bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Estimating the proportion of disease heritability mediated by gene expression levels

Luke J. O'Connor[1-2], Alexander Gusev[3], Xuanyao Liu[2], Po-Ru Loh[2,4], Hilary K. Finucane[2,5], and Alkes L. Price[2,4,6]

[1]Program in Bioinformatics and Integrative Genomics, Harvard Graduate School of Arts and Sciences, Cambridge, MA
[2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA
[3]Dana Farber Cancer Institute, Boston, MA
[4]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA
[5]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA
[6]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA
Correspondence should be addressed to L.J.O (loconnor@g.harvard.edu) or A.L.P. (aprice@hsph.harvard.edu)

March 17, 2017

Abstract

Disease risk variants identified by GWAS are predominantly noncoding, suggesting that gene regulation plays an important role. eQTL studies in unaffected individuals are often used to link disease-associated variants with the genes they regulate, relying on the hypothesis that noncoding regulatory effects are mediated by steady-state expression levels. To test this hypothesis, we developed a method to estimate the proportion of disease heritability mediated by the cis-genetic component of assayed gene expression levels. The method, gene expression co-score regression (GECS regression), relies on the idea that, for a gene whose expression level affects a phenotype, SNPs with similar effects on the expression of that gene will have similar phenotypic effects. In order to distinguish directional effects mediated by gene expression from non-directional pleiotropic or tagging effects, GECS regression operates on pairs of cis SNPs in linkage equilibrium, regressing pairwise products of disease effect sizes on products of cis-eQTL effect sizes. We verified that GECS regression produces robust estimates of mediated effects in simulations. We applied the method to eQTL data in 44 tissues from the GTEx consortium (average N_{eQTL} = 158 samples) in conjunction with GWAS summary statistics for 30 diseases and complex traits (average $N_{\mathrm{GWAS}} = 88K$) with low pairwise genetic correlation, estimating the proportion of SNP-heritability mediated by the cis-genetic component of assayed gene expression in the union of the 44 tissues. The mean estimate was 0.21 (s.e. = 0.01) across 30 traits, with a significantly positive estimate ($p < 0.001$) for every trait. Thus, assayed gene expression in bulk tissues mediates a statistically significant but modest proportion of disease heritability, motivating the development of additional assays to capture regulatory effects and the use of our method to estimate how much disease heritability they mediate.

Introduction

In the past few years, large scale expression quantitative trait locus (eQTL) studies have identified genetic variants associated with the expression levels of thousands of genes [1–6], motivated by the observation that disease risk variants identified by genome-wide association studies (GWAS) are predominantly noncoding [7]. In addition, GWAS risk variants and SNP-heritability (h^2_g; [8, 9]) are enriched near epigenetic marks associated with regulatory function [10–15] and at eQTLs [16–20, 22]. Cis-genetic correlations have been

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

identified between complex traits and gene expression levels of specific genes [23–26], suggesting that eQTL studies can elucidate regulatory disease mechanisms.

However, the interpretation of observed enrichments and cis-genetic correlations between eQTLs and disease risk variants remains in doubt. Enrichment of GWAS signals near eQTLs could arise either from mediation of genetic effects by gene expression levels or from coincidental colocalization of eQTLs and GWAS variants, either due to pleiotropic effects of the same causal SNP or due to tagging effects arising from linkage disequilibrium (LD) between distinct causal SNPs. Indeed, most instances of colocalization between significant autoimmune disease loci and eQTLs are driven by tagging effects, suggesting that coincidental colocalization may be more prevalent than mediation [27]. Cis-genetic correlations can be primarily driven by a single large cis eQTL, so these correlations might also reflect coincidental colocalization [24]. The interpretation of enrichment and cis-genetic correlations depends on the extent that true mediation exists across the genome: if it is prevalent, then a mechanistic basis for these observations is plausible.

Here we develop a method, gene expression co-score regression (GECS regression), to estimate the proportion of disease heritability that is mediated by the cis-genetic component of assayed gene expression. In order to distinguish directional effects mediated by gene expression from non-directional pleiotropic or tagging effects, GECS regression operates on unlinked pairs of SNPs with shared cis genes. It is based on the intuition that if phenotypic effects are mediated by gene expression, then SNPs with similar effects on expression will have similar effects on the phenotype. Because GECS regression operates on GWAS summary statistics, it is broadly applicable to large publicly available data sets [28]. We applied our method to eQTL data in 44 tissues from the GTEx consortium [3] (see URLs) and GWAS summary statistics for 30 genetically distinct complex traits.

Results

Overview of methods

We aim to estimate the proportion of heritability mediated by the cis-genetic component of assayed gene expression levels. In particular, we model *underlying* (random-effect) cis-genetic correlations, which reflect true biological effects, instead of fixed-effect genetic correlations; the latter may be inflated by coincidental colocalization between causal effects (see Online Methods). Our definition of mediation does not precisely correspond to direct causality, as assayed expression may act as an imperfect proxy for truly causal expression levels in some other cell type or cellular context; however, it is unlikely to reflect *reverse* causality (see Online Methods and Supplementary Note). In order to estimate the proportion of heritability mediated by expression, we use an approach that is based on pairs of unlinked SNPs. Specifically, we regress pairwise products of phenotypic effects on products of cis-eQTL effect sizes. Other approaches are unable to recover this parameter, instead allowing contamination from non-mediated effects (see Supplementary Note). On the other hand, our method is not contaminated by non-mediated effects because products of non-mediated phenotypic effects are uncorrelated with products of eQTL effects for unlinked SNPs.

Under a random effects model, the expected value of the product between the marginal trait effect sizes a_i and a_j for a pair of unlinked SNPs (i, j) is:

$$E(a_i a_j | b) = \rho^2 \frac{h_Y^2}{h_{\tilde{Z}}^2} \sum_{\text{shared cis genes } g} b_{ig} b_{jg}. \quad (1)$$

Here ρ^2 is the proportion of heritability mediated by expression, h_Y^2 is the trait SNP-heritability, $h_{\tilde{Z}}^2$ is the mean cis-heritability of expression levels, and b_{ig} is the marginal effect of SNP i on gene g (see Online Methods and Supplementary Note). We refer to the sum over genes of products of cis-eQTL effect sizes as the "GE co-score" of SNPs i and j. Equation (1) does not hold for SNPs in LD; in the presence of positive LD, both $a_i a_j$ and $\sum b_{ig} b_{jg}$ will be positively inflated due to tagging of non-mediated effects. Equation (1) also relies on an independence assumption between the trait effect size and cis-genetic architecture of each gene (see Online Methods).

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Equation (1) suggests an inference method for ρ^2: we regress products of estimated trait effect sizes on estimated GE co-scores for pairs of unlinked SNPs. We refer to this method as GE co-score regression (GECS regression). We define unlinked SNPs using an LD threshold, retaining 75% of SNP pairs that have shared cis genes (with LD computed in the gene expression cohort), and we additionally condition the regression on LD for retained SNP pairs. We use eQTL data from multiple tissues; genes in each tissue are pooled, resulting in an estimate of the heritability mediated by the union of tissues. Because cis-genetic correlations are stronger than environmental correlations across tissues, this approach increases the effective cis-heritability, resulting in increased power (see Online Methods and Supplementary Note). We employ a weighted regression to further increase power. To avoid attenuation bias, we normalize the regression slope using a cross-validation estimate of the variance in true (rather than estimated) GE co-scores. The regression slope is multiplied by the ratio $h^2_{\tilde{Z}}/h^2_Y$ of expression heritability to trait heritability, estimated using LD score regression [34] (see URLs), to produce an estimate of ρ^2. We assess standard errors using a jackknife on 50 contiguous blocks of SNPs, analagous to previous work [34]. To assess statistical significance, we jackknife only the regression numerator, and apply a Z-test. Further details of the method are provided in Online Methods. We have released open source software implementing the method (see URLs).

Simulations

To assess the robustness of our methods, we performed simulations using a subset of real genotypes from the UK Biobank [44] (see URLs) ($N_{\text{GWAS}} = 10,000$; $M = 5,000$ SNPs) and simulated complex trait and gene expression phenotypes ($N_{\text{eQTL}} = 1000$ samples, disjoint from the GWAS samples; 600 genes; 50 tissues). We sampled effects of SNPs on expression, effects of expression on trait, and direct (non-mediated) effects of SNPs on trait from point-normal distributions. We included strong genetic and environmental correlations across tissues. To model coincidental colocalization between eQTLs and non-mediated GWAS variants, SNPs with a large effect on gene expression were more likely to have a nonzero direct effect on the trait (see Online Methods).

We ran GECS regression on the simulated data. When LD pruning was appropriately stringent, our estimates of ρ^2 were approximately unbiased (Figure 1-a). Without LD pruning, more significant inflation due to LD was observed (Figure 1-b). At larger values of ρ^2, modest downwards bias was observed at stringent LD thresholds, due to bias in gene expression heritability estimates obtained using LD score regression. This bias occurs because regression SNPs in the cis window are in LD with non-causal SNPs outside the cis window, which contribute to the LD score of the regression SNP but not to its χ^2 statistic; the amount of bias is small (see Supplementary Note).

A naive fixed-effect estimator of ρ^2 would be the sum of the squares, over genes, of the estimated fixed-effect cis-genetic correlation between that gene's normalized expression level and the trait. To estimate the fixed-effect cis-genetic correlation, we took the dot product between the effect size estimates for each trait and divided by the geometric mean of the true simulated heritability for expression and trait. We evaluated the fixed-effect estimator in simplified simulations with no LD and no correlations among genes (see Online Methods). We hypothesized that, even in the absence of mediation, fixed-effect genetic correlations would arise by chance due to coincidental colocalization between eQTLs and non-mediated effects. We estimated the fixed-effect genetic covariance between each gene and the trait using the covariance between the effect size estimates, summed the squared estimates, and divided by the trait heritability (see Online Methods). Estimates of ρ^2 using this method were close to 1, even when N_{eQTL} and N_{GWAS} were large (so sampling noise did not contribute to the bias). GECS regression estimates, on the other hand, were unbiased (Figure 1-c). We note that in the special case of no LD, our nave fixed-effect estimator is equivalent to cross-trait LD score regression with constrained intercept [35] and equivalent to ρ–HESS [29, 30], implying that those methods are similarly impacted by coincidental colocalization. We are not currently aware of a method that provides an unbiased estimate of the squared random-effect genetic correlation under a sparse genetic architecture.

A more sophisticated approach is to partition trait heritability using eQTLs as an annotation [20, 22], comparing the heritability explained by cis-eQTL SNPs with the heritability explained by other SNPs. We

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

called the most significant 25% of SNPs as cis-eQTLs (similar to the proportion of SNPs that were called as eQTLs in [20, 22]) and tested for enrichment of trait heritability explained by those SNPs using stratified LD score regression [14] (see Online Methods). Because our simulations include colocalization between effects on expression and direct (non-mediated) phenotypic effects, stratified LD score regression detects enrichment even in the absence of mediation (Figure 1-d)); this does not imply that stratified LD score regression or other heritability partitioning methods are flawed, but does imply that they cannot specifically distinguish mediation from colocalization effects. On the other hand, GECS regression had a well-calibrated false-positive rate at $\rho^2 = 0$, with increasing power to detect mediation as ρ^2 increased. We caution that power may depend heavily on the cis-genetic architecture of cis-expression and the genetic architecture of the trait.

