Probabilistic fine-mapping of transcriptome-wide association studies

Nicholas Mancuso[1], Gleb Kichaev[2], Huwenbo Shi[2], Malika Freund[3], Claudia Giambartolomei[1], Alexander Gusev[4], and Bogdan Pasaniuc[1,2,3]

[1]Dept of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, 90024
[2]Bioinformatics Interdepartmental Program, University of California, Los Angeles, 90024
[3]Dept of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, 90024
[4]Dana-Farber Cancer Institute, Boston, 02215

Abstract

Transcriptome-wide association studies (TWAS) using predicted expression have identified thousands of genes whose locally-regulated expression is associated to complex traits and diseases. In this work, we show that linkage disequilibrium (LD) among SNPs induce significant gene-trait associations at non-causal genes as a function of the overlap between eQTL weights used in expression prediction. We introduce a probabilistic framework that models the induced correlation among TWAS signals to assign a probability for every gene in the risk region to explain the observed association signal. Our approach yields credible sets of genes containing the causal gene at a nominal confidence level (e.g., 90%) that can be used to prioritize and select genes for functional assays. Importantly, our approach remains accurate when expression data for causal genes are not available in the casual tissue by leveraging expression prediction from other tissues. We illustrate our approach using an integrative analysis of lipids traits where we correctly identify known causal genes.

Main

Transcriptome-wide association studies (TWAS) using predicted expression have been proposed as an approach to identify genes involved with complex traits and diseases.[1–3] Since TWAS based on predicted expression only relies on the genetic component of expression, it can be viewed as a test for non-zero local genetic correlation between expression and trait.[1,4] Significant genetic correlation is often mis-interpreted as an estimate of the effect of SNPs on trait mediated by the gene of interest. While enticing, this interpretation requires very strong assumptions that are likely violated in empirical data, due to pleiotropic effects of SNPs on trait mediated through other genes.[1–3,5–10] Therefore TWAS has been mostly utilized as a test of association to identify risk regions where eQTLs are likely to be involved in disease risk.

In this work, we show that TWAS gene-trait association statistics at a known risk region are correlated as a function of LD among SNPs and eQTL weights used in the prediction models. This effect is similar to LD-tagging in genome-wide association studies (GWAS) where LD within a region induces associations at tag SNPs (yielding the traditional Manhattan-style plots). Even in the simplest case where a single SNP causally impacts the expression of a gene which in turn causally impacts a trait, LD among SNPs used in the eQTL prediction models induce significant gene-trait associations at nearby non-causal genes in the region. The tagging effect is further exacerbated in the presence of multiple causal SNPs and genes. As an illustrative example consider a risk region with 6 genes where a single SNP is causal for a single gene which impacts trait (no other causal variants are present at this region, see Figure 1). Although genes 3 and 4 in Figure 1 have non-overlapping prediction weights due to different genetic regulation, LD among SNPs

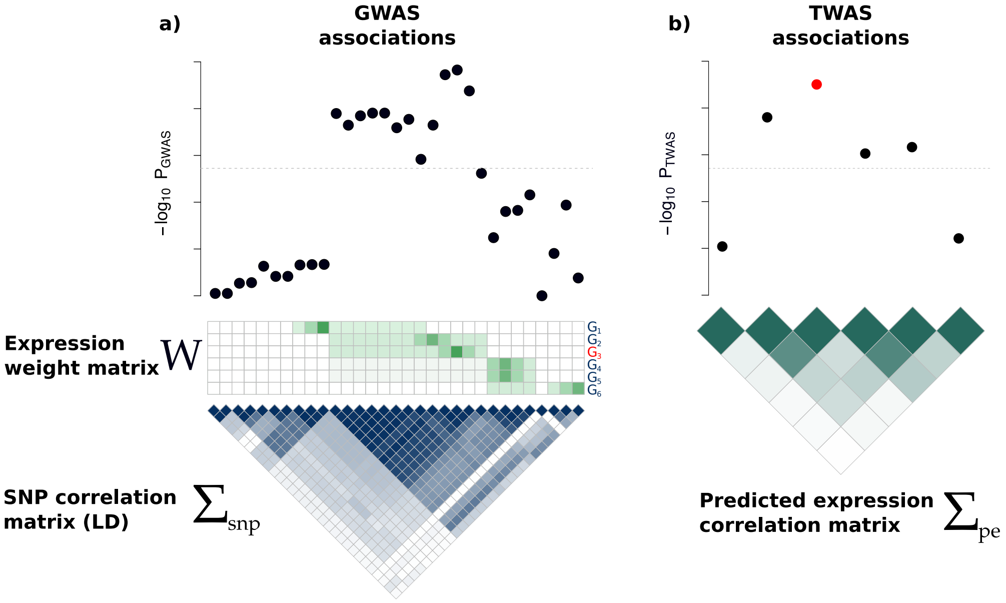
a)
GWAS associations
$-\log_{10} P_{GWAS}$
Expression weight matrix W
G_1
G_2
G_3
G_4
G_5
G_6
SNP correlation matrix (LD) $\sum_{snp}$
b)
TWAS associations
$-\log_{10} P_{TWAS}$
Predicted expression correlation matrix $\sum_{pe}$

Figure 1: Illustration of the induced correlation structure for predicted expression. a) Top: Manhattan plot indicating strength of SNP association with trait. Middle: Expression weight matrix for 6 genes in the same region, with the causal gene in red. Each row corresponds to a gene and each column represents a SNP. Color indicates magnitude of eQTL effect. Bottom: The correlation structure (LD) across SNPs. Darker color indicates stronger correlation. b) Top: TWAS Manhattan plot indicating strength of predicted expression association with trait. Bottom: Induced correlation of predicted exprssion. Darker color indicates stronger correlation between predicted expression levels. Dashed lines indicate the genome-wide (transcriptome-wide) significance threshold.

with non-zero prediction weights induce correlations in the TWAS statistics at genes 3 and 4. This effect is magnified when the expression weights coming from eQTL studies overlap either due to co-regulation or noise in the eQTL weight estimation procedure.

