bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Probabilistic fine-mapping of transcriptome-wide association studies

Nicholas Mancuso[1], Gleb Kichaev[2], Huwenbo Shi[2], Malika Freund[3], 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 while controlling for pleiotropic SNP effects and unmeasured causal expression. Importantly, our approach remains accurate when expression data for causal genes are not available in the causal tissue by leveraging expression prediction from other tissues. 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. We illustrate our approach using an integrative analysis of lipids traits where our approach prioritizes genes with strong evidence for causality.

Introduction

Transcriptome-wide association studies (TWAS) using predicted expression levels have been proposed as an approach to identify novel genomic risk regions and putative risk genes involved for 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.[1,4,5] Significant genetic correlation is often interpreted as an estimate of the effect of SNPs on trait mediated by the gene of interest. However, this interpretation requires very strong assumptions that are likely violated in empirical data due to LD and/or pleiotropic SNP effects.[1–3,6–11] Therefore TWAS has been mostly utilized as a test of association, in contrast to methods that attempt to directly estimate the mediated effect (i.e. Mendelian randomization[3,6–9]).

In this work, we show that the gene-trait association statistics from TWAS 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

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

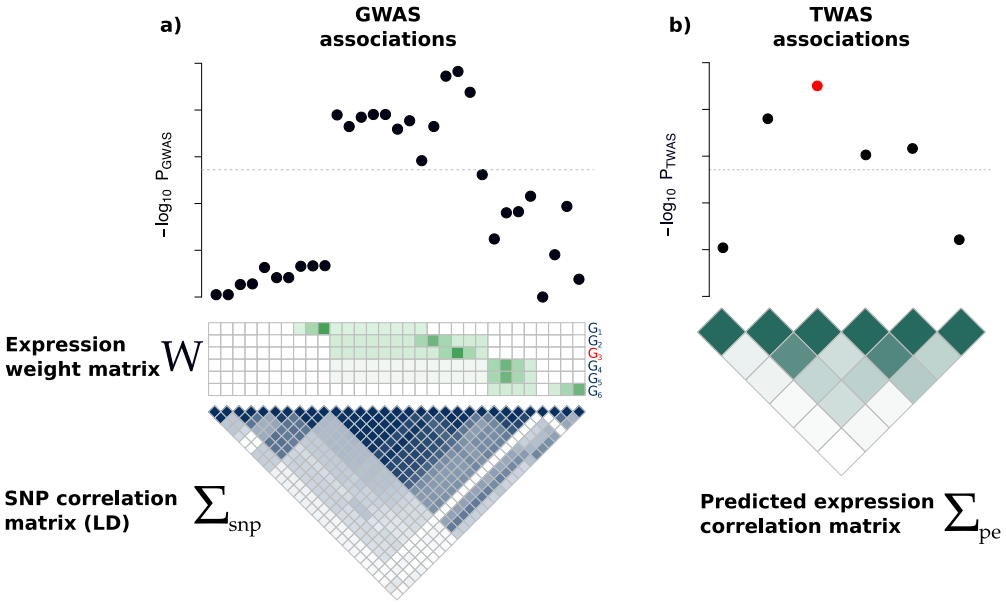

Figure 1: Illustration of the induced correlation structure for predicted expression. a) Top: Manhattan plot indicating strength of SNP association with trait. Middle: Expression weight matrix for 6 genes in the same region, with the causal gene in red. Each row corresponds to a gene and each column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

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 (causal gene in red; no other causal genes are present at this region, see Figure 1). Although genes 3 and 4 in Figure 1 have non-overlapping prediction weights due to different eQTL genetic regulation, LD among SNPs with non-zero prediction weights induce correlations in the TWAS statistics at genes 3 and 4. Estimating the correlation structure between predicted expression among nearby genes enables statistical fine-mapping over gene-trait associations. However, we note several confounding factors need to be addressed for unbiased inference. First, there is a body of evidence supporting horizontal pleiotropic effects from SNPs,[9,11,12] which bias gene-trait association statistics and should be accounted for. Second, it is critical that TWAS fine-mapping approaches be robust if the causal mechanism is not steady-state levels of gene expression. Fine-mapping in these instances without controlling for confounding will result in a misspecified causal model and likely prioritize genes that tag causal mechanisms well.

Here, we propose an approach to perform statistical fine-mapping over the gene-trait association signals while accounting for the correlation structure induced by LD and prediction weights used in the TWAS procedure and simultaneously controlling for certain pleiotropic effects. Our approach, FOCUS (Fine-mapping Of CaUsal gene Sets), takes as input GWAS summary data, expression prediction weights (as estimated from eQTL reference panels), and LD among all SNPs in the risk region, and estimates the probability for

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

any given set of genes to explain the TWAS signal. Our approach extends probabilistic SNP fine-mapping approaches[13–15] to estimate sets of genes that contain the "causal" genes (defined here as the gene responsible for the association signal) at a predefined confidence level (i.e. ρ gene credible set). FOCUS accounts for bias due to missing causal factors by including the *null model* as a possible explanatory factor in the credible set. We perform extensive simulations and show that FOCUS is unbiased in estimating the posterior probabilities and credible sets at a specified certainty when the causal gene is present in the data. When the causal tissue is unavailable and alternative tissues with correlated expression levels are used as a proxy, FOCUS maintains its performance under standard assumptions. FOCUS outputs posterior predictive checks[16] of observed TWAS Z-scores to quantify model agreement given inferred posterior probabilities for causality. Finally, as a demonstration using real GWAS data, we apply FOCUS to four GWASs of lipids levels.[17] We find that FOCUS prioritizes genes with established roles in LDL risk (e.g., *SORT1*).[18]

Results

Methods Overview

To disentangle between causal and tagging gene-trait associations at a TWAS significant region, we analytically derive the covariance structure among 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 a region (TWAS significant and not-significant) using a multivariate Gaussian distribution parameterized by the effect sizes at causal genes ($\boldsymbol{\lambda}_{\text{pe}}$), residual SNP-effects ($\boldsymbol{\lambda}_{\text{snp}}$), and the correlation structure induced by inferred expression weights ($\boldsymbol{\Omega}$) with LD ($\mathbf{V}$) as

$$\mathbf{z}_{\text{twas}} \mid \boldsymbol{\lambda}_{\text{snp}}, \boldsymbol{\lambda}_{\text{pe}}, \boldsymbol{\Omega}, \mathbf{V} \sim \mathcal{N}(\boldsymbol{\Omega}^{\intercal}\mathbf{V}\boldsymbol{\lambda}_{\text{snp}} + \boldsymbol{\Omega}^{\intercal}\mathbf{V}\boldsymbol{\Omega}\boldsymbol{\lambda}_{\text{pe}}, \boldsymbol{\Omega}^{\intercal}\mathbf{V}\boldsymbol{\Omega}),$$

