure S3. Estimating $h^2_{g,local}$ using all SNPs in a locus compared to the top eQTL. A) Estimates of $h^2_{g,local}$ using the described joint estimator versus the top eQTL. Results in the top row are obtained with causal SNPs typed in the data. Results in the bottom row have causal SNPs untyped/pruned from the genotype data. B) Joint estimation of $h^2_{g,local}$ results in better estimates for ρ_{GE}. The dotted line is the identity line. Each point represents the mean estimated ρ_{GE} over 100 simulations. Error bars capture the 95% confidence interval.

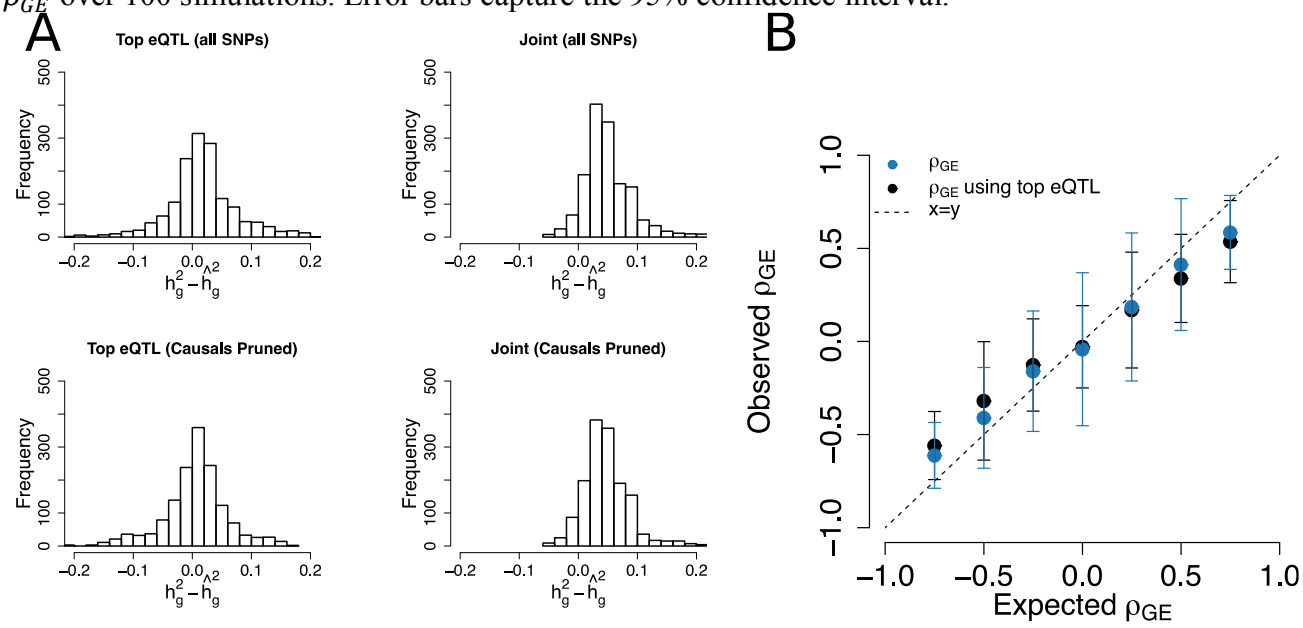


Figure S4. SNP association P-values at reported susceptibility genes not proximal to a genome-wide significant SNP tend to be suggestive. The red dotted line represents genome-wide significance.

