nature genetics

SUPPLEMENTARY INFORMATION
https://doi.org/10.1038/s41588-019-0367-1

In the format provided by the authors and unedited.

Probabilistic fine-mapping of transcriptome-wide association studies

Nicholas Mancuso[1*], Malika K. Freund[2], Ruth Johnson[3], Huwenbo Shi[4], Gleb Kichaev[4], Alexander Gusev[5] and Bogdan Pasaniuc[1,2,4*]

[1]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [2]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [3]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA. [4]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA, USA. [5]Dana-Farber Cancer Institute, Boston, MA, USA. *e-mail: nmancuso@mednet.ucla.edu; pasaniuc@ucla.edu

Supplementary Note: Probabilistic fine-mapping of transcriptome-wide association studies

Nicholas Mancuso[1], Malika K. Freund[2], Ruth Johnson[3], Huwenbo Shi[4], Gleb Kichaev[4], Alexander Gusev[5], Bogdan Pasaniuc[1,2,4]

1. Dept of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095
2. Dept of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095
3. Dept of Computer Science, University of California, Los Angeles, Los Angeles, CA 90095
4. Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, CA 90095
5. Dana-Farber Cancer Institute, Boston, MA 02215

Supplementary Text

Notations

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 $\mathbf{X}^{\mathrm{T}}$) and index rows with a subscript (e.g., $\mathbf{X}_j$). We indicate L block-column partitions for matrix (vector) $\mathbf{X}$ as $\mathbf{X} = (\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(L)})$.

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 $\mathbf{G} \in \mathbb{R}^{\mathrm{n}\times\mathrm{m}}$ as

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

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 $\boldsymbol{\epsilon}$ is random environmental noise with $\mathbb{E}[\tilde{\boldsymbol{\epsilon}}] = \mathbf{0}$ and $\mathbb{V}[\tilde{\boldsymbol{\epsilon}}] = \mathbf{I}_n\tilde{\sigma}_e^2$. We extend the definition by also defining $\mathbf{G}$ as a linear function of underlying genotype and environment, which is governed by $\mathbf{G} = \mathbf{XW} + \mathbf{E}$, 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. The updated model is

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

where $\boldsymbol{\epsilon} = \mathbf{E}\boldsymbol{\alpha} + \tilde{\boldsymbol{\epsilon}}$ is the total contribution from environment, which we parameterize as $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ 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[1], GBLUP[2,3]). Because 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 $\widehat{\mathbf{G}} = \mathbf{X}\boldsymbol{\Omega}$ where $\boldsymbol{\Omega}$ are estimated eQTL effects. The 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,

$$z_j = \frac{1}{\sigma_e\sqrt{n}}\widehat{\mathbf{G}}_j{}^{\mathrm{T}}\mathbf{y} = \frac{1}{\sigma_e\sqrt{n}}(\mathbf{X}\boldsymbol{\Omega})_j^{\mathrm{T}}\mathbf{y} = \frac{1}{\sigma_e\sqrt{n}}\boldsymbol{\Omega}_j^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \frac{1}{\sigma_e\sqrt{n}}\boldsymbol{\Omega}_j^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + \mathbf{XW}\boldsymbol{\alpha} + \boldsymbol{\epsilon})$$

$$= \frac{1}{\sigma_e\sqrt{n}}\left(\boldsymbol{\Omega}_j^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Omega}_j^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{XW}\boldsymbol{\alpha} + \boldsymbol{\Omega}_j^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}\right)$$

$$= \frac{\sqrt{n}}{\sigma_e}\boldsymbol{\Omega}_j^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} + \frac{\sqrt{n}}{\sigma_e}\boldsymbol{\Omega}_j^{\mathrm{T}}\mathbf{V}\mathbf{W}\boldsymbol{\alpha} + \frac{1}{\sigma_e\sqrt{n}}\boldsymbol{\Omega}_j^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}$$

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

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

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

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

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

$$\mathbf{z}_{\mathrm{twas}} | \boldsymbol{\lambda}_{\mathrm{snp}}, \boldsymbol{\lambda}_{\mathrm{pe}}, \boldsymbol{\Omega}, \mathbf{V} \sim N\left(\boldsymbol{\Omega}^{\mathrm{T}}\mathbf{V}\boldsymbol{\lambda}_{\mathrm{snp}} + \boldsymbol{\mathcal{V}}\boldsymbol{\lambda}_{\mathrm{pe}}, \boldsymbol{\mathcal{V}}\right).$$

This formulation asserts that observed marginal TWAS Z-scores are the linear combination of NCPs at causal genes weighted by the covariance structure of predicted expression $\boldsymbol{\mathcal{V}}$ and tagged pleiotropic effects from SNPs $\boldsymbol{\Omega}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta}$. 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 $\boldsymbol{\Omega}$.

Computing the likelihood of $\mathbf{z}_{\mathrm{twas}}$ as described requires knowing $\boldsymbol{\mathcal{V}}$, $\boldsymbol{\lambda}_{\mathrm{snp}}$, and $\boldsymbol{\lambda}_{\mathrm{pe}}$, which are unknown a-priori. First, we can estimate $\boldsymbol{\mathcal{V}}$ using available reference LD panels (e.g., 1000 Genomes[4]) 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 $\boldsymbol{\lambda}_{\mathrm{snp}} = \mathbf{1}_p\lambda_{\mathrm{snp}}$ when conditioned on $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{\lambda}_{\mathrm{pe}}$, which is similar to methods in robust Mendelian Randomization[5-7]. Third, estimating $\boldsymbol{\lambda}_{\mathrm{pe}}$ directly from data is also likely to overfit. To bypass this issue, we treat $\boldsymbol{\lambda}_{\mathrm{pe}}$ as a nuisance parameter and assume that $\boldsymbol{\lambda}_{\mathrm{pe}} | \mathbf{c}, \sigma_c^2 \sim N(0, \mathbf{D}_c)$ where $\mathbf{D}_c = \mathrm{diag}(\frac{n\sigma_c^2}{|\mathbf{c}|} \cdot \mathbf{c})$ is the scaled prior causal effect variance matrix and $\boldsymbol{c}$ is an $m \times 1$ binary vector indicating if ith gene is causal. When there are no causal genes (i.e. $\mathbf{c} = \mathbf{0}$) then $\mathbf{D}_c = \text{diag}(\mathbf{0})$. Incorporating this prior for causal NCPs enables us to integrate out $\boldsymbol{\lambda}_{\text{pe}}$, which results in the variance component model,

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

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 $\boldsymbol{c}$ 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. As an approximate solution, we estimate λ_{snp} using least squares under the null of no causal gene effects on trait. Alternatively, $\boldsymbol{\lambda}_{\text{snp}}$ can be modeled as a random effect and treated as a nuisance parameter (see below).