(see Methods). We control for bias resulting from pleiotropic effects of SNPs by including an intercept term that quantifies the average SNP effect sizes ($\boldsymbol{\lambda}_{\text{snp}}$) tagged by predicted expression ($\boldsymbol{\Omega}^{\intercal}\mathbf{V}\boldsymbol{\lambda}_{\text{snp}}$; see Methods). To allow for genes without prediction models in the relevant tissue (either due to QC and/or low power in eQTL studies), we leverage recent work demonstrating that eQTLs are largely shared across tissues[19] and include prediction models from proxy tissues for such genes (see Methods). We employ a standard Bayesian approach 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 $\boldsymbol{\lambda}_{\text{pe}}$ using a multivariate Gaussian prior (see Methods). We use PIPs to compute ρ-credible gene-sets that contain the causal gene with probability ρ.[14] To account for missing causal mechanisms either due to unpredicted expression or other latent functional mechanisms, we include the null model as a possible outcome in the credible set (see Methods). Lastly, we use a simulation-based procedure to compute posterior predictive checks[16] that measure the FOCUS model's goodness-of-fit given observed TWAS Z-scores.

FOCUS prioritizes causal genes in causal-tissue simulations

To characterize the predicted expression correlation structure and to validate our framework, we used extensive simulations starting from real genotype data to generate expression reference panels and GWAS summary data (see Methods). We confirmed that non-causal genes in risk regions show significant association with

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

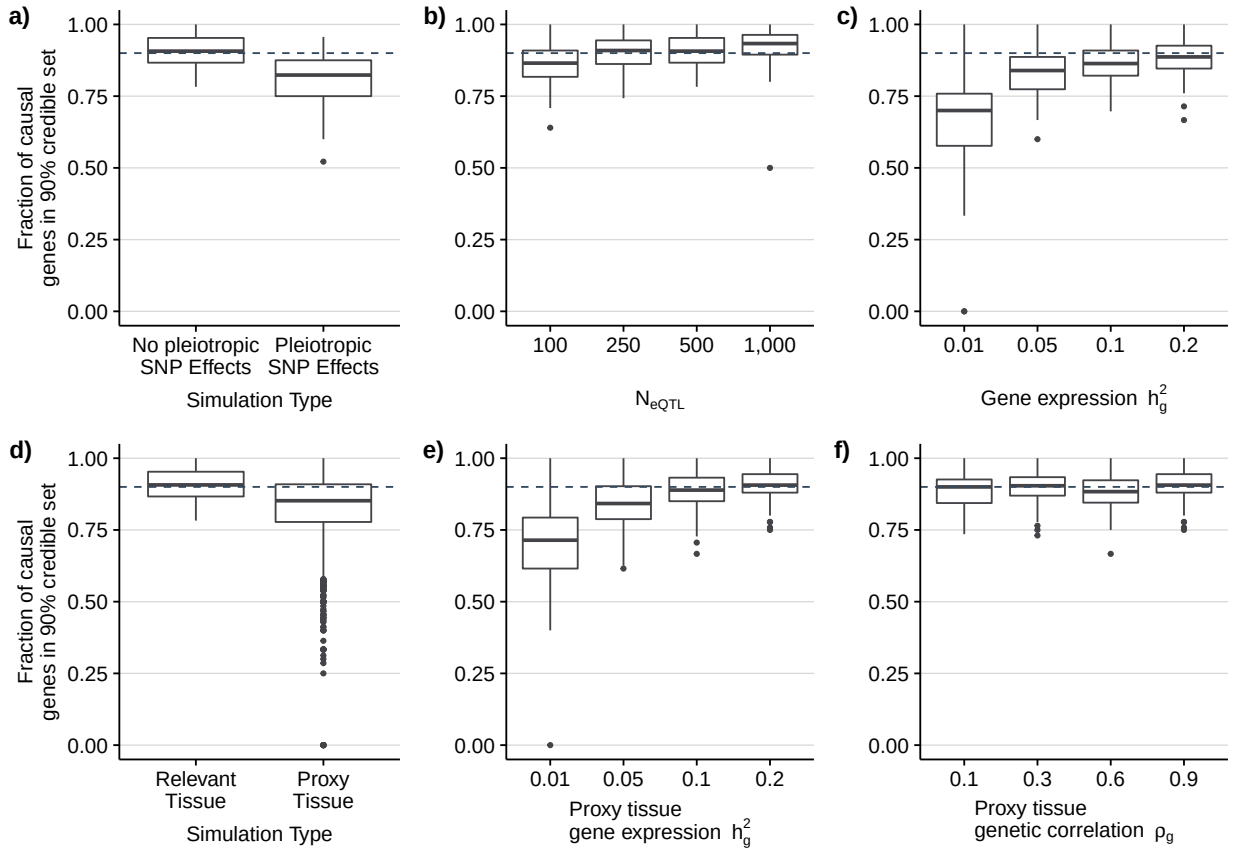

Figure 2: Credible gene-sets are well-calibrated in simulations. Box-plots represent the distribution of the fraction of causal genes captured in the 90% credible set over simulations (see Methods). a) Simulations with and without pleiotropic SNP effects on trait. Prediction models were trained using the relevant (i.e. causal) tissue. b) Calibration as a function of eQTL reference panel sample size. c) Calibration as a function of heritability of causal gene expression. d) Calibration using prediction models trained using proxy tissue measurements. e) Calibration using proxy tissue when heritability of reference gene expression varies compared with fixed $h_g^2 = 0.2$ in the relevant tissue. f) Calibration using proxy tissue when genetic correlation of reference gene expression and gene expression in the relevant tissue varies.

trait as function of LD and eQTL weights (see Supplementary Figure 1), which motivates fine-mapping to prioritize genes causally impacting trait. We simulated complex trait under a variety of architectures to assess the performance of 90%-credible gene-sets computed using FOCUS (see Methods). When the causal gene was assayed in its relevant tissue, we found 90%-credible gene-sets contained 0.91 (S.D. 0.06) of causal genes across simulations on average (see Figure 2). We saw accuracy under general ρ was stable, with credible gene-sets being well-calibrated across various values of ρ (see Supplementary Figure 2). FOCUS models an intercept term to control for pleiotropic SNP effects (i.e. $\boldsymbol{\lambda}_{\text{snp}}$) tagged through predicted expression. In simulations where a fraction of SNPs directly impacted downstream trait, we computed credible sets after estimating an intercept term ($\hat{\lambda}_{\text{snp}}$) and found a decrease in performance (see Figure 2; see Methods). Next, we varied sample size across GWAS and reference eQTL datasets. Intuitively, we found improved performance for FOCUS to detect causal genes as sample size increased (see Figure 2, Supplementary Figure 3). Sample size for the eQTL reference panel affected performance to a larger degree than GWAS sample size, consistent with earlier reports.[1] For example, at $N_{\text{eQTL}} = 100$, we found 90%-credible gene-sets contained