To disentangle between causal and tagging gene-trait associations at a TWAS significant region, we analytically derive the covariance structure among the TWAS statistics as function of LD and eQTL weights used in prediction. Next, we model the entire vector of marginal TWAS association statistics ($\mathbf{z}_{\text{twas}}$) at all genes in the region (TWAS significant and not-significant) using a multivariate Gaussian distribution parametrized by the effect sizes at causal genes ($\boldsymbol{\lambda}$) and the correlation structure induced by expression weights ($\mathbf{W}$) and LD ($\boldsymbol{\Sigma}_{\text{snp}}$) as

$$\mathbf{z}_{\text{twas}} \mid \boldsymbol{\lambda}, \mathbf{W}, \boldsymbol{\Sigma}_{\text{snp}} \sim \mathcal{N}(\mathbf{S}^{-1}\mathbf{W}^{\intercal}\boldsymbol{\Sigma}_{\text{snp}}\mathbf{W}\mathbf{S}^{-1}\boldsymbol{\lambda}, \mathbf{S}^{-1}\mathbf{W}^{\intercal}\boldsymbol{\Sigma}_{\text{snp}}\mathbf{W}\mathbf{S}^{-1}),$$

where $\mathbf{S}$ is a scaling factor (see Methods). To allow for genes without prediction models in the causal tissue (either due to QC and/or low power in eQTL studies), we include prediction models from proxy tissues for such genes (see below). We employ standard Bayesian approaches to compute the marginal posterior inclusion probability (PIP) for each gene in the region to be causal. To avoid overfitting, we integrate out the unknown causal effects $\boldsymbol{\lambda}$ using a a multivariate Gaussian prior (see Methods). Lastly, we use PIPs to compute ρ-credible gene sets that contain the causal gene with probability ρ.[11] Our approach, FOCUS (Fine-mapping Of CaUsal gene Sets), mirrors standard probabilistic GWAS fine-mapping approaches that yield ρ-credible SNP sets.[11–13]

To characterize the predicted expression correlation structure and to validate our framework, we used extensive simulations starting from real genotype data and eQTL weights (see Methods). Under null data where genes have eQTLs but do not impact downstream trait we find FOCUS adequately controls false positives (see Figure S1); for example, 90% credible sets contained the null model in 99% [95% CI 96 – 100%] of simulations. Next, we investigated simulations in which gene expression causally impacts trait (see Meth-

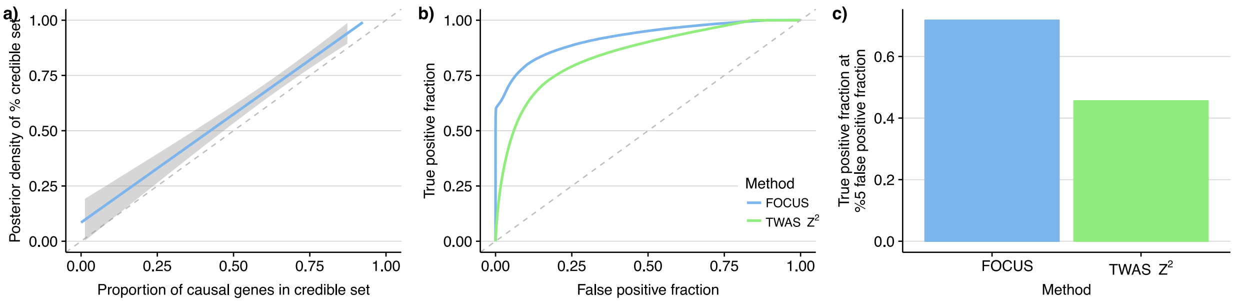

Figure 2: Causal Gene Performance in simulations. a) The total posterior density of computed credibile gene sets closely matches that of the actual proportion of causal genes captured. b) We estimated the empirical ROC curves for simulation results using the posterior probability for FOCUS and ranking based on TWAS χ^2 statistics. Here, "true positive fraction" refers to the fraction of underlying causal genes captured in the causal set. Similarly, "false positive fraction" refers to the fraction of non-causal genes falsely included in the causal set. c) Snapshot of the true positive fraction captured at 5% false positive fraction.

ods). FOCUS accurately captures the causal gene at all posterior densities for the credible set; for example, to capture 90% of causal genes, 47% genes need to be selected in the credible set when ranking on PIPs (see Figure 2). For completeness, we also compared with the simple gene rank from TWAS p-values. We find FOCUS outperforms TWAS p-value ranking in prioritizing and capturing underlying causal genes (see Figure 2); for example, at a false positive rate of 5%, FOCUS identifies ∼50% more causal genes than a simple rank of marginal TWAS statistics. We observed similar results when comparing with with COLOC,[14] a formal colocalization test (see Figure S2; see Supplementary Note). Taken together, we find that FOCUS accurately prioritizes causal genes under a variety of scenarios when expression mediates SNP effects on trait.

