

1 Probabilistic fine-mapping of transcriptome-wide association studies

2 Nicholas Mancuso¹, Gleb Kichaev², Huwenbo Shi², Malika Freund³, Alexander Gusev⁴, and
3 Bogdan Pasaniuc^{1,2,3}

4 ¹Dept of Pathology and Laboratory Medicine, David Geffen School of Medicine, University
5 of California, Los Angeles, 90024

6 ²Bioinformatics Interdepartmental Program, University of California, Los Angeles, 90024

7 ³Dept of Human Genetics, David Geffen School of Medicine, University of California, Los
8 Angeles, 90024

9 ⁴Dana-Farber Cancer Institute, Boston, 02215

11 Abstract

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

23 Introduction

24 Transcriptome-wide association studies (TWAS) using predicted expression levels have been proposed as
25 an approach to identify novel genomic risk regions and putative risk genes involved for complex traits and
26 diseases.¹⁻³ Since TWAS based on predicted expression only relies on the genetic component of expression,
27 it can be viewed as a test for non-zero local genetic correlation between expression and trait.^{1,4,5} Significant
28 genetic correlation is often interpreted as an estimate of the effect of SNPs on trait mediated by the gene of
29 interest. However, this interpretation requires very strong assumptions that are likely violated in empirical
30 data due to LD and/or pleiotropic SNP effects.^{1-3,6-11} Therefore TWAS has been mostly utilized as a test
31 of association, in contrast to methods that attempt to directly estimate the mediated effect (i.e. Mendelian
32 randomization^{3,6-9}).

33 In this work, we show that the gene-trait association statistics from TWAS at a known risk region are
34 correlated as a function of LD among SNPs and eQTL weights used in the prediction models. This effect
35 is similar to LD-tagging in genome-wide association studies (GWAS) where LD within a region induces
36 associations at tag SNPs (yielding the traditional Manhattan-style plots). Even in the simplest case where

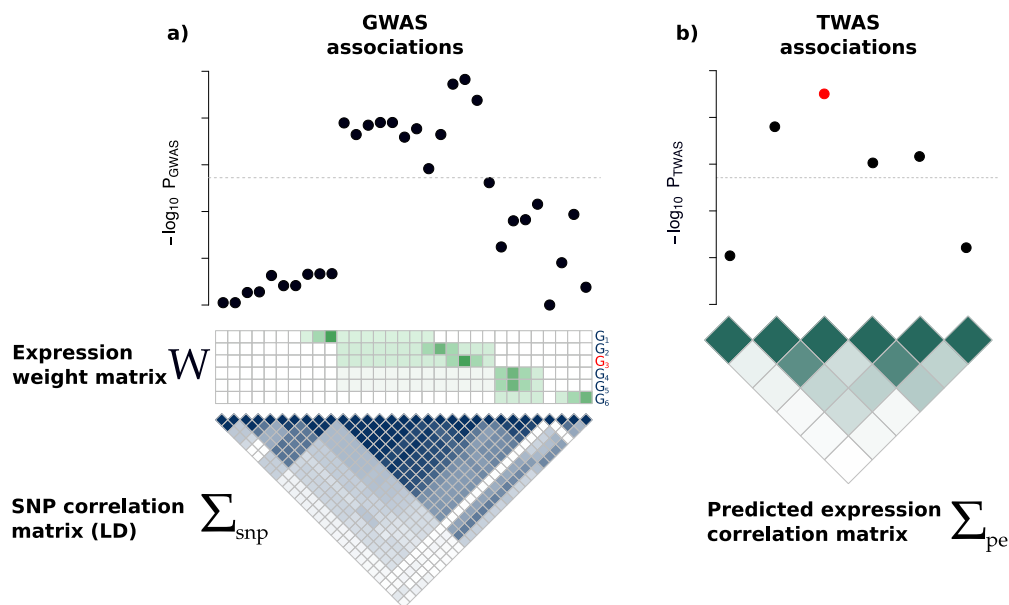


Figure 1: Illustration of the induced correlation structure for predicted expression. a) Top: Manhattan plot indicating strength of SNP association with trait. Middle: Expression weight matrix for 6 genes in the same region, with the causal gene in red. Each row corresponds to a gene and each column 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.

37 a single SNP causally impacts the expression of a gene which in turn causally impacts a trait, LD among
 38 SNPs used in the eQTL prediction models induce significant gene-trait associations at nearby non-causal
 39 genes in the region. The tagging effect is further exacerbated in the presence of multiple causal SNPs and
 40 genes. As an illustrative example, consider a risk region with 6 genes where a single SNP is causal for a single
 41 gene which impacts trait (causal gene in red; no other causal genes are present at this region, see Figure 1).
 42 Although genes 3 and 4 in Figure 1 have non-overlapping prediction weights due to different eQTL genetic
 43 regulation, LD among SNPs with non-zero prediction weights induce correlations in the TWAS statistics at
 44 genes 3 and 4. Estimating the correlation structure between predicted expression among nearby genes enables
 45 statistical fine-mapping over gene-trait associations. However, we note several confounding factors need to
 46 be addressed for unbiased inference. First, there is a body of evidence supporting horizontal pleiotropic
 47 effects from SNPs,^{9, 11, 12} which bias gene-trait association statistics and should be accounted for. Second, it
 48 is critical that TWAS fine-mapping approaches be robust if the causal mechanism is not steady-state levels
 49 of gene expression. Fine-mapping in these instances without controlling for confounding will result in a
 50 misspecified causal model and likely prioritize genes that tag causal mechanisms well.