We assessed the calibration of our jackknife standard errors and p values. We normalized the empirical errors (the signed differences between the true and estimated values) by dividing them by the jackknife standard errors; the standard deviation of the normalized errors was slightly smaller than 1 on average, indicating that estimated standard errors are modestly conservative (Figure S1; Online Methods). When we calculated p values using a Z test, we found that they were conservative under the null (2% positive rate at $\alpha = 0.05$, Figure 1-d).

We investigated the impact on our estimates of varying the cis window size. If the cis window is too small, there are two possible sources of bias. First, downwards bias is introduced because cis-heritable signal is missed. Second, genetic correlations between nearby genes, in conjunction with LD between SNPs that are in cis for each gene, lead to upwards bias due to over-counting; this occurs specifically in the presence of genetic correlations between adjacent genes when the cis window is too small (see Supplementary Note). We varied the GECS regression cis window size while fixing the true cis window size. When the GECS regression cis window size was smaller than the true cis window size, cis-heritable signal was missed, and the estimator was downwardly biased (Figure S2-a). However, estimates stabilized when the GECS regression cis window was as large as the true cis window size (±1Mb). In the presence of correlations between nearby genes (as opposed to only across tissues), estimates of heritability explained per unit of cis-heritability were upwardly biased when GECS regression cis window was smaller than ±1Mb (Figure S2-b).

Application to 30 complex traits

We applied GECS regression to summary statistics for 30 genetically distinct diseases and complex traits (Table S2; Online Methods). These included publicly available summary statistics from large consortia and summary statistics computed from UK Biobank data [44]. UK Biobank summary statistics were computed by applying BOLT-LMM [31] to up to 145,416 unrelated European-ancestry samples. We used gene expression data in 44 tissues from the GTEx consortium [3] (Table S3), computing cis eQTL summary statistics using FastQTL [45]. Expression summary statistics were computed from unrelated individuals consistent with GTEx consortium methods, and we additionally restricted to European samples (85% of individuals).

Estimates of the proportion of trait heritability mediated by the cis-genetic component of assayed gene expression are displayed in Figure 2 and Table 1. The mean estimate was 0.21 (s.e. = 0.01). Estimates were significantly positive ($p < 0.001$) for every trait. Except for multiple sclerosis, which had a large point estimates and a large standard error, estimates of ρ^2 ranged between 0.07 and 0.41. The variation in ρ^2 estimates was due at least in part to differences between phenotypes, rather than sampling error. The variance in ρ^2 estimates was 0.013, compared with a mean squared jackknife standard error of 0.009 (which is likely to be a slight overestimate; Figure S1). A conservative estimate of the variance in true values of ρ^2 across phenotypes, not accounting for positive correlations between noise terms, is the difference between these values, giving a standard deviation of $\sqrt{.004} = .06$ (31% of the mean). Noise terms are more likely to be positively correlated than negatively correlated, as a significant portion of the error comes from estimation of the regression denominator (the variance in GE co-scores), and this portion is highly positively correlated across phenotypes. This value 0.06 is statistically significant, again under the conservative assumption that sampling errors are uncorrelated (χ^2 test $p < 10^{-10}$; see Online Methods). On average, phenotypes with larger estimates also had larger standard errors. This phenomenon arises from the fact that variance in the regression denominator represents a significant portion of the sampling error, while the regression numerator (covariance) contains a large percentage of the variation across phenotypes; therefore, much of the observed

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

error represents a percentage of the point estimate. This phenomenon also explains why estimates can be highly significantly positive (using the numerator as the test statistic) despite relatively large standard errors (Figure 2 and Table 1). Standard errors were also larger for phenotypes with less significant heritability estimates, with a Spearman correlation of 0.40 ($p = 0.02$) between the significance of h_g^2 and the significance of ρ^2.

To assess the robustness of our estimates, we performed five secondary analyses. First, we varied our LD pruning threshold, finding that estimates were stable when we pruned stringently (Figure S3). We also investigated different LD pruning schemes. GECS regression prunes SNP pairs using a composite ranking of estimated LD and estimated *shared* LD (see Online Methods); however, concordant estimates were obtained when pruning using LD only or shared LD only, or when the regression was not conditioned on estimated LD (Table 2). Second, we investigated the impact of our regression weighting scheme, which down-weights SNP pairs whose SNPs have high LD to other regression SNPs; estimates using the weighted regression were concordant with estimates from the unweighted regression (Table 2). Estimates were also concordant if we removed the MAF lower bound ($2f(1-f) > .05$), or if we used a stringent lower bound ($2f(1-f) > .25$; Table 2). Third, to investigate the possibility that we were missing cis-genetic signal by restricting to a cis window of ±1Mb for each gene, we varied the cis window size. Estimates did not increase when we increased the window size to ±2Mb, indicating that little cis-genetic signal is missed (Figure S4). When we reduced the cis window size, we observed an increase in heritability explained and in heritability explained per unit of cis heritability, probably due to LD between cis windows for genetically correlated genes. This phenomenon, which is only expected to occur when significant heritability localizes to the flanking regions of genes, was also observed to a lesser extent in simulations (Figure S2; see Supplementary Note). Fourth, to determine whether the results were dependent on the sample size in the eQTL data set, we split the eQTL tissues into large N and small N subsamples (with 14 and 30 tissues, respectively). Estimates from these subsamples were concordant (Table 2). We also investigated the impact of randomly reducing the number of tissues (Figure S5). Estimates were relatively stable when ≥ 10 tissues were used; standard errors were particularly large when only 1 tissue was used. Fifth, we compared ρ^2 estimates for pairs of summary statistic data sets that were pruned from the primary analysis due to high genetic correlation between traits (Table S4). Most estimates were concordant. Nominally significant differences were observed for 4 out of 15 pairs of traits; only one of these differences ($p = .002$ for the two Years of Education datasets, which are overlapping) was significant after correcting for 15 pairs tested.

Finally, to investigate whether mediation of disease heritability by gene expression levels is tissue-specific, we estimated the proportion of heritability mediated by expression in 13 brain-related tissues, including 10 brain regions and tibial nerve, thyroid, and pituitary gland (Table S3). We assessed the heritability mediated by expression in both brain-related and non-brain-related tissues across nine brain-related traits: anorexia, autism, putamen volume, neuroticism, schizophrenia, years of education, ever smoked, age at menarche and BMI. Averaging across the nine traits, no difference was detected between heritability mediated by brain-related and non-brain-related tissues ($p = 0.76$; Table 3). Instead, both collections of tissues captured the full signal observed in their union ($p = 0.69$ and $p = 0.49$ respectively for the difference between heritability mediated by brain-related tissues and non-brain-related tissues compared with all tissues). These results are consistent with shared cis-eQTL effects and strong cis-genetic correlations across tissues [3, 36, 37].

Discussion

We have introduced GE co-score regression (GECS regression), a method to estimate the proportion of disease heritability mediated by gene expression levels; we have shown that the method can distinguish between mediation and coincidental colocalization effects, unlike other methods. We applied GECS regression to eQTL data in 44 tissues and GWAS summary statistics for 30 genetically distinct diseases and complex traits. We determined that the cis-genetic component of gene expression in the union of these tissues mediates a modest proportion of SNP-heritability, with a mean estimate of 0.21 (s.e. = 0.01). We observed no evidence of substantial tissue-specific cis effects, consistent with previous work [3,36,37]. On the one hand, these results provide validation that eQTL studies do capture regulatory disease mechanisms; on the other

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

hand, they motivate the development of new assays to systematically identify the transcriptional targets of regulatory GWAS variants.

Which other regulatory effects might influence disease? First, some disease heritability may be mediated through mechanisms that could be captured using gene expression data but that were not considered in this study, such as alternative splicing or trans regulatory effects. We anticipate that upcoming GTEx data will include alternative splicing phenotypes, and GECS regression could be applied to this data; on the other hand, larger sample sizes will be needed before GECS regression will produce well-powered estimates using trans eQTL data. Second, regulation might occur in specific cell types or in response to extracellular stimuli, such as glucose availability or immune activation [21]. If this is the case, then eQTL studies in specific cell types under experimental conditions mimicking physiologically relevant contexts would increase heritability explained. Even if cell type- and context-specific data is collected at modest sample sizes initially, GECS regression could be applied to this data together with GTEx data in order to quantify the additional contribution of cell type- and context-specific effects. Third, regulation might be dynamic and transient, occurring sporadically in individual cells in response to internal states that are not synchronized across cells in a tissue. This type of effect might be detected using single-cell eQTL studies in conjunction with sophisticated statistical methods that identify the subset of cells that are in the disease-relevant state, or using hQTL and meQTL data [39–43] if epigenetic changes accompanying transient regulatory events are more stable. Similar to context-specific effects, the proportion of additional heritability mediated by this data could be estimated using GECS regression.

Our approach has several limitations. First, our method requires the assumption that genes with multiple independent eQTLs do not have systematically larger or smaller effect sizes than genes with only one eQTL. Second, our method requires that intermediate phenotypes often have multiple independent QTLs; for molecular phenotypes that rarely have more than one QTL, our method will produce noisy estimates, although the estimates will remain unbiased. Third, our method does not identify which genes are responsible for the signal that is detected; existing methods [23–26] have been used to identify these genes based on fixed-effect genetic correlations, but our results motivate the development of methods testing for the presence of random-effect genetic correlations with individual genes to avoid signals due to coincidental colocalization. Fourth, our method requires large effective eQTL sample sizes, which we have achieved by combining eQTL data across tissues (Figure S5). Fifth, as noted above, power limitations prevent us from applying GECS regression to the trans-genetic component of expression levels, although our estimates will include trans effects that are mediated by cis effects on different genes [38]. Sixth, it would be interesting to quantify the extent to which trans regulation is mediated by cis effects, by using trans expression data instead of disease data, but our method currently requires disjoint samples in the eQTL and GWAS cohorts. Seventh, our method is limited to additive effects, and an entirely different approach would be needed in order to quantify non-additive effects.

In conclusion, we have determined that the cis-genetic component of gene expression mediates a modest proportion of disease and complex trait heritability. These results motivate the continued development of new molecular QTL studies, and the use of GECS regression to quantify their improvement over existing assays in capturing regulatory effects that contribute to human disease.

Acknowledgements

We are grateful to Soumya Raychaudhuri for helpful discussions and to Steven Gazal for assistance with GWAS summary statistics. This research was funded by NIH grants T32 GM074897, T32 HG002295, R01 MH107649, R01 MH101244 and U01 HG009379.