We next performed simulations in which we excluded the simulated causal gene from the analysis; we simulated TWAS associations using real eQTL weights estimated from multiple tissues and masked the causal gene (i.e. tissue-specific eQTL weights with non-zero effect on trait) from the analysis (see Methods). First, we confirmed that in this scenario nearby genes can show significant associations as function of LD and eQTL weights. As hypothesized, we found a strong relationship between the strength of LD and the average TWAS association ($P < 2 \times 10^{-16}$; see Figure S3). Next, we investigated the performance of FOCUS when the causal gene in the correct tissue is missing, but exists in alternative tissues. In real data a gene may act through a tissue that is difficult to assay in large sample sizes, but may have similar cis-regulatory patterns in tissues that are easier to collect (e.g., blood, LCLs). Indeed, several studies[1,4,15,16] established cis-regulated gene expression levels exhibit high genetic correlation across tissues and functional architectures. The intuition in this approach is that the loss in power from using the correlated tissue is offset by the gain in power due to its larger sample size. We found only a minor loss in power using proxy-tissue eQTL weights compared with causal-tissue weights, finding causal genes in the 90% credible set 68% of the time. Collectively, these results demonstrate that FOCUS is relatively robust to model perturbations and performs well when underlying tissue-specific causal genes are represented by proxy tissue eQTL weights.

Having validated our fine-mapping approach in simulations, we illustrate FOCUS by re-analyzing a large-scale GWAS of lipids measurements[17] with eQTL weights from adipose tissue. We assume the causal tissue for expression driving lipids is adipose given its well-understood role in lipids metabolism.[18] To account for missing gene prediction models, we incorporate gene expression models for genes not predictable from the adipose tissue across 47 reference panels measured from 45 tissues; in detail, for a gene without a predicted model in adipose tissue, we include the prediction model with best accuracy across all other tissues (see Table S1; see Methods). Multi-tissue TWAS identifies 449 (276 unique) significant genes at 142 (97 unique) independent 1Mb regions after accounting for the total number of per-trait tests performed ($P < 0.05/15, 276$; see Figures S4-S7; Table 1; Table S2). We applied FOCUS at the 142 1Mb TWAS risk regions (see Methods) to estimate credible sets of genes at each of the loci. First, as a positive control, we examined the 1p13 locus for LDL, as this region harbors risk SNP rs12740374 which has been shown to perturb transcription

Trait	TWAS risk regions	Overlapping gene models	TWAS genes	95% credible set genes	90% credible set genes
HDL	37	1040	133	137	90
LDL	36	930	102	106	89
TC	45	1219	133	160	123
TG	24	700	81	94	73

Table 1: Summary of identified TWAS risk regions. A risk region is defined to be a 1Mb interval overlapping a transcriptome-wide significant gene. The number of overlapping gene models counts the total number of models for predicted gene expression at TWAS risk regions. TWAS genes refers to the total number of transcriptome-wide significant genes at these regions (similarly for genes prioritized by fine mapping).

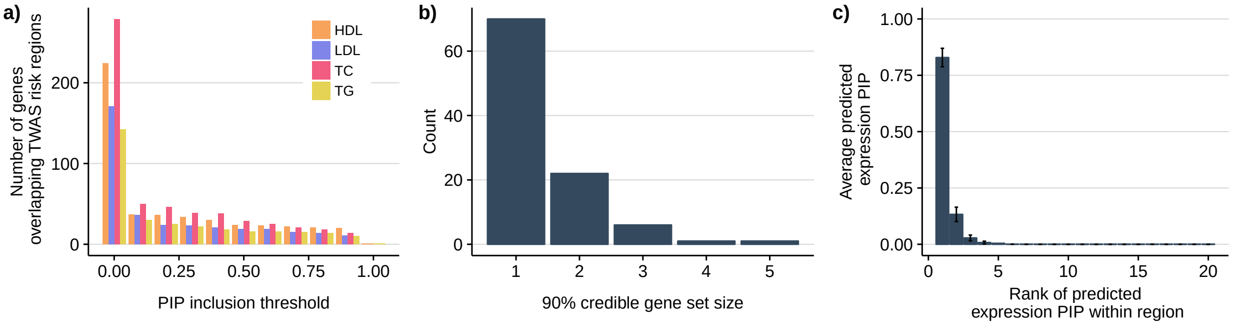

Figure 3: Fine mapping of lipids TWAS risk regions. a) Number of genes with predicted expression at TWAS risk regions for each lipids trait. b) The 90% credible gene set for most lipids TWAS risk regions contained a single gene. c) Average PIP across risk regions according to ranking within each credible set.

of *SORT1* and impact downstream LDL levels.[19] Reassuringly the 90% credible set contained 5 genes which included *SORT1* (see Table 2).

Next, we investigated whether the credible set analysis improved the resolution in causal gene identification on average. We observed 2.2 genes on average within the 90% credible set (see Figure 3; Table S3), a reduction from an average of 3.2 significant TWAS genes per region. For example, at locus 1p36.11 the 90% credible set for HDL contained only *ZDHHC18* compared with 5 TWAS significant genes. We found the average highest PIP across credible sets was ∼80% and decreased exponentially for lower ranked genes (see Figure 3). Interestingly, we also find instances in which the credible set does not contain the top gene. For example, at the region 19p13.11 for LDL the top TWAS gene is *GATAD2A* ($P_{\text{twas}} = 8.65 \times 10^{-13}$; PIP = 69.9%; a gene involved in transcriptional repression[20]), whereas the top fine mapped gene was *CTC-559E9.6* ($P_{\text{twas}} = 1.47 \times 10^{-8}$; PIP = 99.5%; a long intergenic non-protein coding RNA[21]). Computing marginal posterior inclusion probabilities enables us to estimate the expected number of causal genes for lipids traits at risk regions. We applied this to all lipids risk regions and found 1.3 genes on average, which suggests that most regions can be explained by a single gene affecting downstream lipids levels (see Figure S8). We note that we observe a long tail with in the distribution of expected causal genes, with 21% regions harboring greater than 2 causal genes in expectation. Next, we investigated regions whose 90% credible sets contained the null model (i.e. regions with weaker evidence for models of gene expression driving risk). We found only 4/142 instances of the null model captured in credible sets (see Table S3), which suggests the majority of signal at these regions can be explained due to cis-regulated expression levels.