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

56 any given set of genes to explain the TWAS signal. Our approach extends probabilistic SNP fine-mapping
57 approaches^{13–15} to estimate sets of genes that contain the “causal” genes (defined here as the gene responsible
58 for the association signal) at a predefined confidence level (i.e. ρ gene credible set). FOCUS accounts for bias
59 due to missing causal factors by including the *null model* as a possible explanatory factor in the credible set.
60 We perform extensive simulations and show that FOCUS is unbiased in estimating the posterior probabilities
61 and credible sets at a specified certainty when the causal gene is present in the data. When the causal tissue is
62 unavailable and alternative tissues with correlated expression levels are used as a proxy, FOCUS maintains
63 its performance under standard assumptions. FOCUS outputs posterior predictive checks¹⁶ of observed
64 TWAS Z-scores to quantify model agreement given inferred posterior probabilities for causality. Finally, as
65 a demonstration using real GWAS data, we apply FOCUS to four GWASs of lipids levels.¹⁷ We find that
66 FOCUS prioritizes genes with established roles in LDL risk (e.g., *SORT1*).¹⁸

67 Results

68 Methods Overview

69 To disentangle between causal and tagging gene-trait associations at a TWAS significant region, we analyt-
70 ically derive the covariance structure among TWAS statistics as function of LD and eQTL weights used in
71 prediction. Next, we model the entire vector of marginal TWAS association statistics (\mathbf{z}_{twas}) at all genes in
72 a region (TWAS significant and not-significant) using a multivariate Gaussian distribution parameterized by
73 the effect sizes at causal genes ($\boldsymbol{\lambda}_{\text{pe}}$), residual SNP-effects ($\boldsymbol{\lambda}_{\text{snp}}$), and the correlation structure induced by
74 inferred expression weights ($\boldsymbol{\Omega}$) with LD (\mathbf{V}) as

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

75 (see Methods). We control for bias resulting from pleiotropic effects of SNPs by including an intercept term
76 that quantifies the average SNP effect sizes ($\boldsymbol{\lambda}_{\text{snp}}$) tagged by predicted expression ($\boldsymbol{\Omega}^T \mathbf{V} \boldsymbol{\lambda}_{\text{snp}}$; see Methods).
77 To allow for genes without prediction models in the relevant tissue (either due to QC and/or low power
78 in eQTL studies), we leverage recent work demonstrating that eQTLs are largely shared across tissues¹⁹
79 and include prediction models from proxy tissues for such genes (see Methods). We employ a standard
80 Bayesian approach to compute the marginal posterior inclusion probability (PIP) for each gene in the region
81 to be causal. To avoid overfitting, we integrate out the unknown causal effects $\boldsymbol{\lambda}_{\text{pe}}$ using a multivariate
82 Gaussian prior (see Methods). We use PIPs to compute ρ -credible gene-sets that contain the causal gene
83 with probability ρ .¹⁴ To account for missing causal mechanisms either due to unpredicted expression or
84 other latent functional mechanisms, we include the null model as a possible outcome in the credible set
85 (see Methods). Lastly, we use a simulation-based procedure to compute posterior predictive checks¹⁶ that
86 measure the FOCUS model’s goodness-of-fit given observed TWAS Z-scores.

87 FOCUS prioritizes causal genes in causal-tissue simulations

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

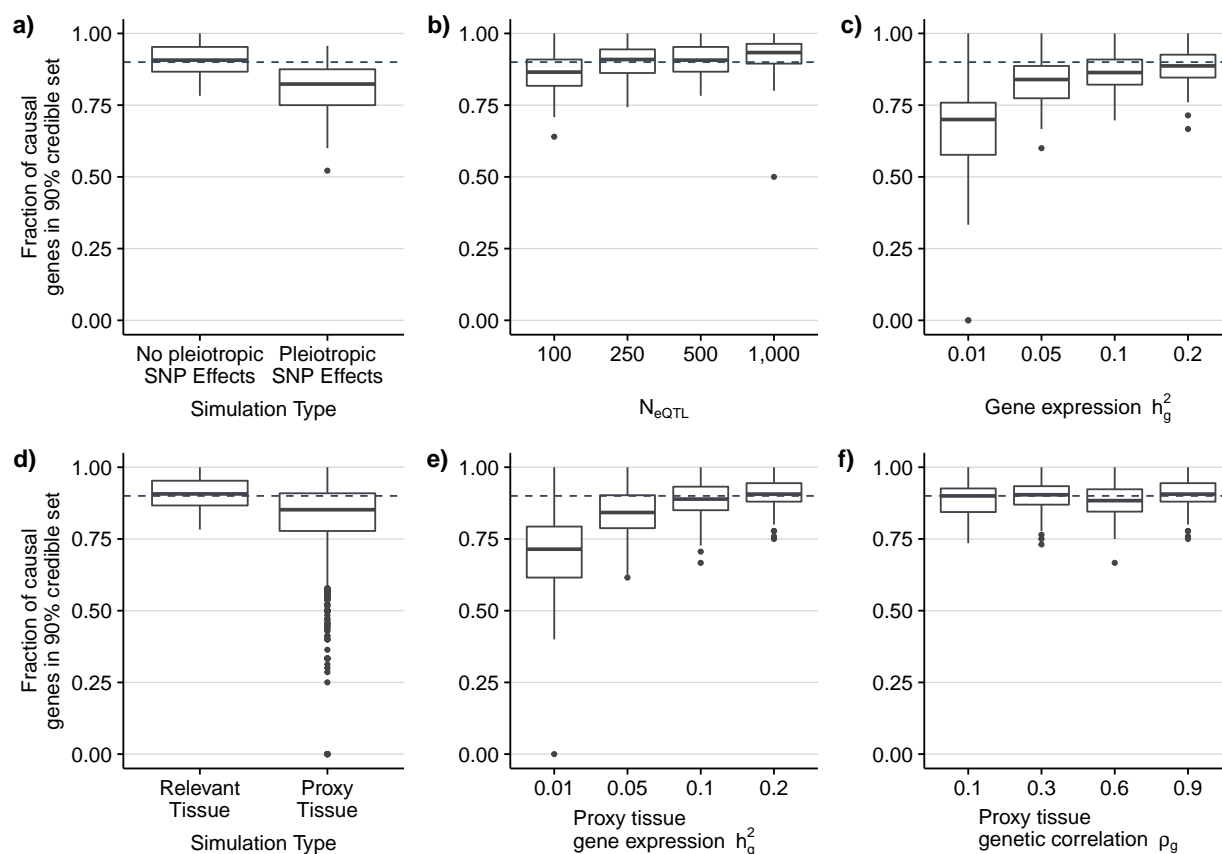


Figure 2: Credible gene-sets are well-calibrated in simulations. Box-plots represent the distribution of the fraction of causal genes captured in the 90% credible set over simulations (see Methods). a) Simulations with and without pleiotropic SNP effects on trait. Prediction models were trained using 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.