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the causal gene in 88% of simulations, which is significantly fewer when compared with 94% for $N_{\text{eQTL}} = 500$ (Mann-Whitney-U $P = 5.46 \times 10^{-9}$). Next, we explored how underlying heritability of expression at causal genes impacts prioritization. Heritability defines the prediction upper bound for SNP-based approaches,[20,21] and we expect performance to improve as non-zero heritability is easier to detect. We confirmed that performance increased with heritability of causal gene expression (see Figure 2). For example, we simulated gene expression having heritability $h_g^2 = 0.01$ and inferred inferred eQTL weights using $N_{\text{eQTL}} = 500$ and found a significant decrease in performance (Mann-Whitney-U $P < 2.2 \times 10^{-16}$). Similarly, we looked at the role of the prior effect-size distribution for predicted gene expression[4] and found performance to remain relatively stable for a wide range of values (see Supplementary Figures 4, 5).

FOCUS remains stable when using proxy tissues

Next, we investigated the performance of FOCUS when the causal gene in the relevant tissue is missing, but is measured in a different tissue (see Methods). 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[1,4,19,22] 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 larger sample size. To simulate gene expression in a proxy tissue, we drew correlated effect sizes at the same eQTLs for the gene expression reference panel (see Methods). Here, we consider a causal gene to be successfully fine-mapped if its corresponding proxy tissue model is in the 90%-credible gene-set. When sample size for eQTL in the relevant- and proxy-tissues are the same, but heritability in proxy tissue is lower than the relevant-tissue, we found a significant loss in accuracy, with 90% credible sets capturing the causal gene 0.83 (S.D. 0.08) of simulations compared with 0.91 (S.D. 0.06) when averaging over values of ρ_g. (Mann-Whitney-U $P = 3.4 \times 10^{-13}$; see Figure 2). This effect was not observed when heritability of proxy tissue gene expression was at least that of expression in the relevant-tissue (Mann-Whitney-U $P = 0.06$). For example, when expression in the relevant tissue was $h_g^2 = 0.2$, but $h_g^2 = 0.01$ in the proxy, we found 90%-credible gene-sets contained the causal gene in significantly fewer simulations (0.79 versus 0.89; Mann-Whitney-U $P < 2.2 \times 10^{-16}$), which suggests that when causal eQTLs are shared across tissues, increased heritability of expression increases power to detect the causal gene. Surprisingly, we found correlation of effect-sizes at shared eQTLs to play no significant role in performance when heritability of expression in the relevant and proxy tissue is kept the same ($h_g^2 = 0.2$; see Figure 2, Supplementary Figure 6). In the case of zero correlation between effect sizes at the same eQTL SNPs, this result can be interpreted as pleiotropic effects on independent molecular traits, which are known to be difficult to differentiate from a causal effect.[1,3,9] 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.

FOCUS is robust to missing causal molecular mechanisms

Our model predicts that predicted expression of nearby genes will be correlated due to linkage between eQTL SNPs, which results in correlated test statistics among gene-trait associations. If predicted expression for the causal gene is not included, nearby genes will likely be prioritized in fine-mapping. This scenario is analogous to absent causal SNPs under SNP fine-mapping approaches. FOCUS controls for this scenario through two mechanisms (see Methods). First, FOCUS explicitly models the null (i.e. no gene-trait relationship for

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

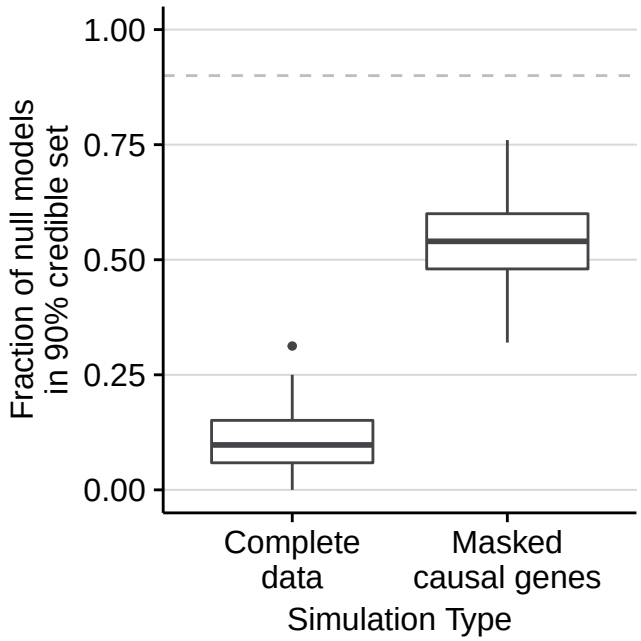

Figure 3: FOCUS credible-sets alleviate bias when the causal gene is missing. "Masked" indicates simulations where the causal genes are pruned before analysis. For comparison, we include results under the complete-data simulation pipeline where causal genes are tested. Box-plots represent the distribution of the proportion of null models captured by 90%-credible gene-sets in simulations.

all nearby genes) as a possible outcome when computing credible gene-sets. We tested the performance of FOCUS in simulations when there is no relationship between expression and trait, and found the null model was contained in the 90%-credible gene-set in 0.98 of our simulations, indicating that FOCUS is accurate under the null. We next performed experiments using simulations where causal gene expression effects downstream trait, but has been pruned from the data before testing. We found the null model in 0.59 (S.D. 0.08) of 90%-credible gene-sets (see Figure 3), which was a significantly greater percentage compared with simulations where the causal gene was present (0.14, S.D. 0.08; Mann-Whitney-U $P < 2.2 \times 10^{-16}$). Second, FOCUS estimates a single intercept term at each region to account for pleiotropic SNP effects on trait. When the predicted causal mechanisms are absent from the data, we expect the intercept to increase in magnitude, as SNP effects mediated through missing mechanisms will be indistinguishable from pleiotropic effects. Indeed, we found an enrichment for significantly non-zero intercept estimates (i.e. $\hat{\lambda}_{\text{snp}}$) when the causal mechanism was absent compared with complete-data simulations (Fisher's exact $P = 2.9 \times 10^{-3}$). Altogether, we find FOCUS is robust in the challenging setting of prioritizing the null model when causal expression is missing.