Gene	Marginal PIP	z_{twas}	Expression Reference
RP5-1160K1.3*	> 0.99	$9.31^\dagger$	Proxy (GTEx:Exposed Lower leg)
GNAI3*	> 0.99	$-6.31^\dagger$	Proxy (YFS:Whole Blood)
CELSR2*	> 0.99	$10.43^\dagger$	Adipose (METSIM)
SORT1*	> 0.99	$-9.46^\dagger$	Proxy (NTR:Whole Blood)
PSRC1*	0.93	$-4.70^\dagger$	Adipose (METSIM)
RP11-20O24.4	0.065	$-7.91^\dagger$	Proxy (GTEx:Prostate)
AMIGO1	< 0.01	0.72	Adipose (METSIM)
ATXN7L2	< 0.01	$7.47^\dagger$	Proxy model (GTEx:Testis)
SYPL2	< 0.01	0.71	Adipose (METSIM)
TMEM167B	< 0.01	4.49	Proxy (CMC:Brain)
AC000032.2	< 0.01	$5.27^\dagger$	Adipose (GTEx)
PSMA5	< 0.01	3.04	Proxy (YFS:Whole Blood)
WDR47	< 0.01	$5.01^\dagger$	Proxy (GTEx:Nerve Tibial)
CYB561D1	< 0.01	-0.29	Proxy (YFS:Whole Blood)
TAF13	< 0.01	-0.55	Adipose (METSIM)
GSTM5	< 0.01	-0.90	Adipose (METSIM)
RP5-1065J22.8	< 0.01	-4.02	Adipose (GTEx)
KIAA1324	< 0.01	4.44	Adipose (METSIM)
AMPD2	< 0.01	0.85	Proxy (GTEx:Testis)
SARS	< 0.01	-0.075	Adipose (METSIM)
GSTM3	< 0.01	2.93	Adipose (METSIM)
CLCC1	< 0.01	1.86	Proxy (CMC:Brain)
GSTM4	< 0.01	-4.65	Adipose (METSIM)
GSTM1	< 0.01	3.46	Adipose (METSIM)
SCARNA2	< 0.01	-3.67	Proxy (GTEx:Skin - Sun Exposed Lower leg)
RP4-735C1.4	< 0.01	-1.35	Adipose (GTEx)
GSTM2	< 0.01	3.29	Adipose (METSIM)

Table 2: Fine mapping of LDL at 1p13. Tissue reference refers to the expression panel used to train predictive models of gene expression. z_{twas} is the association strength with LDL levels. Marginal posterior inclusion probabilities are the non-normalized PIP values for each gene. * indicates gene in 90% credible set. TWAS Z-scores are the association strength under the TWAS test. † indicates gene is transcriptome-wide significant at $P_{\text{twas}} < 0.05/15,276$.

Discussion

In this work we presented FOCUS, a fine-mapping approach that estimates credible sets of causal genes using prediction eQTL weights, LD, and TWAS/GWAS summary statistics. We demonstrated FOCUS adequately controls false positives in null simulations and outperforms naive p-value ranking in identifying causal genes when genes at a region impact downstream trait. We applied FOCUS to four lipids TWASs (e.g., HDL, LDL, triglyceride, and total cholesterol levels) and found *SORT1* correctly identified as a putative causal gene. Interestingly, our real-data results in lipids suggests most regions can be explained by a single causal gene, but significant number of "hotspot" risk regions where multiple causal genes may be influencing trait. Overall, our results highlight the utility of using credible sets in prioritizing causal genes by jointly assigning posterior probabilities, that are both easily interpretable and comparable across genes and regions.

In addition to providing a quantification of the confidence in how many genes need to be validated to identify the causal genes in the region, our probabilistic approach yields several benefits. First, FOCUS naturally allows for multiple causal SNPs and genes while integrating gene-effect sizes using conjugate priors; this is particularly important as recent works have shown that allelic heterogeneity (i.e. multiple causal genes and SNPs at a region) is pervasive in both eQTL and GWAS.[16,22] Second, our approach only requires summary association statistics from linear predictive models associated with complex trait or disease. In this work, we investigate predicted gene expression, but FOCUS could generally be applied to other predicted-molecular traits with an established role in complex trait etiology (e.g., alternatively spliced exons[23,24]). For example, several recent works have supporting evidence for splice variation playing an important role in driving risk of schizophrenia.[25,26]

We showed our approach is well calibrated under various null simulations and robust to perturbations in model assumptions; however, several limitations still exist. First, our model assumes that complex trait or disease risk is a linear function of expression levels as causal genes. Several works have demonstrated that risk prediction using estimated or observed expression levels can outperform standard SNP-based models,[25,27] which supports a linear model of gene expression impacting complex trait or disease risk. However, higher-order models that capture complex regulatory networks of transcription factors and gene expression may also reflect underlying biology. As reference gene expression datasets grow in size, accurately modelling these assumptions may be possible. Second, when the causal gene is untyped in the data, our approach will inflate the posterior probabilities at tagging genes. We attempt to alleviate this scenario by incorporating gene models measured in proxy tissues. Third, we took a tissue-prioritizing approach by preferentially using eQTL weights in adipose tissue given its known role in lipids metabolism[18] for our real-data analysis. This approach may not always be possible for complex traits or diseases with less understood biology. However, recent work has shown that the most relevant (i.e. causal) tissue for complex traits can be accurately estimated using eQTL data.[28] Despite our modelling assumptions and limitations, our approach is a step towards more accurately prioritizing gene sets through our credible set notion.