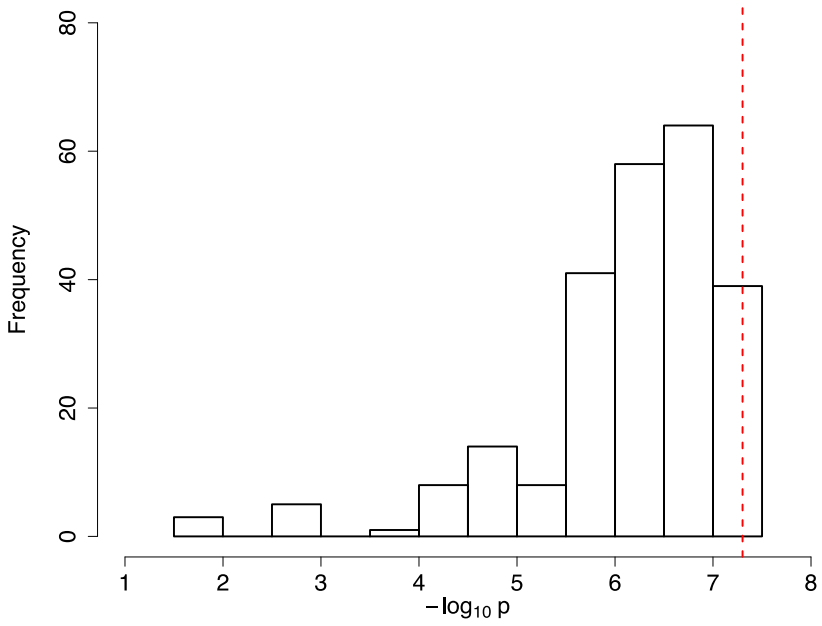


Figure S5. Distance from index SNP to gene transcription start site. Overlap was determined my selecting all SNPs within a 1 Mb flanking region around the gene's TSS. The red curve represents the estimated density assuming a Laplacian distribution of distances from the TSS.

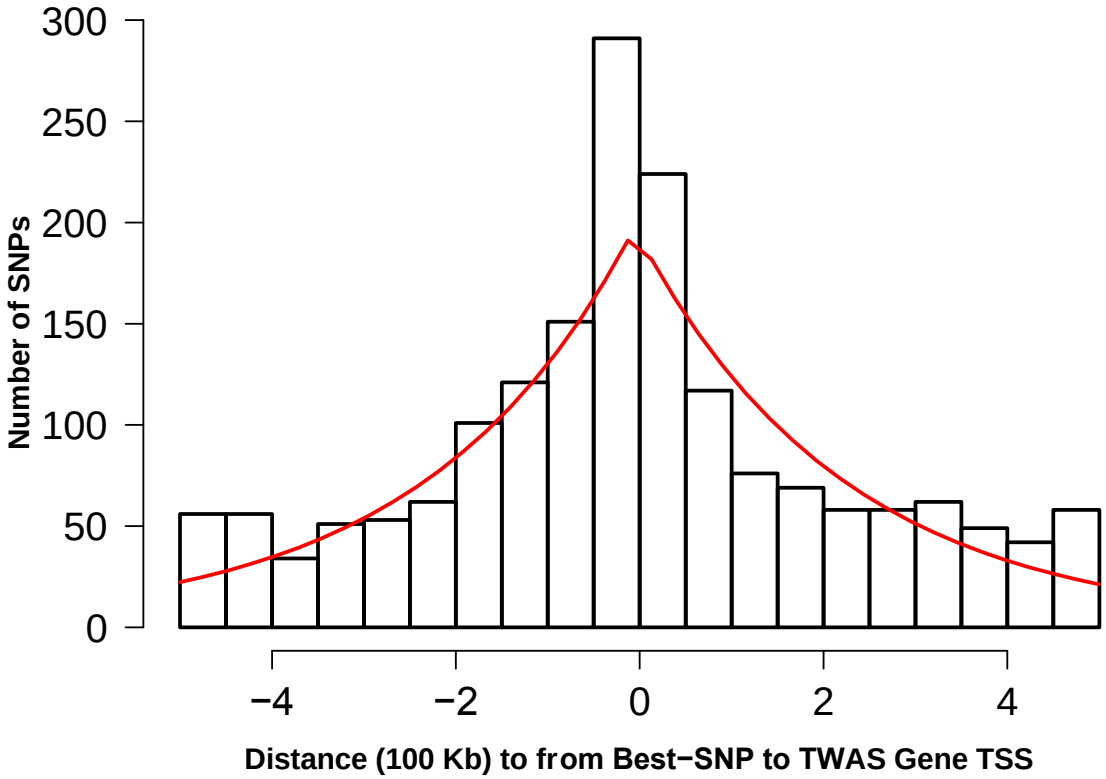


Figure S6. Distribution of association statistics for genes closest to index SNPs versus the top gene. The difference in means was significant under a Welch's t-test. Error-bars capture the lower and upper quartiles, with outliers represented as points.

