Modeling tissue co-regulation to estimate tissue-specific contributions to disease

Tiffany Amariuta[1,2,3,4], Katherine Siewert-Rocks[3,4], Alkes L. Price[3,4,5]

1. Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, CA, USA
2. Department of Medicine, University of California San Diego, La Jolla, CA, USA
3. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
4. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
5. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Abstract

Integrative analyses of genome-wide association studies (GWAS) and gene expression data across diverse tissues and cell types have enabled the identification of putative disease-critical tissues. However, co-regulation of genetic effects on gene expression across tissues makes it difficult to distinguish biologically causal tissues from tagging tissues. While previous work emphasized the potential of accounting for tissue co-regulation, tissue-specific disease effects have not previously been formally modeled. Here, we introduce a new method, tissue co-regulation score regression (TCSC), that disentangles causal tissues from tagging tissues and partitions disease heritability (or covariance) into tissue-specific components. TCSC leverages gene-disease association statistics across tissues from transcriptome-wide association studies (TWAS), which implicate both causal and tagging genes and tissues. TCSC regresses TWAS chi-square statistics (or products of z-scores) on tissue co-regulation scores reflecting correlations of predicted gene expression across genes and tissues. In simulations, TCSC distinguishes causal tissues from tagging tissues while controlling type I error. We applied TCSC to GWAS summary statistics for 78 diseases and complex traits (average N = 302K) and gene expression prediction models for 48 GTEx tissues. TCSC identified 21 causal tissue-trait pairs at 5% FDR, including well-established findings, biologically plausible novel findings (e.g. aorta artery and glaucoma), and increased specificity of known tissue-trait associations (e.g. subcutaneous adipose, but not visceral adipose, and HDL). TCSC also identified 17 causal tissue-trait covariance pairs at 5% FDR. For the positive genetic covariance between BMI and red blood cell count, brain substantia nigra contributed positive covariance while pancreas contributed negative covariance; this suggests that genetic covariance may reflect distinct tissue-specific contributions. Overall, TCSC is a precise method for distinguishing causal tissues from tagging tissues, improving our understanding of disease and complex trait biology.

Introduction

Most diseases are driven by tissue-specific or cell-type-specific mechanisms, thus the inference of causal disease tissues is an important goal[1]. For many polygenic diseases and complex traits, disease-associated tissues have previously been identified via the integration of genome-wide association studies (GWAS) with tissue-level functional data characterizing expression quantitative trait loci (eQTLs)[2-5], gene expression[6-9], or epigenetic features[10-17]. However, it is likely that most disease-associated tissues are not actually causal, due to the high correlation of eQTL effects (resp. gene expression or epigenetic features) across tissues; the correlation of eQTL effects across tissues, i.e. tissue co-regulation, can arise due to shared eQTLs or distinct eQTLs in linkage disequilibrium (LD)[2,18,19,5]. One approach to address this involves comparing eQTL-disease colocalizations across different tissues[2]; however, this approach relies on colocalizations with disease that are specific to a single tissue, and may implicate co-regulated tagging tissues that colocalize with disease. Another approach leverages multi-trait fine-mapping methods to simultaneously evaluate all tissues for colocalization with disease[5]; however, this locus-based approach does not produce genome-wide estimates and it remains the case that many (causal or tagging) tissues may colocalize with disease under this framework. To our knowledge, no previous study has formally modeled genetic co-regulation across tissues to statistically disentangle causal from tagging tissues.

Here, we introduce a new method, tissue co-regulation score regression (TCSC), that disentangles causal tissues from tagging tissues and partitions disease heritability (or genetic covariance of two diseases/traits) into tissue-specific components. TCSC leverages gene-disease association statistics across tissues from transcriptome-wide association studies (TWAS)[20,21,18]. A challenge is that TWAS association statistics include the effects of both co-regulated tissues (see above) and co-regulated genes[18,22]. To address this, TCSC regresses TWAS chi-square statistics (or products of z-scores for two diseases/traits) on tissue co-regulation scores reflecting correlations of predicted gene expression across genes and tissues. TCSC is conceptually related to gene co-regulation score regression (GCSC)[22], a method for identifying disease-enriched gene sets that models gene co-regulation but does not model tissue co-regulation. Distinct from previous methods that analyze each tissue marginally, TCSC jointly models contributions from each tissue to identify causal tissues (analogous to the distinction in GWAS between marginal association and fine-mapping[23]). We validate TCSC using extensive simulations using real genotypes with LD, including comparisons to RTC Coloc[2], RolyPoly[6], LDSC-SEG[7], and CoCoNet[9] (reviewed in [1,24]). We apply TCSC to 78 diseases and complex traits (average N = 302K) and 48 GTEx tissues[19], showing that TCSC recapitulates known biology and identifies biologically plausible novel tissue-trait pairs (or tissue-trait covariance pairs) while attaining increased specificity relative to previous methods.

Results

Overview of TCSC regression

TCSC estimates the disease heritability explained by *cis*-genetic components of gene expression in each tissue when jointly modeling contributions from each tissue; a formal definition of this quantity in terms of SNP-level effects is provided in the **Methods** section. We refer to tissues with nonzero contributions as "causal" tissues (with the caveat that joint-fit effects of gene expression on disease may not reflect biological causality; see **Discussion**). TCSC assumes that gene expression-disease effect sizes are independent and identically distributed (i.i.d.) across genes and tissues (while accounting for the fact that *cis*-genetic components of gene expression are correlated across genes and tissues); violations of this model assumption are explored via simulations below. TCSC leverages the fact that TWAS χ^2 statistics for each gene and tissue include both causal effects of that gene and tissue on disease and tagging effects of *co-regulated* genes and tissues. We define co-regulation based on squared correlations in *cis*-genetic expression, which can arise due to shared causal eQTLs and/or LD between causal eQTLs[18]. TCSC determines that a tissue is causal for disease if genes and tissues with high co-regulation to that tissue have higher TWAS χ^2 statistics than genes and tissues with low co-regulation to that tissue.

In detail, let $h^2_{ge(t')}$ denote the disease heritability explained by the *cis*-genetic component of gene expression in tissue t'. The expected TWAS χ^2 statistic for gene g and tagging tissue t is

$$E\left[\chi^2_{g,t}\right] = N \sum_{t'} l(g,t;t') h^2_{ge(t')}/G_{t'} + 1, \quad (1)$$

where N is GWAS sample size, t' indexes causal tissues, $l(g,t;t')$ are tissue co-regulation scores (defined as $l(g,t;t') = \sum_{g'} r^2\left(\widehat{W}_{g,t}, W_{g',t'}\right)$, where W denotes the *cis*-genetic component of gene expression for a gene-tissue pair across individuals, $\widehat{W}$ denotes the *cis*-predicted expression for a gene-tissue pair, the sum is over genes g'within +/- 1 Mb to gene g), and $G_{t'}$ is the number of significantly *cis*-heritable genes in tissue t'. A derivation of Equation (1) is provided in the **Methods** section. Equation (1) allows us to estimate $h^2_{ge(t')}$ via a multiple linear regression of TWAS χ^2 statistics (for each gene and tagging tissue) on tissue co-regulation scores (**Figure 1**); we note that tissue co-regulation scores reflect $\widehat{W}_{g,t}$ and $W_{g',t'}$ but *estimated* tissue co-regulation scores reflect $\widehat{W}_{g,t}$ and $\widehat{W}_{g',t'}$, necessitating a bias correction step[22] (**Methods**). To facilitate comparisons across diseases/traits, we primarily report the proportion of disease heritability explained by the *cis*-genetic component of gene expression in tissue t' ($\pi_{t'} = h^2_{ge(t')}/h^2_g$), where h^2_g is the common variant SNP-heritability estimated by S-LDSC[13,25,26].

TCSC can also estimate the genetic covariance between two diseases explained by *cis*-genetic components of gene expression in each tissue, using products of TWAS z-scores. In detail, let $\omega_{ge(t')}$ denote the genetic covariance explained by the *cis*-genetic component of gene expression in tissue t' (defined analogously to $h^2_{ge(t')}$; **Methods**). The expected product of TWAS z-scores in disease 1 and disease 2 for gene g and tagging tissue t is

$$E\left[z^1_{gt} \times z^2_{gt}\right] = \sqrt{N_1 N_2} \sum_{t'} l(g,t;t') \omega_{ge(t')}/G_{t'} + \rho N_s/\sqrt{N_1 N_2} \quad (2)$$

where N_1 is GWAS sample size for disease 1, N_2 is GWAS sample size for disease 2, t' indexes causal tissues, $l(g, t; t')$ are tissue co-regulation scores (see above), $G_{t'}$ is the number of significantly *cis*-heritable genes in tissue t' (**Methods**), ρ is the phenotypic correlation between disease 1 and disease 2, and N_s is the number of overlapping GWAS samples between disease 1 and disease 2. Equation (2) allows us to estimate $\omega_{ge(t')}$ via a multiple linear regression of products of TWAS z-scores in disease 1 and disease 2 (for each gene and tagging tissue) on tissue co-regulation scores. We note that the last term in Equation (2) is not known a priori but is accounted for via the regression intercept, analogous to previous work[27]. To facilitate comparisons across diseases/traits, we primarily report the signed proportion of genetic covariance explained by the *cis*-genetic component of gene expression in tissue t' ($\zeta_{t'} = \omega_{ge(t')} / \omega_g$), where ω_g is the common variant genetic covariance estimated by cross-trait LDSC[28].

We restrict gene expression prediction models and TWAS association statistics for each tissue to significantly *cis*-heritable genes in that tissue, defined as genes with significantly positive *cis*-heritability (2-sided $p < 0.01$; estimated using GCTA[29]) and positive adjusted-R^2 in cross-validation prediction. We note that quantitative estimates of the disease heritability explained by the *cis*-genetic component of gene expression in tissue t' ($h^2_{ge(t')}$) are impacted by the number of significantly *cis*-heritable genes in tissue t' ($G_{t'}$), which may be sensitive to eQTL sample size (**Methods**). For each disease (or pair of diseases), we use a genomic block-jackknife with 200 blocks to estimate standard errors on the disease heritability (or covariance) explained by *cis*-genetic components of gene expression in each tissue, and compute 1-sided P-values for nonzero heritability (or 2-sided P-values for nonzero covariance) and false discovery rates (FDR) accordingly; we primarily report causal tissues with FDR < 5%. We use a 1-sided test for nonzero heritability because we are only interested in detecting positive tissue-specific contributions to heritability. Further details, including correcting for bias in tissue co-regulation scores arising from differences between *cis*-genetic vs. *cis*-predicted expression (analogous to GCSC[22]) and utilizing regression weights to improve power, are provided in the **Methods** section. We have publicly released open-source software implementing TCSC regression (see **Code Availability**), as well as all GWAS summary statistics, TWAS association statistics, tissue co-regulation scores, and TCSC output from this study (see **Data Availability**).

Simulations

We performed extensive simulations to evaluate the robustness and power of TCSC, using the TWAS simulator of Mancuso et al.[30] (see **Code Availability**). We used real genotypes from 1000 Genomes European to simulate gene expression values (for each gene and tissue) and complex trait phenotypes, and computed TWAS association statistics for each gene and tissue. In our default simulations, the number of tissues was set to 10. The gene expression sample size (in each tissue) varied from 100 to 1,500 (with the value of 300 corresponding most closely to the GTEx data[19] used in our analyses of real diseases/traits; see below). The number of genes was set to 1,000 across chromosome 1; 100 of the 1,000 genes had nonzero (normally distributed) gene-disease effects in the causal tissue[31]. For each tissue, 500 genes were chosen

to be *cis*-heritable. In the causal tissue and the three most highly genetically correlated tagging tissues, all 100 causal genes were *cis*-heritable. Each *cis*-heritable gene was assigned 5 causal *cis*-eQTLs within 50kb of the gene body, consistent with the upper range of independent eQTLs per gene detected in GTEx[19] and other studies[32-35]. The *cis*-eQTL effect sizes for each gene were drawn from a multivariate normal distribution across tissues to achieve a specified level of co-regulation (see below), the *cis*-heritability of each gene was sampled from an exponential distribution, and neighboring co-regulated genes were assigned the same heritability to maximize gene-gene co-regulation. In each tissue, the average *cis*-heritability (across genes) was set to 0.08 (sd = 0.05, ranging from 0.01 to 0.40) in order to achieve an average estimated *cis*-heritability (*across significantly cis-heritable genes, estimated by GCTA[29]*) varying from 0.11 to 0.31 (across gene expression sample sizes), which matches empirical values from GTEx[19]. The proportions of expressed genes that were significantly *cis*-heritable and the proportion of neighboring genes with significant genetic correlation (of eQTL effects) were also matched to GTEx data[19]. The 10 tissues were split into three tissue categories to mimic biological tissue modules in GTEx[19] (tissues 1-3, tissues 4-6, and tissues 7-10), and average *cis*-genetic correlations between tissues (averaged across genes) were set to 0.795 within the same tissue category, 0.722 between tissue categories, and 0.753 overall[36] (**Methods**). The default GWAS sample size was set to 10,000. The 10 tissues included one causal tissue explaining 100% of trait heritability and nine non-causal tissues; 100% of trait heritability was explained by gene expression. Other parameter values were also explored, including other proportions of trait heritability explained by the causal tissue, other proportions of trait heritability not explained by gene expression, and other values of the number of causal tissues and the number of tagging tissues. Further details of the simulation framework are provided in the **Methods** section. We compared TCSC to four previously published methods: RTC Coloc[2], RolyPoly[6], LDSC-SEG[7], and CoCoNet[9]. We caution that RolyPoly, LDSC-SEG, and CoCoNet do not use eQTL data, and thus the power of TCSC relative to these methods is likely to be highly sensitive to assumptions about the role of gene expression in disease architectures. We caution that the power of TCSC (and other methods) varies greatly with the choice of parameter settings (see below), thus the primary purpose of these simulations was to evaluate the robustness of TCSC relative to other methods.

We first evaluated the bias in TCSC estimates of the disease heritability explained by the *cis*-genetic component of gene expression in tissue t' ($h^2_{ge(t')}$), for both causal and non-causal tissues. For causal tissues, TCSC produced unbiased estimates of $h^2_{ge(t')}$ (**Figure 2A**, **Supplementary Table 1**); this implies that error in eQTL effect size estimates, which impacts TWAS statistics and co-regulation scores, does not bias TCSC estimates for causal tissues. A subtlety is that, as noted above, estimates of $h^2_{ge(t')}$ are impacted by the number of significantly *cis*-heritable genes in tissue t' ($G_{t'}$), which may be sensitive to eQTL sample size. Estimates were conservative when setting $G_{t'}$ to the number of significantly *cis*-heritable genes, and unbiased when setting $G_{t'}$ to the number of true *cis*-heritable genes. For non-causal tissues, TCSC produced estimates of $h^2_{ge(t')}$, that were significantly positive when averaged across all simulations, but not large enough to substantially impact type I error (see below). In this analysis of bias in estimates of $h^2_{ge(t')}$, we could not include a comparison to RTC Coloc,

RolyPoly, LDSC-SEG, or CoCoNet, because these methods do not provide quantitative estimates of $h^2_{ge(t\prime)}$.