Methods

Sampling distribution of marginal TWAS summary statistics

Here we describe the sampling distribution of marginal Z-scores obtained from TWAS. Let expression levels for n individuals at m genes $\mathbf{G} \in \mathbb{R}^{n \times m}$ be defined as a linear function of genotype and environment, which is given by $\mathbf{G} = \mathbf{X}\mathbf{W} + \mathbf{E}$, where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the centered and standardized genotype matrix at p SNPs, $\mathbf{W} \in \mathbb{R}^{p \times m}$ is the eQTL weight matrix, and $\mathbf{E} \in \mathbb{R}^{n \times m}$ is environmental noise. In practice weights $\mathbf{W}$ are unknown and must be estimated (e.g., BSLMM,[29] GBLUP[30,31]) Let predicted expression be defined as $\widehat{\mathbf{G}} = \mathbf{X}\widehat{\mathbf{W}}$ when $\widehat{\mathbf{W}}$ is estimated from data. We standardize $\widehat{\mathbf{G}}$ to have unit variance, represented by $\widetilde{\mathbf{G}} = \widehat{\mathbf{G}}\mathbf{S}^{-1}$, where $\mathbf{S} = \text{diag}(\|\widehat{\mathbf{G}}_1\|, \ldots, \|\widehat{\mathbf{G}}_m\|)$. We model complex trait as linear combination of predicted expression at m genes and an environmental component as $\mathbf{y} = \widetilde{\mathbf{G}}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$ where $\boldsymbol{\alpha}$ is the vector of causal effects at these m genes, and $\boldsymbol{\epsilon} \sim f(0, \sigma_e^2 \mathbf{I}_n)$ is random environmental noise. For notational simplicity we drop the hat from $\widehat{\mathbf{W}}$ and use $\mathbf{W}$ throughout.

We compute the marginal association z_i of gene i on trait $\mathbf{y}$ through a transcriptome-wide association study as,

$$\begin{aligned} z_i &= \frac{1}{\sigma_e\sqrt{n}}\widetilde{\mathbf{G}}_i^{\intercal}\mathbf{y} = \frac{1}{\sigma_e\sqrt{n}}(\mathbf{X}\mathbf{W}\mathbf{S}^{-1})_i^{\intercal}\mathbf{y} = \frac{1}{\sigma_e\sqrt{n}}\mathbf{S}_i^{-1}\mathbf{W}^{\intercal}\mathbf{X}^{\intercal}\mathbf{y} = \frac{1}{\sigma_e\sqrt{n}}\mathbf{S}_i^{-1}\mathbf{W}^{\intercal}\mathbf{X}^{\intercal}(\widetilde{\mathbf{G}}\boldsymbol{\alpha} + \boldsymbol{\epsilon}) \\ &= \frac{1}{\sigma_e\sqrt{n}}\left[\mathbf{S}_i^{-1}\mathbf{W}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\mathbf{W}\mathbf{S}^{-1}\boldsymbol{\alpha} + \mathbf{S}_i^{-1}\mathbf{W}^{\intercal}\mathbf{X}^{\intercal}\boldsymbol{\epsilon}\right] = \frac{\sqrt{n}}{\sigma_e}\mathbf{S}_i^{-1}\mathbf{W}^{\intercal}\boldsymbol{\Sigma}_{\text{snp}}\mathbf{W}\mathbf{S}^{-1}\boldsymbol{\alpha} + \frac{1}{\sigma_e\sqrt{n}}\mathbf{S}_i^{-1}\mathbf{W}^{\intercal}\mathbf{X}^{\intercal}\boldsymbol{\epsilon}. \end{aligned}$$

where $\boldsymbol{\Sigma}_{\text{snp}} = n^{-1}\mathbf{X}^{\intercal}\mathbf{X}$ is the SNP correlation (LD) matrix. The marginal association statistics for all m genes are determined by,

$$\mathbf{z}_{\text{twas}} = \frac{\sqrt{n}}{\sigma_e}\mathbf{S}^{-1}\mathbf{W}^{\intercal}\boldsymbol{\Sigma}_{\text{snp}}\mathbf{W}\mathbf{S}^{-1}\boldsymbol{\alpha} + \frac{1}{\sigma_e\sqrt{n}}\mathbf{S}^{-1}\mathbf{W}^{\intercal}\mathbf{X}^{\intercal}\boldsymbol{\epsilon}.$$

Assuming weights $\mathbf{W}$ and causal gene effects $\boldsymbol{\alpha}$ are fixed, we can compute the expectation and variance of the association statistics as,

$$\begin{aligned} \mathbb{E}[\mathbf{z}_{\text{twas}} \mid \mathbf{W}] &= \mathbb{E}[\frac{\sqrt{n}}{\sigma_e}\mathbf{S}^{-1}\mathbf{W}^{\intercal}\boldsymbol{\Sigma}_{\text{snp}}\mathbf{W}\mathbf{S}^{-1}\boldsymbol{\alpha} \mid \mathbf{W}] + \mathbb{E}[\frac{1}{\sigma_e\sqrt{n}}\mathbf{S}^{-1}\mathbf{W}^{\intercal}\mathbf{X}^{\intercal}\boldsymbol{\epsilon}] = \frac{\sqrt{n}}{\sigma_e}\mathbf{S}^{-1}\mathbf{W}^{\intercal}\boldsymbol{\Sigma}_{\text{snp}}\mathbf{W}\mathbf{S}^{-1}\boldsymbol{\alpha} \\ \mathbb{V}[\mathbf{z}_{\text{twas}} \mid \mathbf{W}] &= \frac{1}{\sigma_e^2 n}\mathbf{S}^{-1}\mathbf{W}^{\intercal}\mathbf{X}^{\intercal}\mathbb{V}[\boldsymbol{\epsilon}]\mathbf{X}\mathbf{W}\mathbf{S}^{-1} = \mathbf{S}^{-1}\mathbf{W}^{\intercal}\boldsymbol{\Sigma}_{\text{snp}}\mathbf{W}\mathbf{S}^{-1}. \end{aligned}$$