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

104 the causal gene in 88% of simulations, which is significantly fewer when compared with 94% for $N_{\text{eQTL}} = 500$
105 (Mann-Whitney-U $P = 5.46 \times 10^{-9}$). Next, we explored how underlying heritability of expression at causal
106 genes impacts prioritization. Heritability defines the prediction upper bound for SNP-based approaches,^{20,21}
107 and we expect performance to improve as non-zero heritability is easier to detect. We confirmed that perfor-
108 mance increased with heritability of causal gene expression (see Figure 2). For example, we simulated gene
109 expression having heritability $h_g^2 = 0.01$ and inferred eQTL weights using $N_{\text{eQTL}} = 500$ and found a
110 significant decrease in performance (Mann-Whitney-U $P < 2.2 \times 10^{-16}$). Similarly, we looked at the role of
111 the prior effect-size distribution for predicted gene expression⁴ and found performance to remain relatively
112 stable for a wide range of values (see Supplementary Figures 4, 5).

113 **FOCUS remains stable when using proxy tissues**

114 Next, we investigated the performance of FOCUS when the causal gene in the relevant tissue is missing, but
115 is measured in a different tissue (see Methods). In real data a gene may act through a tissue that is difficult
116 to assay in large sample sizes, but may have similar cis-regulatory patterns in tissues that are easier to collect
117 (e.g., blood, LCLs). Indeed, several studies^{1,4,19,22} established cis-regulated gene expression levels exhibit
118 high genetic correlation across tissues and functional architectures. The intuition in this approach is that
119 the loss in power from using the correlated tissue is offset by the gain in power due to larger sample size. To
120 simulate gene expression in a proxy tissue, we drew correlated effect sizes at the same eQTLs for the gene
121 expression reference panel (see Methods). Here, we consider a causal gene to be successfully fine-mapped if its
122 corresponding proxy tissue model is in the 90%-credible gene-set. When sample size for eQTL in the relevant-
123 and proxy-tissues are the same, but heritability in proxy tissue is lower than the relevant-tissue, we found a
124 significant loss in accuracy, with 90% credible sets capturing the causal gene 0.83 (S.D. 0.08) of simulations
125 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
126 Figure 2). This effect was not observed when heritability of proxy tissue gene expression was at least that of
127 expression in the relevant-tissue (Mann-Whitney-U $P = 0.06$). For example, when expression in the relevant
128 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
129 in significantly fewer simulations (0.79 versus 0.89; Mann-Whitney-U $P < 2.2 \times 10^{-16}$), which suggests that
130 when causal eQTLs are shared across tissues, increased heritability of expression increases power to detect
131 the causal gene. Surprisingly, we found correlation of effect-sizes at shared eQTLs to play no significant role
132 in performance when heritability of expression in the relevant and proxy tissue is kept the same ($h_g^2 = 0.2$;
133 see Figure 2, Supplementary Figure 6). In the case of zero correlation between effect sizes at the same eQTL
134 SNPs, this result can be interpreted as pleiotropic effects on independent molecular traits, which are known
135 to be difficult to differentiate from a causal effect.^{1,3,9} Collectively, these results demonstrate that FOCUS
136 is relatively robust to model perturbations and performs well when underlying tissue-specific causal genes
137 are represented by proxy tissue eQTL weights.

138 **FOCUS is robust to missing causal molecular mechanisms**

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

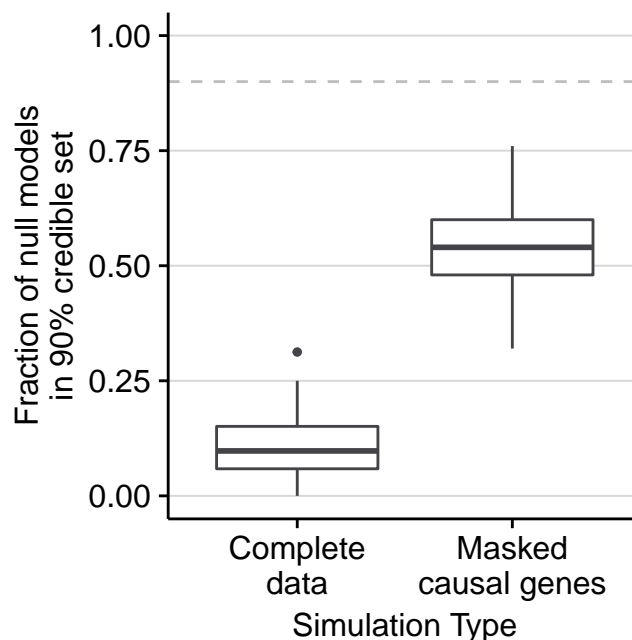


Figure 3: FOCUS credible-sets alleviate bias when the causal gene is missing. “Masked” indicates simulations where the causal genes are pruned before analysis. For comparison, we include results under the complete-data simulation pipeline where causal genes are tested. Box-plots represent the distribution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

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

158 FOCUS improves resolution for fine mapping causal genes

159 Having demonstrated that FOCUS computes well-calibrated credible gene-sets under a wide range of param-
160 eters we next sought to quantify the resolution of credible gene-sets to identify causal genes. In particular,
161 we estimated the average number of genes captured in the credible set. We found 90%-credible gene-sets
162 contained 4.4 genes on average (S.D. 1.9) in the relevant-tissue simulations, which resulted in an average
163 47% of predicted genes per risk region (see Figure 4). We found a similar number of genes in 90%-credible

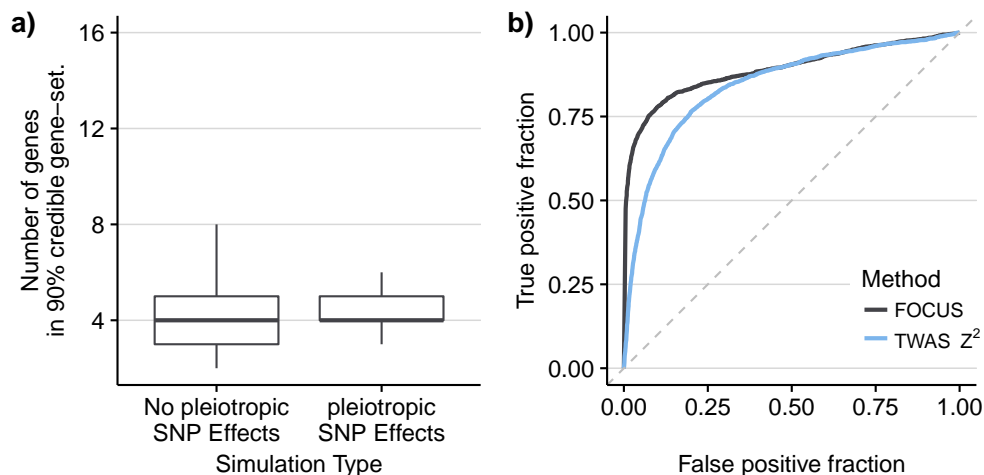


Figure 4: FOCUS accurately prioritizes causal genes in simulations. Box-plots represent the distribution of the total number of causal genes captured in the 90% credible sets over simulations (see Methods). a) Simulations with and without pleiotropic SNP effects on trait. Prediction models were trained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