FOCUS improves resolution for fine mapping causal genes

Having demonstrated that FOCUS computes well-calibrated credible gene-sets under a wide range of parameters we next sought to quantify the resolution of credible gene-sets to identify causal genes. In particular, we estimated the average number of genes captured in the credible set. We found 90%-credible gene-sets contained 4.4 genes on average (S.D. 1.9) in the relevant-tissue simulations, which resulted in an average 47% of predicted genes per risk region (see Figure 4). We found a similar number of genes in 90%-credible

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

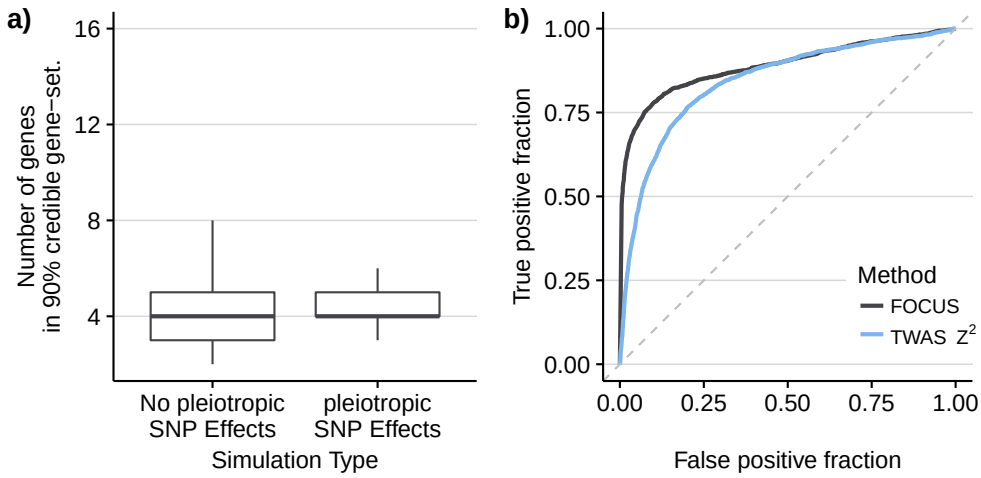

Figure 4: FOCUS accurately prioritizes causal genes in simulations. Box-plots represent the distribution of the total number of causal genes captured in the 90% credible sets over simulations (see Methods). a) Simulations with and without pleiotropic SNP effects on trait. Prediction models were trained using the relevant (i.e. causal) tissue. b) ROC curve computed using PIPs versus TWAS Z^2 for each gene.

gene-sets across simulations when varying model parameters and sample sizes (see Supplementary Figures 7-12). While we advocate the use of credible-sets in practice rather than thresholding on PIPs, for completeness we prioritized genes using PIPs for direct comparison with TWAS p-values (see Methods). We found prioritizing genes using PIPs outperformed TWAS p-value ranking at capturing underlying causal genes (see Figure 4). For example, at a false positive rate of 5%, FOCUS identifies 162% more causal genes than a simple rank of marginal TWAS statistics. A unique feature of FOCUS is that it allows for multiple causal genes at a given region; in this scenario FOCUS attains a a gain of 213% more causal genes compared to that of 132% for single causal regions (see Supplementary Table 2). Overall, FOCUS accurately prioritizes causal genes from non-causal genes with the largest gains when multiple causal genes exist at risk regions.

Application to lipids GWAS data

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 relevant tissue for expression driving lipids is adipose given its well-characterized role.[23–26] To account for missing gene prediction models, we incorporate gene expression models for genes not predictable at current sample sizes from adipose tissue across 45 tissues measured in 47 reference panels. 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 Supplementary Table 1; see Methods). Of the 26,292 known genes in RefSeq (ver 65),[27] we found 12,663 covered in our data with the remaining 2,614 genes not found in RefSeq. Adipose-prioritized TWAS identifies 301 (202 unique) significant genes at 108 (63 unique) independent regions after accounting for the total number of per-trait tests performed ($P < 0.05/15,277$; see Supplementary Figures 13-15; Table 1; Supplementary Table 3). Of the 160 (89 unique) risk regions found through GWAS, 75 (46 unique) overlapped significant TWAS results, which is increased compared with earlier work[29] that found 25% overlap between GWAS and eQTL at risk regions (see Table 1).

Having performed a TWAS across lipids traits, we next sought to prioritize putative causal genes using our

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Lipid Trait	GWAS risk regions	GWAS risk regions with any TWAS genes	TWAS genes at risk regions	Genes in 90%-credible sets
High-density lipoprotein (HDL)	43	18	64	30
Low-density lipoprotein (LDL)	36	20	56	40
Total cholesterol (TC)	51	24	73	53
Triglycerides (TG)	30	13	33	25
Overall (unique)	160 (89)	75 (46)	226 (146)	148 (100)

Table 1: Summary of gene-based fine-mapping in lipids GWAS risk regions. A GWAS risk region is defined to be a LD-block defined by LDetect[28] harboring at least one genome-wide significant SNP ($P < 5 \times 10^{-8}$) reported in ref.[17] A TWAS gene is a gene whose predicted expression reaches transcriptome-wide significance of $P < 0.05/15,277$.

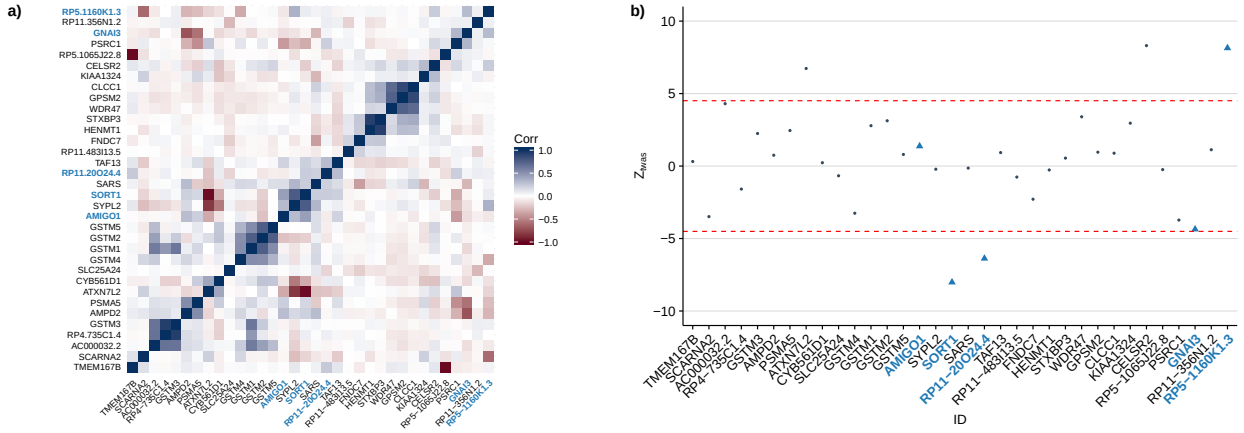

Figure 5: 1p13 locus for LDL. a) Correlation for predicted expression at 1p13 locus. Genes in the 90%-credible set are labeled in light blue. b) TWAS Z-scores at 1p13 locus. Each point represents the association strength for each tested gene. Genes in the 90%-credible gene-set are labeled in light blue. Dashed red lines indicate transcriptome-wide significance threshold.