We use a Bayesian approach similar to fine-mapping methods in GWAS to compute the posterior distribution of our causal genes $\mathbf{c}$:

$$\Pr(\mathbf{c} \mid \mathbf{z}_{\text{twas}}, \lambda_{\text{snp}}, \boldsymbol{\Omega}, \mathbf{V}, n\sigma_c^2) = \frac{\Pr(\mathbf{z}_{\text{twas}}, \mathbf{c} \mid \lambda_{\text{snp}}, \boldsymbol{\Omega}, \mathbf{V}, n\sigma_c^2)}{\Pr(\mathbf{z}_{\text{twas}} \mid \lambda_{\text{snp}}, \boldsymbol{\Omega}, \mathbf{V}, n\sigma_c^2)}$$

$$= \frac{N(\mathbf{z}_{\text{twas}} \mid \boldsymbol{\Omega}^{\text{T}}\mathbf{V}\mathbf{1}_p\lambda_{\text{snp}}, \boldsymbol{\mathcal{V}}\mathbf{D}_c\boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{V}})\Pr(\mathbf{c})}{\sum_{\mathbf{c}' \in \mathcal{C}} \text{N}(\mathbf{z}_{\text{twas}} \mid \boldsymbol{\Omega}^{\text{T}}\mathbf{V}\mathbf{1}_p\lambda_{\text{snp}}, \boldsymbol{\mathcal{V}}\mathbf{D}_{c'}\boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{V}})\Pr(\mathbf{c}')}$$

where $\mathcal{C}$ is the set of all binary strings of length m. We assume a Bernoulli prior for each causal indicator $c_i \sim \text{Bern}(p)$. 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

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

We compute this expression using straightforward enumeration, which is feasible for regions with fewer than 20 gene models. For larger regions, we limit enumeration to at most 5 causal genes. Alternatively, we can compute PIPs using Bayes factors for each model (see below). We model the null as the all-zero configuration (i.e. $\mathbf{c} = \mathbf{0}$). Here, $\Pr(\mathbf{c} = \mathbf{0} \mid \mathbf{z}_{\text{twas}}, \lambda_{\text{snp}}, \boldsymbol{\Omega}, \mathbf{V}, n\sigma_c^2\big)$ captures the probability that none of the predicted expression models included in our analysis explain the observed TWAS Z-scores. Lastly, PIPs enable us to estimate the expected number of causal genes at a risk region m^c as $\mathbb{E}[m^c] = \sum_{i=1}^{m} \text{PIP}\big(c_i = 1 \big| \mathbf{z}_{\text{twas}}, \lambda_{\text{snp}}, \boldsymbol{\Omega}, \mathbf{V}, n\sigma_c^2\big)$.

Efficient Bayes Factors computation with singular-value decomposition

Bayes factors provide an alternative approach to measure the evidence of an alternative model against the null model. We use a derivation similar to that of SNP fine-mapping work[8]. Namely,

$$\text{BF}_{\text{c}} = \frac{N(\mathbf{z} \mid \mathbf{0}, \boldsymbol{\mathcal{V}}\mathbf{D}_c\boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{V}})}{N(\mathbf{z}_{\text{twas}} \mid \mathbf{0}, \boldsymbol{\mathcal{V}})}$$

$$= \frac{N(\mathbf{z} \mid \mathbf{0}, n\sigma_c^2 \boldsymbol{\mathcal{V}}_{cc}\boldsymbol{\mathcal{V}}_{cc} + \boldsymbol{\mathcal{V}}_{cc})}{N(\mathbf{z}_{\text{twas}} \mid \mathbf{0}, \boldsymbol{\mathcal{V}}_{cc})}$$

$$= \frac{|n\sigma_c^2 \boldsymbol{\mathcal{V}}_{cc}\boldsymbol{\mathcal{V}}_{cc} + \boldsymbol{\mathcal{V}}_{cc}|^{-1/2}}{|\boldsymbol{\mathcal{V}}_{cc}|^{-1/2}} \cdot \frac{\exp\left[-\frac{1}{2}\mathbf{z}_c^{\mathbf{T}}(n\sigma_c^2 \boldsymbol{\mathcal{V}}_{cc}\boldsymbol{\mathcal{V}}_{cc} + \boldsymbol{\mathcal{V}}_{cc})^{-\mathbf{1}}\mathbf{z}_c\right]}{\exp\left[-\frac{1}{2}\mathbf{z}_c^{\mathbf{T}}\boldsymbol{\mathcal{V}}_{cc}{}^{-\mathbf{1}}\mathbf{z}_c\right]},$$

where $\boldsymbol{\mathcal{V}}_{cc}$ is the $k \times k$ sub-matrix of $\boldsymbol{\mathcal{V}} = \boldsymbol{\Omega}^{\mathbf{T}}\mathbf{V}\boldsymbol{\Omega}$ restricted to k causal genes indicated by $\mathbf{c}$ (similarly for $\mathbf{z}_c$). Evaluating BF_c requires time $O(k^3)$ due to matrix multiplication and inversion. Computing the inverse of $\boldsymbol{\mathcal{V}}_{cc}$ requires full-rank, which may not be the case in practice when predicted expression is highly correlated. We account for rank-deficiency in $\boldsymbol{\mathcal{V}}_{cc}$ by performing a singular-value decomposition of $\boldsymbol{\mathcal{V}}_{cc}$ and rotating $\mathbf{z}_c$ to an independent basis and keeping positive singular values similar to that noted in ref[9]. Posterior probabilities for a causal configuration $\mathbf{c}$ can be computed by Bayes rule (as noted in the main text) or equivalently with normalized Bayes Factors[8] as,

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

In practice, we found that computing PIPs using the Bayes Factor with SVD decomposition was much more stable in ill-conditioned settings compared with the naive calculation using Bayes rule. If FOCUS is run with the mean term included λ_{snp}, then we replace $\mathbf{z}$ above with the residual after regressing out the effect of $\hat{\lambda}_{\text{snp}}$ estimated under the null $\mathbf{c} = \mathbf{0}$. This approach is conservative but avoids estimating λ_{snp} for each causal configuration $\mathbf{c}$.

Variance component model to control for pleiotropic SNP effects

The default model for FOCUS controls for pleiotropic effects by including a mean term ($\boldsymbol{\lambda}_{\text{snp}} = \mathbf{1}_p\lambda_{\text{snp}}$). We infer λ_{snp} under the null configuration ($\mathbf{c} = \mathbf{0}$), which leads to more conservative PIPs for individual genes. Here, we describe a separate model that accounts for general random pleiotropic effects at a region. If we model $\boldsymbol{\lambda}_{\text{snp}} \sim N(\mathbf{0}, \mathbf{I}_p\sigma_p^2)$ and marginalize out $\boldsymbol{\lambda}_{\text{snp}}$ our new marginal likelihood becomes,

$$\mathbf{z}_{\text{twas}} \mid \lambda_{\text{snp}}, \boldsymbol{\Omega}, \mathbf{V}, \mathbf{c}, n\sigma_c^2, \sigma_p^2 \sim N\left(\mathbf{0}, n\sigma_p^2\boldsymbol{\Omega}^{\text{T}}\mathbf{V}\mathbf{V}\boldsymbol{\Omega} + \boldsymbol{\mathcal{V}}\mathbf{D}_c\boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{V}}\right).$$

We investigated performance of this variance components model when σ_p^2 is known. We found the performance of the variance components model to produce unbiased credible sets but included many more genes on average compared with the mean term model (see **Supplementary Figure 7**). In the general setting σ_p^2 would need to be inferred before computing PIPs (either once conservatively under the null of $\mathbf{c} = \mathbf{0}$ or per individual configuration).