164 gene-sets across simulations when varying model parameters and sample sizes (see Supplementary Figures
165 7-12). While we advocate the use of credible-sets in practice rather than thresholding on PIPs, for complete-
166 ness we prioritized genes using PIPs for direct comparison with TWAS p-values (see Methods). We found
167 prioritizing genes using PIPs outperformed TWAS p-value ranking at capturing underlying causal genes (see
168 Figure 4). For example, at a false positive rate of 5%, FOCUS identifies 162% more causal genes than a
169 simple rank of marginal TWAS statistics. A unique feature of FOCUS is that it allows for multiple causal
170 genes at a given region; in this scenario FOCUS attains a gain of 213% more causal genes compared to
171 that of 132% for single causal regions (see Supplementary Table 2). Overall, FOCUS accurately prioritizes
172 causal genes from non-causal genes with the largest gains when multiple causal genes exist at risk regions.

173 Application to lipids GWAS data

174 Having validated our fine-mapping approach in simulations, we illustrate FOCUS by re-analyzing a large-
175 scale GWAS of lipids measurements¹⁷ with eQTL weights from adipose tissue. We assume the relevant
176 tissue for expression driving lipids is adipose given its well-characterized role.²³⁻²⁶ To account for missing
177 gene prediction models, we incorporate gene expression models for genes not predictable at current sample
178 sizes from adipose tissue across 45 tissues measured in 47 reference panels. In detail, for a gene without
179 a predicted model in adipose tissue, we include the prediction model with best accuracy across all other
180 tissues (see Supplementary Table 1; see Methods). Of the 26,292 known genes in RefSeq (ver 65),²⁷ we
181 found 12,663 covered in our data with the remaining 2,614 genes not found in RefSeq. Adipose-prioritized
182 TWAS identifies 301 (202 unique) significant genes at 108 (63 unique) independent regions after accounting
183 for the total number of per-trait tests performed ($P < 0.05/15,277$; see Supplementary Figures 13-15;
184 Table 1; Supplementary Table 3). Of the 160 (89 unique) risk regions found through GWAS, 75 (46 unique)
185 overlapped significant TWAS results, which is increased compared with earlier work²⁹ that found 25% overlap
186 between GWAS and eQTL at risk regions (see Table 1).

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

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 LDetect²⁸ 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 $P < 0.05/15,277$.

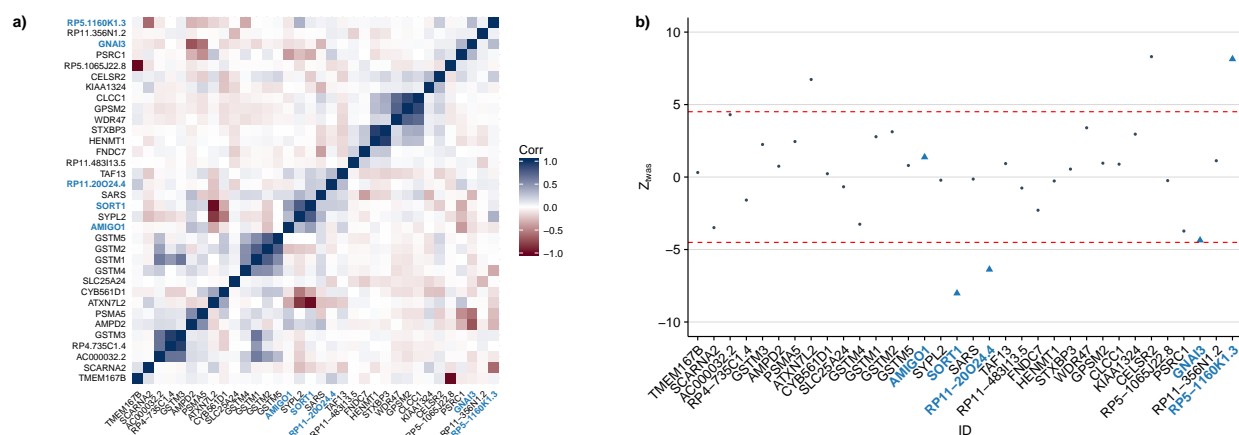


Figure 5: 1p13 locus for LDL. a) Correlation for predicted expression at 1p13 locus. Genes in the 90%-credible set are labeled in light blue. b) TWAS Z-scores at 1p13 locus. Each point represents the association strength for each tested gene. Genes in the 90%-credible gene-set are labeled in light blue.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

188 framework. We applied FOCUS at the 75 GWAS risk regions with evidence for regulatory action on genes
 189 driving lipids levels to compute PIPs and estimate credible sets of genes at each of the regions (see Methods).
 190 We found that observed risk regions can be explained by 1.5 causal genes on average, with 61/75 risk regions
 191 containing fewer than 2 causal genes in expectation (see Supplementary Figure 18). The average maximum
 192 PIP across credible sets was 88% (and decreased exponentially for lower ranked genes; see Supplementary
 193 Figure 17). Together, these results imply that most risk regions can be explained by a single causal gene.
 194 Using computed PIPs, we estimated 90%-credible gene-sets for each risk region and found a significant
 195 reduction in the number of prioritized genes (mean 1.9), compared with transcriptome-wide significant genes
 196 (mean 3.2; one-sided Mann-Whitney-U $P = 7.24 \times 10^{-4}$; Supplementary Figure 17; Supplementary Table
 197 4). As a positive control, we examined the 1p13 locus for LDL, as this region harbors risk SNP rs12740374
 198 which has been shown to perturb transcription of *SORT1* and impact downstream LDL levels.¹⁸ We found
 199 4/34 genes included in the 90% credible set, of which *SORT1* had a posterior probability 95% (see Figure
 200 5).