framework. We applied FOCUS at the 75 GWAS risk regions with evidence for regulatory action on genes driving lipids levels to compute PIPs and estimate credible sets of genes at each of the regions (see Methods). We found that observed risk regions can be explained by 1.5 causal genes on average, with 61/75 risk regions containing fewer than 2 causal genes in expectation (see Supplementary Figure 18). The average maximum PIP across credible sets was 88% (and decreased exponentially for lower ranked genes; see Supplementary Figure 17). Together, these results imply that most risk regions can be explained by a single causal gene. Using computed PIPs, we estimated 90%-credible gene-sets for each risk region and found a significant reduction in the number of prioritized genes (mean 1.9), compared with transcriptome-wide significant genes (mean 3.2; one-sided Mann-Whitney-U $P = 7.24 \times 10^{-4}$; Supplementary Figure 17; Supplementary Table 4). 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 of *SORT1* and impact downstream LDL levels.[18] We found 4/34 genes included in the 90% credible set, of which *SORT1* had a posterior probability 95% (see Figure 5).

Next, we investigated regions whose 90%-credible gene-sets contained the null model (i.e. regions with weaker evidence for models of gene expression driving risk). An instance that contains the null model in its credible set may be partially consistent with observed association between expression levels and trait being due to chance. We found 25/75 instances of the null model captured in credible sets for lipids traits

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(see Supplementary Table 4), which suggests most overlapping GWAS risk regions are more consistent with risk contributed from cis-regulated expression levels, compared with statistical noise explaining observed signal. PIPs output by FOCUS are conditioned on the FOCUS model being correct. If FOCUS's model does not accurately capture the underlying generative process then PIPs will be biased. We used a simulation procedure (see Methods) to quantify model fit for each gene and found the FOCUS model largely agreed with observed data (i.e. TWAS Z-scores; see Supplementary Figure 19).

Discussion

In this work we presented FOCUS, a fine-mapping approach that estimates credible sets of causal genes using prediction eQTL weights, LD, and GWAS summary statistics. We demonstrated FOCUS adequately controls false positives in null simulations and outperforms straightforward p-value ranking in identifying causal genes when genes at a region impact downstream trait. We found 90%-credible gene-sets to be largely stable across a variety of simulations, with the biggest impact in performance due to eQTL reference panel sample size and SNP-heritability of gene expression. 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. 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.[19,30] Second, 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[31,32]). For example, several recent works have supporting evidence for splice variation playing an important role in driving risk of schizophrenia.[33,34]

We showed our approach is well calibrated under various 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 steady-state expression levels at causal genes. Several works have demonstrated that risk prediction using a linear combination of predicted steady-state or observed expression levels can outperform standard SNP-based models,[33,35] 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 data sets grow in size, accurately modeling these assumptions may be possible. Similarly, if risk is mediated through context-specific expression and not steady-state expression levels, then FOCUS will have a loss in performance. Second, while our simulations used GBLUP[36,37] throughout for its straightforward implementation, we recommend a cross-validation approach to select the best fitting linear model (e.g., GBLUP, BSLMM[38]) using the ratio of out-of-sample prediction accuracy normalized by the total SNP-heritability of gene expression, which is implemented in the FUSION framework.[1,33] Third, when the causal gene is

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

untyped in the data, our approach will partially inflate posterior probabilities at tagging genes. We attempt to mitigate this scenario by adding an intercept term to the model and incorporating gene models measured in proxy tissues. We caution that our simulated results using proxy-tissues were performed using a model where causal eQTLs are shared between proxy- and relevant tissues and have correlated effect sizes, which is equivalent to a random-effects model. This assumption may be violated in real data if causal eQTLs are tissue-specific. Recent work, however, has demonstrated that a large number of eQTLs are indeed shared across tissues.[19] Fourth, we took a tissue-prioritizing approach by preferentially using eQTL weights in adipose tissue given its known role in lipids[23–26] 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. likely causal) tissue for complex traits can be accurately estimated using eQTL data.[39] Coupled with estimation of causal tissue, we suggest prioritizing genes with high normalized prediction accuracy in related tissues. We note that our results were strongly dependent on sample size in the eQTL reference panel, which is reflected in expression prediction accuracy. We therefore, recommend prioritizing eQTL data with sample sizes greater than 100 if possible and performing inference on genes with robustly non-zero SNP-heritability. Despite our modeling assumptions and limitations, our approach is a step towards accurately prioritizing gene-sets.

Online Methods

Notation

We denote scalar variables with italicized lower-case letters (e.g., z). Vectors are denoted with bold lower-case letters (e.g., $\mathbf{z}$). Scalar entries for a vector are indexed with a subscript (e.g., jth element of $\mathbf{z}$ is z_j). We denote matrices with bold capital letters (e.g., $\mathbf{X}$, its transpose $\mathbf{X}^\intercal$) and index rows with a subscript (e.g., $\mathbf{X}_j$). We indicate L block-column partitions for matrix (vector) $\mathbf{X}$ as $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(L)})$.

Model and sampling distribution of marginal TWAS summary statistics

We model quantitative trait for n individuals $\mathbf{y}$ by a linear combination of expression levels for m genes $\mathbf{G} \in \mathbb{R}^{n \times m}$ as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\alpha} + \widetilde{\boldsymbol{\epsilon}}$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the centered and variance-standardized genome-wide genotype matrix at p SNPs, $\boldsymbol{\beta}$ are the p pleiotropic effects of $\mathbf{X}$ on $\mathbf{y}$, $\boldsymbol{\alpha}$ is the vector of causal effects for the m genes and $\widetilde{\boldsymbol{\epsilon}}$ is random environmental noise with $\mathbb{E}[\widetilde{\boldsymbol{\epsilon}}] = 0$ and $\mathbb{V}[\widetilde{\boldsymbol{\epsilon}}] = \mathbf{I}_n \widetilde{\sigma}_e^2$. We extend our definition by also defining $\mathbf{G}$ as a linear function of underlying genotype and environment, which is governed by $\mathbf{G} = \mathbf{X}\mathbf{W} + \mathbf{E}$, where $\mathbf{W} \in \mathbb{R}^{p \times m}$ is the eQTL effect-size matrix, and $\mathbf{E} \in \mathbb{R}^{n \times m}$ is environmental noise. Our updated model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{X}\mathbf{W} + \mathbf{E})\boldsymbol{\alpha} + \widetilde{\boldsymbol{\epsilon}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}\mathbf{W}\boldsymbol{\alpha} + \mathbf{E}\boldsymbol{\alpha} + \widetilde{\boldsymbol{\epsilon}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}\mathbf{W}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} = \mathbf{E}\boldsymbol{\alpha} + \widetilde{\boldsymbol{\epsilon}}$ is the total contribution from environment, which we parameterize as $\mathbb{E}[\boldsymbol{\epsilon}] = 0$ and $\mathbb{V}[\boldsymbol{\epsilon}] = \mathbf{I}_n \sigma_e^2$ which is valid provided independence of errors holds. If causal eQTL effect-sizes $\mathbf{W}$ were known, we could prioritize putative susceptibility genes by estimating $\boldsymbol{\alpha}$ using regression. Unfortunately, effect-sizes $\mathbf{W}$ are unknown and must be estimated from data (e.g., BSLMM,[38] GBLUP[36,37]). Because