We next evaluated the type I error of TCSC for non-causal tissues. The type I error of TCSC was approximately well-calibrated, ranging from 5.2% to 6.9% across eQTL sample sizes at a significance threshold of $p = 0.05$ (**Figure 2B**, **Supplementary Table 1**). In comparison, we observed type I errors from 53%-86% for RTC Coloc, 32%-33% for LDSC-SEG, 11%-12% for RolyPoly, and 32%-38% for CoCoNet, substantially greater than the type I error of TCSC (**Supplementary Figure 1**, **Supplementary Table 2**).

We next evaluated the power of TCSC for causal tissues. We determined that TCSC was moderately well-powered to detect causal tissues, with power ranging from 11%-49% across eQTL sample sizes at a nominal significance threshold of $p < 0.05$ (**Figure 2C**) (and 1%-18% at a stringent significance threshold of $p < 0.004$, corresponding to 5% per-trait FDR across tissues in these simulations; **Supplementary Table 1**). As noted above, the power of TCSC varies greatly with the choice of parameter settings (see below), thus the power of TCSC in real-world settings is best evaluated using real trait analysis. As expected, power increased at larger eQTL sample sizes, due to lower standard errors on point estimates of $h^2_{ge(t\prime)}$ (**Figure 2A**). We also evaluated the power of RTC Coloc, RolyPoly, LDSC-SEG, and CoCoNet. For the only other method with type I error less than 15% (RolyPoly), power ranged from 14%-17% across eQTL sample sizes, substantially lower than TCSC (**Supplementary Figure 1**, **Supplementary Table 2**). We also used ROC curves to assess the relationship between the sensitivity (power) and specificity (one minus the false positive rate) of all 5 methods across 1,000 uniformly spaced p-value thresholds. TCSC attained the largest AUC (0.78, vs. 0.54-0.59 for other methods) (**Supplementary Figure 1**).

We similarly evaluated the robustness and power of TCSC when estimating tissue-specific contributions to the genetic covariance between two diseases/traits; we did not compare TCSC to RTC Coloc, RolyPoly, LDSC-SEG, and CoCoNet, which are not applicable to cross-trait analysis. We employed the same simulation framework described above and set the genetic correlation of the two simulated traits to 0.5. We first evaluated the bias in TCSC estimates of the genetic covariance explained by the *cis*-genetic component of gene expression in tissue t' ($\omega_{ge(t\prime)}$), for both causal and non-causal tissues (**Figure 3A**, **Supplementary Table 3**). For causal tissues, TCSC produced unbiased estimates of $\omega_{ge(t\prime)}$ (conservative estimates when setting $G_{t\prime}$ to the number of significantly *cis*-heritable genes, rather than the number of true *cis*-heritable genes), analogous to single-trait simulations. For non-causal tissues, TCSC again produced estimates of $\omega_{ge(t\prime)}$ that were significantly positive when averaged across all simulations, but not large enough to substantially impact type I error. We next evaluated the type I error of cross-trait TCSC for non-causal tissues. TCSC was well-calibrated with type I error ranging from 5.4%-6.7% at $p < 0.05$ (**Figure 3B**). Finally, we evaluated the power of cross-trait TCSC for causal tissues. We determined that cross-trait TCSC was modestly powered at realistic eQTL sample sizes, with power ranging from 8%-27% across eQTL sample sizes at $p < 0.05$ (**Figure 3C**) (and 1-6% power at $p < 0.004$ corresponding to 5% per-trait FDR across tissues in

these simulations; **Supplementary Table 3**); as noted above, the power of TCSC varies greatly with the choice of parameter settings (see below). In ROC curve analysis, TCSC attained an AUC of 0.67 (**Supplementary Figure 1**).

We performed 12 secondary analyses. First, we varied the eQTL sample size across tissues. Specifically, we set the eQTL sample size of the causal tissue to 300 individuals and the eQTL sample sizes of the non-causal tissues to range between 100 and 1,500 individuals. We observed inflated type I error for non-causal tissues (particularly those with larger eQTL sample sizes), implying that large variations in eQTL sample sizes may compromise type I error (**Supplementary Figure 2**). Second, we evaluated the robustness of TCSC when varying the number of expressed genes in the causal tissue under four scenarios: (i) only the 500 *cis*-heritable genes are expressed in the causal tissue, (ii) only 375 *cis*-heritable genes (including all 100 causal genes) are expressed in the causal tissue, (iii) only 225 *cis*-heritable genes (including all 100 causal genes) are expressed in the causal tissue, and (iv) only the 100 causal genes are expressed in the causal tissue. We determined that type I error remained approximately well-calibrated in all scenarios, and that power was dramatically improved and bias for non-causal tissues decreased as the number of tagging genes in the causal tissue decreased (**Supplementary Figures 3-4**); for causal tissues, estimates of $h^2_{ge(t\prime)}$ were upward biased when setting $G_{t\prime}$ to the number of true *cis*-heritable genes and unbiased when setting $G_{t\prime}$ to the number of significantly *cis*-heritable genes across tissues. Third, we varied the true values of $h^2_{ge(t\prime)}$ (or $\omega_{ge(t\prime)}$) for causal tissues. We determined that patterns of bias, type I error, and power were generally robust across different parameter values, although the smallest values resulted in lower power and greater bias for non-causal tissues (**Supplementary Figures 5-6**). Fourth, we varied the number of causal tissues, considering 1, 2, or 3 causal tissues. We observed that the power of TCSC decreased with multiple causal tissues but did not differ greatly between 2 and 3 causal tissues (**Supplementary Figures 7-8**); for causal tissues, estimates of $h^2_{ge(t\prime)}$ were upward biased when setting $G_{t\prime}$ to the number of true *cis*-heritable genes. Fifth, we varied the number of non-causal tissues from 0 to 9. For causal tissues, TCSC estimates were upward biased with fewer tagging tissues but unbiased with more tagging tissues (**Supplementary Figures 9-10**). TCSC type I error and power were generally higher with fewer tagging tissues; this finding does not compromise our real trait analysis, which involve a large number of tissues. Sixth, we modified TCSC to not correct for bias in tissue co-regulation scores arising from differences between *cis*-genetic and *cis*-predicted expression. We determined that removal of bias correction resulted in conservative bias in estimates for causal tissues, increased type I error, and similar power (**Supplementary Figures 11-12**). Seventh, we modified TCSC to apply bias correction to the calculation of all correlations of *cis*-predicted expression contributing to co-regulation scores rather than only those involving the same gene and tissue, which resulted in a decrease in power, anti-conservative bias in estimates for causal tissues, and similar type I error rate (**Supplementary Figures 13-14**). Eighth, we modified TCSC to use bias-corrected co-regulation scores in the calculation of regression weights, which resulted in similar performance to the default setting (**Supplementary Figures 15-16**). We note that regression weights pertain to maximizing signal to noise and not avoiding bias in estimates of $h^2_{ge(t)}$; we continue to not perform bias correction when calculating regression weights,

consistent with GCSC[22]. Ninth, we violated the model assumption that gene-disease effects are independent and identically distributed (i.i.d.) across tissues by including a second causal tissue whose gene-disease effects correlate with varying degree to the gene-disease effects of the original causal tissue (**Supplementary Figures 17-18**). We determined that while this increases noise to TCSC estimates, the estimates are generally unbiased and TCSC is able to powerfully identify the causal tissue, similar to the addition of a causal tissue where there are no shared gene-disease effects (see **Supplementary Figures 7-8**). Tenth, we violated the i.i.d. model assumption by duplicating the causal tissue. We determined that TCSC performs well, (e.g. frequently identifies both tissues as causal and estimates $h^2_{ge(t\prime)}$ for both tissues without bias) despite the violation of model assumption (**Supplementary Figures 19-20**), similar to the previous analysis. Eleventh, we evaluated the robustness of TCSC in the presence of disease heritability that is not mediated via gene expression. We observed that all areas of TCSC performance are affected, with slightly increased type I error rates, decreased power in the case of larger non-mediate heritability, and upward bias in estimates of $h^2_{ge(t\prime)}$ for causal tissues (**Supplementary Figures 21-22**). Finally, we evaluated the robustness of TCSC to variation in the window size used to identify co-regulated genes in the calculation of co-regulation scores and determined that TCSC performance was robust and type I error decreased with larger window sizes (**Supplementary Figures 23-24**). Further details of these secondary analyses are provided in the **Supplementary Note**.

Identifying tissue-specific contributions to 78 diseases and complex traits

We applied TCSC to publicly available GWAS summary statistics for 78 diseases and complex traits (average *N* = 302K; **Supplementary Table 4**) and gene expression data for 48 GTEx tissues[19] (**Table 1**) (see **Data Availability**). The 78 diseases/traits (which include 33 diseases/traits from UK Biobank[37]) were selected to have z-score > 6 for nonzero SNP-heritability (as in previous studies[13,25,38]), with no pair of diseases having squared genetic correlation > 0.1[28] and substantial sample overlap (**Methods**). The 48 GTEx tissues were aggregated into 39 *meta-tissues* (average N = 266, range: *N* = 101-320 individuals, 23 meta-tissues with *N* = 320) in order to reduce variation in eQTL sample size across tissues (**Table 1** and **Methods**); below, we refer to these as "tissues" for simplicity. We constructed gene expression prediction models for an average of 3,993 significantly *cis*-heritable protein-coding genes (as defined above) in each tissue. We primarily report the proportion of disease heritability explained by the *cis*-genetic component of gene expression in tissue t' ($\pi_{t'} = h^2_{ge(t\prime)}/h^2_g$), as well as its statistical significance (using per-trait FDR). We employ a per-trait FDR (as in ref.[39,40]) rather than a global FDR (as in ref.[7]), because power is likely to vary across traits and there are a sufficiently large number of independent quantities estimated per trait ($\pi_{t'}$ jointly estimated across 39 tissues); a global FDR is more appropriate when there are far fewer independent quantities estimated per trait, e.g. due to non-independent, marginal tissue associations in ref.[7].

TCSC identified 21 causal tissue-trait pairs with significantly positive contributions to disease/trait heritability at 5% FDR, spanning 7 distinct tissues and 17 distinct diseases/traits (**Figure 4**, **Supplementary Table 5, Supplementary Figure 25**). Many of the significant findings

recapitulated known biology, including associations of whole blood with blood cell traits such as white blood cell count ($\pi_{t'}$ = 0.21, s.e. = 0.064, $P = 5.7 \times 10^{-4}$) and liver with lipid traits such as LDL ($\pi_{t'}$ = 0.20, s.e. = 0.050, $P = 2.9 \times 10^{-5}$). We obtained independent GWAS summary statistics for 10 traits implicated in 13 significant tissue-trait pairs (**Supplementary Table 4**) and confirmed the same direction of effect for 13 of 13 tissue-trait pairs (including FDR < 5% for 7 of 13 tissue-trait pairs, FDR < 10% for 9 of 13 tissue-trait pairs) (**Supplementary Table 5**); however, FDR < 5% results are expected to include a small number of false positives, and our association of whole blood with major depressive disorder (FDR < 5% in primary analysis; same direction, FR = 84% in independent GWAS data) may be one of these. In our primary analysis, TCSC also identified 5 suggestive tissue-trait pairs with 5% < FDR < 10% (**Figure 4, Supplementary Table 6**).

TCSC also identified several biologically plausible findings not previously reported in the genetics literature. First, aorta artery was associated with glaucoma ($\pi_{t'}$ = 0.15, s.e. = 0.051, $P = 1.3 \times 10^{-3}$). TCSC also identified aorta artery as a causal tissue for diastolic blood pressure (DBP) ($\pi_{t'}$ = 0.078, s.e. = 0.024, $P = 5.1 \times 10^{-4}$), which is consistent with DBP measuring the pressure exerted on the aorta when the heart is relaxed[41]. High blood pressure is a known risk factor for glaucoma[42-46], explaining the role of aorta artery in genetic susceptibility to glaucoma. Second, TCSC identified heart left ventricle (in addition to whole blood) as a causal tissue for platelet count ($\pi_{t'}$ = 0.091, s.e. = 0.031, $P = 1.7 \times 10^{-3}$), consistent with the role of platelets in the formation of blood clots in cardiovascular disease[47-50]. In cardiovascular disease, platelets are recruited to damaged heart vessels after cholesterol plaques rupture, resulting in blood clots due to the secretion of coagulating molecules[51]; antiplatelet drugs have been successful at reducing adverse cardiovascular outcomes[52]. Moreover, the left ventricle serves as a muscle to pump blood throughout the body[53], likely modulating platelet counts and other blood cell counts, creating detectable changes in serum from which platelet counts are measured. Other significant findings are discussed in the **Supplementary Note**, and numerical results for all tissues and diseases/traits analyzed are reported in **Supplementary Table 6**.

TCSC also increased the specificity of known tissue-trait associations. For high density lipoprotein (HDL), previous studies reported that deletion of a cholesterol transporter gene in adipose tissue reduces HDL levels, consistent with the fact that adipose tissues are storage sites of cholesterol and express genes involved in cholesterol transport and HDL lipidation[54,55]. While there are three adipose tissues represented in the GTEx data that we analyzed (subcutaneous, visceral, and breast tissue), TCSC specifically identified subcutaneous adipose ($\pi_{t'}$ = 0.16, s.e. = 0.054, $P = 1.5 \times 10^{-3}$; **Figure 4**), but not visceral adipose or breast tissue ($P > 0.05$; **Supplementary Table 6**), as a causal tissue for HDL. Previous studies have established that levels of adiponectin, a hormone released by adipose tissue to regulate insulin, are significantly positively correlated with HDL[56-58] and more recently, a study has reported that adiponectin levels are associated specifically with subcutaneous adipose tissue and not visceral adipose tissue[59]; thus, the specific role of subcutaneous adipose tissue in HDL may be due to a causal mechanism related to adiponectin. We note that TCSC did not identify liver as a causal tissue for HDL (FDR > 5%), which may be due to limited power in liver due to smaller eQTL sample size. For waist-hip ratio adjusted for BMI (WHRadjBMI), previous studies reported colocalization

of WHRadjBMI GWAS variants with *cis*-eQTLs in subcutaneous adipose, visceral adipose, liver, and whole blood[60], consistent with WHRadjBMI measuring adiposity in the intraabdominal space which is likely regulated by metabolically active tissues[61]. TCSC specifically identified subcutaneous adipose as a suggestive finding ($\pi_{t'}$ = 0.10, s.e. = 0.037, $P = 2.4 \times 10^{-3}$, 5% < FDR < 10%; **Figure 4**), but not visceral adipose, breast, liver, or whole blood ($P > 0.05$; **Supplementary Table 6**), as a causal tissue for WHRadjBMI. The causal mechanism may involve adiponectin secreted from subcutaneous adipose tissue, which is negatively correlated with WHRadjBMI[62]. We note that the *P* value distributions across traits are similar for subcutaneous adipose (median *P* = 0.42) and visceral adipose (median *P* = 0.56) and are comparable to the other 37 analyzed (median *P* = 0.20 – 0.84, **Supplementary Table 7**). For BMI, previous studies have broadly implicated the central nervous system, but did not reveal more precise contributions[63,13,64,65,7,66]. TCSC specifically identified brain cereb. as a suggestive finding ($\pi_{t'}$ = 0.042, s.e. = 0.015, $P = 2.6 \times 10^{-3}$, 5% < FDR < 10%), but not brain cortex or brain limbic (P > 0.05; **Supplementary Table 6**), as a causal tissue for BMI. This finding is consistent with a known role for brain cerebellum in biological processes related to obesity including endocrine homeostasis[67] and feeding control[68]; recently, a multi-omics approach has revealed cerebellar activation in mice upon feeding[69].