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

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

211 Discussion

212 In this work we presented FOCUS, a fine-mapping approach that estimates credible sets of causal genes
213 using prediction eQTL weights, LD, and GWAS summary statistics. We demonstrated FOCUS adequately
214 controls false positives in null simulations and outperforms straightforward p-value ranking in identifying
215 causal genes when genes at a region impact downstream trait. We found 90%-credible gene-sets to be largely
216 stable across a variety of simulations, with the biggest impact in performance due to eQTL reference panel
217 sample size and SNP-heritability of gene expression. We applied FOCUS to four lipids TWASs (e.g., HDL,
218 LDL, triglyceride, and total cholesterol levels) and found *SORT1* correctly identified as a putative causal
219 gene. Interestingly, our real-data results in lipids suggests most regions can be explained by a single causal
220 gene. Overall, our results highlight the utility of using credible sets in prioritizing causal genes by jointly
221 assigning posterior probabilities, that are both easily interpretable and comparable across genes and regions.

222 In addition to providing a quantification of the confidence in how many genes need to be validated to identify
223 the causal genes in the region, our probabilistic approach yields several benefits. First, FOCUS naturally
224 allows for multiple causal SNPs and genes while integrating gene-effect sizes using conjugate priors; this
225 is particularly important as recent works have shown that allelic heterogeneity (i.e. multiple causal genes
226 and SNPs at a region) is pervasive in both eQTL and GWAS.^{19,30} Second, in this work, we investigate
227 predicted gene expression, but FOCUS could generally be applied to other predicted molecular traits with
228 an established role in complex trait etiology (e.g., alternatively spliced exons^{31,32}). For example, several
229 recent works have supporting evidence for splice variation playing an important role in driving risk of
230 schizophrenia.^{33,34}

231 We showed our approach is well calibrated under various simulations and robust to perturbations in model
232 assumptions; however, several limitations still exist. First, our model assumes that complex trait or disease
233 risk is a linear function of steady-state expression levels at causal genes. Several works have demonstrated
234 that risk prediction using a linear combination of predicted steady-state or observed expression levels can
235 outperform standard SNP-based models,^{33,35} which supports a linear model of gene expression impacting
236 complex trait or disease risk. However, higher-order models that capture complex regulatory networks of
237 transcription factors and gene expression may also reflect underlying biology. As reference gene expression
238 data sets grow in size, accurately modeling these assumptions may be possible. Similarly, if risk is mediated
239 through context-specific expression and not steady-state expression levels, then FOCUS will have a loss in
240 performance. Second, while our simulations used GBLUP^{36,37} throughout for its straightforward imple-
241 mentation, we recommend a cross-validation approach to select the best fitting linear model (e.g., GBLUP,
242 BSLMM³⁸) using the ratio of out-of-sample prediction accuracy normalized by the total SNP-heritability
243 of gene expression, which is implemented in the FUSION framework.^{1,33} Third, when the causal gene is

244 untyped in the data, our approach will partially inflate posterior probabilities at tagging genes. We attempt
245 to mitigate this scenario by adding an intercept term to the model and incorporating gene models measured
246 in proxy tissues. We caution that our simulated results using proxy-tissues were performed using a model
247 where causal eQTLs are shared between proxy- and relevant tissues and have correlated effect sizes, which
248 is equivalent to a random-effects model. This assumption may be violated in real data if causal eQTLs are
249 tissue-specific. Recent work, however, has demonstrated that a large number of eQTLs are indeed shared
250 across tissues.¹⁹ Fourth, we took a tissue-prioritizing approach by preferentially using eQTL weights in
251 adipose tissue given its known role in lipids^{23–26} for our real-data analysis. This approach may not always
252 be possible for complex traits or diseases with less understood biology. However, recent work has shown
253 that the most relevant (i.e. likely causal) tissue for complex traits can be accurately estimated using eQTL
254 data.³⁹ Coupled with estimation of causal tissue, we suggest prioritizing genes with high normalized pre-
255 diction accuracy in related tissues. We note that our results were strongly dependent on sample size in
256 the eQTL reference panel, which is reflected in expression prediction accuracy. We therefore, recommend
257 prioritizing eQTL data with sample sizes greater than 100 if possible and performing inference on genes with
258 robustly non-zero SNP-heritability. Despite our modeling assumptions and limitations, our approach is a
259 step towards accurately prioritizing gene-sets.

260 Online Methods

261 Notation

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

266 Model and sampling distribution of marginal TWAS summary statistics

267 We model quantitative trait for n individuals \mathbf{y} by a linear combination of expression levels for m genes
268 $\mathbf{G} \in \mathbb{R}^{n \times m}$ as

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

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

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

274 where $\boldsymbol{\epsilon} = \mathbf{E}\boldsymbol{\alpha} + \tilde{\boldsymbol{\epsilon}}$ is the total contribution from environment, which we parameterize as $\mathbb{E}[\boldsymbol{\epsilon}] = 0$ and
275 $\mathbb{V}[\boldsymbol{\epsilon}] = \mathbf{I}_n \sigma_e^2$ which is valid provided independence of errors holds. If causal eQTL effect-sizes \mathbf{W} were
276 known, we could prioritize putative susceptibility genes by estimating $\boldsymbol{\alpha}$ using regression. Unfortunately,
277 effect-sizes \mathbf{W} are unknown and must be estimated from data (e.g., BSLMM,³⁸ GBLUP^{36,37}). Because

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

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

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

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

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

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

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

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

288 To simplify notation we re-parameterize the causal effects as a non-centrality parameter (NCP) at the causal
 289 genes by $\boldsymbol{\lambda}_{\text{pe}} = \frac{\sqrt{n}\boldsymbol{\alpha}}{\sigma_e}$. We note that $\boldsymbol{\Omega}^\top \mathbf{V}\mathbf{W} \neq \boldsymbol{\Omega}^\top \mathbf{V}\boldsymbol{\Omega}$, but given large-enough sample sizes we expect $\boldsymbol{\Omega}^\top \mathbf{V}\boldsymbol{\Omega}$
 290 to approach $\boldsymbol{\Omega}^\top \mathbf{V}\mathbf{W}^1$. We denote predicted expression covariance as $\boldsymbol{\nu} = \boldsymbol{\Omega}^\top \mathbf{V}\boldsymbol{\Omega}$. The NCP $\boldsymbol{\lambda}_{\text{pe}}$ governs the
 291 statistical power of rejecting the null of no effect of predicted expression on trait ($\boldsymbol{\alpha} = 0$). We parameterize
 292 $\boldsymbol{\beta}$ 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}_{twas} is given
 293 by,

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

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

¹Penalized 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.