Explicit vs implicit simulations using total gene expression for trait

Here we describe two strategies for simulating a continuous trait as a function of total gene expression. Our model for complex trait $\mathbf{y}$ (ignoring pleiotropic effects) is given by,

$$\mathbf{y} = (\mathbf{XW} + \mathbf{E})\boldsymbol{\alpha} + \tilde{\boldsymbol{\epsilon}} = \mathbf{XW}\boldsymbol{\alpha} + \mathbf{E}\boldsymbol{\alpha} + \tilde{\boldsymbol{\epsilon}} = \mathbf{XW}\boldsymbol{\alpha} + \boldsymbol{\epsilon}.$$

The first strategy is to simulate $\mathbf{y}$ using a weighted combination of total expression $(\mathbf{G} = \mathbf{XW} + \mathbf{E})\boldsymbol{\alpha}$ and sampling residual environmental noise $\tilde{\boldsymbol{\epsilon}} \sim N(\mathbf{0}, \mathbf{I}_n \tilde{\sigma}_e^2)$. The second strategy is to simulate $\mathbf{y}$ using a weighted combination of the genetic component of expression $\mathbf{XW}\boldsymbol{\alpha}$ and sample total environmental noise $\mathbf{E}\boldsymbol{\alpha} + \tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_n \sigma_e^2)$. Simulations are parameterized by the total variance in trait explained by predicted (i.e. cis-genetically regulated) gene expression h_{GE}^2. When h_{GE}^2 is fixed, either simulation can sample enough residual noise $\tilde{\boldsymbol{\epsilon}}$ or total noise $\boldsymbol{\epsilon}$ such that h_{GE}^2 is met. We empirically validated this in simulations and found both approaches resulted in similar results (**Supplementary Figure 24**).

Supplementary Tables

Supplementary Table 1. Resolution in identifying causal genes across methods. See attached excel doc.

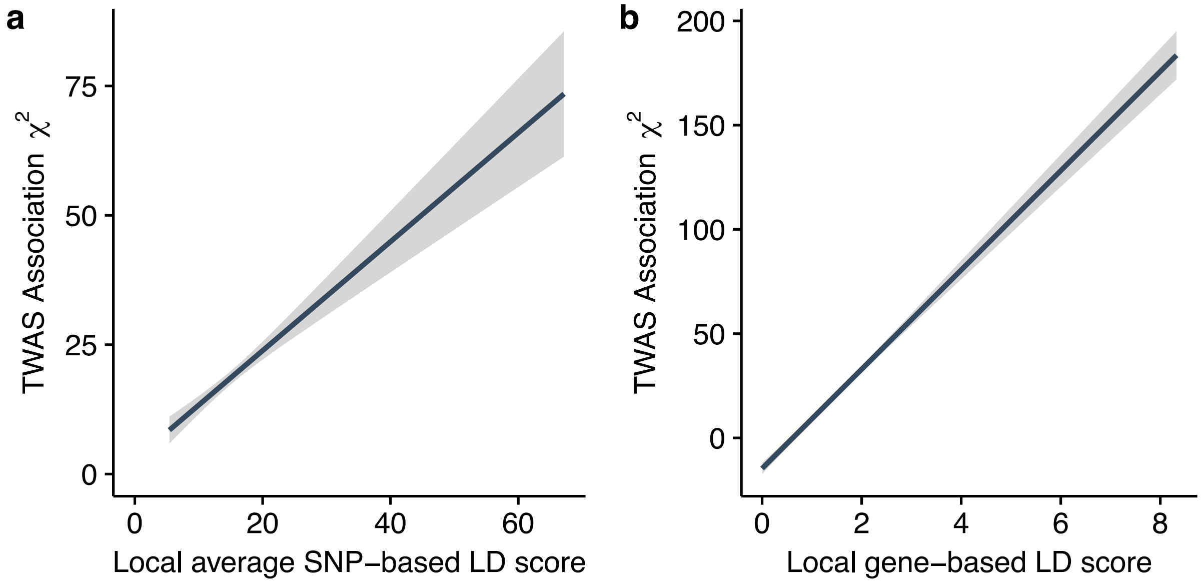
Supplementary Table 2. Summary of expression reference panels. See attached excel doc

Supplementary Table 3. TWAS results for lipids traits. See attached excel doc.

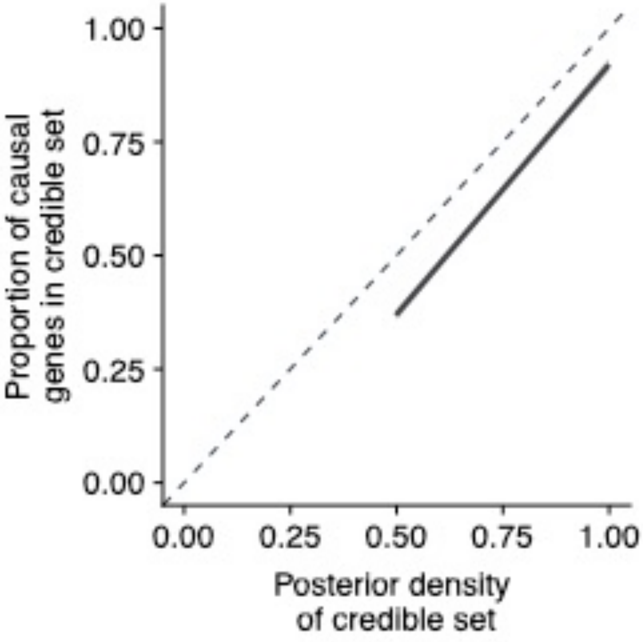
Supplementary Table 4. Fine mapping for lipids traits. See attached excel doc.

Supplementary Table 5. Posterior probabilities for gene models and 90%-credible gene sets at 1p13 locus for LDL. See attached excel doc.