We performed a secondary analysis in which we removed tissues with eQTL sample size less than 320 individuals, as these tissues may often be underpowered (**Figure 2C**). Results are reported in **Supplementary Figure 26** and **Supplementary Table 8.** The number of causal tissue-trait pairs with significantly positive contributions to disease/trait heritability (at 5% FDR) increased from 21 to 23, likely due to a decrease in multiple hypothesis testing burden from removing underpowered tissues. The 23 significant tissue-trait pairs reflect a gain of 8 newly significant tissue-trait pairs (and a loss of 6 formerly significant tissue-trait pairs, of which 5 were lost because the tissue was removed), but estimates of $\pi_{t'}$ for each significant tissue-trait pair were not statistically different from our primary analysis (**Supplementary Table 9**). Notably, among the newly significant tissue-trait pairs, whole blood was associated with hypothyroidism ($\pi_{t'}$ = 0.100, s.e. = 0.032, $P = 8.9 \times 10^{-4}$); we note that thyroid had a quantitatively large but only nominally significant association ($\pi_{t'}$ = 0.452, s.e. = 0.225, *P* = 0.02, FDR = 26%). Esophagus muscularis (rather than lung tissue) was associated with the lung trait FEV1/FVC[70] ($\pi_{t'}$ = 0.167, s.e. = 0.056, $P = 1.4 \times 10^{-3}$). This result may be explained by the fact that smooth muscle in the lung is known to affect FEV1/FVC and influence pulmonary disease pathopysiology[71], and this unobserved causal tissue is likely highly co-regulated with the smooth muscle of the esophagus, which is indeed the site from which the GTEx study sampled the esophagus muscularis tissue[19]. Other newly significant findings are discussed in the **Supplementary Note**, and numerical results for all tissues and diseases/traits are reported in **Supplementary Table 8**.

We also performed a brain-specific analysis in which we applied TCSC to 41 brain traits (average *N* = 226K, **Supplementary Table 10**) while restricting to 13 individual GTEx brain tissues (**Supplementary Table 11**), analogous to previous work[7]. The 41 brain traits reflect a less stringent squared genetic correlation threshold of 0.25; we relaxed our threshold so that we would have a substantial number of brain traits to analyze, as many would were excluded

under the original threshold of 0.1. The 13 GTEx brain tissues were analyzed without merging tissues into meta-tissues, and irrespective of eQTL sample size (range: N = 101-189 individuals); we expected power to be limited due to the eQTL small sample sizes and substantial co-regulation among individual brain tissues. TCSC identified 8 brain tissue-brain trait pairs at 5% FDR (**Supplementary Figure 27, Supplementary Table 12**). For ADHD, TCSC identified brain hippocampus as a causal tissue ($\pi_{t'}$ = 0.127, s.e. = 0.045, $P = 2.5 \times 10^{-3}$), consistent with the correlation between hippocampal volume and ADHD diagnosis in children[72]. A recent ADHD GWAS identified a locus implicating the *FOXP2* gene[73], which has been reported to regulate dopamine secretion in mice[74]; hippocampal activation results in the firing of dopamine neurons[75]. For BMI, TCSC identified brain amygdala ($\pi_{t'}$ = 0.054, s.e. = 0.023, $P = 8.3 \times 10^{-3}$) and brain cerebellum ($\pi_{t'}$ = 0.039, s.e. = 0.016, $P = 7.0 \times 10^{-3}$) as causal tissues, consistent with previous work linking the amygdala to obesity and dietary self-control[76], although no previous study has implicated the amygdala in genetic regulation of BMI. As for brain cerebellum, previous research has implicated the cerebellar function in dietary behavior, rather than strictly regulation motor control function[67-69]. We note that the brain-specific analysis is expected to have greater power to identify tissue-trait pairs than the analysis of **Figure 4** due to the smaller number of total tissues in the model (as simulations show higher power for TCSC when there are fewer tagging tissues; **Supplementary Figure 9**). Other significant findings are discussed in the **Supplementary Note**, and numerical results for all brain tissues and brain traits analyzed are reported in **Supplementary Table 12**.

Comparisons of TCSC to other methods

We compared TCSC to two previous methods, RTC Coloc[2] and LDSC-SEG[7], that identify disease-critical tissues using gene expression data. RTC Coloc identifies disease-critical tissues based on tissue specificity of eQTL-GWAS colocalizations. LDSC-SEG identifies disease-critical tissues based on heritability enrichment of specifically expressed genes. We included RTC Coloc in these comparisons because it is the only other method that analyzes eQTL data and included LDSC-SEG because we believe it is the most widely used method. We note that RTC Coloc and LDSC-SEG analyze each tissue marginally, whereas TCSC jointly models contributions from each tissue to identify causal tissues (analogous to the distinction in GWAS between marginal association and fine-mapping[23]). Thus, we hypothesized that RTC Coloc and LDSC-SEG may output multiple highly statistically significant associated tissues for a given trait, whereas TCSC may output a single causal tissue with weaker statistical evidence of causality. To assess whether TCSC indeed attains higher specificity, we evaluated the results of each method both for causal tissues identified by TCSC and for the most strongly co-regulated tagging tissue (based on Spearman ρ for estimated eQTL effect sizes, averaged across genes, from ref.[19]). Our primary analyses focused on 7 traits with at least one tissue-trait association for each of the three methods (**Methods**).

Results for the 7 traits are reported in **Figure 5** and **Supplementary Table 13**; results for all 17 diseases/traits with causal tissue-trait associations identified by TCSC (**Figure 4**) are reported in **Supplementary Figure 28** and **Supplementary Table 14**, and complete results for all diseases/traits and tissues included in these comparisons are reported in **Supplementary Table**

15. We reached three main conclusions. First, for a given disease/trait, RTC Coloc typically implicates a broad set of tissues (not just strongly co-regulated tissues) (**Figure 5A**); for example, for WBC count, RTC Coloc implicated 8 of 10 tissues in **Figure 5**. This is consistent with our simulations, in which RTC Coloc suffered a high type I error rate and had a substantially lower AUC than TCSC (**Supplementary Figure 1**). Second, for a given disease/trait, LDSC-SEG typically implicates a small set of strongly co-regulated tissues (**Figure 5B**); for WBC count, LDSC-SEG implicated 3 of 8 tissues in **Figure 5**, consisting of whole blood and spleen (which are strongly co-regulated) plus breast tissue. This is consistent with our simulations, in which LDSC-SEG suffered a substantial type I error rate and had a substantially lower AUC than TCSC (**Supplementary Figure 1**). Third, for a given disease/trait, TCSC typically implicates one causal tissue (**Figure 5C**); for WBC count, TCSC implicated only whole blood as a causal tissue, with even the most strongly co-regulated tagging tissue reported as non-significant. This is consistent with our simulations, in which TCSC attained moderate power to identify causal tissues with approximately well-calibrated type I error. However, we caution that the higher specificity of TCSC in identifying unique causal tissues may be accompanied by incomplete power to identify secondary causal tissues; accordingly, we observed less significant (lower $-\log_{10}$P-value and lower $-\log_{10}$FDR) results for causal tissues in **Figure 5C** than in **Figure 5A** and **Figure 5B**) (**Supplementary Table 14**). We also observed similar patterns when comparing TCSC to RTC Coloc and LDSC-SEG in the brain-specific analysis of **Supplementary Figure 27** (**Supplementary Figure 29**, **Supplementary Table 16, Supplementary Note**). Based on simulations, we expect that RTC Coloc and LDSC-SEG both attain higher power at the cost of higher false positives.

Identifying tissue-specific contributions to the genetic covariance between two diseases/traits

We applied cross-trait TCSC to 262 pairs of disease/traits (**Supplementary Table 17**) and gene expression data for 48 GTEx tissues[19] (**Table 1**) (see **Data Availability**). Of 3,003 pairs of the 78 disease/traits analyzed above, the 262 pairs of diseases/traits were selected based on significantly nonzero genetic correlation ($p < 0.05 / 3{,}003$; see **Methods**). The 48 GTEx tissues were aggregated into 39 meta-tissues, as before (**Table 1** and **Methods**). We primarily report the signed proportion of genetic covariance explained by the *cis*-genetic component of gene expression in tissue t' ($\zeta_{t'} = \omega_{ge(t')} / \omega_g$), as well as its statistical significance (using per-trait FDR). We note that the direction of effect of tissue-specific contributions to the genetic covariance between two traits may be in the opposite direction of the global covariance between two traits, analogous to how local contributions to genome-wide genetic correlation may be in the opposite direction of the genome-wide genetic correlation[77-80].

TCSC identified 17 causal tissue-trait covariance pairs with significant contributions to trait covariance at 5% FDR, spanning 12 distinct tissues and 13 distinct trait pairs (**Figure 6A**, **Supplementary Table 18**). For 16 of the 17 causal tissue-trait covariance pairs, the causal tissue was non-significant for *both* constituent traits in the single-trait analysis of **Supplementary Table 8**. Findings that recapitulated known biology included both examples involving a tissue-trait pair that was significant in the single-trait analysis (marked by an underline in **Figure 6A**, **Figure 4**) and examples in which both tissue-trait pairs were non-significant in the single-trait

analysis (**Supplementary Table 6**). Consistent with the significant contribution of liver to LDL heritability in the single-trait analysis, TCSC identified a suggestive positive contribution of liver to the genetic covariance of LDL and total cholesterol ($\zeta_{t'}$ = 0.090, s.e. = 0.029, $P = 1.0 \times 10^{-3}$, 5% < FDR < 10%), and consistent with the positive contributions of whole blood to eosinophil count heritability and to platelet count heritability in the single-trait analysis, TCSC identified a significant positive contribution of whole blood to the genetic covariance of eosinophil count and platelet count ($\zeta_{t'}$ = 0.30, s.e. = 0.10, $P = 2.3 \times 10^{-3}$). TCSC also identified 15 suggestive tissue-trait covariance pairs with 5% < FDR < 10% (**Figure 6A**, **Supplementary Table 19**).

TCSC identified several biologically plausible findings not previously reported in the genetics literature. First, brain substantia nigra had a significantly positive contribution to the genetic covariance of BMI and red blood cell count (RBC count) ($\zeta_{t'}$ = 0.28, s.e. = 0.084, $P = 4.6 \times 10^{-4}$), while pancreas had a significantly negative contribution ($\zeta_{t'}$ = -0.25, s.e. = 0.079, $P = 8.7 \times 10^{-4}$). In the brain, energy metabolism is regulated by oxidation and previous work has shown that red blood cells play a large role in these metabolic processes as oxygen sensors[81]; in addition, previous studies have reported differences in the level of oxidative enzymes in red blood cells between individuals with high BMI and low BMI[82,83], suggesting that genes regulating oxidative processes might have pleiotropic effects on RBC count and BMI. In the pancreas, pancreatic inflammation (specifically acute pancreatitis) is associated with reduced levels of red blood cells, or anemia[84], while pancreatic fat is associated with metabolic disease and increased BMI[85]. Once again, the contrasting results for brain substantia nigra and pancreas suggest that genetic covariance may reflect distinct tissue-specific contributions. Second, brain substantia nigra had a significantly negative contribution to the genetic covariance of age at first birth and height ($\zeta_{t'}$ = -0.11, s.e. = 0.032, $P = 4.5 \times 10^{-4}$). Previous work in *C. elegans* reported that fecundity is positively regulated by dopamine[86,87], which is produced in the substantia nigra[88]. Therefore, it is plausible that reproductive outcomes related to fecundity, such as age at first birth, are also regulated by dopamine via the substantia nigra. Dopamine also plays a role in regulating the levels of key growth hormones such as IGF-1 and IGF-BP3[89] and has been previously shown to be associated with height[90]. Third, pituitary had a significantly negative contribution to the genetic covariance of vitamin D and WHR | BMI ($\zeta_{t'}$ = -0.19, s.e. = 0.057, $P = 4.5 \times 10^{-4}$). Irregularities in pituitary development are associated with decreased vitamin D levels and decreased IGF-1 levels, the latter of which is integral for bone development and is directly proportional to body proportion phenotypes such as WHR | BMI[91-93]. Fourth, LCLs had a suggestive negative contribution to the genetic covariance of eosinophil count and white blood cell count ($\zeta_{t'}$ = -0.081, s.e. = 0.028, $P = 1.8 \times 10^{-3}$, 5% < FDR < 10%, in contrast to the suggestive positive contribution of whole blood: $\zeta_{t'}$ = 0.32, s.e. = 0.12, $P = 2.4 \times 10^{-3}$, 5% < FDR < 10%). This is plausible as previous studies have reported the suppression of proliferation of lymphocytes (the white blood cell hematopoietic lineage from which LCLs are derived) by molecules secreted from eosinophils[94-96]. The contrasting results for whole blood and LCLs suggest that genetic covariance may reflect distinct tissue-specific contributions. Other significant findings are discussed in the **Supplementary Note**. Numerical results for all tissues and disease/trait pairs analyzed are reported in **Supplementary Table 19**.

As noted above, for 16 of the 17 causal tissue-trait covariance pairs, the causal tissue was non-significant for both constituent traits. We sought to formally assess whether differences in tissue-specific contributions to genetic covariance vs. constituent trait heritability were statistically significant. Specifically, for each causal tissue-trait covariance pair, we estimated the differences between the tissue-specific contribution to covariance ($\zeta_{t'}$) and the tissue-specific contributions to heritability for each constituent trait ($\pi_{t'}$) (and estimated standard errors by jackknifing differences across the genome). We note that $\zeta_{t'}$ and $\pi_{t'}$ are both signed proportions and are therefore on the same scale, thus the scenario in which these two quantities are equal is a natural and parsimonious null. We identified five tissue-trait covariance pairs for which these differences were statistically significant at 5% FDR for both constituent traits and $\pi_{t'}$ was non-significant for both constituent traits (marked by double asterisks in **Figure 6A**, **Supplementary Table 20**). For BMI and RBC count, negative contribution of pancreas (**Figure 6B**) and the positive contribution of brain substantia nigra (**Figure 6C**) to genetic covariance were each larger than the respective contributions of those tissues to BMI and RBC count heritability, which were non-significant. Other examples are discussed in the **Supplementary Note**. Numerical results for all tissues and trait pairs are reported in **Supplementary Table 20**. These findings were consistent with simulations we performed in which TCSC frequently detected tissue-specific contributions to covariance while failing to detect tissue-specific contributions to heritability for *both* traits, both in our original simulation framework and in a new simulation framework in which tissue-specific contributions to covariance were greater than contributions to heritability (**Supplementary Table 21**).