URLs

GE co-score regression software will publicly released prior to publication at `http://github.com/lukejoconnor/gecs`. Description of GTEx data and publicly available data, `http://gtexportal.org` [3]; LD score re-

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

gression software, `http://github.com/bulik/ldsc` [14, 34, 35]; description of UK Biobank data, `http://www.ukbiobank.ac.uk/` [44].

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

References

[1] Lappalainen, Tuuli, et al. "Transcriptome and genome sequencing uncovers functional variation in humans." Nature 501.7468 (2013): 506-511.

[2] Wright, Fred A., et al. "Heritability and genomics of gene expression in peripheral blood." Nature genetics 46.5 (2014): 430-437.

[3] Ardlie, Kristin G., et al. "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans." Science 348.6235 (2015): 648-660.

[4] Zhang, Xiaoling, et al. "Identification of common genetic variants controlling transcript isoform variation in human whole blood." Nature genetics 47.4 (2015): 345-352.

[5] Hore, Victoria, et al. "Tensor decomposition for multiple-tissue gene expression experiments." Nature Genetics 48.9 (2016): 1094-1100.

[6] Zhernakova, D. V., et al. "Identification of context-dependent expression quantitative trait loci in whole blood." Nature genetics 49.1 (2017): 139.

[7] Hindorff, Lucia A., et al. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proceedings of the National Academy of Sciences 106.23 (2009): 9362-9367.

[8] Yang, Jian et al. "Common SNPs explain a large proportion of the heritability for human height." Nature Genetics 42 (2010): 565-9 .

[9] Yang, Jian, et al. "Genome partitioning of genetic variation for complex traits using common SNPs." Nature genetics 43.6 (2011): 519-525.

[10] Maurano, Matthew T., et al. "Systematic localization of common disease-associated variation in regulatory DNA." Science 337.6099 (2012): 1190-1195.

[11] Trynka, Gosia, et al. "Chromatin marks identify critical cell types for fine mapping complex trait variants." Nature genetics 45.2 (2013): 124-130.

[12] Gusev, Alexander, et al. "Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases." The American Journal of Human Genetics 95.5 (2014): 535-552.

[13] Farh, Kyle Kai-How, et al. "Genetic and epigenetic fine mapping of causal autoimmune disease variants." Nature 518.7539 (2015): 337-343.

[14] Finucane, Hilary K., et al. "Partitioning heritability by functional annotation using genome-wide association summary statistics." Nature genetics 47.11 (2015): 1228-1235.

[15] Won, Hong-Hee, et al. "Disproportionate contributions of select genomic compartments and cell types to genetic risk for coronary artery disease." PLoS Genet 11.10 (2015): e1005622.

[16] Giambartolomei, Claudia, et al. "Bayesian test for colocalisation between pairs of genetic association studies using summary statistics." PLoS Genet 10.5 (2014): e1004383.

[17] Ding, Jun, et al. "Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals." The American Journal of Human Genetics 87.6 (2010): 779-789.

[18] Nicolae, Dan L., et al. "Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS." PLoS Genet 6.4 (2010): e1000888.

[19] Hu, Xinli, et al. "Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets." The American Journal of Human Genetics 89.4 (2011): 496-506.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

[20] Davis, Lea K., et al. "Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture." PLoS Genet 9.10 (2013): e1003864.

[21] Hu, Xinli, et al. "Regulation of gene expression in autoimmune disease loci and the genetic basis of proliferation in CD4+ effector memory T cells." PLoS Genet 10.6 (2014): e1004404.

[22] Torres, Jason M., et al. "Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait." The American Journal of Human Genetics 95.5 (2014): 521-534.

[23] Gamazon, Eric R., et al. "A gene-based association method for mapping traits using reference transcriptome data." Nature genetics 47.9 (2015): 1091-1098.

[24] Gusev, Alexander, et al. "Integrative approaches for large-scale transcriptome-wide association studies." Nature genetics 48 (2016): 245-252.

[25] Zhu, Zhihong, et al. "Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets." Nature genetics 48 (2016):481:487.

[26] Mancuso, Nicholas, et al. "Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits." The American Journal of Human Genetics 100.3 (2017): 473-487.

[27] Chun, Sung, et al. "Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types." Nature Genetics (2017).

[28] Pasaniuc, Bogdan, and Alkes L. Price. "Dissecting the genetics of complex traits using summary association statistics." Nature Reviews Genetics 18 (2017): 117-127.

[29] Shi, Huwenbo, Gleb Kichaev, and Bogdan Pasaniuc. "Contrasting the genetic architecture of 30 complex traits from summary association data." The American Journal of Human Genetics 99.1 (2016): 139-153.

[30] Shi, Huwenbo, et al. "Local genetic correlation gives insights into the shared genetic architecture of complex traits." bioRxiv (2016): 092668.

[31] Loh, Po-Ru, et al. "Efficient Bayesian mixed-model analysis increases association power in large cohorts." Nature genetics 47.3 (2015): 284-290.

[32] Haseman, J. K., and R. C. Elston. "The investigation of linkage between a quantitative trait and a marker locus." Behavior genetics 2.1 (1972): 3-19.

[33] Golan, David, Eric S. Lander, and Saharon Rosset. "Measuring missing heritability: inferring the contribution of common variants." Proceedings of the National Academy of Sciences 111.49 (2014): E5272-E5281.

[34] Bulik-Sullivan, Brendan K., et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies." Nature genetics 47.3 (2015): 291-295.

[35] Bulik-Sullivan, Brendan, et al. "An atlas of genetic correlations across human diseases and traits." Nature genetics 47 (2015): 1236-1241.

[36] Price, Alkes L., et al. "Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals." PLoS Genet 7.2 (2011): e1001317.

[37] Liu, Xuanyao, et al (in press)."Functional architectures of local and distal gene expression regulation in multiple human tissues" American Journal of Human Genetics (2017).

[38] Pierce, Brandon L., et al. "Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians." PLoS Genet 10.12 (2014): e1004818.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

[39] McVicker, Graham, et al. "Identification of genetic variants that affect histone modifications in human cells." Science 342.6159 (2013): 747-749.

[40] Grubert, Fabian, et al. "Genetic control of chromatin states in humans involves local and distal chromosomal interactions." Cell 162.5 (2015): 1051-1065.

[41] Waszak, Sebastian M., et al. "Population variation and genetic control of modular chromatin architecture in humans." Cell 162.5 (2015): 1039-1050.

[42] Allum, Fiona, et al. "Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants." Nature communications 6 (2015).

[43] Chen, Lu, et al. "Genetic drivers of epigenetic and transcriptional variation in human immune cells." Cell 167.5 (2016): 1398-1414.

[44] Sudlow, Cathie, et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age." PLoS medicine 12.3 (2015): e1001779.

[45] Ongen, Halit, et al. "Fast and efficient QTL mapper for thousands of molecular phenotypes." Bioinformatics (2015): btv722.

[46] Stegle, Oliver, et al. "Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses." Nature protocols 7.3 (2012): 500-507.

[47] Chang, Christopher C., et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets." Gigascience 4.1 (2015): 7.

[48] Yang, Jian, et al. "GCTA: a tool for genome-wide complex trait analysis." The American Journal of Human Genetics 88.1 (2011): 76-82.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Tables

Phenotype	ρ^2	std err	Z score	h_g^2	std err
Anorexia	0.2	0.05	8.3	0.2	0.03
Autism spectrum	0.35	0.14	11.5	0.39	0.08
Putamen volume	0.32	0.1	8.8	0.37	0.08
Neuroticism	0.12	0.02	11.6	0.1	0.01
Schizophrenia	0.08	0.02	11.2	0.5	0.03
Years of Education	0.08	0.01	8.9	0.13	0.01
Waist-hip ratio (BMI adjusted)	0.14	0.02	7.5	0.14	0.02
HDL	0.41	0.16	3.7	0.06	0.03
LDL	0.32	0.07	5.4	0.09	0.03
Triglycerides	0.13	0.04	9.6	0.15	0.04
Smoking status	0.09	0.01	11.9	0.14	0.01
Blood pressue (Systolic)	0.12	0.02	11.8	0.2	0.01
HbA1C	0.27	0.09	5.4	0.1	0.02
Type 2 Diabetes	0.24	0.08	7.3	0.09	0.02
Celiac Disease	0.35	0.16	6.9	0.3	0.1
Crohns Disease	0.19	0.04	8.8	0.51	0.1
Lupus	0.28	0.08	6.3	0.43	0.1
Multiple Sclerosis	0.53	0.33	5.8	0.1	0.05
Primary Biliary Cirrhosis	0.35	0.16	9	0.46	0.11
Rheumatoid Arthritis	0.31	0.11	8.6	0.15	0.04
Ulcerative Colitis	0.28	0.1	9.3	0.24	0.05
Asthma	0.16	0.03	8.7	0.07	0.01
Eczema	0.17	0.04	12.3	0.08	0.01
Age at menarche	0.1	0.01	6.8	0.3	0.02
Age at menopause	0.19	0.04	10.1	0.19	0.03
Lung FEV1/FVC ratio	0.11	0.02	11.9	0.25	0.02
Lung forced expiratory volume (FVC)	0.11	0.01	10.4	0.25	0.02
Bone mineral density (heel)	0.13	0.04	5.8	0.29	0.03
BMI	0.09	0.01	8.8	0.3	0.02
Height	0.1	0.02	9.5	0.53	0.04

Table 1: Estimates of heritability mediated by expression for 30 genetically distinct phenotypes. We report point estimates of ρ^2 and h_g^2 for each trait, together with standard errors and significance of ρ^2 estimates.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

	Mean estimated ρ^2	Standard error
Default parameters	.21	.01
Pruning on LD only	.21	.01
Pruning on shared LD only	.21	.02
No LD conditioning	.23	.03
No regression weighting	.22	.02
No MAF lower bound	.21	.01
Stringent MAF lower bound	.21	.01
Large N_{eQTL} tissues	.25	.01
Small N_{eQTL} tissues	.22	.01

Table 2: Estimated heritability mediated by expression for different modifications to GECS regression, on average across non-redundant phenotypes. Standard errors were estimated by jackknifing. Different MAF ranges were used for estimation of GECS regression slope, but the same SNP set (heterozygosity $\geq .05$) was retained for heritability estimation.