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

inferring eQTL effect-sizes genome-wide is challenging, models typically focus only on *cis*- or local-SNPs at each gene. Let predicted expression be defined as $\widehat{\mathbf{G}} = \mathbf{X}\boldsymbol{\Omega}$ when $\boldsymbol{\Omega}$ is estimated from data. Our local-SNP model for L independent genetic regions is given by,

$$\mathbf{y} = \sum_{k=1}^{L} \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)} + \sum_{k=1}^{L} \mathbf{X}^{(k)}\mathbf{W}^{(k)}\boldsymbol{\alpha}^{(k)} + \boldsymbol{\epsilon}.$$

Here we describe the sampling distribution of marginal TWAS Z-scores obtained from an association test. For simplicity, we focus our attention to genes in a single genomic region and drop the $(\cdot)$ notation. Specifically, we compute the marginal association z_j of gene j with $\mathbf{y}$ through a transcriptome-wide association study as,

$$\begin{aligned} z_j &= \frac{1}{\sigma_e\sqrt{n}}\widehat{\mathbf{G}}_j^\intercal\mathbf{y} = \frac{1}{\sigma_e\sqrt{n}}(\mathbf{X}\boldsymbol{\Omega})_j^\intercal\mathbf{y} = \frac{1}{\sigma_e\sqrt{n}}\boldsymbol{\Omega}_j^\intercal\mathbf{X}^\intercal\mathbf{y} = \frac{1}{\sigma_e\sqrt{n}}\boldsymbol{\Omega}_j^\intercal\mathbf{X}^\intercal(\mathbf{X}\boldsymbol{\beta} + \mathbf{X}\mathbf{W}\boldsymbol{\alpha} + \boldsymbol{\epsilon}) \\ &= \frac{1}{\sigma_e\sqrt{n}}\left[\boldsymbol{\Omega}_j^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Omega}_j^\intercal\mathbf{X}^\intercal\mathbf{X}\mathbf{W}\boldsymbol{\alpha} + \boldsymbol{\Omega}_j^\intercal\mathbf{X}^\intercal\boldsymbol{\epsilon}\right] \\ &= \frac{\sqrt{n}}{\sigma_e}\boldsymbol{\Omega}_j^\intercal\mathbf{V}\boldsymbol{\beta} + \frac{\sqrt{n}}{\sigma_e}\boldsymbol{\Omega}_j^\intercal\mathbf{V}\mathbf{W}\boldsymbol{\alpha} + \frac{1}{\sigma_e\sqrt{n}}\boldsymbol{\Omega}_j^\intercal\mathbf{X}^\intercal\boldsymbol{\epsilon}. \end{aligned}$$

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

$$\mathbf{z}_{\text{twas}} = \frac{\sqrt{n}}{\sigma_e}\boldsymbol{\Omega}^\intercal\mathbf{V}\boldsymbol{\beta} + \frac{\sqrt{n}}{\sigma_e}\boldsymbol{\Omega}^\intercal\mathbf{V}\mathbf{W}\boldsymbol{\alpha} + \frac{1}{\sigma_e\sqrt{n}}\boldsymbol{\Omega}^\intercal\mathbf{X}^\intercal\boldsymbol{\epsilon}.$$

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

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

To simplify notation we re-parameterize the causal effects as a non-centrality parameter (NCP) at the causal genes by $\boldsymbol{\lambda}_{\text{pe}} = \frac{\sqrt{n}\boldsymbol{\alpha}}{\sigma_e}$. We note that $\boldsymbol{\Omega}^\intercal\mathbf{V}\mathbf{W} \neq \boldsymbol{\Omega}^\intercal\mathbf{V}\boldsymbol{\Omega}$, but given large-enough sample sizes we expect $\boldsymbol{\Omega}^\intercal\mathbf{V}\boldsymbol{\Omega}$ to approach $\boldsymbol{\Omega}^\intercal\mathbf{V}\mathbf{W}$[1]. We denote predicted expression covariance as $\boldsymbol{\mathcal{V}} = \boldsymbol{\Omega}^\intercal\mathbf{V}\boldsymbol{\Omega}$. The NCP $\boldsymbol{\lambda}_{\text{pe}}$ governs the statistical power of rejecting the null of no effect of predicted expression on trait ($\boldsymbol{\alpha} = 0$). We parameterize $\boldsymbol{\beta}$ similarly as $\boldsymbol{\lambda}_{\text{snp}} = \frac{\sqrt{n}}{\sigma_e}\boldsymbol{\beta}$. If we assume $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n\sigma_e^2)$, then our sampling distribution for $\mathbf{z}_{\text{twas}}$ is given by,

$$\mathbf{z}_{\text{twas}} \mid \lambda_{\text{snp}}, \boldsymbol{\lambda}_{\text{pe}}, \boldsymbol{\Omega}, \mathbf{V} \sim \mathcal{N}(\boldsymbol{\Omega}^\intercal\mathbf{V}\boldsymbol{\lambda}_{\text{snp}} + \boldsymbol{\mathcal{V}}\boldsymbol{\lambda}_{\text{pe}}, \boldsymbol{\mathcal{V}}).$$

This formulation asserts that observed marginal TWAS Z-scores are the linear combination of NCPs at causal genes convoluted through the covariance structure of predicted expression $\boldsymbol{\mathcal{V}}$ and tagged pleiotropic effects from SNPs $\boldsymbol{\Omega}^\intercal\mathbf{V}\boldsymbol{\beta}$. Likewise, the resulting covariance structure $\boldsymbol{\mathcal{V}}$ is the the product of the underlying LD structure of SNPs $\mathbf{V}$ and the weight matrix learned from expression data $\boldsymbol{\Omega}$.

[1] Penalized regression may exhibit bias, but does so at the benefit of further reducing the mean-squared error, which is a measure of closeness to underlying parameters.