Discussion

We developed a new method, tissue co-regulation score regression (TCSC), that disentangles causal tissues from tagging tissues and partitions disease heritability (or genetic covariance of two diseases/traits) into tissue-specific components. We applied TCSC to 78 diseases and complex traits and 48 GTEx tissues, identifying 21 tissue-trait pairs (and 17 tissue-trait covariance pairs) with significant tissue-specific contributions. TCSC identified biologically plausible novel tissue-trait pairs, including associations of aorta artery with glaucoma, esophagus muscularis with FEV1/FVC, and heart left ventricle with platelet count. TCSC also identified biologically plausible novel tissue-trait covariance pairs, including a negative contribution of LCLs to the covariance of eosinophil count and white blood cell count (in contrast to the positive contribution of whole blood) and a positive contribution of brain substantia nigra and a negative contribution of pancreas to the covariance of BMI and red blood cell count; in particular, our findings suggest that genetic covariance may reflect distinct tissue-specific contributions.

TCSC differs from previous methods in jointly modeling contributions from each tissue to disentangle causal tissues from tagging tissues (analogous to the distinction in GWAS between marginal association and fine-mapping[23]). We briefly discuss several other methods that use eQTL or gene expression data to identify disease-associated tissues. RTC Coloc identifies disease-associated tissues based on tissue specificity of eQTL-GWAS colocalizations[2]; this study made a valuable contribution in emphasizing the importance of tissue co-regulation,

but did not model tissue-specific effects, such that RTC Coloc may implicate many tissues (**Figure 5A**). LDSC-SEG identifies disease-critical tissues based on heritability enrichment of specifically expressed genes[7]; this distinguishes a focal tissue from the set of all tissues analyzed, but does not distinguish closely co-regulated tissues (**Figure 5B**). MaxCPP models contributions to heritability enrichment of fine-mapped eQTL variants across tissues or meta-tissues[4]; although this approach proved powerful when analyzing eQTL effects that were meta-analyzed across all tissues, it has limited power to identify disease-critical tissues: fine-mapped eQTL annotations for blood (resp. brain) were significant conditional on annotations constructed using all tissues only when meta-analyzing results across a large set of blood (resp. brain) traits (Fig. 4 of ref.[4]). eQTLenrich compares eQTL enrichments of disease-associated variants across tissues[3]; this approach produced compelling findings for eQTL that were aggregated across tissues, but tissue-specific analyses often implicated many tissues (Fig. 1d of ref.[3]). MESC estimates the proportion of heritability causally mediated by gene expression in assayed tissues[97]; this study made a valuable contribution in its strict definition and estimation of mediated effects (see below), but did not jointly model distinct tissues and had limited power to distinguish disease-critical tissues (Fig. 3 of ref.[97]). CAFEH leverages multi-trait fine-mapping methods to simultaneously evaluate all tissues for colocalization with disease[5]; however, this locus-based approach does not produce genome-wide estimates and it remains the case that many (causal or tagging) tissues may colocalize with disease under this framework. Likewise, methods for identifying tissues associated to disease/trait covariance do not distinguish causal tissues from tagging tissues[98,99].

We note several limitations of our work. First, TCSC requires tissue-specific eQTL data (thus requiring genotype/gene expression data in substantial sample size), whereas some methods (LDSC-SEG[7], RolyPoly[6], and CoCoNet[9]) only require gene expression data in limited sample size. However, TCSC attains lower type I error and higher AUC than those methods in our simulations (**Supplementary Figure 1**); and its results are generally consistent in independent GWAS data (**Supplementary Table 5**), although all methods likely produce some false positives. Moreover, methods that only use gene expression data exclude contributions to disease from genes that are ubiquitously expressed but have cell-type-specific functionality or cell-type-specific genetic regulation such as transcription factors, which are widely believed to orchestrate large transcriptional programs important to disease[100]. Second, joint-fit effects of gene expression on disease may not reflect biological causality; if a causal tissue or cell type is not assayed[101], TCSC may identify a co-regulated tissue (e.g. a tissue whose cell type composition favors a causal cell type) as causal or may identify a set of co-regulated tissues that collectively tag the causal tissue as causal. We anticipate that this limitation will become less severe as potentially causal tissues, cell types and contexts are more comprehensively assayed. Third, TCSC does not achieve a strict definition or estimation of mediated effects; this is conceptually appealing and can, in principle, be achieved by modeling non-mediated effects, but may result in limited power to distinguish disease-critical tissues[97]. Fourth, TCSC has low power at small eQTL sample sizes; in addition, TCSC estimates are impacted by the number of significantly *cis*-heritable genes in a focal tissue, which can lead to conservative bias at small eQTL sample sizes. We anticipate that these limitations will become less severe as eQTL sample sizes increase. Fifth, TCSC is susceptible to large variations in eQTL sample size, which may

compromise type I error; therefore, there is a tradeoff between maximizing the number of tissues analyzed and limiting the variation in eQTL sample size. Sixth, TCSC assumes that causal gene expression-disease effects are independent across tissues; this assumption may become invalid for tissues and cell types assayed at high resolution. However, we verified via simulations that TCSC performs well when this model assumption is violated (**Supplementary Figures 17-20**). Seventh, TCSC does not formally model measurement error in tissue co-regulation scores, but instead applies a heuristic bias correction. We determined that the bias correction generally performs well in simulations. Eighth, TCSC does not produce locus-specific estimates or identify causal tissues at specific loci. However, genome-wide results from TCSC may be used as a prior for locus-based methods (analogous to GWAS fine-mapping with functional priors[102]). Ninth, TCSC performs less well in the presence of disease heritability that is not mediated through gene expression (**Supplementary Figures 21-22**). Tenth, we did not apply TCSC to single-cell RNA-seq (scRNA-seq) data, which represents a promising new direction as scRNA-seq sample sizes increase[103-105,35]; we caution that scRNA-seq data may require new eQTL modeling approaches[103]. Finally, we focused our cross-trait analyses on relatively independent traits from the single-trait analysis, to enable comparisons with single-trait results (**Figure 6B, 6C**); cross-trait analysis of more strongly genetically correlated traits is a future direction of high interest. Despite these limitations, TCSC is a powerful and generalizable approach for modeling tissue co-regulation to estimate tissue-specific contributions to disease.

Code Availability

TCSC software including a quick start tutorial: https://github.com/TiffanyAmariuta/TCSC/
Mancuso Lab TWAS Simulator: https://github.com/mancusolab/twas_sim.
FUSION software: http://gusevlab.org/projects/fusion/.

Data Availability

We have made 78 GWAS summary statistics and 41 brain-specific summary statistics publicly available at https://github.com/TiffanyAmariuta/TCSC/tree/main/sumstats, TWAS association statistics publicly available at https://alkesgroup.broadinstitute.org/TCSC/TWAS_sumstats/, tissue co-regulation scores publicly available at https://github.com/TiffanyAmariuta/TCSC/tree/main/coregulation_scores, and TCSC output publicly available at https://github.com/TiffanyAmariuta/TCSC/tree/main/results.

Acknowledgements

We thank Huwenbo Shi, Martin Zhang, and Benjamin Strober for helpful discussions. This work was funded by NIH grants U01 HG009379, R01 MH101244, R37 MH107649, R01 HG006399, R01 MH115676 and U01 HG012009.

Methods

TCSC regression

TCSC leverages the fact that the TWAS χ^2 statistic for a gene-tissue pair includes the direct effects of the gene on the disease as well as the tagging effects of co-regulated tissues and genes with shared eQTLs or eQTLs in LD. Thus, genes that are co-regulated across many tissues will tend to have higher χ^2 statistics than genes regulated in a single tissue. TCSC determines that a tissue causally contributes to disease if genes with high co-regulation to the tissue have higher TWAS χ^2 statistics than genes with low co-regulation to the tissue.

We model the genetic component of gene expression as a linear combination of SNP-level effects:

$$W_{jgt'} = \sum_m X_{jm}\beta_{gt'm}, \quad (3)$$

where W_{jgt} is the *cis*-genetic component of gene expression in individual *j* for gene *g* and tissue *t'*, X_{jm} is the standardized genotype of individual *j* for SNP *m*, and β_{gtm} is the standardized effect of the m^{th} SNP on the *cis*-genetic component of gene expression of gene *g* in tissue *t'*. We define the *cis*-genetic component of gene expression W_{jgt} to have mean 0 and variance 1 and $\beta_{gt'm}$ to have mean 0 and variance $\frac{1}{M_g}$, where M_g is the number of *cis* variants for gene *g*.

TCSC assumes that true gene-disease effects are identically distributed (i.i.d.) across genes and tissues while accounting for the fact that *cis*-genetic components of gene expression (and *cis*-genetic predictions of gene expression) are correlated[1] (see **Supplementary Figures 17-20** for simulations where gene expression-trait effect sizes are not i.i.d. across genes and tissues; TCSC performs well despite violations of model assumptions). The high correlation of *cis*-eQTLs across tissues leads to tagging from co-regulated tissues[2]. We model phenotype as a linear combination of genetic components of gene expression across genes in different tissues:

$$Y_j = \sum_{t'} \sum_g W_{jgt}\, \alpha_{gt'} + \epsilon_j, \quad (4)$$

where Y_j is the (binary or continuous-valued) phenotype of individual *j*, α_{gt} is the standardized effect size of the *cis*-genetic component of gene expression on disease and ϵ_j is the component of phenotype not explained by cis-genetic components of gene expression. We emphasize that we model disease as a function of the unobserved true *cis*-genetic component of gene expression $W_{jgt'}$, *not* the genetically predicted value $\widehat{W}_{jgt'}$ obtained from gene expression prediction models. Equation (4) can be rewritten in terms of SNP-level effects:

$$Y_j = \sum_i X_{ji}\beta'_i + \sum_{t'} \sum_g \sum_m X_{jm}\beta_{gt'm}\, \alpha_{gt} + \epsilon_j, \quad (5)$$

where β'_i are direct SNP-disease effects not mediated through gene expression.

We define the disease heritability explained by *cis*-genetic expression across all tissues as follows:

$$h^2_{ge} = var(\sum_{t\prime} \sum_g \sum_m X_{jm} \beta_{gt\prime m}\, \alpha_{gt}) \quad (6)$$

Because W_{jgt} has mean 0 and variance 1 and α_{gt} are assumed to be i.i.d. across genes and tissues (see above), Equation (5) implies that:

$$h^2_{ge} = \sum_{t\prime} \sum_g \alpha^2_{gt\prime}\,, \quad (7)$$

analogous to the relationship between SNP effect sizes and SNP-heritability[27]: $h^2_g = var(\sum_i X_j\, \beta_i)$. We emphasize that the respective terms in Equation (5) for each tissue t' are independent as $\alpha_{gt\prime}$ are assumed to be i.i.d. across genes and tissues. It follows that the disease heritability explained by a particular tissue t' is

$$h^2_{ge(t')} = var\left(\sum_g \sum_m X_{jm} \beta_{gt\prime m}\, \alpha_{gt\prime}\right), \quad (8)$$

which given that $W_{jgt\prime}$ has mean 0 and variance 1 and $\alpha_{gt\prime}$ is i.i.d. across genes, reduces to:

$$h^2_{ge(t')} = \sum_g \alpha^2_{gt\prime}. \quad (9)$$

Equation (7) and Equation (9) imply that $h^2_{ge} = \sum_{t\prime} h^2_{ge(t\prime)}$. Now, let $\alpha_{gt\prime}$ be a random variable drawn from a normal distribution with mean zero and tissue-specific variance $var(\alpha_{gt'}) = \tau_{t\prime}$. Then

$$h^2_{ge(t')} = \sum_g var(\alpha_{gt'}) = \tau_{t\prime} G_{t\prime}, \quad (10)$$

where $G_{t\prime}$ is the number of significantly *cis*-heritable genes in the model. In simulations, we demonstrate that when there are similar numbers of *cis*-heritable genes across tissues, setting $G_{t\prime}$ to the total number of unique *cis*-heritable genes produces unbiased estimates in TCSC for the causal tissue; however, when there are varying numbers of *cis*-heritable genes across tissues (fewer in the causal tissue), this produces upward biased estimates (**Supplementary Figures 3-4**) and thus setting $G_{t\prime}$ to the number of significantly *cis*-heritable genes in tissue t' is recommended. With this variance term, we can define a polygenic model that relates TWAS χ^2 statistics to co-regulation scores, which explicitly model the covariance structure of the χ^2 statistics. This strategy is analogous to modeling the dependence of GWAS χ^2 statistics on LD scores[27].

In a TWAS, the estimated value of the gene-disease effect size $\alpha_{gt'}$ is proportional to the correlation of the *cis*-genetic components of gene expression and their true gene-disease effect sizes for nearby genes across tissues, analogous to GCSC[22]:

$$E[\hat{\alpha}_{gt}] = \sum_{t'} \sum_{g'} r(\widehat{W}_{gt}, W_{g\prime t\prime}) \alpha_{g\prime t'} + \epsilon_g, \quad (11)$$

where $r(\widehat{W}_{gt}, W_{g't'})$ is the estimated correlation in *cis*-genetic predicted expression between gene *g in tissue t* and genes *g'* in tissue *t'*. ϵ_g is the component of phenotype not explained by *cis*-genetic components of gene expression, with mean 0 and variance σ_e^2 / N.