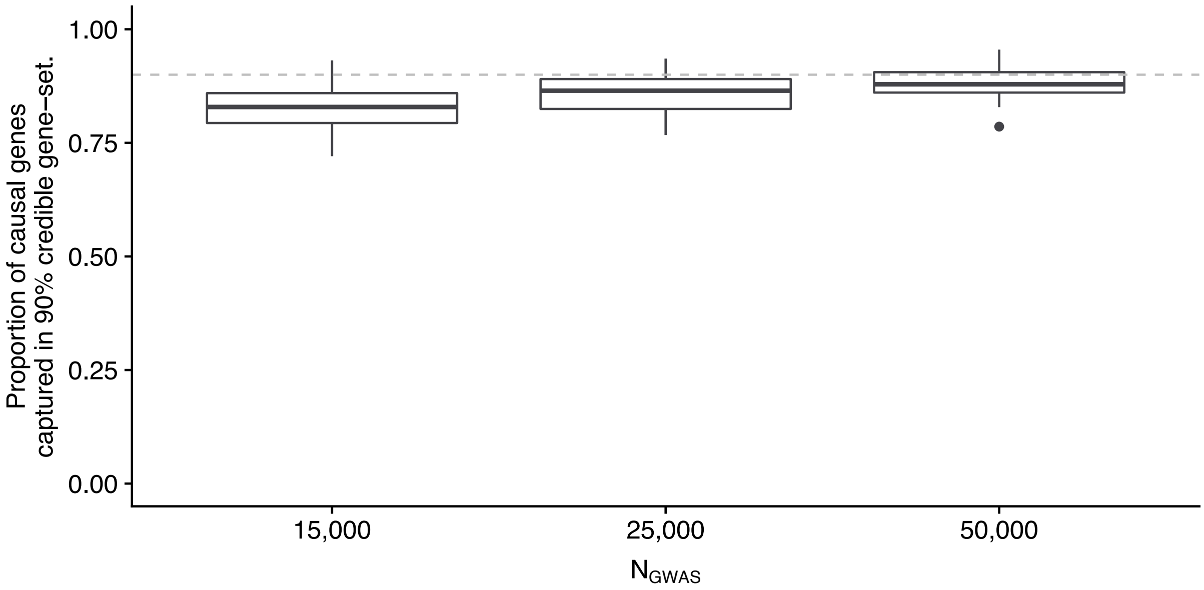
Supplementary Figures



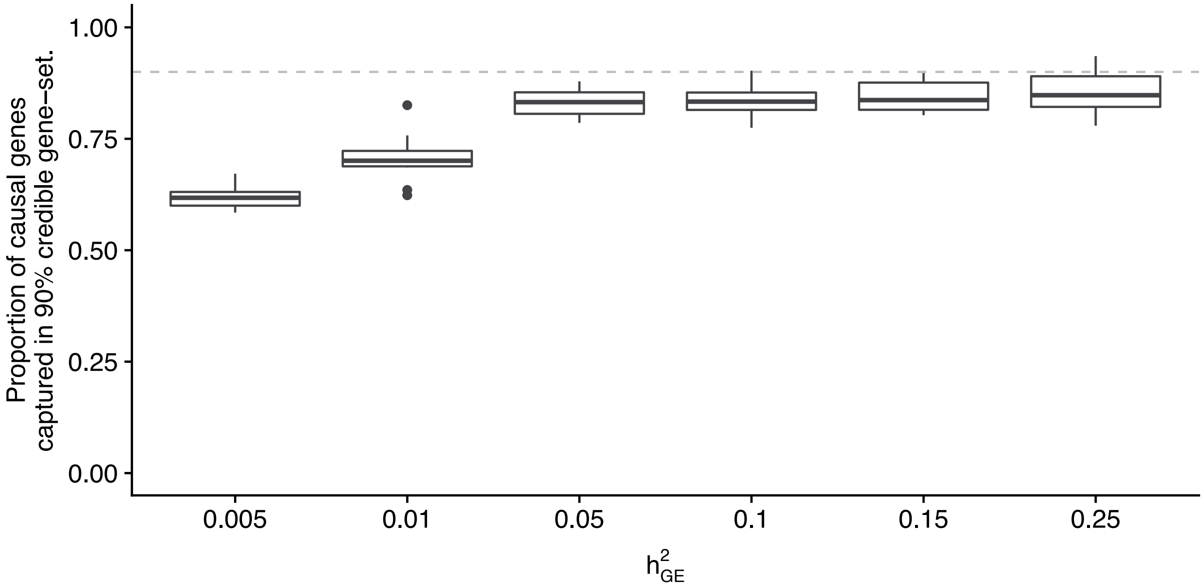
Supplementary Figure 1. Genes with large average LD tag causal gene TWAS associations. Using TWAS results from our simulation pipeline (i.e. $N_{GWAS} = 50{,}000, N_{eQTL} = 500, h^2_{GE} = 0.1$, gene expression $h^2_g = 0.2$), we computed the correlation between TWAS association strength (χ^2) at non-causal genes with two measures of regional correlation. a) The average LD-score for SNPs defining a gene model (i.e. p^{-1} trace($\mathbf{VV}$)). b) The "gene-based" correlation score considering the LD at a region along with eQTL weights (i.e. $(\boldsymbol{\mathcal{V}}\boldsymbol{\mathcal{V}})_{i,i}$). The dark line represents the best-fit regression line and the gray area represents the 95% predictive interval.



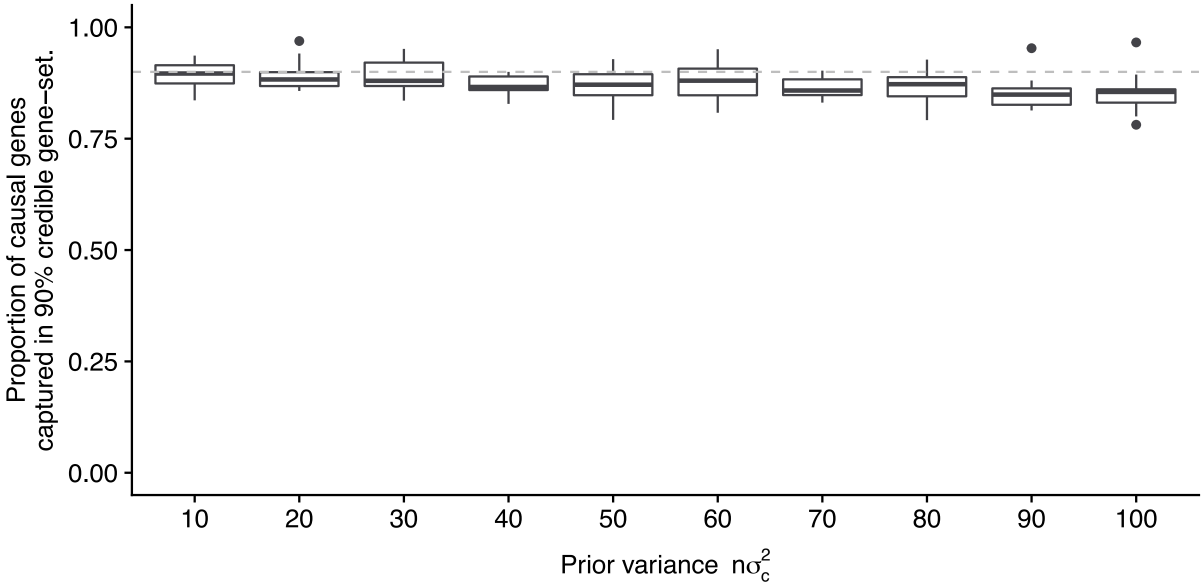
Supplementary Figure 2. Credible gene-sets as a function of $\boldsymbol{\rho}$. We simulated 100 complex traits as a function of underlying gene expression (i.e. $N_{GWAS} = 50{,}000$, $N_{eQTL} = 500$, $h_{GE}^2 = 0.1$, gene expression $h_g^2 = 0.2$). We ran FOCUS on the TWAS summary data using $n\sigma_c^2 = 80$ and computed ρ-credible sets. The posterior density of each credible set was computed as the normalized sum of posterior probability for all genes in the credible set. The solid line indicates the regression best-fit line and the shaded area is the 95% prediction interval. The dashed line indicates the identity.