To simplify notation we re-parameterize the causal effects as a non-centrality parameter (NCP) at the causal genes by $\boldsymbol{\lambda} = \frac{\sqrt{n}}{\sigma_e}\boldsymbol{\alpha}$ and denote predicted expression correlation as $\boldsymbol{\Sigma}_{\text{pe}} = \mathbf{S}^{-1}\mathbf{W}^{\intercal}\boldsymbol{\Sigma}_{\text{snp}}\mathbf{W}\mathbf{S}^{-1}$. The NCP $\boldsymbol{\lambda}$ governs the statistical power of rejecting the null of no effect of predicted expression on trait ($\boldsymbol{\alpha} = 0$). If we assume $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I}_n)$, then our sampling distribution for $\mathbf{z}_{\text{twas}}$ is given by,

$$\mathbf{z}_{\text{twas}} \mid \boldsymbol{\lambda}, \boldsymbol{\Sigma}_{\text{pe}} \sim \mathcal{N}(\boldsymbol{\Sigma}_{\text{pe}}\boldsymbol{\lambda}, \boldsymbol{\Sigma}_{\text{pe}}).$$

This formulation asserts that observed marginal TWAS Z-scores are the linear combination of NCPs at causal genes convoluted through the correlation structure of predicted expression $\boldsymbol{\Sigma}_{\text{pe}}$. Likewise, the resulting correlation structure $\boldsymbol{\Sigma}_{\text{pe}} = \mathbf{W}^{\intercal}\boldsymbol{\Sigma}_{\text{snp}}\mathbf{W}$ is the the product of the underlying LD structure of SNPs $\boldsymbol{\Sigma}_{\text{snp}}$ and the weight matrix learned from expression data $\mathbf{W}$. Computing the likelihood of $\mathbf{z}_{\text{twas}}$ as described requires knowing $\boldsymbol{\lambda}$ and $\boldsymbol{\Sigma}_{\text{pe}}$, which are unknown a-priori; however, we can estimate $\boldsymbol{\Sigma}_{\text{pe}}$ using available reference LD panels (e.g., 1000 Genomes[32]) and estimated expression weights $\widehat{\mathbf{W}}$. This procedure should be unbiased when GWAS individuals are sampled randomly from the same population. This assumption may not be met when diseased individuals are ascertained on for increased statistical power. Estimating $\boldsymbol{\lambda}$ directly from the data is likely to overfit, as our model is over specified. To bypass this issue, we treat $\boldsymbol{\lambda}$ as a nuisance parameter and assume that $\boldsymbol{\lambda} \mid \mathbf{c}, \sigma_c^2 \sim \mathcal{N}(0, \mathbf{D}_{\mathbf{c}})$ where

$$\mathbf{D}_{\mathbf{c}} = \text{diag}(n\sigma_c^2 \cdot \mathbf{c}) + \text{diag}(\delta \cdot (1 - \mathbf{c})),$$

σ_c^2 is the prior causal effect variance, $\mathbf{c}$ is a binary vector indicating if ith gene is causal, and $\delta = 10^{-6}$ is small noise is to ensure that $\mathbf{D}_{\mathbf{c}}$ is full rank. Incorporating this prior for causal NCPs enables us to integrate out $\boldsymbol{\lambda}$, which results in the variance component model,

$$\mathbf{z}_{\text{twas}} \mid \boldsymbol{\Sigma}_{\text{pe}}, \mathbf{c}, \sigma_c^2 \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\text{pe}} + \boldsymbol{\Sigma}_{\text{pe}}\mathbf{D}_{\mathbf{c}}\boldsymbol{\Sigma}_{\text{pe}}).$$

Under this model the variance in $\mathbf{z}_{\text{twas}}$ is due to uncertainty from finite sample size ($\boldsymbol{\Sigma}_{\text{pe}}$) as well as uncertainty in the underlying causal NCPs ($\boldsymbol{\Sigma}_{\text{pe}}\mathbf{D}_{\mathbf{c}}\boldsymbol{\Sigma}_{\text{pe}}$). In principle, we can estimate σ_c^2; however, this comes at a significant computation cost, as estimation would need to be performed for each causal configuration $\mathbf{c}$. To mitigate this we use set $n\sigma_c^2 = 13$, which is similar to what we observe genome-wide in real data.

Equipped with our likelihood model for $\mathbf{z}_{\text{twas}}$, we take a Bayesian approach similar to fine-mapping methods in GWAS to compute the posterior distribution of our causal genes $\mathbf{c}$.

$$\Pr(\mathbf{c} \mid \mathbf{z}_{\text{twas}}, \mathbf{\Sigma}_{\text{pe}}, \sigma_c^2) = \frac{\Pr(\mathbf{z}_{\text{twas}}, \mathbf{c} \mid \mathbf{\Sigma}_{\text{pe}}, \sigma_c^2)}{\Pr(\mathbf{z}_{\text{twas}})} = \frac{\mathcal{N}(\mathbf{z}_{\text{twas}} \mid 0, \mathbf{\Sigma}_{\text{pe}} + \mathbf{\Sigma}_{\text{pe}}\mathbf{D}_{\mathbf{c}}\mathbf{\Sigma}_{\text{pe}}) \Pr(\mathbf{c})}{\sum_{\mathbf{c}' \in \mathcal{C}} \mathcal{N}(\mathbf{z}_{\text{twas}} \mid 0, \mathbf{\Sigma}_{\text{pe}} + \mathbf{\Sigma}_{\text{pe}}\mathbf{D}_{\mathbf{c}'}\mathbf{\Sigma}_{\text{pe}}) \Pr(\mathbf{c}')}$$