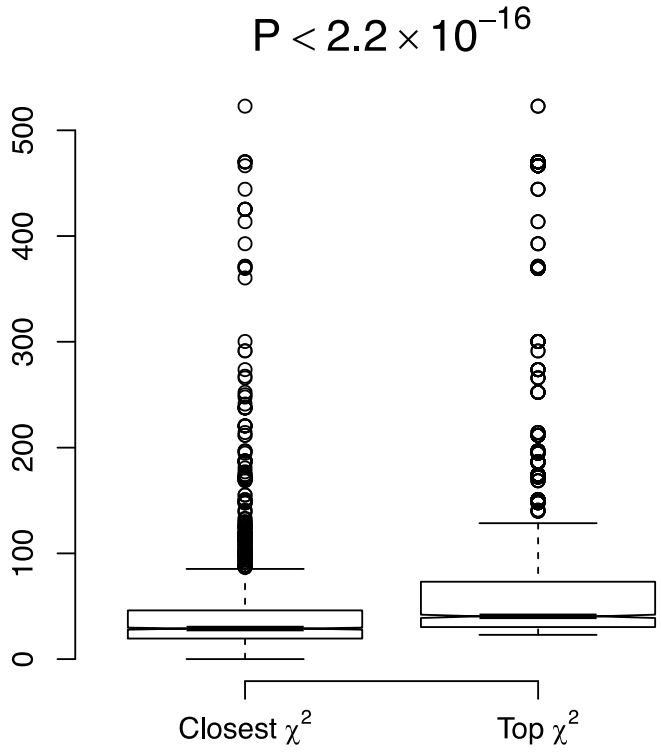


Figure S7. Number of genes associated with multiple traits.

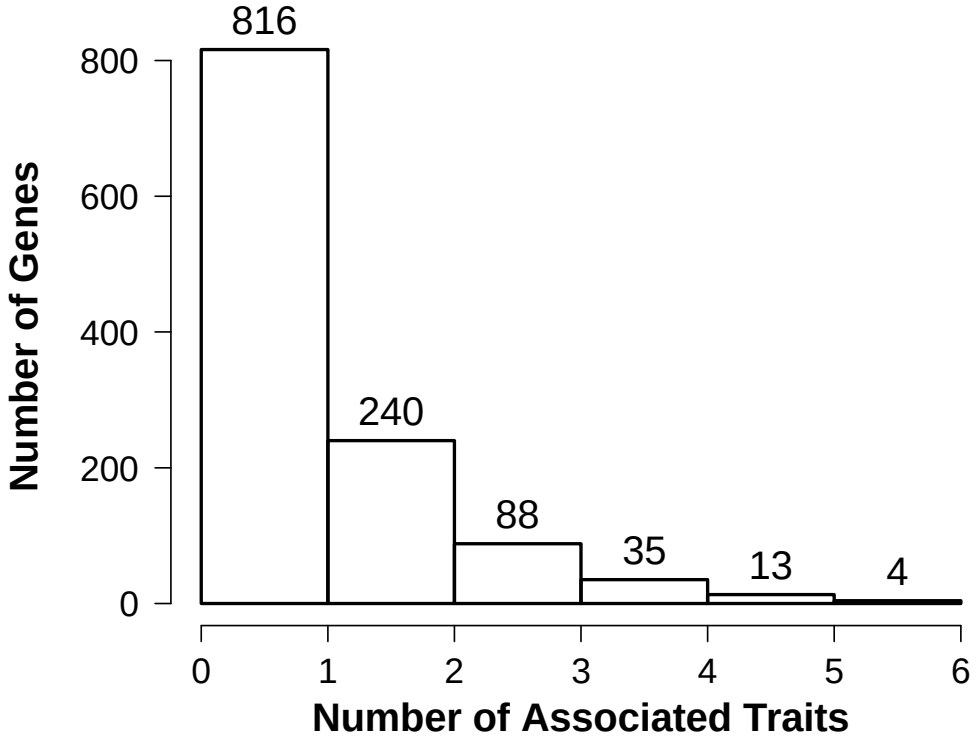


Figure S8. Comparing ρ_{GE} estimates computed using the top eQTL versus the entire locus. Estimates of ρ_{GE} in real data using the top eQTL are highly consistent with original estimates. The blue line represents the regression line fitted to the data.