The value of the TWAS χ^2 is proportional to the squared estimated disease-gene effect size and the GWAS sample size *N* as follows:

$$\chi^2_{gt} = N\hat{\alpha}^2_{gt} \tag{12}$$

Using the equations (9) and (10), we can write the expectation of TWAS χ^2 as follows:

$$\mathrm{E}\left[\chi^2_{gt}\right] = \mathrm{E}\left[N\hat{\alpha}^2_{gt}\right] \tag{13}$$

$$= N\mathrm{E}\left[\left(\sum_{t'}\sum_{g'}\hat{r}_{gg't'}\,\alpha_{g't'} + \epsilon_g\right)^2\right] \tag{14}$$

$$= N\sum_{t'}\sum_{g'}\mathrm{E}\left[\hat{r}^2_{gg't'}\right]\mathrm{E}\left[\alpha^2_{g't'}\right] + N\mathrm{E}\left[\epsilon^2_g\right] \tag{15}$$

$$\approx N\sum_{t'}\sum_{g'}(r^2_{gg't'} + \tfrac{1}{N})h^2_{ge(t')}/G_{t'} + N\sigma_e^2/N \tag{16}$$

$$= N\sum_{t'}\sum_{g'}(r^2_{gg't'} + \tfrac{1}{N})\tau_{t'} + \sigma_e^2 \tag{17}$$

$$= N\sum_{t'}\sum_{g'}r^2_{gg't'}\tau_{t'} + \sum_{t'}\sum_{g'}\tau_{t'} + \sigma_e^2 \tag{18}$$

$$= N\sum_{t'}\sum_{g'}r^2_{gg't'}\tau_{t'} + \sum_{t'}var(\alpha_{g't'}) + \sigma_e^2 \tag{19}$$

$$= N\sum_{t'}l(g,t;t')\tau_{t'} + 1 \tag{20}$$

$$= N\sum_{t'}l(g,t;t')h^2_{ge(t')}/G_{t'} + 1 \tag{1}$$

To go from Equation (15) to Equation (16) we use the following relationship from the derivation of LDSC[13]:

$$\mathrm{E}\left[\hat{r}^2_{gg't'}\right] \approx r^2_{gg't'} + \frac{1}{N} \tag{21}$$

We go from Equation (19) to Equation (20) because the variance of the phenotype Y_j is $\sum_{t'}var(\alpha_{g't'}) + \sigma_e^2$ and is equal to one. We also introduce the notation that $\sum_{g'}r^2_{gg't'}$ are the tissue and gene co-regulation scores $l(g,t;\mathrm{t}')$, see below. We are interested in estimating $\tau_{t'}$, the per-gene disease heritability explained by the cis-genetic component of gene expression in tissue t'. From the derivation, the genome-wide tissue-specific contribution to disease heritability is estimated as

$$h^2_{ge(t')} = G_{t'}\tau_{t'}. \quad (22)$$

For the analysis of tissue-specific contributions to the covariance between two diseases, we can extend TCSC by using products of TWAS z-scores. Following the polygenic model described above, the expected product of TWAS z-scores in disease 1 and disease 2 for gene g and tagging tissue t is

$$E\left[z^1_{gt} \times z^2_{gt}\right] = \sqrt{N_1 N_2}\sum_{t'} l(g,t;t')\omega_{ge(t')}/G_{t'} + \rho N_s/\sqrt{N_1 N_2} \quad (23)$$

where N_1 is GWAS sample size for disease 1, N_2 is GWAS sample size for disease 2, t' indexes causal tissues, $l(g,t;t')$ are tissue co-regulation scores (see below), $\omega_{ge(t')}$ is the genetic covariance explained by the *cis*-genetic component of gene expression in tissue t', $G_{t'}$ is the number of significantly *cis*-heritable genes in tissue t' (see below), ρ is the phenotypic correlation between disease 1 and disease 2, and N_s is the number of overlapping GWAS samples between disease 1 and disease 2. The last term represents the intercept[28], and while we use a free intercept in the multivariate regression on co-regulation scores, the estimation of this term only plays a role in the estimation of regression weights (see below).

For estimates of $h^2_{ge(t')}$ and $\omega_{ge(t')}$, we use a free intercept; the estimation of $\rho N_s\sqrt{N_1 N_2}$ serves only to inform the heteroscedasticity weights (see below) and is not used in the multivariate TCSC regression to estimate $\omega_{ge(t')}$. To estimate standard errors, we use a genomic block jackknife over 200 genomic blocks with an equal number of genes in each. The standard deviation is computed as the square root of the weighted variance across the jackknife estimates (where the weight of each block is equal to the sum of the regression weights for the genes in that block) multiplied by 200 blocks. We expect that the jackknife standard error will be conservative relative to the empirical standard error across estimates due to variation in causal signal across loci[106].

Estimating tissue co-regulation scores and correcting for bias

We define the co-regulation score of gene g with tissues t and t' as

$$l(g,t;t') = \sum_{g'} r^2\left(\widehat{W}_{g,t}, W_{g',t'}\right), \quad (24)$$

where W denotes the *cis*-genetic component of gene expression for a gene-tissue pair across individuals, $\widehat{W}$ denotes the *cis*-predicted expression for a gene-tissue pair, and genes g' are within +/- 1 Mb of the focal gene g. TCSC corrects for bias in tissue co-regulation scores arising from differences between *cis*-genetic vs. *cis*-predicted expression (analogous to GCSC[22]). We apply bias correction to co-regulation scores in the special case when $g = g'$ and $t = t'$. While co-regulation scores aim to estimate $r^2\left(\widehat{W}_{g,t}, W_{g',t'}\right)$, the squared correlation of the predicted *cis*-genetic component of expression of gene g and tissue t (corresponding to the TWAS $\chi^2_{g,t}$ statistic) with the *actual cis*-genetic component of gene expression of gene g' in tissue t', when $g = g'$ and $t = t'$, the estimated value of $r^2\left(\widehat{W}_{g,t}, W_{g',t'}\right)$ will always equals one because the

estimate is based on $r^2\left(\widehat{W}_{g,t}, \widehat{W}_{g',t'}\right)$. However, this implies that predictions of the cis-genetic component of expression are perfectly accurate, which is unlikely to be the case. Therefore, the estimated value of $r^2\left(\widehat{W}_{g,t}, W_{g',t'}\right)$ if left to equal one will cause co-regulation scores to be systematically inflated.

Therefore, when $g = g'$ and $t = t'$, we set

$$r^2\left(\widehat{W}_{g,t}, W_{g',t'}\right) = R^2 / h^2_{GCTA}, \quad (25)$$

where R^2 is the cross-validation prediction statistic of the gene expression model for gene g in tissue t and h^2_{GCTA} is the GCTA-estimated *cis*-heritability of gene expression for gene g in tissue t. The quotient R^2 / h^2_{GCTA} is the accuracy of the gene expression prediction model, which reflects the upper bound on how much the *cis*-predicted expression can be correlated with the true *cis*-genetic component of gene expression. While we only consider genes with h^2_{GCTA} p < 0.01, the uncertainty in h^2_{GCTA} estimates should be modest and therefore not greatly impact our bias correction. We note that TWAS tests the null hypothesis that a specific weighted linear combination of SNPs is not associated with disease (and does not test the null hypothesis that the *cis*-genetic component of gene expression is not associated with disease).

TCSC regression weights

TCSC uses three sets of regression weights to increase power (analogous to GCSC[22]). The first regression weight is inversely proportional to $L(g,t)$, the total co-regulation score of each gene-tissue pair summed across tissues t':

$$L(g,t) = \sum_{t'} \sum_{g'} r^2\left(\widehat{W}_{g,t}, W_{g',t'}\right) \quad (26)$$

(without applying bias correction; see above), which allows TCSC to properly account for redundant contributions of co-regulated genes to TWAS χ^2 statistics.

The second regression weight is inversely proportional to $T(g,t)$, the number of tissues in which a gene is significantly *cis*-heritable:

$$T(g,t) = \sum_{t' \in g,t' significantly\ cis-heritable} 1, \quad (27)$$

thereby up-weighting signal from genes that are regulated in a limited number of tissues and preventing TCSC from attributing more weight to genes that are co-regulated across many tissues.

The third regression weight is inversely proportional to $H_{h^2}(g,t)$, the heteroscedasticity of χ^2 statistics, and is computed differently for estimates of $h^2_{ge(t')}$ than for estimates of $\omega_{ge(t')}$ (analogous to GCSC[22] and cross-trait LDSC[28], respectively).

For estimates of $h^2_{ge(t\prime)}$, we estimate $H_{h^2}(g,t)$ in two steps. First, we make a crude estimate of heritability explained by predicted expression (h^2_{ge}) as follows:

$$\mu_\chi = N\mu_L h^2_{ge} + 1, \tag{28}$$

where μ_χ is the mean χ^2 statistic:

$$\mu_\chi = \frac{\sum_{t\prime}\sum_{g\prime}\chi^2_{g\prime,t\prime}}{\sum_{t\prime}G_{t\prime}}, \tag{29}$$

where *N* is the GWAS sample size, g' iterates over significantly *cis*-heritable genes and t' iterates over tissues, and μ_L is the mean value of total co-regulation across tissues t',

$$\mu_L = \frac{\sum_{t\prime}\sum_{g\prime}L(g\prime,t\prime)}{\sum_{t\prime}G_{t\prime}}. \tag{30}$$

Then, we compute the heteroscedasticity for each significantly *cis*-heritable gene-tissue pair as

$$H_{h^2}(g,t) = \left(NL(g,t)h^2_{ge} + 1\right)^2. \tag{31}$$

Finally, we combine the three regression weights as follows:

$$Weight_{h^2}(g,t) = \frac{1}{L(g,t)T(g,t)H_{h^2}(g,t)}. \tag{32}$$

For estimates of $\omega_{ge(t')}$, we estimate $H_\omega(g,t)$ in two steps. First, we regress the products of TWAS z-scores on total tissue co-regulation scores, $L(g,t)$, using regression weights, $Weight_\omega(g,t)$, computed as follows:

$$Weight_\omega(g,t) = \frac{1}{L(g,t)T(g,t)H_\omega(g,t)} \tag{33}$$

where $H_\omega(g,t)$ is *first* estimated as follows:

$$H_\omega(g,t) = \left(N_1L(g,t)h^2_{ge}(trait\ 1) + 1\right)\left(N_2L(g,t)h^2_{ge}(trait\ 2) + 1\right) + \left(\frac{\sqrt{N_1N_2}\omega_{ge}L(g,t)}{\sum_{t\prime}\frac{G_{t\prime}}{T\prime}}\right)^2, \tag{34}$$

where $h^2_{ge}(trait\ 1)$ is the crude heritability estimate for trait 1 and $h^2_{ge}(trait\ 2)$ is the crude heritability estimate for trait 2, ω_{ge} is estimated as $\frac{\sum_{t\prime}\sum_{g\prime}z^{(1)}_{g\prime t\prime}z^{(2)}_{g\prime t\prime}}{\sqrt{N_1N_2}}$, N_1 is the sample size of the first GWAS, N_2 is the sample size of the second GWAS, and T' is the total number of tissues in the regression.

Second, we use the regression intercept to estimate the product ρN_s:

$$\rho N_s = intercept * \sqrt{N_1N_2}, \tag{35}$$

where ρ represents the phenotypic correlation between trait 1 and 2 and N_s represents the number of shared samples between GWAS 1 and 2. We also use the coefficient of the regression to update our estimate of ω_{ge}, such that we may update the heteroscedasticity weight as follows:

$$H_\omega(g,t) = \left(N_1L(g,t)h^2_{ge}(trait\ 1) + 1\right)\left(N_2L(g,t)h^2_{ge}(trait\ 2) + 1\right) +$$

$$\left(\frac{\sqrt{N_1N_2}\omega_{ge}L(g,t)}{\sum_{t'}\frac{G_{t'}}{T'}} + \rho N_s\sqrt{N_1N_2}\right)^2. \quad (36)$$

Finally, we combine the three regression weights as follows:

$$Weight_{\omega}(g,t) = \frac{1}{L(g,t)T(g,t)H_{\omega}(g,t)}. \quad (37)$$

Simulating TCSC

We employed a widely used TWAS simulation framework (Mancuso Lab TWAS Simulator, see **Code Availability**) to assess the power, bias, and calibration of TCSC in the presence of co-regulation across genes and tissues. We simulated a genome in which there are 1,000 protein-coding genes from chromosome 1, of which 100 (10%) are causal[31]. Each primary simulation consists of 10 tissues, of which at least one is causal, defined as having nonzero gene- disease effect sizes. We create a covariance structure among tissues mimicking empirical GTEx data. We use a previously published method to estimate the causal cross-tissue correlation of eQTL effect sizes which is 0.75[36]. We observe that not all GTEx tissues are equally correlated to one another. We estimate three different cross-tissue eQTL correlation quantities: (1) average correlation across all pairs of tissues = 0.75, (2) average correlation across similar tissues = 0.80, e.g. brain (13 in GTEx) or adipose (2 in GTEx) tissues, and (3) average correlation across dissimilar tissues, e.g. pairs of brain and adipose tissues = 0.74. To represent these biological modules, we let simulated tissues 1-3 have higher correlation of true eQTL effects to one another than to other tissues; likewise for tissues 4-6 and 7-10. We set covariance parameters, described below, such that the similar tissues had an average eQTL correlation of 0.789 across genes, dissimilar tissues have an average eQTL correlation of 0.737, and the average eQTL correlation across any pair of tissues is 0.751. We use real genotypes from European individuals in the 1000 Genomes Project to define the pairwise SNP LD structure which is used to simulate genotypes, gene expression traits, and complex traits/diseases. We simulate each gene having 5 true *cis*-eQTLs, based on the upper bound of empirical data from GTEx[19] and others[35], as well as the value used in other TWAS simulation methods[34]. Between pairs of co-regulated tissues, the same gene shares 3 *cis*-eQTLs. Between pairs of co-regulated genes in the same tissue, 3 *cis*-eQTLs are shared. The minimum allowed *cis*-heritability of a gene is 0.01 in our simulations. *Cis*-heritability is approximated as the sum of squared true *cis*-eQTL effect sizes, as done previously[22]. Effect sizes for the 3 shared eQTLs across tissues are sampled from a multivariate normal distribution with mean 0 and a variance-covariance matrix. We define the variance and covariance terms of this matrix such that (1) the proportion of genes detected as significantly *cis*-heritable by GCTA at a given sample size and (2) the average *cis* heritability of detected genes at a given sample size match empirical observations from GTEx data at sample sizes N = 100, 200, 300 and 500. As a result, the diagonal of the variance-covariance matrix, e.g. the variance term, is set to 0.075, and the off-diagonal elements are set to the product of the variance term and the desired correlation for each tissue pair, described above.

For each of 1,000 independent simulations per analysis, we simulate a GWAS (N = 10,000) by creating a complex trait which is the summation of the genetic components of causal gene expression (in the causal tissue). We use simulated genotypes based on the LD

structure of 1000 Genomes. Gene-disease effect sizes are drawn from a normal distribution with mean 0 and variance 1. In cross-trait TCSC analysis, effect sizes across genes between the two traits are correlated with default R_g = 0.5. To simulate a GWAS trait, we first compute the genetic component of each gene, which is the product of GWAS cohort genotypes and eQTL effects, such that we have 100 gene-specific traits. We then add noise to each gene-specific trait such that the total variance of the phenotype explained by the five eQTLs from the causal tissue is equal to a specified value; the value of $h^2_{ge(t)}$ in primary simulations is 10%. Then, we multiply each gene-specific trait by the causal gene-disease effect size, consistent with the additive generative model of gene-level effects on trait (see above). Finally, we take the sum across all gene-specific traits to make one complex trait, where the total variance of the trait explained by gene effects from the causal tissue is $h^2_{ge(t)}$, e.g. 10%.

We simulate an eQTL cohort of various gene expression sample sizes (N = 100, 200, 300, 500, 1000, 1500) using simulated genotypes based on the LD structure of 1000 Genomes. We simulate total gene expression in the eQTL cohort by adding a desired amount of noise to the genetic component of gene expression, e.g. the product of individual genotypes and true eQTL effect sizes, with variance equal to one minus the gene expression heritability, which is the sum of squared eQTL effects. Next, we fit gene expression prediction models by regressing the total gene expression on eQTL cohort genotypes of *cis* variants using lasso regularization, a standard approach used in TWAS. We define significantly *cis*-heritable genes as genes with GCTA heritability *P* value < 0.01[21] and heritability estimate > 0, and adjusted-R^2 > 0 in cross-validation prediction.