Here we assume a simple prior for our causal indicator vector where $\mathbf{c}_i \sim \text{Bernoulli}(p)$. In practice we set $p = 1/m$, which is equivalent with a prior expectation of one causal gene driving risk at a given region. This assumption is likely violated when signal for $\mathbf{z}_{\text{twas}}$ is low, and we recommend only including regions with at least one transcriptome-wide significant Z-score. We compute the marginal posterior inclusion probability (PIP) for the ith gene as

$$\text{PIP}(c_i = 1 | \mathbf{z}_{\text{twas}}, \mathbf{\Sigma}_{\text{pe}}) = \sum_{\mathbf{c}' \in \mathcal{C} : c_i' = 1} \Pr(\mathbf{c}' | \mathbf{z}_{\text{twas}}, \mathbf{\Sigma}_{\text{pe}}),$$

where $\mathcal{C}$ is the set of all binary strings of length m. PIPs offer a flexible mechanism to create gene sets for functional followup. While various approaches exist to define followup sets, we use a simple approach that takes the top k' genes until 90% of the posterior density is explained.

Simulations

We simulated TWAS association statistics using real eQTL weights packaged with the FUSION software (see URLs) and estimated $\mathbf{\Sigma}_{\text{pe}}$ by predicting expression into the 489 European samples from the 1000 Genomes project.[32] To simulate TWAS association statistics at a region we sampled $\mathbf{z}_{\text{twas}} \sim \mathcal{N}(\mathbf{\Sigma}_{\text{pe}}\boldsymbol{\lambda}, \mathbf{\Sigma}_{\text{pe}})$ where $\boldsymbol{\lambda} \sim \mathcal{N}(0, \mathbf{D}_{\mathbf{c}})$ for a causal configuration $c_i \sim \text{Bernoulli}(p = 1/m)$ given $\sigma_c^2 = 40$ and m being the number of tissue-specific gene models in the region. For our initial simulations where gene expression influences downstream trait, we restricted gene models to a single tissue by randomly sampling one tissue. This ensures that the $\mathbf{\Sigma}_{\text{pe}}$ reflects the correlation across different genes, rather than same gene but across tissues. To simulate regions with masked causal genes (i.e. gene with non-zero λ in the causal tissue), we sampled $\mathbf{z}_{\text{twas}}$ from the above model, but did not report the causal gene in the output. To simulate a null region that harbors eQTL signal but does not contribute to downstream trait, we set $c_i = 0$ for all $i \in \{1, \ldots, m\}$, which results in the model $\mathbf{z}_{\text{twas}} \sim \mathcal{N}(0, \mathbf{\Sigma}_{\text{pe}})$. To test how well the causal gene is tagged in non-causal tissues, we kept all tissue-specific gene models to estimate $\mathbf{\Sigma}_{\text{pe}}$ at a region, sampled $\mathbf{z}_{\text{twas}}$ association statistics, and then masked the tissue-specific gene model with non-zero λ.

Datasets

We downloaded publically available summary statistics for lipids measurements GWAS.[17] We filtered sites that were not bi-allelic, were ambiguous (i.e. alelle 1 is reverse complement with allele 2), or had MAF less than 0.01. To perform TWAS on each of the lipids traits we used the software FUSION (see URLs). FUSION takes a summary-based approach to TWAS and requires as input GWAS summary statistics (i.e. SNP Z-scores) and eQTL weights. We downloaded publically available expression weight data as part of the FUSION package. Reference LD was estimated in 1000 Genomes[32] using 489 European individuals. Quality control, cis-heritability of expression, and model fitting have been described elsewhere.[1,4,25] We prioritized adipose for our TWAS approach and used other reference panels as to act as proxy for adipose. That is, for all possible tissue-specific gene models in a region we first test predicted expression using adipose gene models. Then for the remaining genes found only in proxy tissue models, we select those with the best prediction accuracy (i.e. out-of-sample R^2). This resulted in 15,276 unique genes. Risk regions for FOCUS are $\sim$ 1Mb regions that contain at least one transcriptome-wide significant gene-trait association ($P_{\text{twas}} < 0.05/15,276$).

URLS

FOCUS: `http://github.com/bogdanlab/focus/`

FUSION: http://gusevlab.org/projects/fusion/

Lipids GWAS: http://lipidgenetics.org/

References

[1] A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. Penninx, R. Jansen, E. de Geus, DI. Boomsma, FA. Wright, PF. Sullivan, E. Nikkola, M. Alvarez, M. Civelek, AJ. Lusis, T. Lehtimäki, E. Raitoharju, M. Kähönen, I. Seppälä, OT. Raitakari, J. Kuusisto, M. Laakso, AL. Price, P. Pajukanta, and B. Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 2016.

[2] Eric R. Gamazon, Heather E. Wheeler, Kaanan P. Shah, Sahar V. Mozaffari, Keston Aquino-Michaels, Robert J. Carroll, Anne E. Eyler, Joshua C. Denny, G. TEx Consortium, Dan L. Nicolae, Nancy J. Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*, 47(9):1091–1098, 2015.

[3] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R. Robinson, Joseph E. Powell, Grant W. Montgomery, Michael E. Goddard, Naomi R. Wray, Peter M. Visscher, and Jian Yang. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nat Genet*, advance online publication, 2016.

[4] Nicholas Mancuso, Huwenbo Shi, Pagé Goddard, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *The American Journal of Human Genetics*, 100(3):473–487, 2017.

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

[6] D. A. Lawlor, R. M. Harbord, J. A. Sterne, N. Timpson, and S. G. Davey. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med*, 27, 2008.

[7] B. L. Pierce and S. Burgess. Efficient design for mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am J Epidemiol*, 178, 2013.

[8] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International Journal of Epidemiology*, 44(2):512–525, 2015.

[9] Michael Wainberg, Nasa Sinnott-Armstrong, David Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, Johan LM Bjorkegren, Manuel A Rivas, et al. Vulnerabilities of transcriptome-wide association studies. *bioRxiv*, page 206961, 2017.