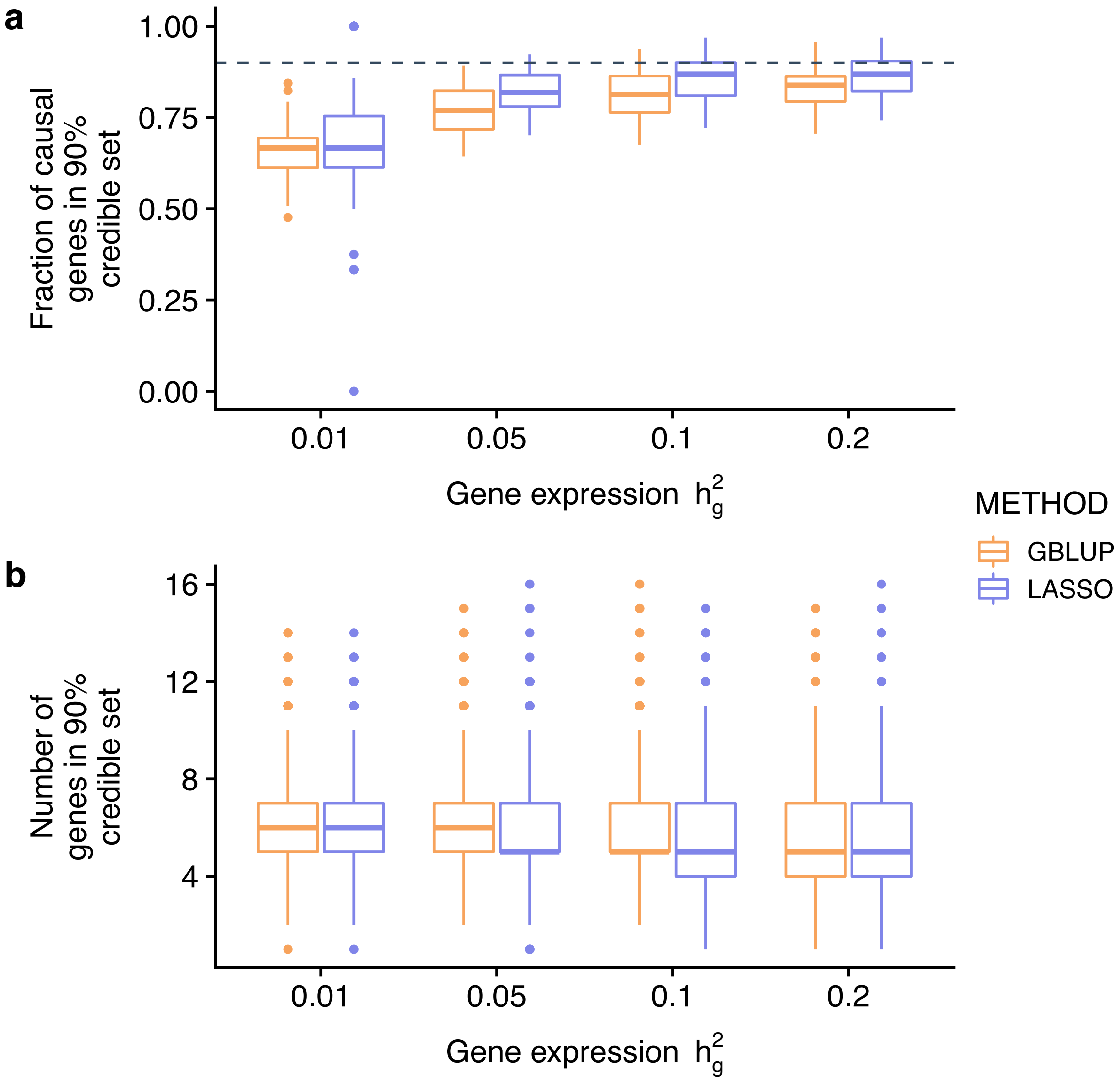
Supplementary Figure 3. Influence of GWAS sample size in prioritizing causal genes. We varied N_{GWAS} in our simulation pipeline keeping other parameters fixed (i.e. $N_{eQTL} = 500$, $h^2_{GE} = 0.1$, gene expression $h^2_g = 0.2$). We ran FOCUS on the TWAS summary data using $n\sigma^2_c = 80$ and computed 90%-credible gene sets. Box-plots indicate the median, upper and lower quartiles, and 1.5 inter-quartile range of the proportion of causal genes captured in 90% credible gene-sets. Outliers are shown as points. The dashed line indicates 90%.



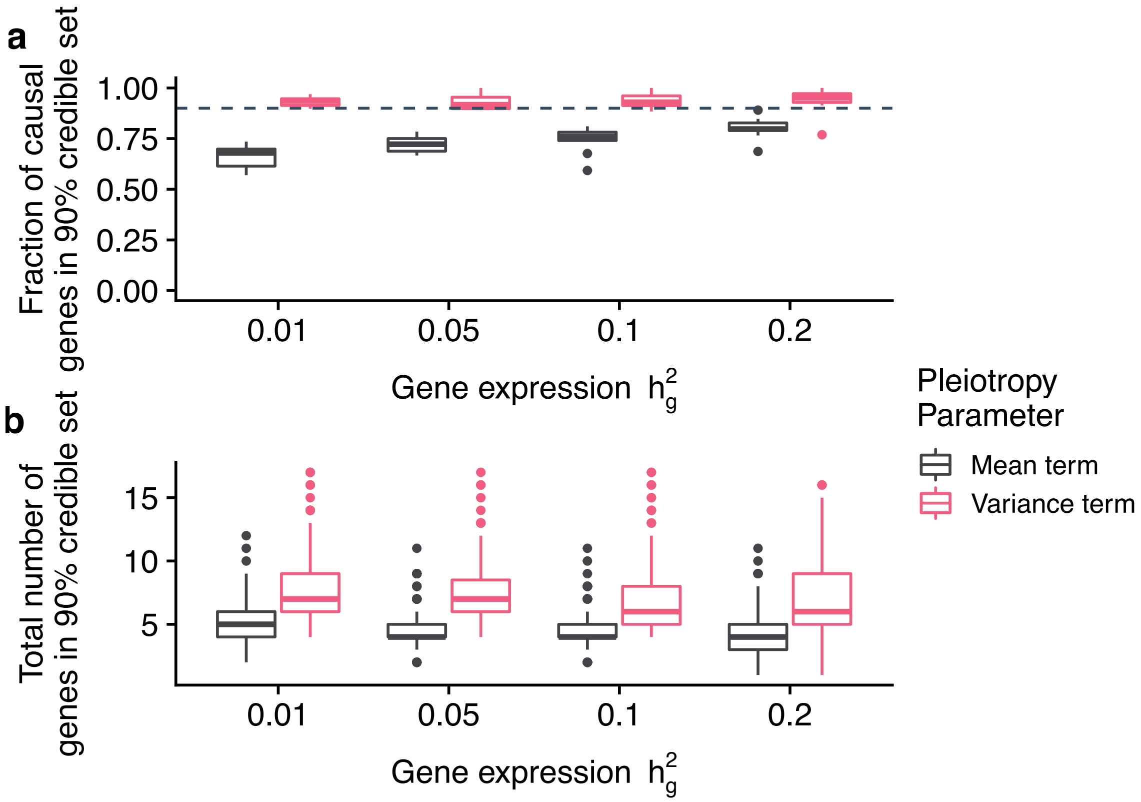
Supplementary Figure 4. Influence of trait-variance explained from predicted gene expression in prioritizing causal genes. We varied h^2_{GE}in our simulation pipeline keeping other parameters fixed (i.e. $N_{GWAS} = 50{,}000$, $N_{eQTL} = 500$, gene expression $h^2_g = 0.2$). We ran FOCUS on the TWAS summary data using $n\sigma^2_c = 80$ and computed ρ-credible sets. Box-plots indicate the median, upper and lower quartiles, and 1.5 inter-quartile range of the proportion of causal genes captured in 90% credible gene-sets. Outliers are shown as points. The dashed line indicates 90%.