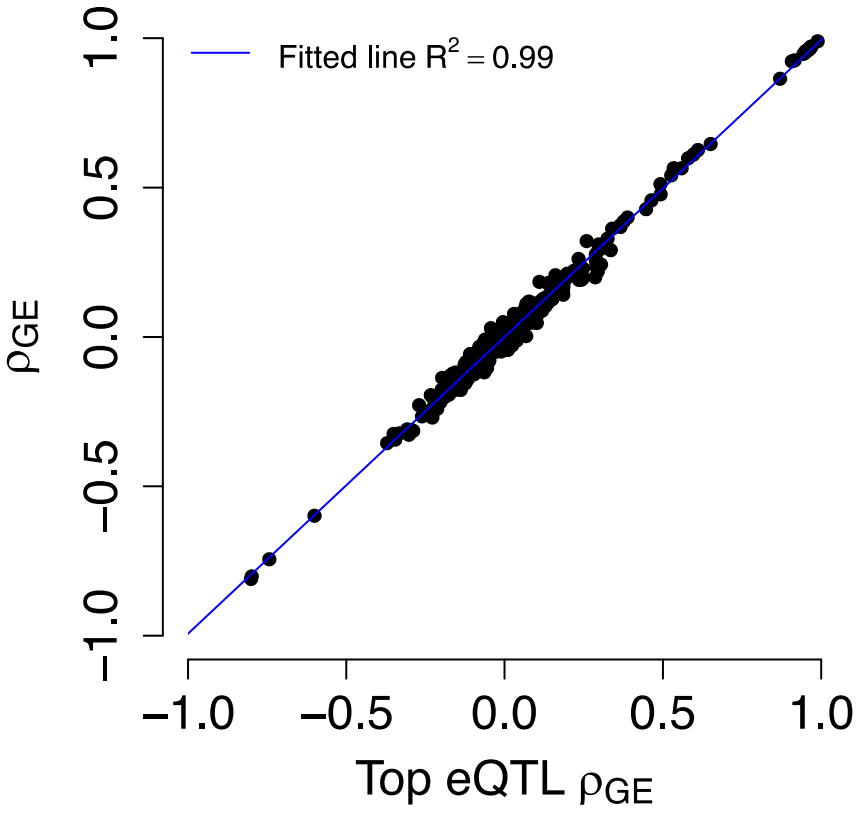


Figure S9. Molecular function analysis of TWAS genes for all traits. We only list functions that are significant ($P < 0.05$) after Bonferroni correction.