	Mean estimated ρ^2	Standard error
Brain-related tissues and cognitive-related traits	.17	.02
Non-brain-related tissues and cognitive-related traits	.18	.02
All tissues and cognitive-related traits	.17	.01

Table 3: Estimated heritability mediated by expression for nine brain-related traits: anorexia, autism, putamen volume, neuroticism, schizophrenia, years of education, ever smoked, age at menarche and BMI. There was no significant difference between any of the three sets of tissues.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Figures

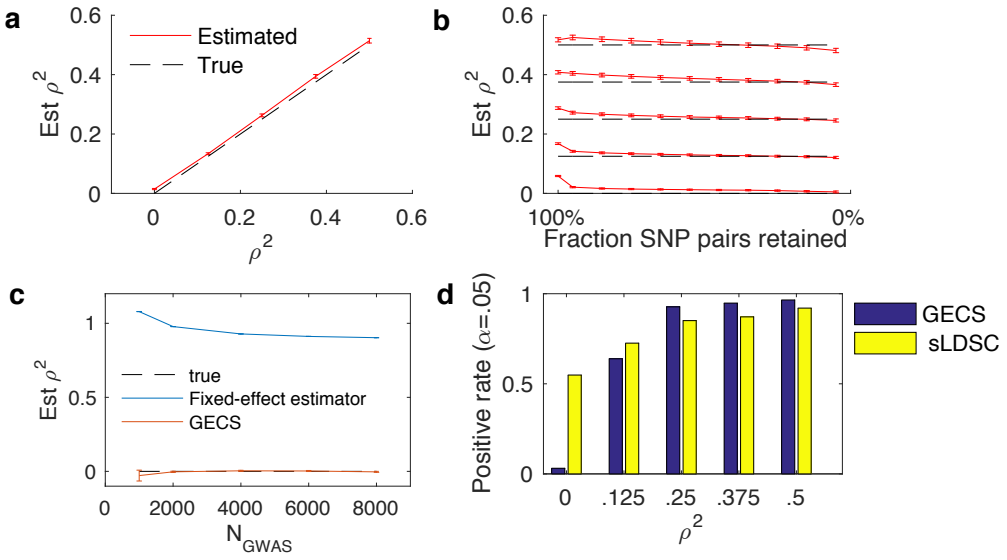


Figure 1: Performance of GE score regression in simulations using real genotypes and simulated gene expression and phenotypes. Simulations included strong colocalization between eQTLs and non-mediated disease SNPs. (a) Estimates were approximately unbiased when 75% of SNP pairs were retained. (b) Estimates were inflated when no LD pruning was applied, but LD pruning reduced the bias. Dashed horizontal lines indicate true values of ρ^2. (c) Estimates of ρ^2 using GECS regression and a naive fixed-effect estimator at different eQTL and GWAS sample sizes, in a simplified simulation setup including colocalization but no mediation. (d) Comparison of power to detect an effect between GECS regression and stratified LD score regression (sLDSC), using eQTLs as an annotation. Error bars indicate standard error based on 288 simulations (a,b,d) or 40,000 simulations (c). Numerical results are reported in Table S1.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

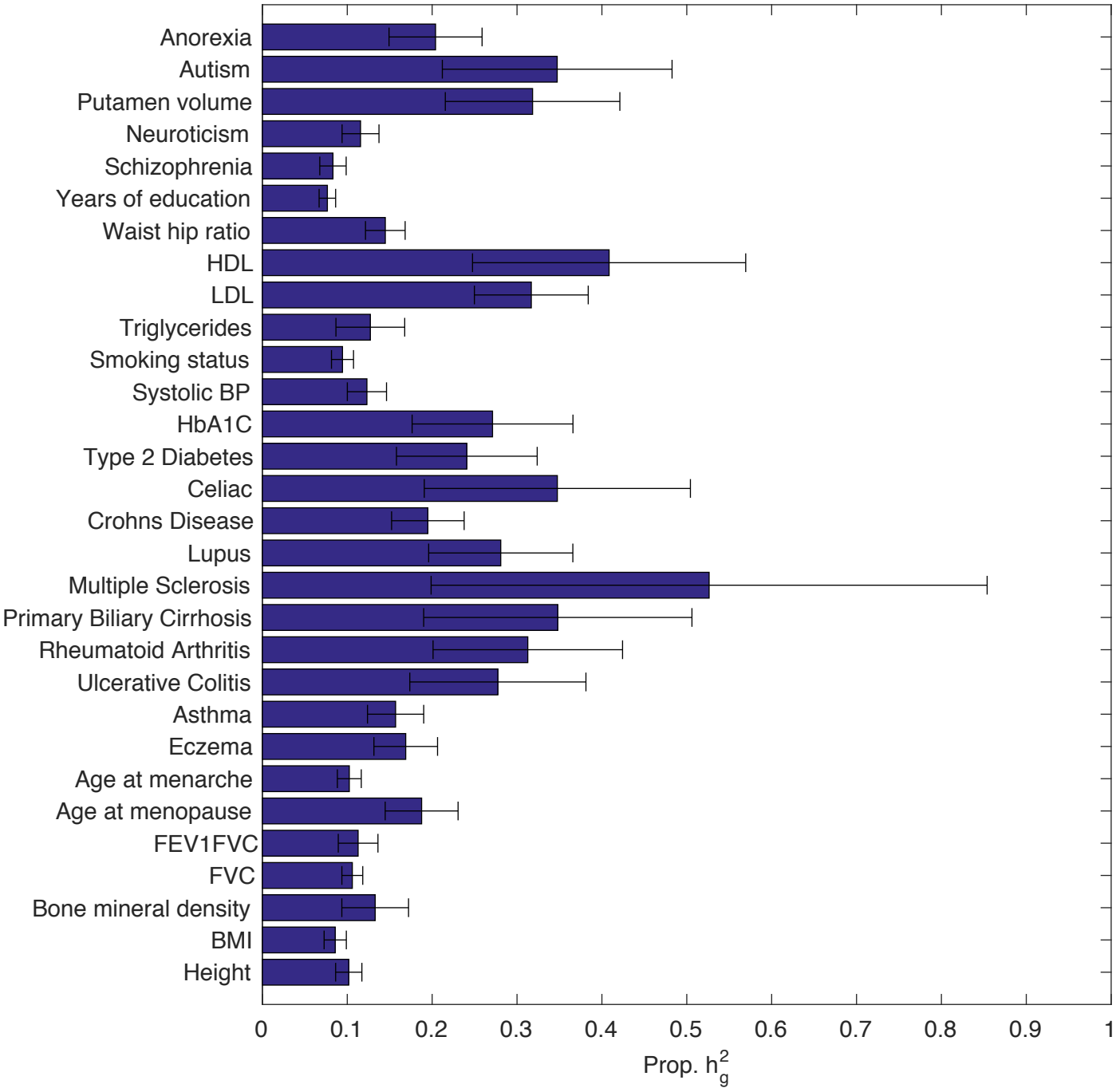


Figure 2: Estimates of the proportion of trait heritability mediated by the cis-genetic component of assayed gene expression for 30 diseases and complex traits. Related traits are grouped, and order is alphabetical within groups. Error bars represent jackknife standard errors. Numerical results are reported in Table 1.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Online Methods

Definition of heritability mediated by expression

We model a quantitative trait Y:

$$Y = X\gamma + \tilde{Z}\alpha + \epsilon, \quad \tilde{Z} = X\beta_{\tilde{Z}} \tag{2}$$

where X is the random genotype vector, γ is the random vector of direct SNP effects on the trait, $\tilde{Z}$ is the cis-genetic component of gene expression, α is the random vector of gene effects on the trait, ϵ is the environmental component of Y, and $\beta_{\tilde{Z}}$ is the random $M \times G$ matrix of cis-eQTL effects. Effects of the trans-genetic component of expression are incorporated into $X\gamma$, and effects of the environmental component of expression are incorporated into ϵ. We do not assume that SNP and gene effect sizes are drawn from a specific family of distributions. The proportion of heritability mediated by expression is defined as

$$\rho^2 := \frac{1}{E_{\alpha,\beta,\gamma}(h_Y^2)} E_{\alpha,\beta}(Var(\tilde{Z}\alpha|\alpha,\beta)), \tag{3}$$

where $h_Y^2 = Var(X\gamma + \tilde{Z}\alpha|\alpha,\beta,\gamma)$ is the SNP-heritability of Y.

We use the term *mediation* because random-effect cis-genetic correlations require mechanistic rather than stochastic explanations. In particular, the random-effects parameter ρ^2 differs from the expected maximum prediction accuracy of the genetic component of Y from $\tilde{Z}$ *fixing* γ:

$$\rho^2_{fixed} := \frac{1}{h_Y^2} \max_{a \in \mathbb{R}^G} corr(Y, \tilde{Z}a|\beta,\alpha,\gamma)^2. \tag{4}$$

The arg max above is not equal to α, since coincidental genetic correlations, which arise after fixing γ due to the finite number of SNPs, would be included. Thus, $E(\rho^2_{fixed}) \geq \rho^2$ (with equality when $\rho^2 = 1$). In contrast, SNP-heritability can be defined as $h_Y^2 = \max_{b \in \mathbb{R}^M} corr(Y, Xb)^2$ because the arg max equals the true causal effect size vector.

Mediation is not the same as causality; for instance, expression of a gene in one tissue might act as a proxy for expression of that gene in different, truly causal cell type and context. However, cis-genetic correlations are unlikely to reflect *reverse* causality (effects of trait on expression), as we can obtain an upper bound on the contribution of reverse causality to ρ^2 for a polygenic trait (see Supplementary Note).

Estimation using GECS regression

In order to estimate ρ^2, we exploit a relation between the covariance between SNP effect sizes on expression and the covariance between SNP effect sizes on trait. This approach is related to Haseman-Elston Regression [32,33] (see Supplementary Note). If a SNP has a causal effect β_i on the expression level of a gene, and that gene has an effect α on the trait, then the effect of the SNP on the trait is:

$$\delta_i = \beta_i\alpha + \gamma_i, \tag{5}$$

where γ_i is the direct (non-mediated) effect. Now, given the eQTL effects of two SNPs, we have:

$$\begin{aligned} E(\delta_i\delta_j|\beta_i,\beta_j) =& E(\beta_i\alpha + \gamma_i|\beta_i,\beta_j)E(\beta_j\alpha + \gamma_j|\beta_i,\beta_j) \\ =& \beta_i\beta_j E(\alpha^2), \end{aligned} \tag{6}$$

where we have assumed that α is independent of β_i, β_j; this assumption is discussed below. Under the independence assumption, $E(\alpha^2)$ is proportional to ρ^2 (see Supplementary Note). Replacing causal effects with marginal effects and allowing for multiple genes, we obtain (for SNPs not in LD):

$$E(a_i a_j|b) = E(\alpha^2) \sum_{\text{shard cis genes } g} b_{ig} b_{jg}, \tag{7}$$

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

where a_i is the marginal effect of SNP i on Y (including both mediated and non-mediated effects), and b_{ig} is the marginal effect of SNP i on the expression level of gene g. By relating the average gene effect size, $E(\alpha^2)$, with the total amount of mediation, ρ^2, we recover equation (1); for a full derivation, see the Supplementary Note.

Motivated by equation (1), we estimate ρ^2 by regressing products of trait effect sizes on sums over genes of the products of eQTL effect sizes. The GECS regression estimate is equal to the regression coefficient times the mean expression heritability, divided by the trait heritability.

Key Assumptions. GECS regression requires certain assumptions that, if they were violated, would lead to a biased estimate of ρ^2.