298 Computing the likelihood of \mathbf{z}_{twas} as described requires knowing \mathbf{V} , λ_{snp} , and λ_{pe} , which are unknown
 299 a-priori. First, we can estimate \mathbf{V} using available reference LD panels (e.g., 1000 Genomes⁴⁰) and inferred
 300 expression weights $\mathbf{\Omega}$. Second, while we can estimate β from data, it will typically be the case that $p \gg m$,
 301 which limits inference. To account for this, we make the simplifying assumption that $\lambda_{\text{snp}} = \mathbf{1}_p \lambda_{\text{snp}}$ when
 302 conditioned on \mathbf{V} and λ_{pe} , which is similar to methods in robust Mendelian Randomization.^{9,11,12} Third,
 303 estimating λ_{pe} directly from data is also likely to overfit. To bypass this issue, we treat λ_{pe} as a nuisance
 304 parameter and assume that $\lambda_{\text{pe}} | \mathbf{c}, \sigma_c^2 \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_{\mathbf{c}})$ where $\mathbf{D}_{\mathbf{c}} = \text{diag}(\frac{n\sigma_c^2}{|\mathbf{c}|} \cdot \mathbf{c})$ is the scaled prior causal effect
 305 variance and \mathbf{c} is a binary vector indicating if i th gene is causal. Incorporating this prior for causal NCPs
 306 enables us to integrate out λ_{pe} , which results in the variance component model,

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

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

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

$$\begin{aligned} \Pr(\mathbf{c} | \mathbf{z}_{\text{twas}}, \lambda_{\text{snp}}, \mathbf{\Omega}, \mathbf{V}, \mathbf{c}, n\sigma_c^2) &= \frac{\Pr(\mathbf{z}_{\text{twas}}, \mathbf{c} | \lambda_{\text{snp}}, \mathbf{\Omega}, \mathbf{V}, \mathbf{c}, n\sigma_c^2)}{\Pr(\mathbf{z}_{\text{twas}} | \mathbf{\Omega}, \mathbf{V}, n\sigma_c^2)} \\ &= \frac{\mathcal{N}(\mathbf{z}_{\text{twas}} | \mathbf{\Omega}^T \mathbf{V} \mathbf{1}_p \lambda_{\text{snp}}, \mathbf{V} + \mathbf{V} \mathbf{D}_{\mathbf{c}} \mathbf{V}) \Pr(\mathbf{c})}{\sum_{\mathbf{c}' \in \mathcal{C}} \mathcal{N}(\mathbf{z}_{\text{twas}} | \mathbf{\Omega}^T \mathbf{V} \mathbf{1}_p \lambda_{\text{snp}}, \mathbf{V} + \mathbf{V} \mathbf{D}_{\mathbf{c}'} \mathbf{V}) \Pr(\mathbf{c}')} \end{aligned}$$

314 where \mathcal{C} is the set of all binary strings of length m . We assume a Bernoulli prior for each causal indicator
 315 $c_i \sim \text{Bernoulli}(p)$. In practice, we set $p = 1 \times 10^{-3}$. This assumption is likely violated when signal for \mathbf{z}_{twas}
 316 is low, and we recommend only including regions with at least one transcriptome-wide significant gene. We
 317 compute the marginal posterior inclusion probability (PIP) for the i th gene as

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

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

321 Model validation using the posterior predictive distribution

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

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

329 For trial $t \in [T]$

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

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

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

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

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

335 Simulations

336 We simulated TWAS association statistics starting from real genotype data and gene definitions. To sim-
 337 ulate genotype samples, we first partitioned genotype data for 489 individuals of European ancestry in
 338 1000Genomes⁴⁰ into independent LD blocks as defined by LDetect.²⁸ We annotated LD-blocks with all
 339 genes in RefSeq²⁷ whose transcription start site was flanked by region boundaries. To simulate GWAS and
 340 expression reference panel genotypes we sampled standardized genotypes using the multivariate Gaussian ap-
 341 proximation $\mathbf{x} \sim \mathcal{N}(0, \mathbf{V})$ where \mathbf{V} is LD estimated from the 1000Genomes samples. For both GWAS panel
 342 and eQTL reference panel, we simulated gene expression of each gene in the LD-block annotation list by se-
 343 lecting 1 or 2 causal SNPs preferentially located near 100kb of the TSS and then computed $\mathbf{G} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$ where
 344 \mathbf{X} is the $n \times p$ centered and standardized genotype matrix, \mathbf{w} are the causal eQTL effects, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1 - h_g^2)$
 345 is random environmental noise. To simulate expression in two correlated tissues, we sample eQTL effects at
 346 shared causals under a bi-variate Gaussian distribution as $(\mathbf{w}_{\cdot,1}, \mathbf{w}_{\cdot,2}) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} h_{g,1}^2 & \rho_g \\ \rho_g & h_{g,2}^2 \end{bmatrix}\right)$ where $h_{g,\cdot}^2$ is
 347 the SNP-heritability for gene expression in tissue \cdot , and ρ_g is genetic correlation. We repeated this for a total
 348 of 25 randomly sampled LD-blocks. We simulated complex trait for the GWAS panel as a linear combination
 349 of the genetic components of expression at causal genes. We first sampled causal genes at each LD-block with
 350 probability $1/m_i$ where m_i is the number of genes at block i . Then we computed $\mathbf{y} = \sum_i (\mathbf{X}^{(i)} \mathbf{W}^{(i)}) \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}$
 351 where $\boldsymbol{\alpha}_i \sim \mathcal{N}(0, \mathbf{D}_c)$ are causal effects for genes in the i th LD-block, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1 - h_{\text{GE}}^2)$. Next, we
 352 performed an association scan for $(\mathbf{y}, \mathbf{X}^{(1)}, \dots, \mathbf{X}^{(25)})$ and computed SNP-trait Z-scores \mathbf{z}_{gwas} using Wald
 353 statistics from linear regression. To perform a TWAS we fitted weights $\boldsymbol{\Omega}^{(1)}, \dots, \boldsymbol{\Omega}^{(25)}$ for the expression
 354 reference panel using GBLUP^{36,37} which were used to compute \mathbf{z}_{twas} . We then performed fine-mapping using
 355 the FOCUS algorithm on simulated \mathbf{z}_{twas} vectors. Unless stated otherwise, simulation parameters were set
 356 to $N_{\text{gwas}} = 50,000$, $N_{\text{eQTL}} = 500$, expression $h_g^2 = 0.2$ and trait $h_{\text{GE}}^2 = 0.2$ (i.e. variance explained in trait
 357 due to genetic component of gene expression⁴). For proxy-tissue simulations, we used values of proxy-tissue
 358 expression $h_g^2 \in \{0.01, 0.05, 0.1, 0.2\}$ and $\rho_g \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