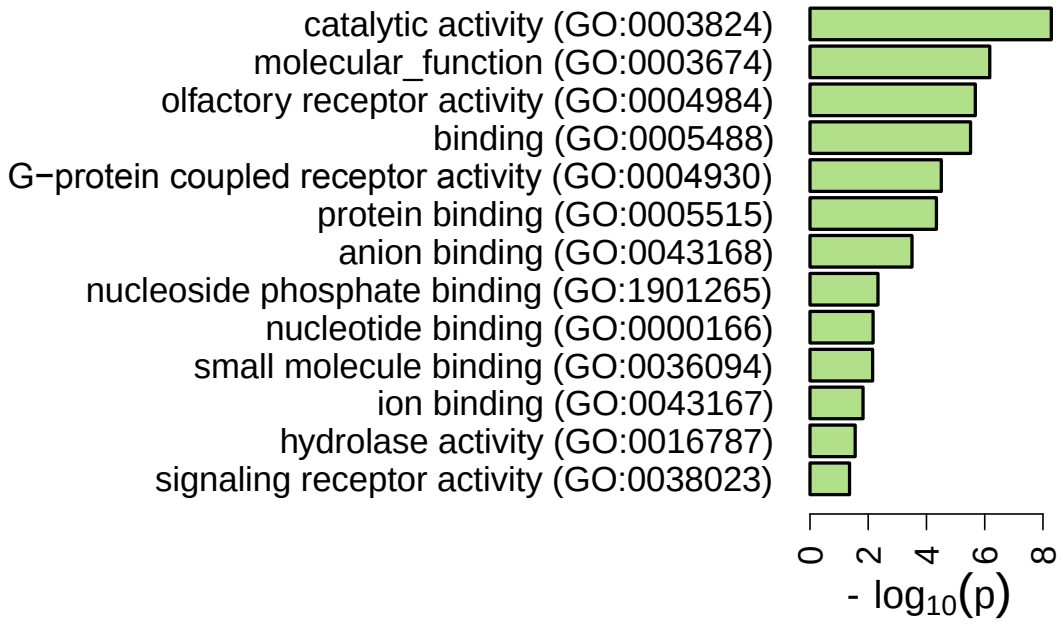
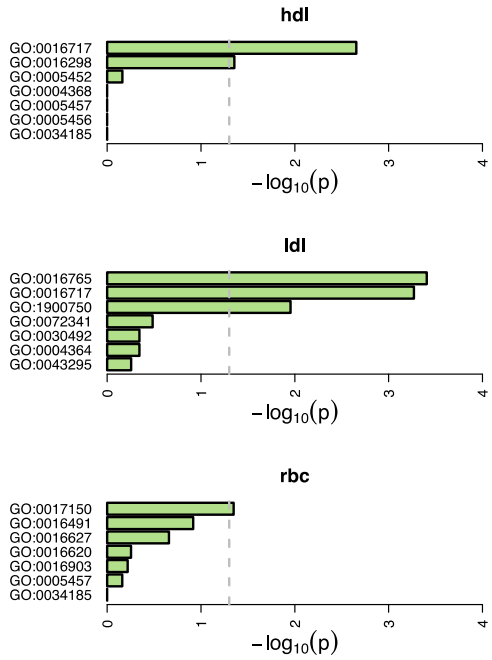
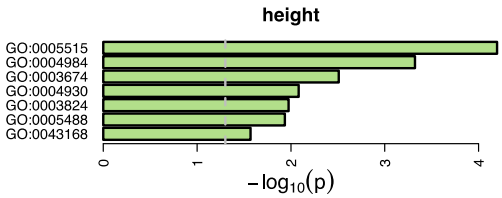
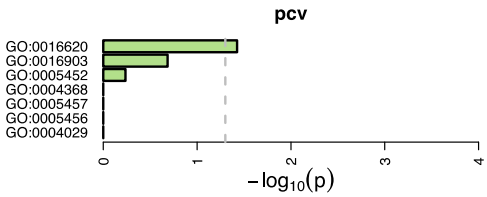


Figure S10. Molecular function analysis of TWAS genes. We only list functions that are significant ($P < 0.05$) after Bonferroni correction.







GO Terms	
GO:0003674	molecular_function
GO:0003824	catalytic activity
GO:0004029	aldehyde dehydrogenase (NAD) activity
GO:0004364	glutathione transferase activity
GO:0004930	G−protein coupled receptor activity
GO:0004984	olfactory receptor activity
GO:0005488	binding
GO:0005515	protein binding
GO:0016298	lipase activity
GO:0016491	oxidoreductase activity
GO:0016620	oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor
GO:0017150	tRNA dihydrouridine synthase activity
GO:0034185	apolipoprotein binding
GO:0043168	anion binding
GO:0043295	glutathione binding
GO:1900750	oligopeptide binding

Figure S11. Biological process analysis of TWAS genes for all traits. We only list functions that are significant ($P < 0.05$) after Bonferroni correction.