- **No LD or shared LD between SNP pairs** GE co-score regression is biased in the presence of LD or shared LD between SNP pairs. LD between SNP pairs induces a correlation between the noise terms in the effect size estimates for these SNPs (both the effects on expression and the effects on disease). This phenomenon leads to inflation in both the GE coscore regression numerator and the GE coscore regression denominator, and the magnitude of this effect is sample size dependent. Therefore, when the eQTL cohort is much smaller than the GWAS cohort, inadequate LD pruning can result in *downward* rather than upward bias in the estimate of ρ^2, if the GE coscore regression denominator is more inflated than the numerator. This accounts for the increase in estimates of ρ^2 when comparing 100% of SNPs retained in Figure 1-b with 95% of SNPs retained at a true ρ^2 value of 0.5.

  Shared LD refers to LD with a third SNP; one measure of shared LD is $\sum_k (r_{ik} r_{jk})^2$. In principle, SNPs with no LD could nonetheless have shared LD. In the presence of shared LD between i, j and k, the marginal effects of SNPs i and j contain a contribution from the causal effect of SNP k, in proportion to r_{ik} and r_{jk} respectively. This phenomenon occurs both for effects on trait and for effects on expression, regardless of whether the effects on trait are mediated by the effects on expression. Therefore, equation (1) requires that we exclude SNP pairs with shared LD. In practice, it is sufficient that SNPs have only a small amount of shared LD.

- **Independence between cis-genetic architecture and effects of expression on trait** In the derivation of equation (1) (see Supplementary Note), we assume that effects of gene expression on trait are drawn independently of eQTL effects: that is, genes with a certain type of genetic architecture do not have larger or smaller phenotypic effects on average. This assumption is needed because GECS regression implicitly weights some genes in the regression more heavily than others (that is, some genes contribute more to the variance in GE co-scores). Ideally, each gene would be weighted in proportion to its heritability, and indeed, if every eQTL for a given gene is doubled, then that gene will be weighted twice as heavily. However, weighting also depends on the number of independent eQTLs for a gene. Genes with only one large eQTL, due to either one causal variant or several causal variants in LD with each other, receive little weight in the regression due to LD pruning. Genes with multiple independent eQTLs are weighted more heavily, and the regression implicitly extrapolates from these genes to single-eQTL genes; thus, the estimator is biased if implicitly up-weighted genes have larger or smaller effect sizes on average than down-weighted genes.

Additional details

- **GE coscores.** Estimated GE coscores $\hat{\mathcal{G}}$ were computed for each subsample and from the whole sample: $\hat{\mathcal{G}}_{ij} = \sum_g \hat{b}_{ig} \hat{b}_{jg}$.

- **LD estimates from expression data.** In-sample LD was also measured for all regression SNPs sharing a cis gene, in each subsample, using PLINK [47]. For pruning, the shared LD between SNPs i, j was estimated as $\sum_k max(r_{ik}^{(1)2}, r_{ik}^{(2)2}) max(r_{jk}^{(1)2}, r_{jk}^{(2)2})$, where $r_{ij}^{(p)}$ is the LD between SNPs i and j in subsample p. The LD between SNPs i, j was estimated as $\hat{r}_{ij} = (r_{ij}^{(1)2} + r_{ij}^{(2)2})/2$. For a discussion of the choice of estimator, see the Supplementary Note.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

- **Conditioning on LD.** To remove effects from subtle LD below the threshold, we residualized estimate GE co-scores on estimated LD: $\hat{\mathcal{G}}'_{ij} = \hat{\mathcal{G}}_{ij} - c\hat{r}_{ij}$.

- **Regression weights.** For each regression SNP, we computed the regression-SNP LD score, $\ell_i = \sum_k r^2_{ik}$, using in-sample LD values. For each SNP pair i, j, the regression weight was $w_{ij} = 1/(\ell_i \ell_j)$.

- **LD pruning.** In order to ascertain SNP pairs having no shared LD and also no LD, SNP pairs having at least one shared cis gene were sorted by estimated shared LD and sorted by LD. The maximum of the two indices was then used to re-sort the SNP pairs, and the 25% of SNP pairs with highest combined LD were discarded (set $w_{ij} = 0$).

- **Regression numerator.** The regression numerator was:

$$\sum_{(i,j)} w_{ij} \hat{\mathcal{G}}'_{ij} \hat{\alpha}_i \hat{\alpha}_j,$$

  where $\hat{\alpha}_i$ is the estimated marginal effect size of SNP i on the trait and $\hat{\mathcal{G}}_{ij}$ is the estimated GE coscore from the whole expression sample.

- **Regression denominator.** To eliminate attenuation bias, the regression denominator was not $\sum w_{ij} \hat{\mathcal{G}}^2_{ij}$ but rather

$$\sum_{(i,j)} w_{ij} \hat{\mathcal{G}}^{(1)'}_{ij} \hat{\mathcal{G}}^{(2)'}_{ij},$$

  where $\hat{\mathcal{G}}^{(p)}_{ij}$ is the estimated GE coscore from subsample p.

- **Disease heritability estimation.** To estimate disease heritability and average cis heritability of gene expression levels, we use LD score regression [14, 34].

- **Cis heritability estimation.** To estimate the average cis-heritability of gene expression levels, we apply cross-trait LD score regression [35] with no intercept to the subsampled summary statistics (treating expression in the two cohorts as separate traits); this procedure is similar to running cross-trait LD score regression with constrained intercept [35], except that it is robust to inflation or deflation in the LD score intercept arising from finite-sample effects. We use LD scores computed between cis regression SNPs and all reference SNPs within a 1cM window. In principle, it is preferable to compute LD scores between cis regression SNPs and cis reference SNPs only (see [37]), but in practice there is little difference (see Supplementary Note).

- **Block jackknife.** We estimated the standard error on the estimate of ρ^2 by jackknifing on 50 blocks of contiguous SNPs, similar to LD score regression [34]. The same jackknife blocks were used for all traits. To estimate the standard error on the mean value of ρ^2 across phenotypes, we jackknifed the mean. To assess significance of the estimates, we estimated the standard error of only the regression numerator (i.e., the covariance between GE coscores and products of trait Z scores), and applied a Z test to the regression numerator divided by its standard error.

- **Trait selection.** From the full set of 42 GWAS data sets (Table S2), pairs of data sets with estimated genetic correlation greater than 0.5 (estimated using cross-trait LD score regression [35]) were greedily pruned, beginning with the data sets having the highest genetic correlation. In each case, we retained the data set with higher power as measured by the significance of $\hat{h}^2_g$ using LD score regression [34].

Simulations

Primary simulation For the primary simulation experiments, segments of chromosome 1 spanning about 1% of the genome were used, with cis windows of 601 SNPs (~ 2Mb), $M = 5000$ SNPs, and $G = 60$ partially heritable genes in $T = 50$ tissues (average $h^2_{cis} = .25$, $h^2_{trans} = .25$). Sample size was $N_{\mathrm{eQTL}} = 1000$ for the

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

expression cohort, $N_{\text{GWAS}} = 10,000$ for the GWAS cohort, and $N_{\text{reference}} = 500$ for the reference panel cohort (used to compute LD scores); cohorts were disjoint. Entries of the $M \times GT$ eQTL effect size matrix were drawn initially from a point-normal distribution with on average 10 causal SNPs per cis locus and the same proportion of causal SNPs in trans. Effect sizes for cis SNPs were larger than effect sizes for trans SNPs, so that on expectation the sum of squared effect sizes within the cis region was equal to the sum of squared effect sizes in the trans region for each gene; as a result, on average half of gene expression heritability was in cis. To model correlations across tissues, the eQTL matrix was multiplied by a $GT \times GT$ matrix A, where $A_{ii} = 1$, $A_{ij} = 0.8$ if i, j correspond to the same gene in different tissues, and $A_{ij} = 0$ otherwise; afterwards, there was correlation 0.8 between expression levels of the same gene in different tissues. We drew a heritability value for each gene-tissue pair from a $Beta(3, 3)$ distribution (which has mean 0.5). Then, the columns of the eQTL matrix were scaled so that the variance of the genetic component of each gene-tissue pair corresponded to these heritability values. Gaussian noise was added with covariance matrix B, where $B_{ii} = 1 - h_i^2$, $B_{ij} = 0.2\sqrt{(1-h_i^2)(1-h_j^2)}$ if i and j correspond to the same gene in different tissues, and $B_{ij} = 0$ otherwise. Effect sizes of expression on trait were drawn from a point-Normal distribution (20% of genes causal), scaled so that the variance of the product of these effect sizes with the gene expression matrix corresponded to the simulated value of ρ^2 (3). Effect sizes of SNPs directly on the trait were also drawn from a point-normal distribution, where the probability of SNP i being causal depended on its maximum eQTL effect size:

$$p_i = min\Big(1, p_0 \frac{max_g|\beta_{ig}|}{\frac{1}{N}\sum_i max_g|\beta_{ig}|}\Big), \tag{8}$$

where we used $p_0 = 0.2$. This results in a large amount of colocalization.

Secondary simulations For simulations with different cis window sizes, we also modeled genetic correlations between adjacent genes (with overlapping but distinct cis windows) instead of across tissues. For simulations that included a comparison to the fixed-effect estimator, we used simulated, normally distributed genotypes with no LD. There were 500 SNPs and 100 genes, with disjoint cis windows of 5 SNPs per gene, and 100% of gene expression variance was determined by these 5 SNPs. Effects of expression on trait and of SNPs directly on trait were modeled similarly to the main simulations.

Fixed-effect estimator Our fixed-effect estimator of ρ^2 was the sum over genes of the squared estimated genetic covariance between the trait and the gene, divided by the total heritability; the genetic covariance was estimated as the sum over cis SNPs of the products of the estimated effect sizes on each trait. Because cis-heritability of gene expression was 1, we did not normalize by estimated expression heritability.

Partitioned LD score regression We ran partitioned LD score regression [14] on the simulated data, using LD scores derived from 500 individuals (not overlapping with the GWAS or eQTL cohorts). SNPs were ranked by their maximum χ^2 statistic over cis genes, and the top 25% of SNPs were called as eQTLs. Then, we regressed trait χ^2 statistics on the vector of LD scores as well as the binary vector of eQTL calls, with intercept fixed at 1. We estimated the standard error of these estimates using 20 jackknife blocks, and obtained p values using a Z test. This approach is similar, in practice, to the approach of computing LD scores to called eQTLs; however, since eQTLs are called based on marginal, rather than causal, effect sizes, the latter approach assigns overly large scores to regions with more LD, since SNPs in these regions are more likely to be called as eQTLs (however, both of these approaches produce similar results).