Then we estimate co-regulation scores at each different eQTL sample size by predicting gene expression into a cohort of 500 individuals, to approximate the size of the European sample of 1000 Genomes (*N* = 489). Using significantly *cis*-heritable genes from each tissue at a given sample size, we estimate gene and tissue co-regulation scores $l(g, t; t')$ as described above, including bias correction. In simulations, *cis* genes are defined as genes within the same 1 Mb block.

Then we apply TWAS to individual-level simulated GWAS data and gene expression prediction models. We predict gene expression into each of the 10,000 GWAS cohort individuals across all significantly *cis*-heritable genes for each tissue. We regress each complex trait on predicted gene expression to obtain TWAS z-scores. Finally, we run TCSC by regressing TWAS χ^2 statistics, or products of TWAS z-scores, on bias-corrected gene and tissue co-regulation scores.

Simulating other tissue-disease association methods

We simulated four tissue-trait association methods: RTC Coloc[2], LDSC-SEG[7], RolyPoly[6], or CoCoNet[9]. First, we simulated RTC Coloc method[2] by leveraging our existing TCSC simulation framework such that both methods could be compared via application to same simulated data. We used the same simulated GWAS cohort of 10,000 individuals as in our TCSC simulations and then followed the steps of the RTC Coloc method as published. Briefly, we perform a genome-wide association study using our simulated complex trait and the genotypes of our simulated GWAS cohort and select null variants with similar LD properties. Then, we simulate an eQTL cohort consisting of total gene expression and genotypes, using the same underlying true eQTL effect sizes as for TCSC simulations. Then, we perform colocalization analysis of GWAS variants

with eQTLs, across 10 tissues at 6 different eQTL sample sizes, to obtain the regulatory trait concordance (RTC) score. This is repeated for the set of null variants. Next, we perform colocalization analysis of eQTL variants between pairs of tissues to obtain tissue-sharing RTC scores, and similarly repeat this for null variants. GWAS-eQTL RTC scores are divided by tissue-sharing RTC scores summed across variants. Tissue-specific enrichment is computed as the ratio of this quotient to the null quotient. The enrichment *P* value is obtained using a Wilcox test comparing the values of the quotient to the values of the null quotient.

Second, we simulated the three methods that utilize GWAS data and total expression across tissues: LDSC-SEG[7], RolyPoly[6], and CoCoNet[9]. To this end, we retained the full GWAS summary statistics from the RTC Coloc analysis above. We separately simulated total expression across tissues in which the 100 causal genes in addition to 200 randomly selected genes were positively differentially expressed in the causal tissue and the two tagging tissues in the same simulated "module" as the causal tissue, e.g. with higher genetic correlation of gene regulatory effects. We also selected 100 random non-causal genes to be negatively differentially expressed in the causal tissue and the other two module tissues. For the remaining 7 tagging tissues, we randomly selected 300 genes to be positively differentially expressed, some of which at random will be causal genes, and let the remaining 700 genes be negatively differentially expressed. Then, as previously done[7], we calculated the t-statistics for the specific expression of each gene in each tissue. While we have modules of tissues that are more highly correlated to one another, these within-module tissues were excluded from the calculation of t-statistics, as previously done[7]. Finally, we created SNP-based annotations for each tissue, across 1000 simulations, and across 6 sample sizes, in which SNPs within +/- 100 kb of a specifically expressed gene is assigned a value of 1 and 0 otherwise, as previously done[7]. Then, we calculated LD scores and partitioned the heritability of our simulated complex traits. For the simulations of RolyPoly and CoCoNet, we installed the following R packages: rolypoly and CoCoNet and used the simulated data above to run each method. While CoCoNet does not technically use GWAS summary statistics, but rather gene-based "outcome variables", we used the label of causal or non-causal for each gene in each tissue of our simulations as the outcome variable.

Gene expression prediction models and tissue co-regulation scores in GTEx data

We downloaded GTEx v8 gene expression data for 49 tissues. We excluded tissues with fewer than 100 samples, e.g. kidney cortex (n = 69). We retained only European samples for each tissue, as labeled by GTEx via PCA of genotypes. We constructed gene expression models for two scenarios: (1) subsampling to 320 individuals including meta-analyzed tissues (**Table 1**) or (2) using all European samples per tissue. We recommend meta-analyzing gene expression prediction models across tissues in the case of tissues with low eQTL sample size (e.g. < 320 samples) and high pairwise genetic correlation (e.g. > 0.93). We determined in simulations that TCSC is sensitive to eQTL sample size differences, such that a tagging tissue with larger sample size than a causal tissue can produce false positive results; the subsampling approach was designed to mitigate this issue. For the subsampling procedure, we first set aside tissues with more than 320 samples; we chose 320 based on the average GTEx tissue sample size (*N* = 271) and robustness of TCSC in simulations at *N* = 300. Then, we grouped tissues with genetic correlation, e.g. marginal effect size correlation as reported by GTEx, with $R_g > 0.93$, an arbitrary

threshold that produced biologically plausible groups of related tissues, separating groups of brain tissues based on cranial compartment. We meta-analyzed gene expression prediction models for these grouped tissues in order to achieve a total sample size of 320 individuals where each tissue contributed an approximately equal number of samples, using an inverse-variance weighted meta-analysis across genes that were significantly *cis*-heritable in two or more constituent tissues. The prediction weights of genes that were significantly *cis*-heritable in a single constituent tissue were left unmodified.

To construct gene expression prediction models, we applied FUSION[21] (**Code Availability**) to individual-level GTEx data by regressing measured gene expression on genotypes of common variants (MAF > 0.05) and covariates provided by GTEx[19]. FUSION uses several different regression models: single eQTL, elastic net, lasso, and BLUP and the following covariates: sex, 5 genotyping principal components, PEER factors[107], and assay type. We defined significantly *cis*-heritable genes as protein-coding genes with GCTA heritability $p < 0.01$[21], heritability estimate > 0, and adjusted-$R^2 > 0$ in cross-validation prediction.

We used gene expression prediction models of significantly *cis*-heritable genes to predict expression into 489 European individuals from 1000 Genomes[108]. We then estimated tissue co-regulation scores using Equation (24) and Equation (25), where *cis*-predicted gene expression is used to estimate the *cis*-genetic component of gene expression.

GWAS summary statistics and TWAS association statistics

We collected GWAS summary statistics from 78 relatively independent heritable complex diseases and traits (average N = 302K) with heritability z-score > 6. We estimated the heritability of all summary statistics and genetic correlation of all pairs of summary statistics. We excluded traits with heritability z-score < 6, using S-LDSC with the baseline-LD v2.2 model[13,25,26] and as done previously[25]. We excluded one of each pair of traits that are *both* genetically correlated and have significantly overlapping samples. Specifically, for any pair of non-UK Biobank traits with an estimated sample overlap greater than the following threshold -- squared cross-trait LDSC intercept / (trait 1 S-LDSC intercept * trait 2 S-LDSC intercept) > 0.1[28] -- the trait with the larger SNP heritability z-score was retained. For any pair of UK Biobank traits with a squared genetic correlation > 0.1, the trait with the larger SNP heritability z-score was retained[38]. In total, this procedure resulted in 78 sets of relatively independent GWAS summary statistics. We limited all analyses (including cross-trait analyses) to the 78 relatively independent traits in order to avoid redundant findings across single-trait (and cross-trait) analyses. For the brain-specific analysis, we first selected brain-related diseases and complex traits, e.g. psychiatric disorders and behavioral phenotypes, excluding multi case-control studies and case vs case studies. Then, we applied our standard filters as described above, but relaxing the threshold of squared genetic correlation to 0.25.

We used FUSION[21] (**Code Availability**) to compute TWAS association statistics for each pair of signed GWAS summary statistics and each significantly *cis*-heritable gene-tissue pair, across the two scenarios described above. We further removed genes within the MHC (chromosome 6, 29 Mb - 33 Mb) and TWAS $\chi^2 > 80$ or $\chi^2 > 0.001N$, where N is the GWAS sample size, as previously used for quality control in the heritability analysis of GWAS summary statistics[13]. TCSC scales linearly with the number of genes and quadratically with the number of

tissues. After all input datasets are created and processed, running TCSC on a single real GWAS trait with 39 tissues takes about two minutes.

RTC Coloc and LDSC-SEG analysis of GWAS summary statistics and GTEx tissues

We downloaded supplementary tables for the RTC coloc method[2] and for LDSC-SEG[7]. For traits in our set of 78 GWAS summary statistics that were not analyzed by the LDSC-SEG study and for traits that are inherently brain-related (as these traits require a different procedure for generating tissue-specific gene sets), we ran LDSC-SEG ourselves. To this end, we downloaded LD scores for GTEx tissues and specifically expressed gene set SNP-level annotations (https://alkesgroup.broadinstitute.org/LDSCORE/LDSC_SEG_ldscores/) and ran LDSC-SEG as previously described[7]. For brain-related traits, we additionally ran a brain-specific analysis using LDSC-SEG, also as previously described[7]. Briefly, specifically expressed genes were determined via a t-test of the sentinel brain tissue against all other brain tissues, rather than against all other non-brain GTEx tissues, as done in the primary analysis of the LDSC-SEG study. For traits in our set that were not analyzed by the RTC Coloc study, of which there were few, we did not apply their method, as it was too computationally intensive to apply to real trait data.

Tables

Meta-tissue	Constituent Tissue(s)	Sample Size
Adipose Subcutaneous	Adipose Subcutaneous	320
Adipose Visceral Omentum	Adipose Visceral Omentum	320
Adrenal Gland	Adrenal Gland	200
Aorta Artery	Aorta Artery	320
† Brain Basal Ganglia	Putamen, Caudate, Nucleus Accumbens	320
† Brain Cereb.	Cerebellum, Cerebellar Hemisphere	320
† Brain Cortex	Frontal, Anterior, Cingulate	320
† Brain Limbic	Amygdala, Hippocampus, Hypothalamus	320
Brain Spinal Cord	Brain Spinal Cord	115
Brain Substantia Nigra	Brain Substantia Nigra	101
Breast Mammary Gland	Breast Mammary Gland	320
Coronary Artery	Coronary Artery	180
Cultured Fibroblasts	Cultured Fibroblasts	320
Esophagus Mucosa	Esophagus Mucosa	320
Esophagus Muscularis	Esophagus Muscularis	320
Heart Atrial Appendage	Heart Atrial Appendage	320
Heart Left Ventricle	Heart Left Ventricle	320
LCLs	LCLs	116
Liver	Liver	183
Lung	Lung	320
Minor Salivary Gland	Minor Salivary Gland	118
Muscle Skeletal	Muscle Skeletal	320
Ovary	Ovary	140
Pancreas	Pancreas	252
Pituitary	Pituitary	220
Prostate	Prostate	186
Skin (sun exposed)	Skin (sun exposed)	320
Skin (sun unexposed)	Skin (sun unexposed)	320
† Sigmoid Intestine	Sigmoid Colon, Gastroesophageal Junction	320
Spleen	Spleen	185
Stomach	Stomach	269
Tibial Artery	Tibial Artery	320
Tibial Nerve	Tibial Nerve	320
Testis	Testis	277
Thyroid	Thyroid	320
† Transverse Intestine	Transverse Colon, Small Intestine	320
Uterus	Uterus	108
Vagina	Vagina	122
Whole Blood	Whole Blood	320

Table 1. GTEx meta-tissues and constituent tissues analyzed. For each meta-tissue we list the constituent tissue(s) and total sample size. Daggers denote meta-tissues with more than one constituent tissue; for these meta-tissues, each constituent tissue has equal sample size up to rounding error (an exception is the transverse intestine meta-tissue, which includes 176 transverse colon samples and all 144 small intestine samples).

Figures

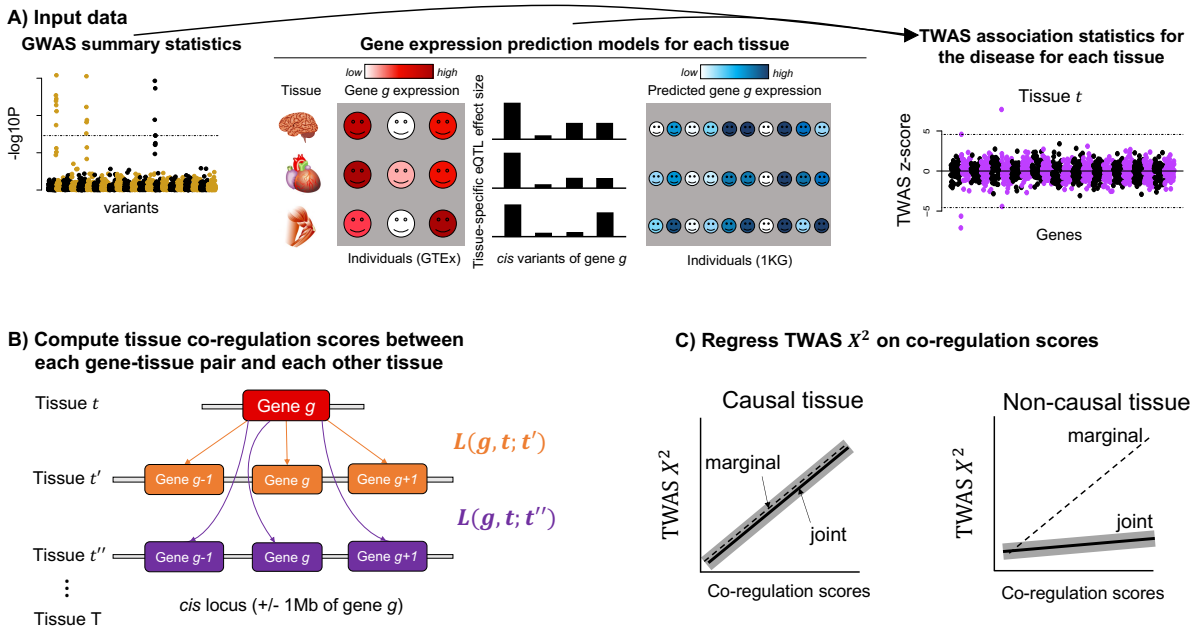

Figure 1. Overview of TCSC regression. (A) Input data to TCSC includes (1) GWAS summary statistics for a disease and (2) gene expression prediction models for each tissue, which are used to produce (3) TWAS summary statistics for the disease for each tissue. (B) TCSC computes tissue co-regulation scores $L(g,t;t')$ for each gene-tissue pair (g,t) with potentially causal tissues t'. (C) TCSC regresses TWAS chi-squares on tissue co-regulation scores to estimate tissue-specific contributions to disease. The shadow indicates the standard error of the TCSC estimate (joint models only).