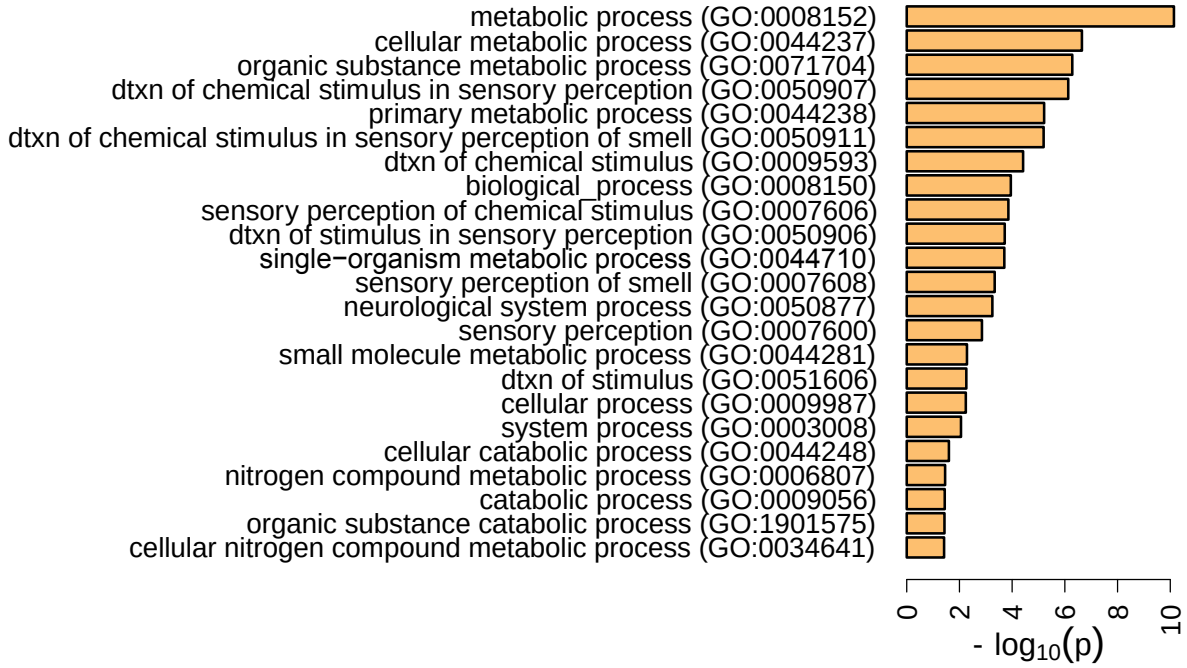
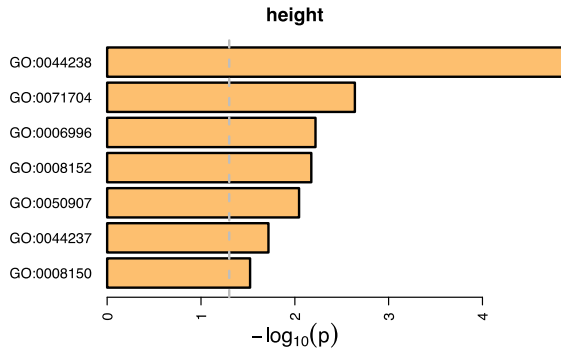
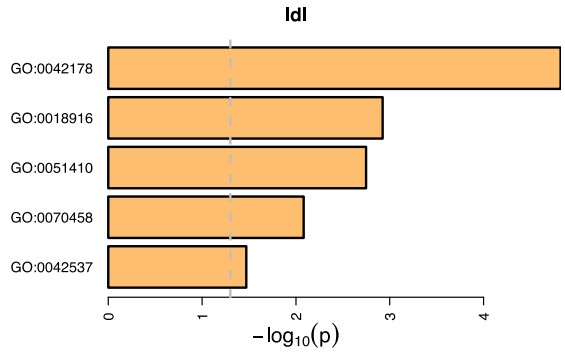
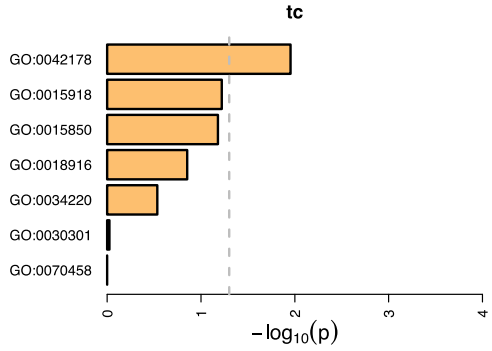


Figure S12. Biological process analysis of TWAS genes for height, LDL, and total cholesterol. We only list functions that are significant ($P < 0.05$) after Bonferroni correction.







GO Terms	
GO:0006996	organelle organization
GO:0008150	biological_process
GO:0008152	metabolic process
GO:0015850	organic hydroxy compound transport
GO:0015918	sterol transport
GO:0018916	nitrobenzene metabolic process
GO:0030301	cholesterol transport
GO:0042178	xenobiotic catabolic process
GO:0042537	benzene−containing compound metabolic process
GO:0044237	cellular metabolic process
GO:0044238	primary metabolic process
GO:0050907	detection of chemical stimulus involved in sensory perception
GO:0051410	detoxification of nitrogen compound
GO:0070458	cellular detoxification of nitrogen compound
GO:0071704	organic substance metabolic process