Standard error assessment In order to assess the jackknife estimate of the standard errors, we computed point estimates of $\hat{\rho^2}$ of ρ^2 as well as jackknife estimates $\hat{\sigma}$ of the standard deviation of the estimate. We computed normalized errors, $(\hat{\rho^2} - \rho^2)/\hat{\sigma}$, and reported the root mean squared normalized error. This metric enables us to determine whether error bars are expected to give accurate confidence intervals; if instead we computed the average estimated standard error with the empirical standard error, it is possible the estimated standard error would be unbiased, and yet confidence intervals derived from estimated standard errors would be badly miscalibrated.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

GTEx data set

For every gene and every tissue, effect size estimates $\hat{b}$ were produced for all regression SNPs within 1Mb of the TSS ($G \sim 24,000$ genes $\times$ $T = 44$ tissues; $M \sim 170,000$ regression SNPs), both for the full expression cohort and for two random subsamples containing half of the cohort. Pooling of tissues increases power (see Supplementary Note). We used FastQTL [45], restricted to unrelated individuals, and corrected for sex, genotyping platform, 3 PCs, and 15-35 PEER factors [46], consistent with GTEx consortium methods [3]. We additionally restricted to European samples.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Supplementary Note

Reverse causality

We can obtain an upper bound on the contribution of reverse causality to ρ^2 for a polygenic trait. Related intuition is discussed in reference [24]. The worst-case hypothetical scenario is a fully-heritable trait T that explains the entire trans-genetic component of gene expression, in addition to some of the cis-genetic component for every gene. For gene g, the covariance between $\tilde{Z}_g$ and T is

$$Cov(\tilde{Z}_g, T) = \theta_g h^2_{T_g}, \tag{9}$$

where θ_g is the effect of T on Z_g and $h^2_{T_g}$ is the SNP-heritability of T localizing to the cis region of SNP g. Now, the marginal effect of gene g on T (due to reverse causality) is $Cov(\tilde{Z}_g, T)/Var(\tilde{Z}_g)$, and

$$\rho^2 \leq \sum_g Var(\tilde{Z}_g)(\frac{Cov(\tilde{Z}_g, T)}{Var(\tilde{Z}_g)})^2 = \sum_g (\theta_g h^2_{T_g})^2 / Var(\tilde{Z}_g). \tag{10}$$

That is, if we regress y on a vector X, the variance explained is at most the sum of the squared marginal effect sizes, normalized by the variance in each X_i. The reason this number cannot be large is that if T is a polygenic trait, then the trans-genetic component of T must explain much more than the cis-genetic component. If the heritability of T is spread evenly across the genome, then we have $h^2_{T_g} \approx 2Mb/3200Mb$. $\theta_g^2/Var(\tilde{Z}_g)$ is at most the total heritability of gene g divided by its cis heritability; if we take $\theta_g^2/Var(\tilde{Z}_g) = 2$ and $G = 20,000$ genes, then we obtain an upper bound on ρ^2 of

$$\rho^2 \leq 20,000 \times (2/3200)^2 \times 2 \approx 0.02. \tag{11}$$

If we wish to bound the contribution of reverse causality to our overall results, rather than to an individual trait, then we can revise the assumption that T explains the entire trans-genetic component of gene expression; this study considered 30 genetically distinct traits, and it is impossible that all of them have this property. If instead each trait individually explains $1/30^{\text{th}}$ of this trans-genetic component, then we have $\theta_g^2/Var(\tilde{Z}_g) = 2/30$, and we have $\rho^2 \leq 0.001$. This type of argument also extends to causal scenarios where a small number of intermediaries affect both the trait and gene expression. If there are k such intermediaries, then a similar bound is obtained for ρ^2, differing by a factor of k.

Alternate approaches

We considered several alternative approaches, but these would not produce reliable estimates of ρ^2, because they fail to distinguish between mediation and colocalization.

- **Comparison of χ^2 statistics.** One possible approach is to regress phenotypic χ^2 statistics on sums of eQTL χ^2 statistics, possibly with an LD score covariate. This would be an attractive method for measuring colocalization, and under the assumption that all colocalization is driven by mediation, it would provide an estimate of ρ^2.

- **Aggregation of fixed-effect estimates.** Suppose we were given a vector s of fixed effect cis-genetic correlations, i.e. for gene k, $s_k = Corr(X\beta_k, Y)$. A natural approach would be to aggregate these estimates; for example, in the absence of cis-genetic correlations between genes, an estimator of ρ^2 would be $\hat{\rho^2} = ||s||^2$. This estimator is used in Figure 1-c. However, this estimator ignores the difference between the fixed-effect quantity s and the random-effect quantity α. Moreover, the difference between $||s||^2$ and $||\alpha||^2$ depends on the amount of colocalization between eQTLs and non-mediated effects, with a larger difference in the presence of stronger colocalization; thus, this approach also fails to distinguish between mediation and colocalization.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

- **Imputation of a gene expression relatedness matrix.** A third approach is to impute a gene expression relatedness matrix ("GERM") using existing methods [23, 24], which could be used as the input to a standard heritability estimator [31, 33, 48]. Unfortunately, the GERM will "tag" the genetic relatedness matrix (GRM), leading to upward bias in the estimate. Even if the GRM is used as a covariate, colocalization will allow the GERM to tag *important* genetic variants, and there will still be inflation.

Derivation of moment condition

Here, we derive equation (1) under the assumption that gene effect sizes are drawn independently of the eQTL entries β and the LD matrix R:

$$\alpha \perp R, \beta. \tag{12}$$

Violations of this assumption are discussed below. The marginal effect of a SNP on the phenotype is

$$g_i := cov(X_i, Y) = \sum_{\text{linked SNPs } k} r_{ik}\gamma_k + \sum_{\text{cis genes } g} b_{ig}\alpha_g, \tag{13}$$

where $B = R\beta = (b_{ig})$ is the marginal eQTL matrix. The covariance between marginal SNP effect sizes g is

$$E(g_i g_j | R, B, \gamma, \alpha) = \sum_{\text{linked SNPs } k} r_{ik} r_{jk} \gamma_k^2 + \sum_{\text{cis genes } g} b_{ig} b_{jg} \alpha_g^2. \tag{14}$$

Now we marginalize out γ and α, using the fact that γ is uncorrelated (on expectation) with β:

$$E(g_i g_j | R, B) = \sum_{\text{linked SNPs } k} r_{ik} r_{jk} E(\gamma_k^2 | R, B) + \sum_{\text{cis genes } g} b_{ig} b_{jg} E(\alpha_g^2 | R, B). \tag{15}$$

Replacing true effect sizes with their estimates from a finite sample, we have

$$E(z_i z_j | R, B) = r_{ij} + N[\sum_{\text{linked SNPs } k} r_{ik} r_{jk} E(\gamma_k^2 | R, B) + \sum_{\text{cis genes } g} b_{ig} b_{jg} E(\alpha_g^2 | R, B)]. \tag{16}$$

We wish to avoid assuming that $E(\gamma_k^2 | R, B) = const$, i.e. that non-mediated effects do not colocalize with eQTLs. Therefore, we remove pairs (i, j) of linked SNPs, so that $r_{ik} r_{jk} \approx 0$. The term r_{ij} also vanishes. By assumption (12), $E(\alpha_g^2 | R, B) = const$. Defining σ_α^2 as this constant, we have that

$$E(z_i z_j | R, B) = N \mathcal{G}_{ij} \sigma_\alpha^2, \quad \mathcal{G}_{ij} = \sum_{\text{cis genes } g} b_{ig} b_{jg}. \tag{17}$$

The independence assumption (12) also implies that

$$\rho^2 = G \frac{h_Z^2}{h_Y^2} \sigma_\alpha^2, \tag{18}$$

and we recover (1).

Some types of violations of the independence assumption, but not others, lead to violations of (1) and to bias in the estimator. Only genes with more than one independent eQTL contribute to $\mathcal{G}$, so if genes with more polygenic cis-genetic architectures have larger or smaller effect sizes on average, the moment condition is violated and bias is introduced in the estimator. In effect, GECS regression extrapolates from genes with multiple eQTLs to genes with only one eQTL. In contrast, violations of (12) mediated by differences in cis heritability do not lead to violations of (1), even though they do lead to violations of both (16) and (18); this is because more-heritable genes contribute more heavily to $\mathcal{G}$, and they contribute proportionally to ρ^2. This can be seen by considering the effect of adding a gene with zero cis heritability. Similarly, if cis-genetically correlated genes have correlated effect sizes, it leads to compensatory violations of (16) and (18), but not (1). This can be seen by considering the effect on of adding a perfectly-correlated copy of every gene in the genome, with an identical effect size to its original, while halving all effect sizes.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Power and number of tissues

GECS regression improves power by concatenating tissues, treating the same gene in k tissues as k distinct genes. When cis-genetic correlations across tissues are strong and environmental correlations are weak, this approach increases power by increasing the effective heritability of each gene. In particular, if the tissues are perfectly cis-genetically correlated, and there is no other covariance between the tissues (due to environment or trans-genetic correlations), then the effective cis-heritability is

$$h^2_{eff} = \frac{Th^2_{cis}}{(T-1)h^2_{cis}+1}. \tag{19}$$

To see this, it is useful to construct a vector $b^* = \hat{R}\beta$, where $\hat{R}$ is the in-sample LD matrix and β is the *true* causal effect size vector. Note that b^* differs from both the true marginal effect size vector $b = R\beta$ and the estimated marginal effect size vector $\hat{b}$. b^* has the property that in the absence of environmental noise ($h^2_{cis} = 1$), $\hat{b} = b^*$. Similarly, given a large number of perfectly cis-genetically correlated tissues from each individual in the eQTL cohort, the average across tissues of the estimated effect size vectors converges to b^*.

We can write the estimated effect vector for SNP i and tissue t as $\hat{b}_{it} = b^*_i + \epsilon_{it}$, where

$$Var(\epsilon_{it}) \approx \frac{1-h^2_{cis}}{N_{\text{eQTL}}}. \tag{20}$$

As a proxy for GECS regression power, we can use

$$\frac{E(A|\hat{R})^2}{Var(A|\hat{R})}, \quad A = g_i g_j \sum_{t=1}^{T} \hat{b}_{it}\hat{b}_{jt}. \tag{21}$$

where g_i is the effect size of SNP i on the phenotype; we use g_i instead of $\hat{g}_i$ to isolate the component of power that is dependent on the eQTL cohort (i.e., when the GWAS cohort is large). We further condition on the in-sample eQTL cohort LD matrix $\hat{R}$ in order to isolate the component of power that is determined by the number of tissues and by the cis-heritability.