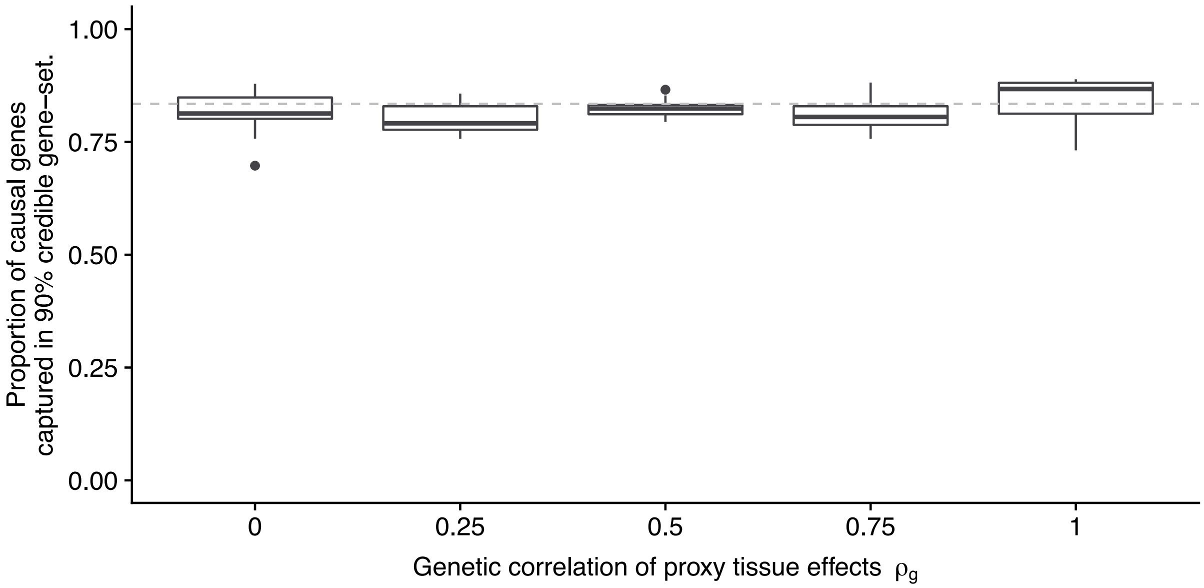
Supplementary Figure 5. FOCUS is stable to various settings of prior variance. We used our simulation pipeline keeping all parameters fixed (i.e. $N_{GWAS} = 50{,}000$, $N_{eQTL} = 500$, $h_{GE}^2 = 0.1$, gene expression $h_g^2 = 0.2$). We ran FOCUS on the TWAS summary data varying $n\sigma_c^2$ and computed 90%-credible gene sets. Box-plots indicate the median, upper and lower quartiles, and 1.5 inter-quartile range of the proportion of causal genes captured in 90% credible gene-sets. Outliers are shown as points. The dashed line indicates 90%.



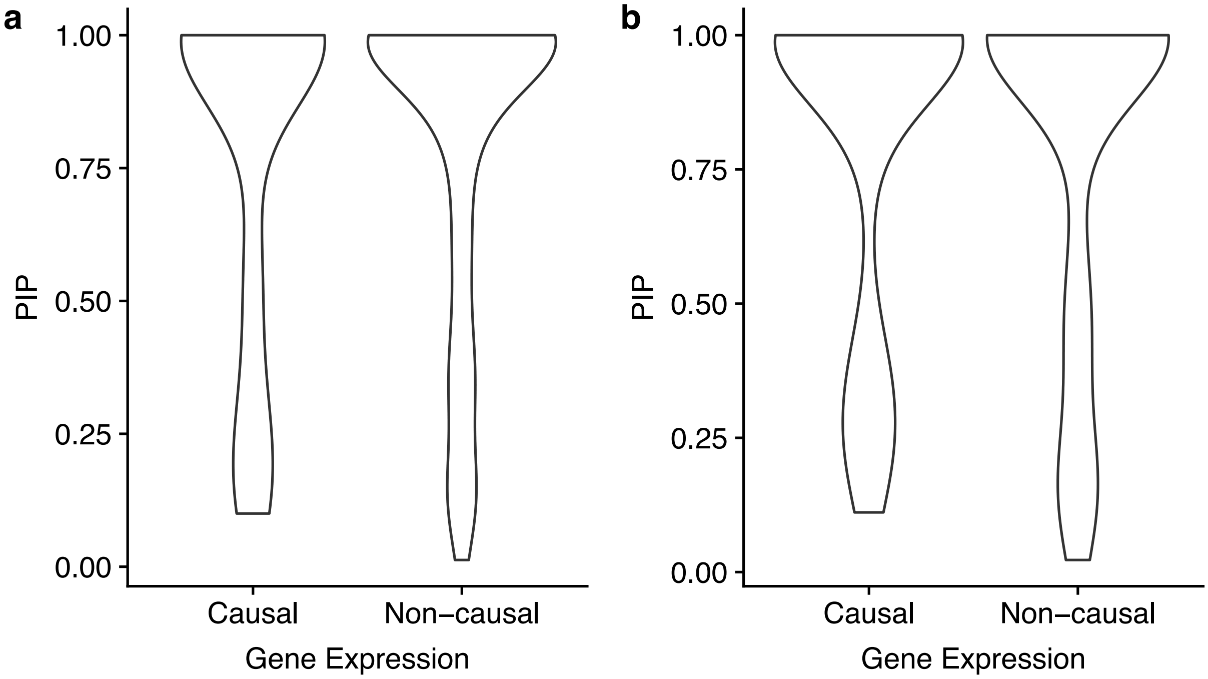
Supplementary Figure 6. Performance of FOCUS using GBLUP versus LASSO for gene expression prediction. We modified our simulation pipeline to train predictive models for gene expression using LASSO in addition to GBLUP. **a)** Distribution of the fraction of genes that are causal for downstream trait captured in the 90%-credible gene set. **b)** Distribution of the size of the 90%-credible gene sets across all simulations. Box-plots indicate the median, upper and lower quartiles, and 1.5 inter-quartile range. Outliers are shown as points. The dashed line indicates 90%.