Trait	Short Name	Sample Size	Number of SNPs	Trait measurement	Trait Group
Age at Menarche	AM	132989	1821879	Quantitative	Metabolic
Body Mass Index	BMI	226814	1859666	Quantitative	Anthropometric
College	COL	126559	1792881	Dichotomous	Social
Crohn's Disease	CD	51874	4822932	Dichotomous	Immune-related
Education Years	EY	126559	1788888	Quantitative	Social
Fasting Glucose	FG	46186	1824182	Quantitative	Metabolic
Fasting Insulin	FI	46186	1822388	Quantitative	Metabolic
Femoral Neck BMD	FN	53236	4637340	Quantitative	Anthropometric
Forearm BMD	FA	53236	4725343	Quantitative	Anthropometric
Hemoglobin	HB	51496	1894024	Quantitative	Hematopoietic
HBA1C	HBA1C	46368	1870395	Quantitative	Hematopoietic
Height	HEIGHT	241286	1854761	Quantitative	Anthropometric
High Density Lipoprotein	HDL	95572	1805617	Quantitative	Metabolic
HOMA-B	HOMA-B	46186	1820938	Quantitative	Metabolic
HOMA-IR	HOMA-IR	46186	1821061	Quantitative	Metabolic
Inflammatory Bowel Disease	IBD	65643	4823603	Dichotomous	Immune-related
Low Density Lipoprotein	LDL	90811	1803637	Quantitative	Metabolic
Lumbar Spine	LS	53236	4636561	Quantitative	Anthropometric
MCH Concentration	MCHC	47157	1893281	Quantitative	Hematopoietic
Mean Cell Hemoglobin	MCH	43733	1892019	Quantitative	Hematopoietic
Mean Cell Volume	MCV	48689	1893769	Quantitative	Hematopoietic
Number of Platelets	PLT	66867	1954590	Quantitative	Hematopoietic
Packed Cell Volume	PCV	45125	1893412	Quantitative	Hematopoietic
Red Blood Cell Count	RBC	45500	1892553	Quantitative	Hematopoietic
Rheumatoid Arthritis	RA	58284	4265540	Dichotomous	Immune-related
Schizophrenia	SCZ	74626	4772186	Dichotomous	Neurological
Total Cholesterol	TC	95802	1805676	Quantitative	Metabolic
Triglycerides	TG	92007	1803908	Quantitative	Metabolic
Type 2 Diabetes	T2D	61857	1806359	Dichotomous	Metabolic
Ulcerative Colitis	UC	47746	4823578	Dichotomous	Immune-related

Table S1. Summary of the 30 GWAS data.

Expression Weights	Causal variants	h^2_{GE}	SE
GBLUP	Typed	0.30	0.01
GBLUP	Untyped	0.27	0.01
True	Typed	0.50	0.01

Table S3. Simulation results for h^2_{GE} estimates. We simulated 100 complex traits as a linear function of gene expression at 50 loci (see Material and Methods). We re-ran simulations with the causal variants for expression untyped in the genotyping data. We present the mean h^2_{GE}estimate along with the standard error across all simulation runs.

Gene	Chr	Tx Start	Tx End	Current Index SNP	SNP P	New Index SNP	BP	SNP P
SDCCAG8	chr1	243419306	243663393	rs12080886	5.73E-07	rs2992632	243503764	3.245E-11
ABCB9 *MPHOSPH9*	chr12	123405497	123451056	rs7980687	1.59E-06	rs10773002	123746961	7.742E-18
STK24	chr13	99102452	99174379	rs17574378	1.52E-07	rs9556958	99100046	1.208E-11
EIF3CL *SULT1A1* *RP11-1348G14.4* *TUFM* *MIR4721* *SH2B1* *NFATC2IP*	chr16	28390902	28415206	rs8049439	1.52E-07	rs8049439	28837515	6.992E-11

Table S7. TWAS predicted susceptibility loci for Education Years. Reported TWAS susceptibility loci for Education Years that did not overlap a genome-wide significant SNP within ±0.5Mb of transcription start-site in the Rietveld et al. Science 2013 study ($N = 126{,}559$) were proximal to genome-wide significant SNPs found in the much larger Okbay et al. Nature 2016 study ($N = 293{,}723$).