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Computing the likelihood of $\mathbf{z}_{\text{twas}}$ as described requires knowing $\boldsymbol{\mathcal{V}}$, $\boldsymbol{\lambda}_{\text{snp}}$, and $\boldsymbol{\lambda}_{\text{pe}}$, which are unknown a-priori. First, we can estimate $\boldsymbol{\mathcal{V}}$ using available reference LD panels (e.g., 1000 Genomes[40]) and inferred expression weights $\boldsymbol{\Omega}$. Second, while we can estimate $\boldsymbol{\beta}$ from data, it will typically be the case that $p \gg m$, which limits inference. To account for this, we make the simplifying assumption that $\boldsymbol{\lambda}_{\text{snp}} = \mathbf{1}_p \lambda_{\text{snp}}$ when conditioned on $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{\lambda}_{\text{pe}}$, which is similar to methods in robust Mendelian Randomization.[9,11,12] Third, estimating $\boldsymbol{\lambda}_{\text{pe}}$ directly from data is also likely to overfit. To bypass this issue, we treat $\boldsymbol{\lambda}_{\text{pe}}$ as a nuisance parameter and assume that $\boldsymbol{\lambda}_{\text{pe}} \mid \mathbf{c}, \sigma_c^2 \sim \mathcal{N}(\mathbf{0}, \mathbf{D_c})$ where $\mathbf{D_c} = \text{diag}(\frac{n\sigma_c^2}{|\mathbf{c}|} \cdot \mathbf{c})$ is the scaled prior causal effect variance and $\mathbf{c}$ is a binary vector indicating if ith gene is causal. Incorporating this prior for causal NCPs enables us to integrate out $\boldsymbol{\lambda}_{\text{pe}}$, which results in the variance component model,

$$\mathbf{z}_{\text{twas}} \mid \lambda_{\text{snp}}, \boldsymbol{\Omega}, \mathbf{V}, \mathbf{c}, n\sigma_c^2 \sim \mathcal{N}(\boldsymbol{\Omega}^\intercal \mathbf{V} \mathbf{1}_p \lambda_{\text{snp}}, \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{V}} \mathbf{D_c} \boldsymbol{\mathcal{V}}).$$

Under this model the variance in $\mathbf{z}_{\text{twas}}$ is due to uncertainty from finite sample size ($\boldsymbol{\mathcal{V}}$) as well as uncertainty in the underlying causal NCPs ($\boldsymbol{\mathcal{V}} \mathbf{D_c} \boldsymbol{\mathcal{V}}$). In principle, we can estimate σ_c^2 using Empirical Bayes; however, this comes at a significant computation cost, as estimation would need to be performed for each causal configuration $\mathbf{c}$ across risk regions. To mitigate this hindrance, we set $n\sigma_c^2 = 40$, which is similar to what we observe at transcriptome-wide significant regions.

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 $\mathbf{c}$,

$$\begin{aligned}
\Pr(\mathbf{c} \mid \mathbf{z}_{\text{twas}}, \lambda_{\text{snp}}, \boldsymbol{\Omega}, \mathbf{V}, \mathbf{c}, n\sigma_c^2) &= \frac{\Pr(\mathbf{z}_{\text{twas}}, \mathbf{c} \mid \lambda_{\text{snp}}, \boldsymbol{\Omega}, \mathbf{V}, \mathbf{c}, n\sigma_c^2)}{\Pr(\mathbf{z}_{\text{twas}} \mid \boldsymbol{\Omega}, \mathbf{V}, n\sigma_c^2)} \\
&= \frac{\mathcal{N}(\mathbf{z}_{\text{twas}} \mid \boldsymbol{\Omega}^\intercal \mathbf{V} \mathbf{1}_p \lambda_{\text{snp}}, \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{V}} \mathbf{D_c} \boldsymbol{\mathcal{V}}) \Pr(\mathbf{c})}{\sum_{\mathbf{c}' \in \mathcal{C}} \mathcal{N}(\mathbf{z}_{\text{twas}} \mid \boldsymbol{\Omega}^\intercal \mathbf{V} \mathbf{1}_p \lambda_{\text{snp}}, \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{V}} \mathbf{D}_{\mathbf{c}'} \boldsymbol{\mathcal{V}}) \Pr(\mathbf{c}')}
\end{aligned}$$

where $\mathcal{C}$ is the set of all binary strings of length m. We assume a Bernoulli prior for each causal indicator $\mathbf{c}_i \sim \text{Bernoulli}(p)$. In practice, we set $p = 1 \times 10^{-3}$. 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 gene. We compute the marginal posterior inclusion probability (PIP) for the ith gene as

$$\text{PIP}(c_i = 1 \mid \mathbf{z}_{\text{twas}}, \lambda_{\text{snp}}, \boldsymbol{\Omega}, \mathbf{V}, n\sigma_c^2) = \sum_{\mathbf{c}' \in \mathcal{C} : c_i' = 1} \Pr(\mathbf{c}' \mid \mathbf{z}_{\text{twas}}, \lambda_{\text{snp}}, \boldsymbol{\Omega}, \mathbf{V}, n\sigma_c^2).$$

Alternatively, we can compute PIPs using Bayes factors for each model (see Supplementary Note). PIPs offer a flexible mechanism to generate gene-sets for functional followup. We use an approximate approach that takes the top k' genes until a percentage ρ of the normalized-posterior mass is explained.

Model validation using the posterior predictive distribution

To test the validity of the FOCUS model at GWAS risk regions in real data, we use a posterior predictive sampling procedure.[16] This approach alternates between sampling causal configurations $\mathbf{c}$ from the posterior distribution and sampling Z-scores $\mathbf{z}^*_{\text{twas}}$ from the generative distribution after conditioning on the causal configuration. This enables us to compare the distribution of simulated data with our observed statistics $\mathbf{z}_{\text{twas}}$. When our observed data $\mathbf{z}_{\text{twas}}$ are not fit within reasonable bounds of the simulated data we can be more confident that the FOCUS model and computed PIPs are inconsistent with the actual data generating

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

process. Specifically, at each risk region with m genes we perform the following:

For trial $t \in [T]$

1. For gene $i \in [m]$

   Sample causal status $c_i \sim \text{Bernoulli}(p_i = \text{PIP}(c_i = 1 \mid \mathbf{z}_{\text{twas}}, \lambda_{\text{snp}}, \mathbf{\Omega}, \mathbf{V}, n\sigma_c^2))$

2. Sample simulated Z-scores $\mathbf{z}^*_{\text{twas}} \sim \mathcal{N}(\mathbf{\Omega}^\intercal \mathbf{V} \mathbf{1}_p \hat{\lambda}_{\text{snp}}, \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{V}} \mathbf{D_c} \boldsymbol{\mathcal{V}})$

3. Output $(t, \mathbf{c}, \mathbf{z}^*_{\text{twas}})$

We compute a posterior Z-score (and p-value) of model fit for the ith gene as $Z_{\text{post},i} = \frac{\text{mean}(\mathbf{z}^*_{\text{twas},i}) - \mathbf{z}_{\text{twas},i}}{\text{sd}(\mathbf{z}^*_{\text{twas},i})}$.