359 Datasets

360 We downloaded publicly available summary statistics for lipids measurements GWAS.¹⁷ We filtered sites
 361 that were not bi-allelic, were ambiguous (i.e. allele 1 is reverse complement with allele 2), or had MAF
 362 less than 0.01. To perform TWAS on each of the lipids traits we used the software FUSION (see URLs).
 363 FUSION takes a summary-based approach to TWAS and requires as input GWAS summary statistics (i.e.
 364 SNP Z-scores) and eQTL weights. We downloaded publicly available expression weight data as part of the

365 FUSION package. Reference LD was estimated in 1000 Genomes⁴⁰ using 489 European individuals. Quality
366 control, cis-heritability of expression, and model fitting have been described elsewhere.^{1,4,33} We prioritized
367 adipose for our TWAS approach and used other reference panels as to act as proxy for adipose. That is,
368 for all possible tissue-specific gene models in a region we first test predicted expression using adipose gene
369 models. Then for the remaining genes found only in proxy tissue models, we select those with the best
370 prediction accuracy (i.e. out-of-sample R^2 normalized by complete-data h_g^2 estimates). This resulted in
371 15,277 unique genes. Risk regions for FOCUS are \approx 1Mb regions obtained from LDetect²⁸ that contain at
372 least one genome-wide significant SNP ($P_{\text{gwas}} < 5 \times 10^{-8}$).

373 URLs

374 FOCUS: <http://github.com/bogdanlab/focus/>

375 FUSION: <http://gusevlab.org/projects/fusion/>

376 Lipids GWAS: <http://lipidgenetics.org/>

377 References

- 378 [1] A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. Penninx, R. Jansen, E. de Geus, DI. Boomsma,
379 FA. Wright, PF. Sullivan, E. Nikkola, M. Alvarez, M. Civelek, AJ. Lusi, T. Lehtimäki, E. Raitoharju,
380 M. Kähönen, I. Seppälä, OT. Raitakari, J. Kuusisto, M. Laakso, AL. Price, P. Pajukanta, and B. Pasa-
381 niuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*,
382 2016.
- 383 [2] Eric R. Gamazon, Heather E. Wheeler, Kaanan P. Shah, Sahar V. Mozaffari, Keston Aquino-Michaels,
384 Robert J. Carroll, Anne E. Eyler, Joshua C. Denny, G. TEx Consortium, Dan L. Nicolae, Nancy J. Cox,
385 and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome
386 data. *Nat Genet*, 47(9):1091–1098, 2015.
- 387 [3] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R. Robinson, Joseph E. Powell, Grant W.
388 Montgomery, Michael E. Goddard, Naomi R. Wray, Peter M. Visscher, and Jian Yang. Integration of
389 summary data from gwas and eqtl studies predicts complex trait gene targets. *Nat Genet*, advance
390 online publication, 2016.
- 391 [4] Nicholas Mancuso, Huwenbo Shi, Pagé Goddard, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc.
392 Integrating gene expression with summary association statistics to identify genes associated with 30
393 complex traits. *The American Journal of Human Genetics*, 100(3):473–487, 2017.
- 394 [5] Huwenbo Shi, Nicholas Mancuso, Sarah Spendlove, and Bogdan Pasaniuc. Local genetic correlation
395 gives insights into the shared genetic architecture of complex traits. *The American Journal of Human*
396 *Genetics*, 101(5):737–751, 2017.
- 397 [6] George Davey Smith and Shah Ebrahim. ‘mendelian randomization’: can genetic epidemiology con-
398 tribute to understanding environmental determinants of disease? *International Journal of Epidemiology*,
399 32(1):1–22, 2003.

- 400 [7] D. A. Lawlor, R. M. Harbord, J. A. Sterne, N. Timpson, and S. G. Davey. Mendelian randomization:
401 using genes as instruments for making causal inferences in epidemiology. *Stat Med*, 27, 2008.
- 402 [8] B. L. Pierce and S. Burgess. Efficient design for mendelian randomization studies: subsample and
403 2-sample instrumental variable estimators. *Am J Epidemiol*, 178, 2013.
- 404 [9] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid in-
405 struments: effect estimation and bias detection through egger regression. *International Journal of*
406 *Epidemiology*, 44(2):512–525, 2015.
- 407 [10] Michael Wainberg, Nasa Sinnott-Armstrong, David Knowles, David Golan, Raili Ermel, Arno Ru-
408 usalepp, Thomas Quertermous, Ke Hao, Johan LM Bjorkegren, Manuel A Rivas, et al. Vulnerabilities
409 of transcriptome-wide association studies. *bioRxiv*, page 206961, 2017.
- 410 [11] Richard Barfield, Helian Feng, Alexander Gusev, Lang Wu, Wei Zheng, Bogdan Pasaniuc, and Peter
411 Kraft. Transcriptome-wide association studies accounting for colocalization using egger regression.
412 *Genetic epidemiology*, 2018.
- 413 [12] Jack Bowden, George Davey Smith, Philip C. Haycock, and Stephen Burgess. Consistent estimation in
414 mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic*
415 *Epidemiology*, 40(4):304–314, 2016.
- 416 [13] Julian B Maller, Gilean McVean, Jake Byrnes, Damjan Vukcevic, Kimmo Palin, Zhan Su, Joanna MM
417 Howson, Adam Auton, Simon Myers, Andrew Morris, et al. Bayesian refinement of association signals
418 for 14 loci in 3 common diseases. *Nature genetics*, 44(12):1294–1301, 2012.
- 419 [14] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying
420 causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 2014.
- 421 [15] Gleb Kichaev, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L. Price,
422 Peter Kraft, and Bogdan Pasaniuc. Integrating functional data to prioritize causal variants in statistical
423 fine-mapping studies. *PLoS Genet*, 10(10):e1004722, 2014.
- 424 [16] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin.
425 *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.
- 426 [17] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianos,
427 Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, Cristen J Willer, et al.
428 Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707, 2010.
- 429 [18] Kiran Musunuru, Alanna Strong, Maria Frank-Kamenetsky, Noemi E. Lee, Tim Ahfeldt, Katherine V.
430 Sachs, Xiaoyu Li, Hui Li, Nicolas Kuperwasser, Vera M. Ruda, James P. Pirruccello, Brian Muchmore,
431 Ludmila Prokunina-Olsson, Jennifer L. Hall, Eric E. Schadt, Carlos R. Morales, Sissel Lund-Katz,
432 Michael C. Phillips, Jamie Wong, William Cantley, Timothy Racie, Kenechi G. Ejebe, Marju Orho-
433 Melander, Olle Melander, Victor Koteliansky, Kevin Fitzgerald, Ronald M. Krauss, Chad A. Cowan,
434 Sekar Kathiresan, and Daniel J. Rader. From noncoding variant to phenotype via sort1 at the 1p13
435 cholesterol locus. *Nature*, 466(7307):714–719, Aug 2010.
- 436 [19] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204,
437 2017.