First,

$$E(A|\hat{R}) = Tg_i g_j b^*_i b^*_j, \tag{22}$$

where we have used that i and j are not in LD (so ϵ_{it} is uncorrelated with ϵ_{jt}). Next,

$$\begin{aligned} Var(A|\hat{R}) =& (g_i g_j)^2 \sum_{t=1}^{T} Var(\hat{b}_{it}\hat{b}_{jt}) \\ \approx& T(g_i g_j)^2 (b^{*2}_{it} + b^{*2}_{jt} + \frac{1-h^2_{cis}}{N_{\text{eQTL}}}) \frac{1-h^2_{cis}}{N_{\text{eQTL}}} \end{aligned} \tag{23}$$

where we have used independence of noise terms across tissues in the first line and equation (20) together with the assumption of no LD between SNPs i and j in the second line. Disregarding terms of size $1/N^2_{eQTL}$ and assuming $g_i g_j \neq 0$ and $h^2_{cis} \neq 1$, we have

$$\frac{E(A|\hat{R})^2}{Var(A|\hat{R})} = \frac{Tb^{*2}_i b^{*2}_j}{(b^{*2}_{it} + b^{*2}_{jt})(1-h^2_{cis})/N_{\text{eQTL}}}. \tag{24}$$

Taking $\frac{b^{*2}_i b^{*2}_j}{b^{*2}_{it}+b^{*2}_{jt}} \propto h^2_{cis}$, we have

$$\frac{E(A|\hat{R})^2}{Var(A|\hat{R})} \propto \frac{Th^2_{cis}}{1-h^2_{cis}}. \tag{25}$$

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Finally, we solve

$$\frac{Th^2_{cis}}{1-h^2_{cis}} = \frac{h^2_{eff}}{1-h^2_{eff}} \tag{26}$$

to obtain equation (19).

Additional sources of bias

Bias in GECS regression due to expression heritability flanking the cis window GECS regression only considers cis-eQTL effects, defined using some cis radius. If the chosen cis radius is too small, the primary effect is that mediation will be missed for variants outside of the window (Figure S2). However, there is also a competing effect, which occurs as a result of LD between cis SNPs in conjunction with genetic correlations between genes. To illustrate this effect, consider an extreme example: the genome comprises 2 perfectly correlated genes and 20 SNPs, with SNPs $1, 2, ..., 10$ being in cis for gene A and SNPs $11, 12, ...20$ being in cis for gene B. SNP 1 is in perfect LD with SNP 11, SNP 2 is in perfect LD with SNP 12, and so on. If we apply GECS regression, the estimator will be upwardly biased by a factor of 2, since genetic effects on gene A perfectly tag effects on gene B, and this will be reflected in the GECS regression numerator but not in the GECS regression denominator (the variance in GE coscores). In contrast, if the cis radius is extended so that every SNP is cis for both genes, then there is no longer any bias, since the tagging is also reflected in the GECS regression denominator. Thus, this type of bias requires genetic correlations between genes with cis windows that do not perfectly overlap (i.e., not only between the same gene in different tissues), and it also requires LD between SNPs that are in cis for one gene or the other but not both. It is expected to disappear when the cis window is sufficiently large, as SNPs on the boundaries of the cis region (which might have LD with SNPs outside the cis region) will explain little expression heritability.

Bias in cis-heritability estimation due to LD between cis and trans SNPs We estimate cis-heritability of gene expression levels using LD score regression on cis SNPs with LD scores computed to both cis and trans SNPs. In principle, it is preferable to use LD scores between regression SNPs and cis reference SNPs and trans reference SNPs separately [37]; the inclusion of non-cis reference SNPs introduces a small amount of downward bias in the cis-heritability estimates. If SNPs flanking the cis region explain zero heritability, then estimates are biased toward zero by a factor of approximately $\bar{\ell}_{\text{cis}}/\bar{\ell}$, where $\bar{\ell}_{\text{cis}}$ is the average LD score between regression SNPs and cis reference SNPs, and $\bar{\ell}$ is the average LD score between regression SNPs and all reference SNPs. For a cis window size of ±1Mb, this bias is small (at most a few percent), since cis SNPs have more LD with other cis SNPs than with trans SNPs; it is more substantial when a smaller cis window size is used.

Relationship with Haseman-Elston Regression

GECS regression regression is related to Haseman-Elston (HE) Regression [32, 33]. HE regression relates the covariance between the phenotypes of two individuals with the covariance between their genotype vectors (i.e., their genetic relatedness):

$$E(Y_{\text{person 1}}Y_{\text{person 2}}) \propto E(X_{\text{person 1}} \cdot X_{\text{person 2}}). \tag{27}$$

This follows from

$$Y_{\text{person i}} = X_{\text{person i}}\beta + \epsilon_{\text{person i}} \tag{28}$$

where β is random and, crucially, $E(\epsilon_1\epsilon_2) = 0$. Similarly, GECS regression relates the covariance between the phenotypic *effects* of two SNPs with the covariance between their eQTL vectors (i.e., the GE coscore):

$$E(g_{\text{SNP 1}}g_{\text{SNP 2}}) \propto E(b_{\text{SNP 1}} \cdot b_{\text{SNP 2}}), \tag{29}$$

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

and this follows from

$$g_{\text{SNP i}} = b_{\text{SNP i}}\alpha + \gamma_{\text{SNP i}}, \tag{30}$$

where $E(\gamma_1\gamma_2) = 0$ specifically for SNPs with no shared LD.

Choice of LD estimator

We estimate LD using in-sample LD for the two subsamples that are used to estimate the GE coscore regression denominator. We have found that when out-of-sample LD is used, estimates are inflated due to residual LD. It is sufficient to remove in-sample LD for the eQTL cohort only (and not the GWAS cohort) because even if there is residual LD in the GWAS cohort (owing to LD mismatch or finite sample size), this residual LD will not correlate with residual LD in the eQTL cohort. It is necessary to use in-sample LD for the cross-validation subsamples, since if we ascertain SNP pairs with low total LD, then this ascertainment procedure introduces dependencies between the cross-validation subsamples, leading to downward bias in the GE coscore regression denominator.

Supplementary Tables

Table S1: Performance of GE score regression in simulations using real genotypes and simulated gene expression and phenotypes (see Excel file). Simulations included strong colocalization between eQTLs and non-mediated disease SNPs. (a) Estimates were approximately unbiased when 75% of SNP pairs were retained. (b) Estimates were inflated when no LD pruning was applied, but LD pruning reduced the bias. Dashed horizontal lines indicate true values of ρ^2. (c) Estimates of ρ^2 using GECS regression and a naive fixed-effect estimator at different eQTL and GWAS sample sizes, in a simplified simulation setup including colocalization but no mediation. (d) Comparison of power to detect an effect between GECS regression and partitioned LD score regression, using eQTLs as an annotation.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Phenotype	Reference / Dataset	N
Anorexia	Boraska et al., 2014 Mol Psych	32143
Autism Spectrum	PGC Cross-Disorder Group, 2013 Lancet	10263
Putamen volume	Franke et al., 2016 Nat Genet	11840
Neuroticism	Okbay et al., 2016 Nat Genet	170911
Schizophrenia	SCZ Working Group of the PGC, 2014 Nature	70100
Years of Education	Okbay et al., 2016 Nature	328917
Waist-hip ratio (BMI adjusted)	UK Biobank	145375
HDL	Teslovich et al., 2010 Nature	97750
LDL	Teslovich et al., 2010 Nature	93354
Triglycerides	Teslovich et al., 2010 Nature	94461
Smoking status	UK Biobank	145227
Blood pressure (Systolic)	UK Biobank	134011
HbA1C	Soranzo et al., 2010 Diabetes	46368
Type 2 Diabetes	Morris et al., 2012 Nat Genet	60786
Celiac Disease	Dubois et al., 2010 Nat Genet	15283
Crohn's Disease	Jostins et al., 2012 Nature	20883
Lupus	Bentham et al., 2015 Nat Genet	14267
Multiple sclerosis	IMS Genetics Consortium, 2011 Nature	27148
Primary Biliary Cirrhosis	Cordell et al., 2015 Nat Commun	13239
Rheumatoid Arthritis	Okada et al., 2014 Nature	37681
Ulcerative Colitis	Jostins et al., 2012 Nature	27432
Asthma	UK Biobank	145416
Eczema	UK Biobank	145416
Age at Menarche	UK Biobank	74944
Age at Menopause	UK Biobank	44410
Lung FEV1/FVC ratio	UK Biobank	123935
Lung Forced Expiratory Volume (FVC)	UK Biobank	123935
Bone mineral density (heel)	UK Biobank	141441
BMI	Speliotes et al., 2010 Nat Genet	122033
Height	Lango Allen et al., 2010 Nature	131547
Bipolar Disorder	BIP Working Group of the PGC, 2011 Nat Genet	16731
Blood pressure (Diastolic)	UK Biobank	134011
BMI	UK Biobank	145209
College Education	UK Biobank	144204
Coronary Artery Disease	Schunkert et al., 2011 Nat Genet	77210
Depressive symptoms	Okbay et al., 2016 Nat Genet	161460
Ever Smoked	TAG Consortium, 2010 Nat Genet	74035
Fasting Glucose	Manning et. al., 2012 Nat Genet	58074
Height	UK Biobank	145368
Hypertension	UK Biobank	145379
Subjective wellbeing	Okbay et al., 2016 Nat Genet	298420
Years of Education	Rietveld et al., 2013 Science	126559

Table S2: GWAS datasets analyzed. The first 30 phenotypes (above horizontal line) were included in the primary analysis, and the remaining phenotypes were excluded from the primary analysis due to high genetic correlation. Of the first 30 phenotypes, related traits are grouped and order is alphabetical within groups. To avoid bias due to regression SNP selection, GWAS using specialty chip data were excluded, similar to [14].

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Tissue	N_{eQTL}
Adipose Subcutaneous	298
Adipose Visceral	185
Adrenal Gland	126
Artery Aorta	197
Artery Coronary	118
Artery Tibial	285
Brain Anterior*	72
Brain Caudate*	100
Brain Cerebellar*	89
Brain Cerebellum*	103
Brain Cortex*	96
Brain Frontal*	92
Brain Hippocampus*	81
Brain Hypothalamus*	81
Brain Nucleus*	93
Brain Putamen*	82
Breast Mammary	183
Cells EBV-transformed	114
Cells Transformed	272
Colon Sigmoid	124
Colon Transverse	169
Esophagus Gastroesophageal	127
Esophagus Mucosa	241
Esophagus Muscularis	218
Heart Atrial	159
Heart Left	190
Liver	97
Lung	278
Muscle Skeletal	361
Nerve Tibial*	256
Ovary	85
Pancreas	149
Pituitary*	87
Prostate	87
Skin Not	196
Skin Sun	302
Small Intestine	77
Spleen	89
Stomach	170
Testis	157
Thyroid*	278
Uterus	70
Vagina	79
Whole Blood	338

Table S3: GTEx tissues included in the analysis, and sample size. Tissues marked with an asterisk were included in the brain and brain-related tissues analysis (Table 3).