Simulations

We simulated TWAS association statistics starting from real genotype data and gene definitions. To simulate genotype samples, we first partitioned genotype data for 489 individuals of European ancestry in 1000Genomes[40] into independent LD blocks as defined by LDetect.[28] We annotated LD-blocks with all genes in RefSeq[27] whose transcription start site was flanked by region boundaries. To simulate GWAS and expression reference panel genotypes we sampled standardized genotypes using the multivariate Gaussian approximation $\mathbf{x} \sim \mathcal{N}(0, \mathbf{V})$ where $\mathbf{V}$ is LD estimated from the 1000Genomes samples. For both GWAS panel and eQTL reference panel, we simulated gene expression of each gene in the LD-block annotation list by selecting 1 or 2 causal SNPs preferentially located near 100kb of the TSS and then computed $\mathbf{G} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$ where $\mathbf{X}$ is the $n \times p$ centered and standardized genotype matrix, $\mathbf{w}$ are the causal eQTL effects, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1 - h^2_g)$ is random environmental noise. To simulate expression in two correlated tissues, we sample eQTL effects at shared causals under a bi-variate Gaussian distribution as $(\mathbf{w}_{\cdot,1} \mathbf{w}_{\cdot,2}) \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}h^2_{g,1} & \rho_g \\ \rho_g & h^2_{g,2}\end{bmatrix}\right)$ where $h^2_{g,\cdot}$ is the SNP-heritability for gene expression in tissue $\cdot$, and ρ_g is genetic correlation. We repeated this for a total of 25 randomly sampled LD-blocks. We simulated complex trait for the GWAS panel as a linear combination of the genetic components of expression at causal genes. We first sampled causal genes at each LD-block with probability $1/m_i$ where m_i is the number of genes at block i. Then we computed $\mathbf{y} = \sum_i (\mathbf{X}^{(i)} \mathbf{W}^{(i)}) \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}$ where $\boldsymbol{\alpha}_i \sim \mathcal{N}(0, \mathbf{D_c})$ are causal effects for genes in the $i$the LD-block, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1 - h^2_{\text{GE}})$. Next, we performed an association scan for $(\mathbf{y}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(25)})$ and computed SNP-trait Z-scores $\mathbf{z}_{\text{gwas}}$ using Wald statistics from linear regression. To perform a TWAS we fitted weights $\mathbf{\Omega}^{(1)}, \ldots, \mathbf{\Omega}^{(25)}$ for the expression reference panel using GBLUP[36,37] which were used to compute $\mathbf{z}_{\text{twas}}$. We then performed fine-mapping using the FOCUS algorithm on simulated $\mathbf{z}_{\text{twas}}$ vectors. Unless stated otherwise, simulation parameters were set to $N_{\text{gwas}} = 50,000$, $N_{\text{eQTL}} = 500$, expression $h^2_g = 0.2$ and trait $h^2_{\text{GE}} = 0.2$ (i.e. variance explained in trait due to genetic component of gene expression[4]). For proxy-tissue simulations, we used values of proxy-tissue expression $h^2_g \in \{0.01, 0.05, 0.1, 0.2\}$ and $\rho_g \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

Datasets

We downloaded publicly available summary statistics for lipids measurements GWAS.[17] We filtered sites that were not bi-allelic, were ambiguous (i.e. allele 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 publicly available expression weight data as part of the

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

FUSION package. Reference LD was estimated in 1000 Genomes[40] using 489 European individuals. Quality control, cis-heritability of expression, and model fitting have been described elsewhere.[1,4,33] 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 normalized by complete-data h^2_g estimates). This resulted in 15,277 unique genes. Risk regions for FOCUS are $\approx$ 1Mb regions obtained from LDetect[28] that contain at least one genome-wide significant SNP ($P_{\text{gwas}} < 5 \times 10^{-8}$).

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] Huwenbo Shi, Nicholas Mancuso, Sarah Spendlove, and Bogdan Pasaniuc. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *The American Journal of Human Genetics*, 101(5):737–751, 2017.

[6] 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.

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[7] 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.

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

[9] 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.

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

[11] Richard Barfield, Helian Feng, Alexander Gusev, Lang Wu, Wei Zheng, Bogdan Pasaniuc, and Peter Kraft. Transcriptome-wide association studies accounting for colocalization using egger regression. *Genetic epidemiology*, 2018.

[12] Jack Bowden, George Davey Smith, Philip C. Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314, 2016.

[13] 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.

[14] 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.

[15] 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.

[16] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

[17] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, Cristen J Willer, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707, 2010.

[18] 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.

[19] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[20] Alexander Gusev, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics*, 95(5):535–552, 2014.

[21] Naomi R Wray, Jian Yang, Ben J Hayes, Alkes L Price, Michael E Goddard, and Peter M Visscher. Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics*, 14(7):507–515, 2013.

[22] 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.

[23] Brian R Krause and Arthur D Hartman. Adipose tissue and cholesterol metabolism. *Journal of lipid research*, 25(2):97–110, 1984.

[24] Soazig Le Lay, Stéphane Krief, Céline Farnier, Isabelle Lefrère, Xavier Le Liepvre, Raymond Bazin, Pascal Ferré, and Isabelle Dugail. Cholesterol, a cell size-dependent signal that regulates glucose metabolism and gene expression in adipocytes. *Journal of Biological Chemistry*, 276(20):16904–16910, 2001.

[25] 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.

[26] Willeke de Haan, Alpana Bhattacharjee, Piers Ruddle, Martin H Kang, and Michael R Hayden. Abca1 in adipocytes regulates adipose tissue lipid content, glucose tolerance, and insulin sensitivity. *Journal of lipid research*, 55(3):516–523, 2014.

[27] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2015.

[28] Tomaz Berisa and Joseph K Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2):283, 2016.

[29] Sung Chun, Alexandra Casparino, Nikolaos A Patsopoulos, Damien C Croteau-Chonka, Benjamin A Raby, Philip L De Jager, Shamil R Sunyaev, and Chris Cotsapas. Limited statistical evidence for shared genetic effects of eqtls and autoimmune-disease-associated loci in three major immune-cell types. *Nature Genetics*, 49(4):600–605, 2017.

[30] 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.

[31] 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.

bioRxiv preprint doi: https://doi.org/10.1101/236869; this version posted June 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[32] 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.

[33] Alexander Gusev, Nicholas Mancuso, Hyejung Won, Maria Kousi, Hilary K Finucane, Yakir Reshef, Lingyun Song, Alexias Safi, Steven McCarroll, Benjamin M Neale, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature genetics*, 50(4):538, 2018.

[34] 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.

[35] 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.

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

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

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

[39] 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.

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