- 438 [20] Alexander Gusev, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi
439 Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, et al. Partitioning heritability of regulatory
440 and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics*,
441 95(5):535–552, 2014.
- 442 [21] Naomi R Wray, Jian Yang, Ben J Hayes, Alkes L Price, Michael E Goddard, and Peter M Visscher.
443 Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics*, 14(7):507–515, 2013.
- 444 [22] Xuanyao Liu, Hilary K Finucane, Alexander Gusev, Gaurav Bhatia, Steven Gazal, Luke O’Connor,
445 Brendan Bulik-Sullivan, Fred A Wright, Patrick F Sullivan, Benjamin M Neale, et al. Functional
446 architectures of local and distal regulation of gene expression in multiple human tissues. *The American
447 Journal of Human Genetics*, 100(4):605–616, 2017.
- 448 [23] Brian R Krause and Arthur D Hartman. Adipose tissue and cholesterol metabolism. *Journal of lipid
449 research*, 25(2):97–110, 1984.
- 450 [24] Soazig Le Lay, Stéphane Krief, Céline Farnier, Isabelle Lefrère, Xavier Le Liepvre, Raymond Bazin, Pas-
451 cal Ferré, and Isabelle Dugail. Cholesterol, a cell size-dependent signal that regulates glucose metabolism
452 and gene expression in adipocytes. *Journal of Biological Chemistry*, 276(20):16904–16910, 2001.
- 453 [25] Anders H Berg, Terry P Combs, and Philipp E Scherer. Acrp30/adiponectin: an adipokine regulating
454 glucose and lipid metabolism. *Trends in Endocrinology & Metabolism*, 13(2):84–89, 2002.
- 455 [26] Willeke de Haan, Alpana Bhattacharjee, Piers Ruddle, Martin H Kang, and Michael R Hayden. Abca1
456 in adipocytes regulates adipose tissue lipid content, glucose tolerance, and insulin sensitivity. *Journal
457 of lipid research*, 55(3):516–523, 2014.
- 458 [27] Nuala A O’Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufu, Diana Haddad, Rich McVeigh,
459 Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence
460 (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic
461 acids research*, 44(D1):D733–D745, 2015.
- 462 [28] Tomaz Berisa and Joseph K Pickrell. Approximately independent linkage disequilibrium blocks in
463 human populations. *Bioinformatics*, 32(2):283, 2016.
- 464 [29] Sung Chun, Alexandra Casparino, Nikolaos A Patsopoulos, Damien C Croteau-Chonka, Benjamin A
465 Raby, Philip L De Jager, Shamil R Sunyaev, and Chris Cotsapas. Limited statistical evidence for shared
466 genetic effects of eqtls and autoimmune-disease-associated loci in three major immune-cell types. *Nature
467 Genetics*, 49(4):600–605, 2017.
- 468 [30] Farhad Hormozdiari, Anthony Zhu, Gleb Kichaev, Chelsea J-T Ju, Ayellet V Segrè, Jong Wha J
469 Joo, Hyejung Won, Sriram Sankararaman, Bogdan Pasaniuc, Sagiv Shifman, et al. Widespread allelic
470 heterogeneity in complex traits. *The American Journal of Human Genetics*, 100(5):789–802, 2017.
- 471 [31] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney Mc-
472 Cormick, Christian D Haudenschild, Kenneth B Beckman, Jianxin Shi, Rui Mei, et al. Characterizing
473 the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. *Genome research*,
474 24(1):14–24, 2014.

- 475 [32] Yang I Li, Bryce van de Geijn, Anil Raj, David A Knowles, Allegra A Petti, David Golan, Yoav Gilad,
476 and Jonathan K Pritchard. Rna splicing is a primary link between genetic variation and disease. *Science*,
477 352(6285):600–604, 2016.
- 478 [33] Alexander Gusev, Nicholas Mancuso, Hyejung Won, Maria Kousi, Hilary K Finucane, Yakir Reshef,
479 Lingyun Song, Alexias Safi, Steven McCarroll, Benjamin M Neale, et al. Transcriptome-wide association
480 study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature genetics*,
481 50(4):538, 2018.
- 482 [34] SS Kaalund, EN Newburn, Tuo Ye, R Tao, C Li, A Deep-Soboslay, MM Herman, TM Hyde, DR Wein-
483 berger, BK Lipska, et al. Contrasting changes in drd1 and drd2 splice variant expression in schizophre-
484 nia and affective disorders, and associations with snps in postmortem brain. *Molecular psychiatry*,
485 19(12):1258–1266, 2014.
- 486 [35] Urko M Marigorta, Lee A Denson, Jeffrey S Hyams, Kajari Mondal, Jarod Prince, Thomas D Walters,
487 Anne Griffiths, Joshua D Noe, Wallace V Crandall, Joel R Rosh, et al. Transcriptional risk scores link
488 gwas to eqtls and predict complications in crohn’s disease. *Nature Genetics*, 49(10):1517–1521, 2017.
- 489 [36] D Habier, RL Fernando, and JCM Dekkers. The impact of genetic relationship information on genome-
490 assisted breeding values. *Genetics*, 177(4):2389–2397, 2007.
- 491 [37] Paul M VanRaden. Efficient methods to compute genomic predictions. *Journal of dairy science*,
492 91(11):4414–4423, 2008.
- 493 [38] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear
494 mixed models. *PLoS Genet*, 9(2):e1003264, 2013.
- 495 [39] Halit Ongen, Andrew A Brown, Olivier Delaneau, Nikolaos Panousis, Alexandra C Nica, Emmanouil T
496 Dermizakis, GTEx Consortium, et al. Estimating the causal tissues for complex traits and diseases.
497 *bioRxiv*, page 074682, 2016.
- 498 [40] Consortium The Genomes Project. A global reference for human genetic variation. *Nature*,
499 526(7571):68–74, 2015.