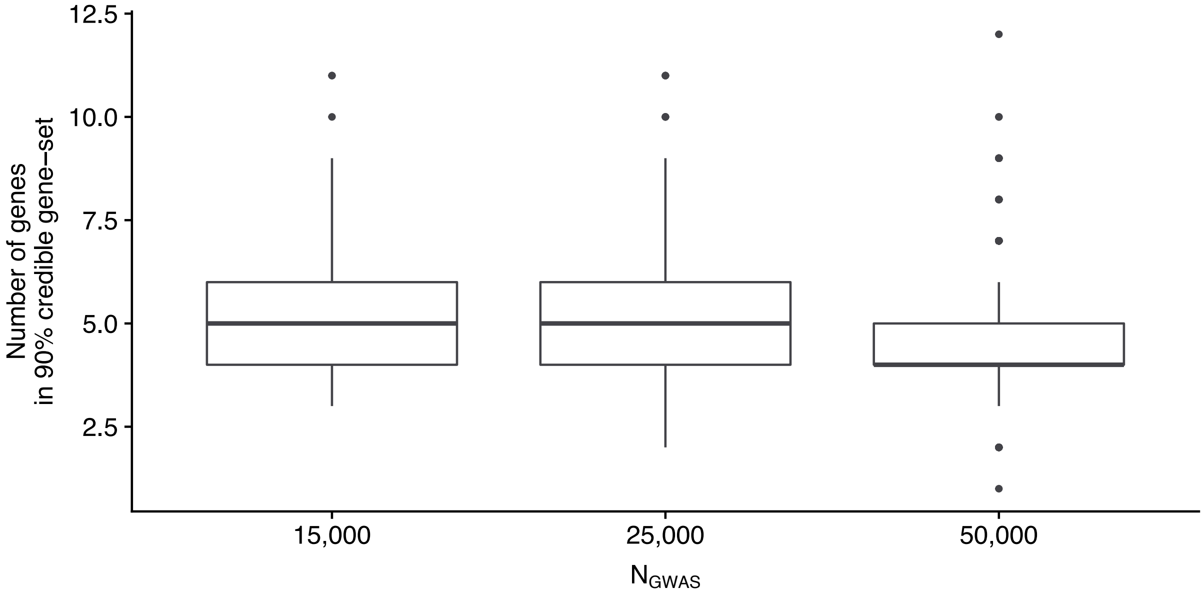
Supplementary Figure 7. Performance of FOCUS using fixed-effects versus random-effects to model pleiotropic effects of SNPs on downstream trait. We compared the FOCUS model that accounts for pleiotropic effects with a fixed mean term versus a model that controls for pleiotropy using a variance component (i.e. variance term). This latter model is equivalent to modeling pleiotropic effects as random. Results shown are for simulations where the prior variance term for pleiotropic effects are known. **a)** Distribution for the fraction of causal genes captured in 90%-credible gene sets across simulations. **b)** Distribution of the number of genes in the 90%-credible gene sets. Box-plots indicate the median, upper and lower quartiles, and 1.5 inter-quartile range. Outliers are shown as points. The dashed line indicates 90%.



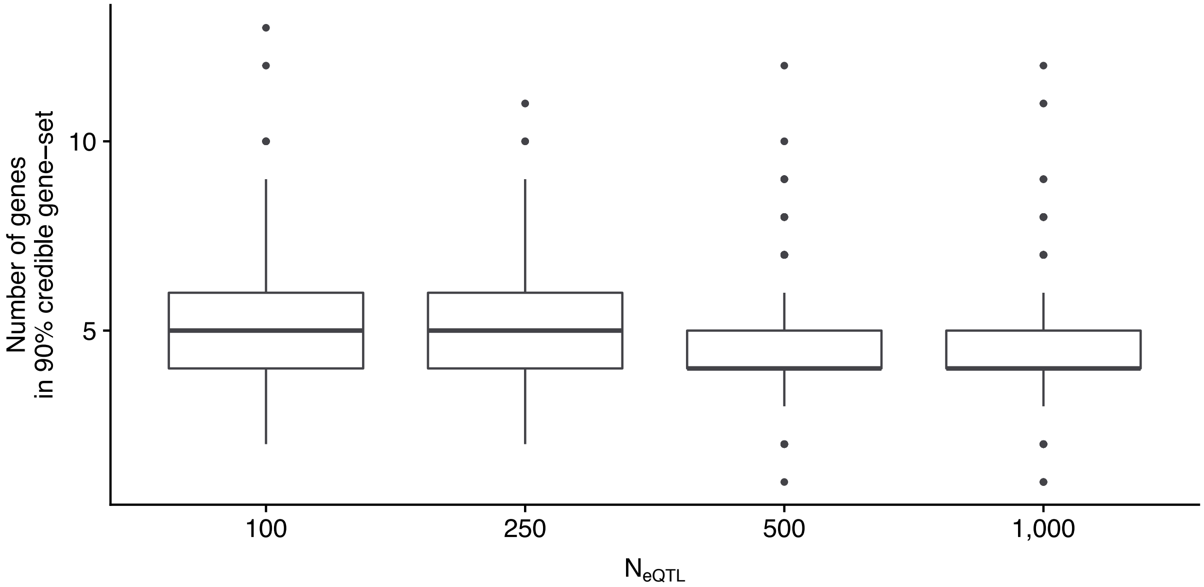
Supplementary Figure 8. FOCUS maintains performance when using expression in correlated proxy tissues. We used our simulation pipeline keeping all parameters fixed (i.e. $N_{GWAS} = 50{,}000$, $N_{eQTL} = 500$, $h_{GE}^2 = 0.1$, gene expression $h_g^2 = 0.2$), but here we used proxy-tissue gene expression data (also $h_g^2 = 0.2$) for the eQTL reference panel. We ran FOCUS on the TWAS summary data using $n\sigma_c^2 = 80$ and computed 90%-credible gene sets. Box-plots indicate the median, upper and lower quartiles, and 1.5 inter-quartile range of the proportion of causal genes captured in 90% credible gene-sets. Outliers are shown as points. The dashed line indicates the proportion captured using simulations in the relevant tissue.



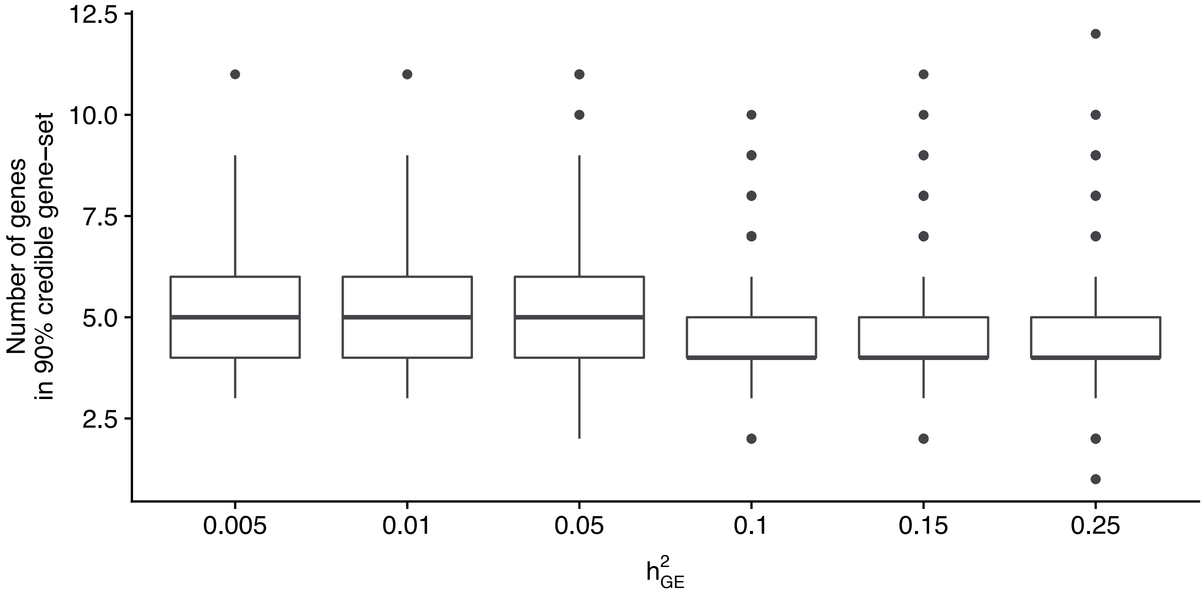
Supplementary Figure 9. Genes sharing a single regulatory variant exhibit similar posterior inclusion probability. We simulated expression for a pair of genes regulated by the same eQTL across 100 instances, where only one gene mediates risk for downstream trait. We find that the computed marginal PIPs are similar for the causal and non-causal gene. **a)** Results when the prior probability for a gene to be causal is set to $p = {}^{1}/_{2}$ (i.e. 1 gene per region in prior expectation). **b)** Results for simulations when the prior probability for a gene to be causal is set to $p = 1 \times 10^{-3}$.