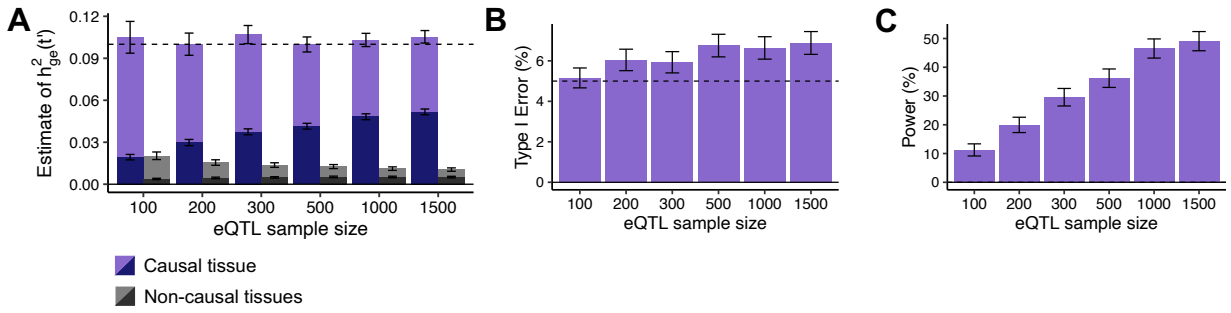

Figure 2. Robustness and power of TCSC regression in simulations. (A) Bias in estimates of disease heritability explained by the *cis*-genetic component of gene expression in tissue t' ($h^2_{ge(t\prime)}$) for causal (light and dark purple) and non-causal (light and dark gray) tissues, across 1,000 simulations per eQTL sample size. Light purple (resp. gray) indicates that $G_{t\prime}$ was set to the total number of true *cis*-heritable genes across tissues, dark purple (resp. gray) indicates that $G_{t\prime}$ was set to the number of significantly *cis*-heritable genes detected in each tissue. The dashed line indicates the true value of $h^2_{ge(t\prime)}$ for causal tissues. (B) Percentage of estimates of $h^2_{ge(t\prime)}$ for non-causal tissues that were significantly positive at $p < 0.05$, across 1,000 simulations per eQTL sample size. The type I error for TCSC ranged from 5.2% to 6.9%. In comparison, we observed type I errors from 53%-86% for RTC Coloc, 32%-33% for LDSC-SEG, 11%-12% for RolyPoly, and 32%-38% for CoCoNet (**Supplementary Figure 1**, **Supplementary Table 2**). (C) Percentage of estimates of $h^2_{ge(t\prime)}$ for causal tissues that were significantly positive at $p < 0.05$, across 1,000 simulations per eQTL sample size. Error bars denote 95% confidence intervals. Numerical results are reported in **Supplementary Table 1**.

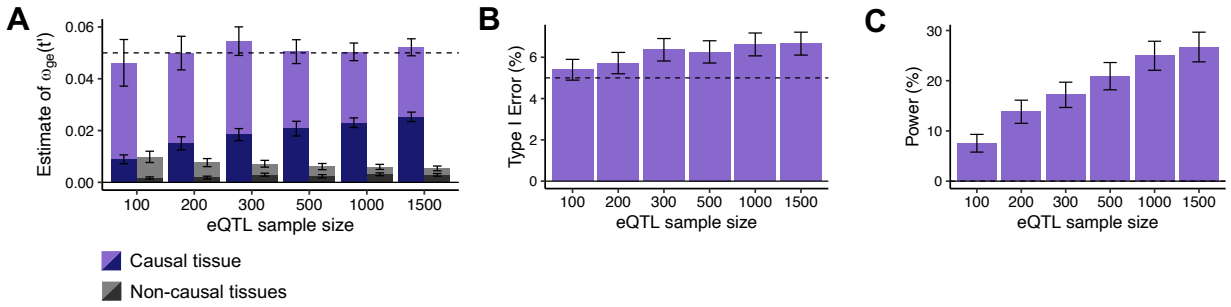

Figure 3. Robustness and power of cross-trait TCSC in simulations. (A) Bias in estimates of genetic covariance explained by the *cis*-genetic component of gene expression in tissue t' ($\omega_{ge(t\prime)}$) for causal (light and dark purple) and non-causal (light and dark gray) tissues, across 1,000 simulations per eQTL sample size. Light purple (resp. gray) indicates that $G_{t\prime}$ was set to the total number of true *cis*-heritable genes across tissues, dark purple (resp. gray) indicates that $G_{t\prime}$ was set to the number of significantly *cis*-heritable genes detected in each tissue. The dashed line indicates the true value of $\omega_{ge(t\prime)}$ for causal tissues. (B) Percentage of estimates of $\omega_{ge(t\prime)}$ for non-causal tissues that were significantly positive at $p < 0.05$, across 1,000 simulations per eQTL sample size. (C) Percentage of estimates of $\omega_{ge(t\prime)}$ for causal tissues that were significantly positive at $p < 0.05$, across 1,000 simulations per eQTL sample size. Error bars denote 95% confidence intervals. Numerical results are reported in **Supplementary Table 3**.

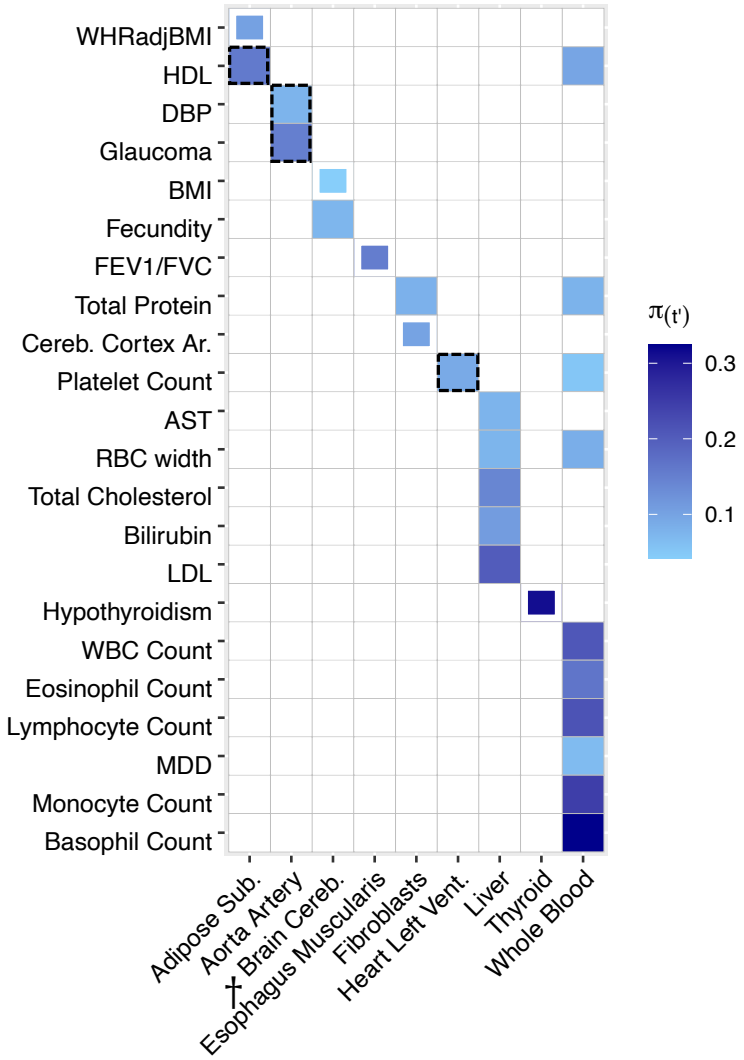

Figure 4. TCSC estimates tissue-specific contributions to disease and complex trait heritability. We report estimates of the proportion of disease heritability explained by the *cis*-genetic component of gene expression in tissue t' ($\pi_{t'}$). We report tissue-trait pairs with FDR of 10% or lower, where full boxes denote FDR of 5% or lower and partial boxes denote FDR between 5% and 10%. Dashed boxes denote results that are highlighted in the main text. Tissues are ordered alphabetically. Daggers denote meta-tissues with more than one constituent tissue. Diseases/traits are ordered with respect to causal tissues. Numerical results are reported in **Supplementary Tables 5 and Supplementary Table 6** (for all traits). WHRadjBMI: waist-hip-ratio adjusted for body mass index. HDL: high-density lipoprotein. DBP: diastolic blood pressure. BMI: body mass index. FEV1/FVC: forced expiratory volume in one second divided by forced vital capacity. Cereb. Cortex Ar.: cerebral cortex surface area. AST: aspartate aminotransferase. LDL: low-density lipoprotein. WBC Count: white blood cell count. MDD: major depressive disorder.

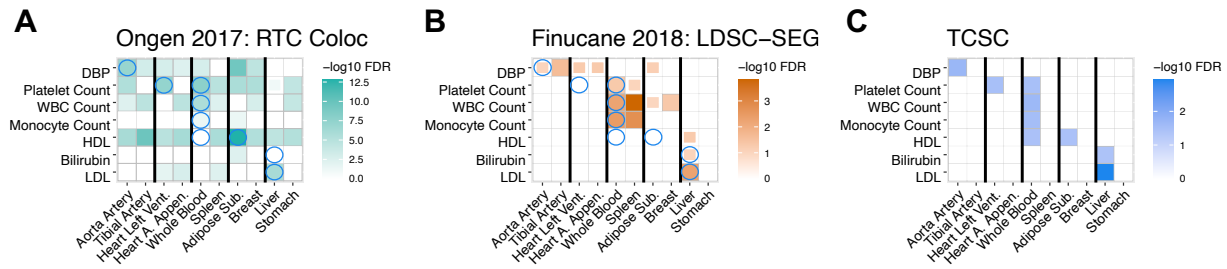

Figure 5. Comparison of disease-critical tissues identified by RTC Coloc, LDSC-SEG and TCSC. We report -$\log_{10}$FDR values for (A) RTC Coloc, (B) LDSC-SEG, (C) TCSC, across 7 traits with at least one significantly associated tissues (at FDR 5%) for each of the three methods and 10 tissues consisting of the causal tissues identified by TCSC and the most strongly co-regulated tagging tissues, ordered consecutively. We report tissue-trait pairs with FDR of 10% or lower, where full boxes denote FDR of 5% or lower and partial boxes denote FDR between 5% and 10%. Blue circles in panels (A) and (B) denote the causal tissue-trait pairs identified by TCSC. Numerical results are reported in **Supplementary Table 13**.

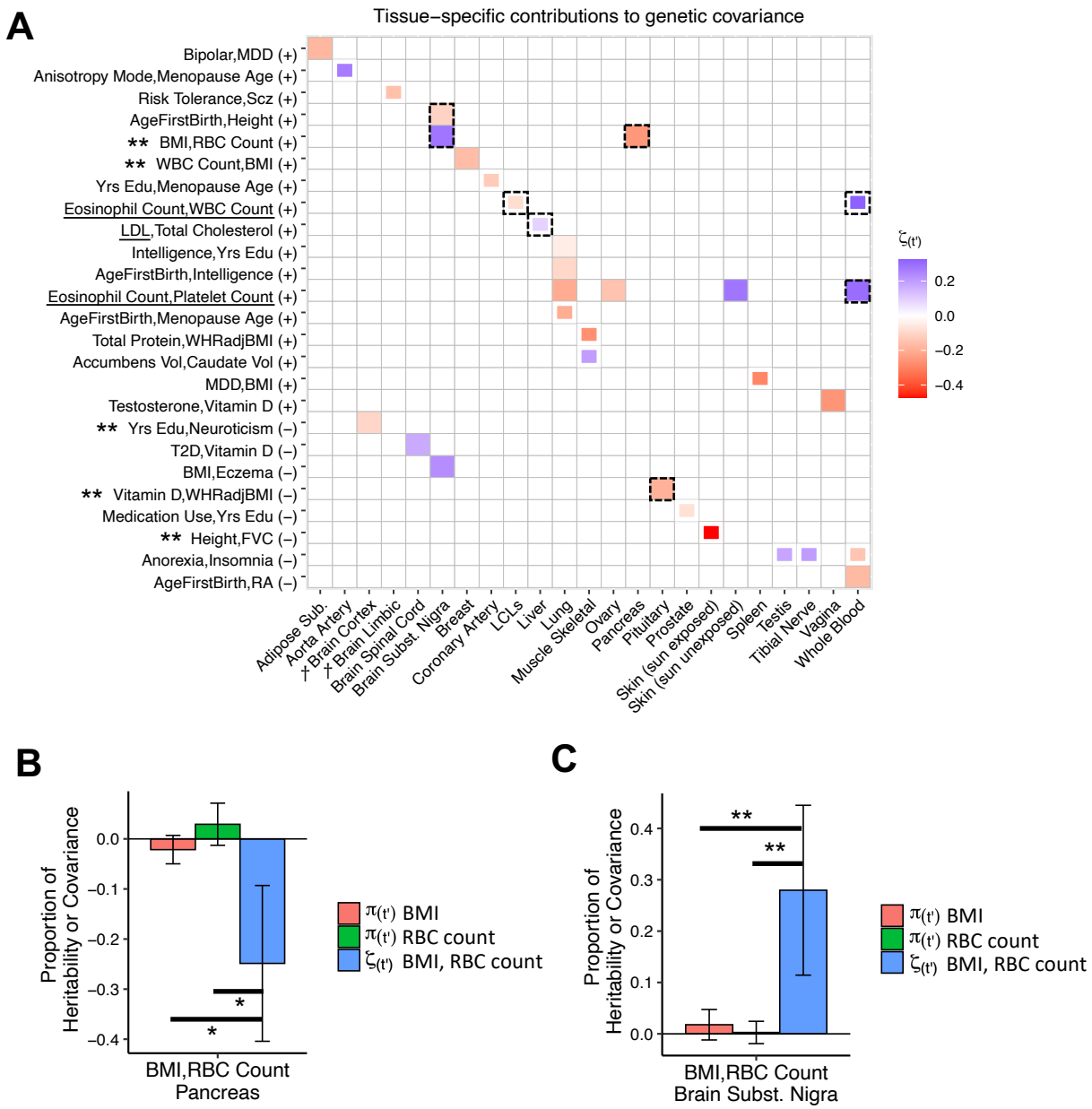

Figure 6. Cross-trait TCSC estimates tissue-specific contributions to the genetic covariance of two diseases/traits. (A) We report estimates of the signed proportion of genetic covariance explained by the *cis*-genetic component of gene expression in tissue t' ($\zeta_{t'}$). We report tissue-trait covariance pairs with FDR of 10% or lower, where full boxes denote FDR of 5% or lower and partial boxes denote FDR between 5% and 10%. Dashed boxes denote results that are highlighted in the main text. Tissues are ordered alphabetically. Daggers denote meta-tissues with more than one constituent tissue. Trait pairs are ordered by positive (+) or negative (-) genetic covariance, and further ordered with respect to causal tissues. Underlined traits are those for which TCSC identified a causal tissue in **Figure 4**: for eosinophil count, WBC count, and platelet count the causal tissue was whole blood, and for LDL the causal tissue was liver. Double asterisks denote trait pairs for which the differences between the tissue-specific contribution to

covariance and the tissue-specific contributions to heritability were significant for both constituent traits and the tissue-specific contributions to heritability were non-significant for both constituent traits. Numerical results are reported in **Supplementary Tables 18-19**. BMI: body mass index. RBC Count: red blood cell count. WBC Count: white blood cell count. LDL: low-density lipoprotein. Yrs Edu: years of education. WHRadjBMI: waist-hip-ratio adjusted for body mass index. Accumbens Vol: brain accumbens volume. Caudate Vol: brain caudate volume. MDD: major depressive disorder. Scz: Schizophrenia. T2D: type 2 diabetes. FVC: forced vital capacity. RA: rheumatoid arthritis. (B) For BMI and red blood cell count (RBC Count), we report estimates of the proportion of trait heritability for each trait and signed proportion of genetic covariance explained by the *cis*-genetic component of gene expression in pancreas. Lines with asterisks denote significant differences at 10% FDR between respective estimates, assessed by jackknifing the differences. (C) For BMI and red blood cell count (RBC Count), we report estimates of the proportion of trait heritability for each trait and proportion of genetic covariance explained by the *cis*-genetic component of gene expression in the brain substantia nigra. Lines with double asterisks denote significant differences at 5% FDR between respective estimates, assessed by jackknifing the differences. Numerical results are reported in **Supplementary Table 21**.