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Phenotype 1	ρ^2	std err	Phenotype 2	ρ^2	std err	$p_{\text{difference}}$	genet corr	std err
Depressive symptoms	0.17	0.07	Neuroticism	0.12	0.02	0.03	0.78	0.05
Depressive symptoms	0.17	0.07	Subjective wellbeing	0.35	0.34	0.16	-1.18	0.14
Neuroticism	0.12	0.03	Subjective wellbeing	0.35	0.34	0.07	-0.99	0.09
Bipolar Disorder	0.25	0.09	Schizophrenia	0.12	0.02	0.02	0.79	0.06
Coronary Artery Disease	0.77	0.66	Hypertension	0.15	0.05	0.33	0.94	0.24
Diastolic	0.13	0.03	Hypertension	0.15	0.05	0.03	0.74	0.03
Ever smoked	0.21	0.1	Smoking status	0.13	0.02	0.1	0.96	0.09
Diastolic	0.13	0.03	Systolic	0.12	0.04	0.81	0.75	0.03
Hypertension	0.15	0.05	Systolic	0.12	0.04	0.05	0.72	0.03
Fasting Glucose	0.24	0.2	Type 2 Diabetes	0.27	0.1	0.43	0.74	0.14
Years of Ed. - Okbay	0.09	0.02	Years of Ed. - Rietveld	0.2	0.04	0.002	1.07	0.03
BMI - Speliotes	0.1	0.02	BMI - UKB	0.1	0.03	0.59	0.97	0.03
Years of Ed. - Okbay	0.09	0.02	College Ed. - UKB	0.12	0.02	0.04	0.97	0.02
Years of Ed. - Rietveld	0.2	0.04	College Ed. - UKB	0.12	0.02	0.14	0.95	0.05
Height - Allen	0.13	0.03	Height - UKB	0.12	0.03	0.92	1.02	0.02

Table S4: Estimated heritability mediated by expression for pairs of genetically correlated traits or studies. P values were obtained by jackknifing on the difference between the estimates.

Table S5: Calibration of jackknife standard errors (see Excel file). Signed empirical errors were divided by the jackknife standard errors, and the standard deviation of the normalized errors is reported for different values of ρ^2.

Table S6: Effect of cis window size on estimated ρ^2 in simulations (see Excel file). (a) Estimates of ρ^2 for different window sizes, either in the presence of genetic correlations across tissues or in the presence of genetic correlations between adjacent genes. (b) Estimates of ρ^2 per unit of cis heritability, either in the presence of genetic correlations across tissues or in the presence of genetic correlations between adjacent genes.

Table S7: Stability of estimated ρ^2 across LD thresholds (see Excel file). Estimates of ρ^2 were obtained for different LD thresholds and averaged across phenotypes. Increasingly stringent thresholds were placed on the amount of shared LD for SNP pairs in the regression, and standard errors were calculated by jackknifing on the mean.

Table S8: Effect of cis window size on estimated ρ^2 (see Excel file). (a) Heritability mediated by expression was variable for small window size values, but stable for larger window sizes. (b) Heritability mediated by expression, per unit of expression heritability, was larger when the cis window size was smaller than ±1Mb.

Table S9: Dependence on number of tissues (see Excel file). Estimates of ρ^2 were obtained using random samples of tissues and averaged across non-redundant phenotypes. Error bars indicate jackknife standard errors (which do not include variability due to true differences among tissues).

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Supplementary Figures

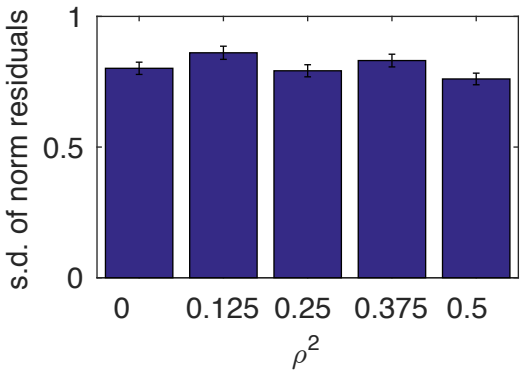


Figure S1: Calibration of jackknife standard errors. Signed empirical errors were divided by the jackknife standard errors, and the standard deviation of the normalized errors is reported for different values of ρ^2. Error bars indicate standard errors on the sample standard deviation, based on 288 simulations. Numerical results are reported in Table .

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

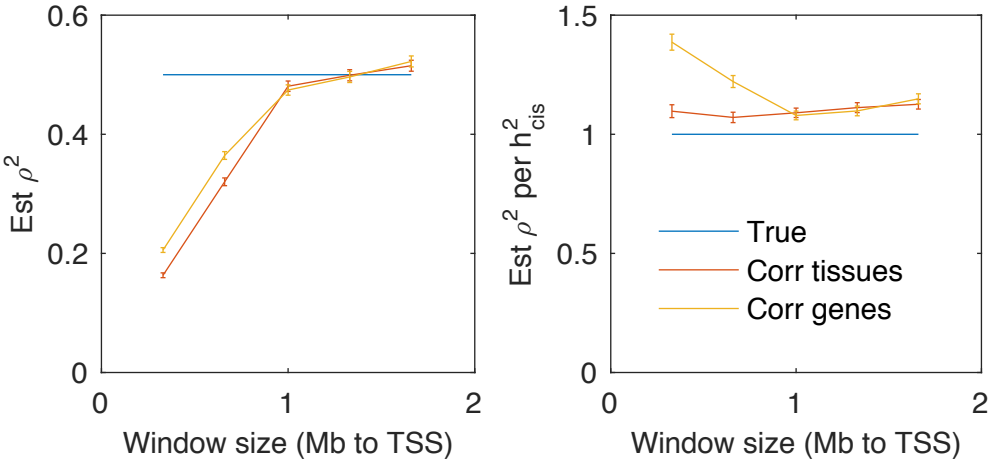


Figure S2: Effect of cis window size on estimated ρ^2 in simulations. (a) Estimates of ρ^2 for different window sizes, either in the presence of genetic correlations across tissues or in the presence of genetic correlations between adjacent genes. (b) Estimates of ρ^2 per unit of cis heritability, either in the presence of genetic correlations across tissues or in the presence of genetic correlations between adjacent genes. Error bars indicate standard error based on 576 simulations. Numerical results are reported in Table S6.

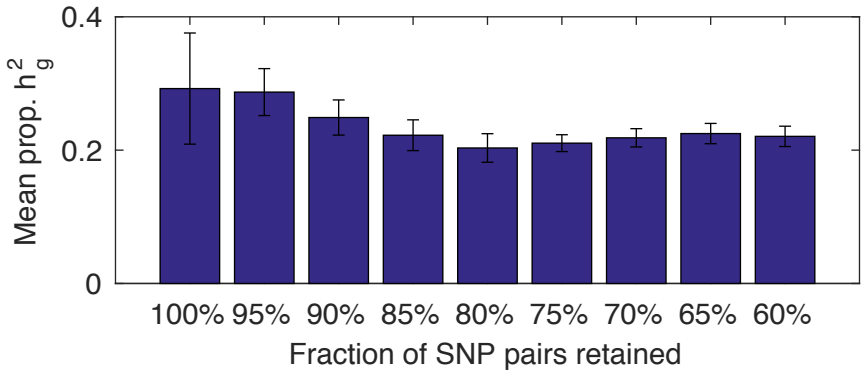


Figure S3: Stability of estimated ρ^2 across LD thresholds. Estimates of ρ^2 were obtained for different LD thresholds and averaged across phenotypes. Increasingly stringent thresholds were placed on the amount of shared LD for SNP pairs in the regression, and standard errors were calculated by jackknifing on the mean. Numerical results are reported in Table S7.

bioRxiv preprint doi: https://doi.org/10.1101/118018; this version posted March 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

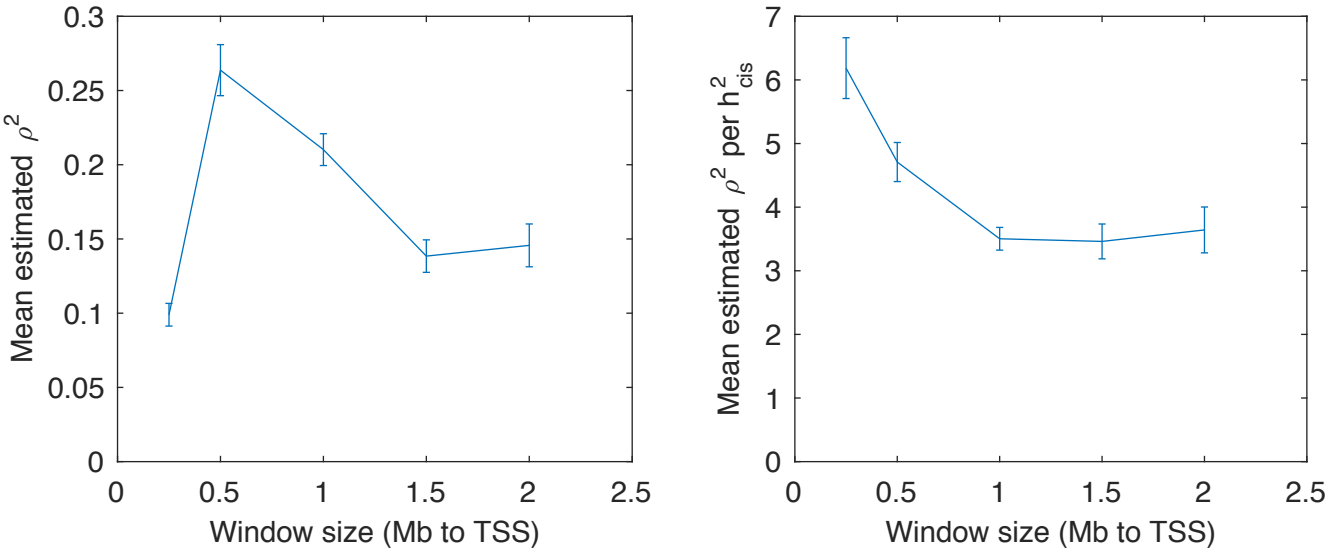


Figure S4: Effect of cis window size on estimated ρ^2. (a) Heritability mediated by expression was variable for small window size values, but stable for larger window sizes. (b) Heritability mediated by expression, per unit of expression heritability, was larger when the cis window size was smaller than ±1Mb. Numerical results are reported in Table S8.

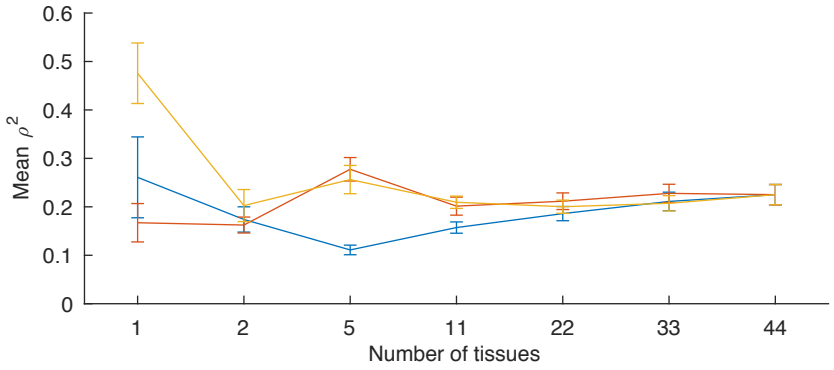


Figure S5: Dependence on number of tissues. Estimates of ρ^2 were obtained using random samples of tissues and averaged across non-redundant phenotypes. Error bars indicate jackknife standard errors (which do not include variability due to true differences among tissues). Numerical results are reported in Table S9.