[10] Richard Barfield, Helian Feng, Alexander Gusev, Lang Wu, Wei Zheng, Bogdan Pasaniuc, and Peter Kraft. Assessing the genetic effect mediated through gene expression from summary eqtl and gwas data. *bioRxiv*, page 223263, 2017.

[11] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 2014.

[12] Julian B Maller, Gilean McVean, Jake Byrnes, Damjan Vukcevic, Kimmo Palin, Zhan Su, Joanna MM Howson, Adam Auton, Simon Myers, Andrew Morris, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics*, 44(12):1294–1301, 2012.

[13] Gleb Kichaev, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L. Price, Peter Kraft, and Bogdan Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet*, 10(10):e1004722, 2014.

[14] Claudia Giambartolomei, Damjan Vukcevic, Eric E. Schadt, Lude Franke, Aroon D. Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLOS Genetics*, 10(5):1–15, 05 2014.

[15] Xuanyao Liu, Hilary K Finucane, Alexander Gusev, Gaurav Bhatia, Steven Gazal, Luke O'Connor, Brendan Bulik-Sullivan, Fred A Wright, Patrick F Sullivan, Benjamin M Neale, et al. Functional architectures of local and distal regulation of gene expression in multiple human tissues. *The American Journal of Human Genetics*, 100(4):605–616, 2017.

[16] Alexis Battle, Christopher D Brown, Barbara E Engelhardt, Stephen B Montgomery, GTEx Consortium, et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

[17] Consortium Global Lipids Genetics. Discovery and refinement of loci associated with lipid levels. *Nat Genet*, 45(11):1274–1283, 2013.

[18] Anders H Berg, Terry P Combs, and Philipp E Scherer. Acrp30/adiponectin: an adipokine regulating glucose and lipid metabolism. *Trends in Endocrinology & Metabolism*, 13(2):84–89, 2002.

[19] Kiran Musunuru, Alanna Strong, Maria Frank-Kamenetsky, Noemi E. Lee, Tim Ahfeldt, Katherine V. Sachs, Xiaoyu Li, Hui Li, Nicolas Kuperwasser, Vera M. Ruda, James P. Pirruccello, Brian Muchmore, Ludmila Prokunina-Olsson, Jennifer L. Hall, Eric E. Schadt, Carlos R. Morales, Sissel Lund-Katz, Michael C. Phillips, Jamie Wong, William Cantley, Timothy Racie, Kenechi G. Ejebe, Marju Orho-Melander, Olle Melander, Victor Koteliansky, Kevin Fitzgerald, Ronald M. Krauss, Chad A. Cowan, Sekar Kathiresan, and Daniel J. Rader. From noncoding variant to phenotype via sort1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714–719, Aug 2010.

[20] Marc Brackertz, Joern Boeke, Ru Zhang, and Rainer Renkawitz. Two highly related p66 proteins comprise a new family of potent transcriptional repressors interacting with mbd2 and mbd3. *Journal of Biological Chemistry*, 277(43):40958–40966, 2002.

[21] Mammalian Gene Collection (MGC) Program Team et al. Generation and initial analysis of more than 15,000 full-length human and mouse cdna sequences. *Proceedings of the National Academy of Sciences*, 99(26):16899–16903, 2002.

[22] Farhad Hormozdiari, Anthony Zhu, Gleb Kichaev, Chelsea J-T Ju, Ayellet V Segrè, Jong Wha J Joo, Hyejung Won, Sriram Sankararaman, Bogdan Pasaniuc, Sagiv Shifman, et al. Widespread allelic heterogeneity in complex traits. *The American Journal of Human Genetics*, 100(5):789–802, 2017.

[23] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney McCormick, Christian D Haudenschild, Kenneth B Beckman, Jianxin Shi, Rui Mei, et al. Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. *Genome research*, 24(1):14–24, 2014.

[24] Yang I Li, Bryce van de Geijn, Anil Raj, David A Knowles, Allegra A Petti, David Golan, Yoav Gilad, and Jonathan K Pritchard. Rna splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, 2016.

[25] Alexander Gusev, Nick Mancuso, Hilary K. Finucane, Yakir Reshef, Lingyun Song, Alexias Safi, Edwin Oh, Steven McCaroll, Benjamin Neale, Roel Ophoff, Michael C. Donovan, Nicholas Katsanis, Gregory E. Crawford, Patrick F. Sullivan, Bogdan Pasaniuc, and Alkes L. Price. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *bioRxiv*, 2016.

[26] SS Kaalund, EN Newburn, Tuo Ye, R Tao, C Li, A Deep-Soboslay, MM Herman, TM Hyde, DR Weinberger, BK Lipska, et al. Contrasting changes in drd1 and drd2 splice variant expression in schizophrenia and affective disorders, and associations with snps in postmortem brain. *Molecular psychiatry*, 19(12):1258–1266, 2014.

[27] Urko M Marigorta, Lee A Denson, Jeffrey S Hyams, Kajari Mondal, Jarod Prince, Thomas D Walters, Anne Griffiths, Joshua D Noe, Wallace V Crandall, Joel R Rosh, et al. Transcriptional risk scores link gwas to eqtls and predict complications in crohn's disease. *Nature Genetics*, 49(10):1517–1521, 2017.

[28] Halit Ongen, Andrew A Brown, Olivier Delaneau, Nikolaos Panousis, Alexandra C Nica, Emmanouil T Dermitzakis, GTEx Consortium, et al. Estimating the causal tissues for complex traits and diseases. *bioRxiv*, page 074682, 2016.

[29] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet*, 9(2):e1003264, 2013.

[30] D Habier, RL Fernando, and JCM Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397, 2007.

[31] Paul M VanRaden. Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11):4414–4423, 2008.

[32] Consortium The Genomes Project. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.