References

1. Hekselman, I. & Yeger-Lotem, E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat Rev Genet* **21**, 137-150 (2020).
2. Ongen, H. *et al.* Estimating the causal tissues for complex traits and diseases. *Nat Genet* **49**, 1676-1683 (2017).
3. Gamazon, E.R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat Genet* **50**, 956-967 (2018).
4. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat Genet* **50**, 1041-1047 (2018).
5. Arvanitis, M., Tayeb, K., Strober, B.J. & Battle, A. Redefining tissue specificity of genetic regulation of gene expression in the presence of allelic heterogeneity. *Am J Hum Genet* **109**, 223-239 (2022).
6. Calderon, D. *et al.* Inferring Relevant Cell Types for Complex Traits by Using Single-Cell Gene Expression. *Am J Hum Genet* **101**, 686-699 (2017).
7. Finucane, H.K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621-629 (2018).
8. Bryois, J. *et al.* Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. *Nat Genet* **52**, 482-493 (2020).
9. Shang, L., Smith, J.A. & Zhou, X. Leveraging gene co-expression patterns to infer trait-relevant tissues in genome-wide association studies. *PLoS Genet* **16**, e1008734 (2020).
10. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-5 (2012).
11. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* **45**, 124-30 (2013).
12. Pickrell, J.K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet* **94**, 559-73 (2014).
13. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).
14. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
15. Backenroth, D. *et al.* FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-Specific Functional Effects of Noncoding Variation: Methods and Applications. *Am J Hum Genet* **102**, 920-942 (2018).
16. Amariuta, T. *et al.* IMPACT: Genomic Annotation of Cell-State-Specific Regulatory Elements Inferred from the Epigenome of Bound Transcription Factors. *Am J Hum Genet* **104**, 879-895 (2019).
17. Boix, C.A., James, B.T., Park, Y.P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300-307 (2021).
18. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* **51**, 592-599 (2019).
19. Consortium, G.T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).

20. Gamazon, E.R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091-8 (2015).
21. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245-52 (2016).
22. Siewert-Rocks, K.M., Kim, S.S., Yao, D.W., Shi, H. & Price, A.L. Leveraging gene co-regulation to identify gene sets enriched for disease heritability. *Am J Hum Genet* **109**, 393-404 (2022).
23. Schaid, D.J., Chen, W. & Larson, N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491-504 (2018).
24. Zhu, H., Shang, L. & Zhou, X. A Review of Statistical Methods for Identifying Trait-Relevant Tissues and Cell Types. *Front Genet* **11**, 587887 (2020).
25. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat Genet* **49**, 1421-1427 (2017).
26. Gazal, S., Marquez-Luna, C., Finucane, H.K. & Price, A.L. Reconciling S-LDSC and LDAK functional enrichment estimates. *Nat Genet* **51**, 1202-1204 (2019).
27. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
28. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-41 (2015).
29. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
30. Mancuso, N. *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. *Nat Genet* **51**, 675-682 (2019).
31. Gazal, S. *et al.* Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat Genet* **54**, 827-836 (2022).
32. Hormozdiari, F. *et al.* Widespread Allelic Heterogeneity in Complex Traits. *Am J Hum Genet* **100**, 789-802 (2017).
33. Abell, N.S. *et al.* Multiple causal variants underlie genetic associations in humans. *Science* **375**, 1247-1254 (2022).
34. Li, Z. *et al.* METRO: Multi-ancestry transcriptome-wide association studies for powerful gene-trait association detection. *Am J Hum Genet* **109**, 783-801 (2022).
35. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
36. Liu, X. *et al.* Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues. *Am J Hum Genet* **100**, 605-616 (2017).
37. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
38. Gazal, S. *et al.* Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat Genet* **50**, 1600-1607 (2018).
39. Reshef, Y.A. *et al.* Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nat Genet* **50**, 1483-1493 (2018).
40. Zhang, M.J. *et al.* Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat Genet* **54**, 1572-1580 (2022).

41. Homan, T.D., Bordes, S. & Cichowski, E. Physiology, Pulse Pressure. in *StatPearls* (Treasure Island (FL), 2022).
42. Kass, M.A. *et al.* The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Arch Ophthalmol* **120**, 701-13; discussion 829-30 (2002).
43. Zhao, D., Cho, J., Kim, M.H. & Guallar, E. The association of blood pressure and primary open-angle glaucoma: a meta-analysis. *Am J Ophthalmol* **158**, 615-27 e9 (2014).
44. Levine, R.M., Yang, A., Brahma, V. & Martone, J.F. Management of Blood Pressure in Patients with Glaucoma. *Curr Cardiol Rep* **19**, 109 (2017).
45. De Moraes, C.G., Cioffi, G.A., Weinreb, R.N. & Liebmann, J.M. New Recommendations for the Treatment of Systemic Hypertension and their Potential Implications for Glaucoma Management. *J Glaucoma* **27**, 567-571 (2018).
46. Leeman, M. & Kestelyn, P. Glaucoma and Blood Pressure. *Hypertension* **73**, 944-950 (2019).
47. Gregg, D. & Goldschmidt-Clermont, P.J. Cardiology patient page. Platelets and cardiovascular disease. *Circulation* **108**, e88-90 (2003).
48. Coppinger, J.A. *et al.* Characterization of the proteins released from activated platelets leads to localization of novel platelet proteins in human atherosclerotic lesions. *Blood* **103**, 2096-104 (2004).
49. Gawaz, M., Langer, H. & May, A.E. Platelets in inflammation and atherogenesis. *J Clin Invest* **115**, 3378-84 (2005).
50. Davi, G. & Patrono, C. Platelet activation and atherothrombosis. *N Engl J Med* **357**, 2482-94 (2007).
51. Badimon, L., Padro, T. & Vilahur, G. Atherosclerosis, platelets and thrombosis in acute ischaemic heart disease. *Eur Heart J Acute Cardiovasc Care* **1**, 60-74 (2012).
52. Meadows, T.A. & Bhatt, D.L. Clinical aspects of platelet inhibitors and thrombus formation. *Circ Res* **100**, 1261-75 (2007).
53. Berman, M.N., Tupper, C. & Bhardwaj, A. Physiology, Left Ventricular Function. in *StatPearls* (Treasure Island (FL), 2022).
54. Chung, S., Sawyer, J.K., Gebre, A.K., Maeda, N. & Parks, J.S. Adipose tissue ATP binding cassette transporter A1 contributes to high-density lipoprotein biogenesis in vivo. *Circulation* **124**, 1663-72 (2011).
55. McGillicuddy, F.C., Reilly, M.P. & Rader, D.J. Adipose modulation of high-density lipoprotein cholesterol: implications for obesity, high-density lipoprotein metabolism, and cardiovascular disease. *Circulation* **124**, 1602-5 (2011).
56. Zoccali, C. *et al.* Adiponectin, metabolic risk factors, and cardiovascular events among patients with end-stage renal disease. *J Am Soc Nephrol* **13**, 134-141 (2002).
57. Ryo, M. *et al.* Adiponectin as a biomarker of the metabolic syndrome. *Circ J* **68**, 975-81 (2004).
58. Toth, P.P. Adiponectin and high-density lipoprotein: a metabolic association through thick and thin. *Eur Heart J* **26**, 1579-81 (2005).
59. Van Linthout, S. *et al.* Impact of HDL on adipose tissue metabolism and adiponectin expression. *Atherosclerosis* **210**, 438-44 (2010).

60. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187-196 (2015).
61. Emdin, C.A. *et al.* Genetic Association of Waist-to-Hip Ratio With Cardiometabolic Traits, Type 2 Diabetes, and Coronary Heart Disease. *JAMA* **317**, 626-634 (2017).
62. Smith, J., Al-Amri, M., Sniderman, A. & Cianflone, K. Leptin and adiponectin in relation to body fat percentage, waist to hip ratio and the apoB/apoA1 ratio in Asian Indian and Caucasian men and women. *Nutr Metab (Lond)* **3**, 18 (2006).
63. Farooqi, I.S. Defining the neural basis of appetite and obesity: from genes to behaviour. *Clin Med (Lond)* **14**, 286-9 (2014).
64. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).
65. Medic, N. *et al.* Increased body mass index is associated with specific regional alterations in brain structure. *Int J Obes (Lond)* **40**, 1177-82 (2016).
66. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649 (2018).
67. Onat, F. & Cavdar, S. Cerebellar connections: hypothalamus. *Cerebellum* **2**, 263-9 (2003).
68. Zhu, J.N. & Wang, J.J. The cerebellum in feeding control: possible function and mechanism. *Cell Mol Neurobiol* **28**, 469-78 (2008).
69. Low, A.Y.T. *et al.* Reverse-translational identification of a cerebellar satiation network. *Nature* **600**, 269-273 (2021).
70. Soler Artigas, M. *et al.* Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. *Nat Commun* **6**, 8658 (2015).
71. Chung, K.F. The role of airway smooth muscle in the pathogenesis of airway wall remodeling in chronic obstructive pulmonary disease. *Proc Am Thorac Soc* **2**, 347-54; discussion 371-2 (2005).
72. Plessen, K.J. *et al.* Hippocampus and amygdala morphology in attention-deficit/hyperactivity disorder. *Arch Gen Psychiatry* **63**, 795-807 (2006).
73. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet* **51**, 63-75 (2019).
74. Enard, W. *et al.* A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. *Cell* **137**, 961-71 (2009).
75. Floresco, S.B., Todd, C.L. & Grace, A.A. Glutamatergic afferents from the hippocampus to the nucleus accumbens regulate activity of ventral tegmental area dopamine neurons. *J Neurosci* **21**, 4915-22 (2001).
76. Kim, M.S. *et al.* Prefrontal Cortex and Amygdala Subregion Morphology Are Associated With Obesity and Dietary Self-control in Children and Adolescents. *Front Hum Neurosci* **14**, 563415 (2020).
77. Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet* **7**, e1002254 (2011).
78. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. *Am J Hum Genet* **101**, 737-751 (2017).

79. Zhang, Y. *et al.* SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. *Genome Biol* **22**, 262 (2021).
80. Werme, J., van der Sluis, S., Posthuma, D. & de Leeuw, C.A. An integrated framework for local genetic correlation analysis. *Nat Genet* **54**, 274-282 (2022).
81. Wei, H.S. *et al.* Erythrocytes Are Oxygen-Sensing Regulators of the Cerebral Microcirculation. *Neuron* **91**, 851-862 (2016).
82. Olusi, S.O. Obesity is an independent risk factor for plasma lipid peroxidation and depletion of erythrocyte cytoprotectic enzymes in humans. *Int J Obes Relat Metab Disord* **26**, 1159-64 (2002).
83. Ozata, M. *et al.* Increased oxidative stress and hypozincemia in male obesity. *Clin Biochem* **35**, 627-31 (2002).
84. Druml, W., Laggner, A.N., Lenz, K., Grimm, G. & Schneeweiss, B. Pancreatitis in acute hemolysis. *Ann Hematol* **63**, 39-41 (1991).
85. Sakai, N.S., Taylor, S.A. & Chouhan, M.D. Obesity, metabolic disease and the pancreas-Quantitative imaging of pancreatic fat. *Br J Radiol* **91**, 20180267 (2018).
86. Schafer, W.R. & Kenyon, C.J. A calcium-channel homologue required for adaptation to dopamine and serotonin in Caenorhabditis elegans. *Nature* **375**, 73-8 (1995).
87. Weinshenker, D., Garriga, G. & Thomas, J.H. Genetic and pharmacological analysis of neurotransmitters controlling egg laying in C. elegans. *J Neurosci* **15**, 6975-85 (1995).
88. Triarhou, L.C. Introduction. Dopamine and Parkinson's disease. *Adv Exp Med Biol* **517**, 1-14 (2002).
89. Zielonka, M. *et al.* Dopamine-Responsive Growth-Hormone Deficiency and Central Hypothyroidism in Sepiapterin Reductase Deficiency. *JIMD Rep* **24**, 109-13 (2015).
90. Comings, D.E. *et al.* The dopamine D2 receptor (DRD2) as a major gene in obesity and height. *Biochem Med Metab Biol* **50**, 176-85 (1993).
91. Ameri, P. *et al.* Interactions between vitamin D and IGF-I: from physiology to clinical practice. *Clin Endocrinol (Oxf)* **79**, 457-63 (2013).
92. Wacker, M. & Holick, M.F. Vitamin D - effects on skeletal and extraskeletal health and the need for supplementation. *Nutrients* **5**, 111-48 (2013).
93. Delecroix, C., Brauner, R. & Souberbielle, J.C. Vitamin D in children with growth hormone deficiency due to pituitary stalk interruption syndrome. *BMC Pediatr* **18**, 11 (2018).
94. Peterson, C.G., Skoog, V. & Venge, P. Human eosinophil cationic proteins (ECP and EPX) and their suppressive effects on lymphocyte proliferation. *Immunobiology* **171**, 1-13 (1986).
95. Nakagome, K. *et al.* IL-5-induced hypereosinophilia suppresses the antigen-induced immune response via a TGF-beta-dependent mechanism. *J Immunol* **179**, 284-94 (2007).
96. Onyema, O.O. *et al.* Eosinophils downregulate lung alloimmunity by decreasing TCR signal transduction. *JCI Insight* **4**(2019).
97. Yao, D.W., O'Connor, L.J., Price, A.L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat Genet* **52**, 626-633 (2020).
98. Leyden, G.M. *et al.* Harnessing tissue-specific genetic variation to dissect putative causal pathways between body mass index and cardiometabolic phenotypes. *Am J Hum Genet* **109**, 240-252 (2022).

99. Thom, C.S., Wilken, M.B., Chou, S.T. & Voight, B.F. Body mass index and adipose distribution have opposing genetic impacts on human blood traits. *Elife* **11**(2022).
100. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **175**, 598-599 (2018).
101. Umans, B.D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet* **37**, 109-124 (2021).
102. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat Genet* **52**, 1355-1363 (2020).
103. Nathan, A. *et al.* Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature* **606**, 120-128 (2022).
104. Perez, R.K. *et al.* Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
105. Soskic, B. *et al.* Immune disease risk variants regulate gene expression dynamics during CD4(+) T cell activation. *Nat Genet* **54**, 817-826 (2022).
106. Weissbrod, O. *et al.* Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat Genet* **54**, 450-458 (2022).
107. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* **6**, e1000770 (2010).
108. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).