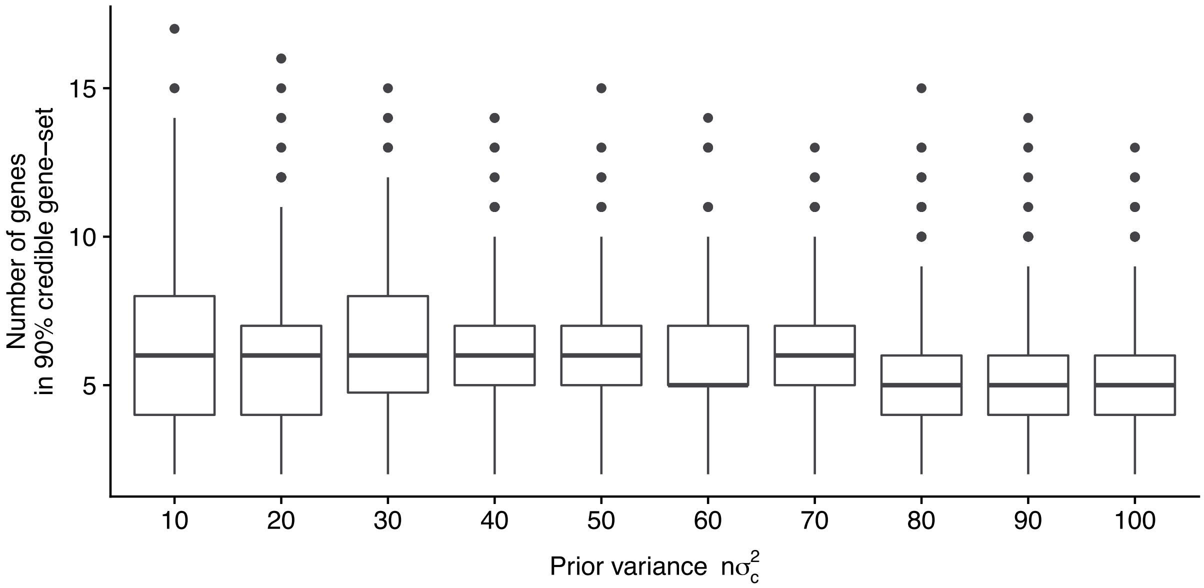
Supplementary Figure 10. Size of 90%-credible gene-sets as a function of GWAS size. We varied N_{GWAS} in our simulation pipeline keeping other parameters fixed (i.e. $N_{eQTL} = 500$, $h_{GE}^2 = 0.1$, gene expression $h_g^2 = 0.2$). We ran FOCUS on the TWAS summary data using $n\sigma_c^2 = 80$ and computed 90%-credible gene sets. Box-plots indicate the median, upper and lower quartiles, and 1.5 inter-quartile range of the number of causal genes captured in 90% credible gene-sets. Outliers are shown as points.



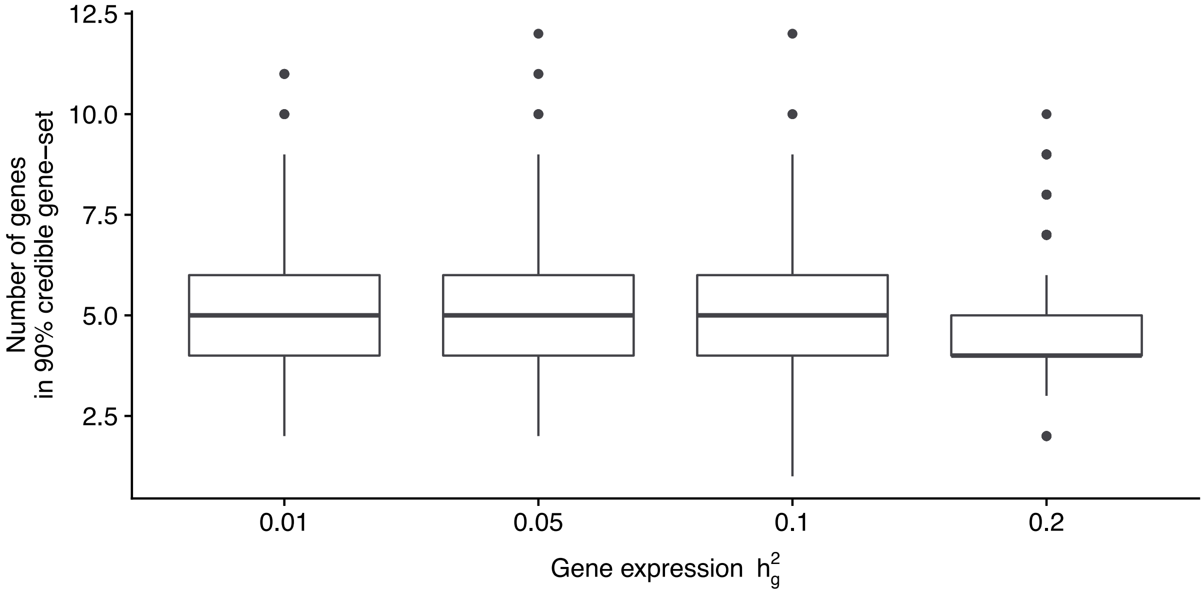
Supplementary Figure 11. Size of 90%-credible gene-sets as a function of eQTL reference panels size. We varied N_{eQTL} in our simulation pipeline keeping other parameters fixed (i.e. $N_{GWAS} = 50{,}000$, $h^2_{GE} = 0.1$, gene expression $h^2_g = 0.2$). We ran FOCUS on the TWAS summary data using $n\sigma^2_c = 80$ and computed 90%-credible gene sets. Box-plots indicate the median, upper and lower quartiles, and 1.5 inter-quartile range of the number of causal genes captured in 90% credible gene-sets. Outliers are shown as points.



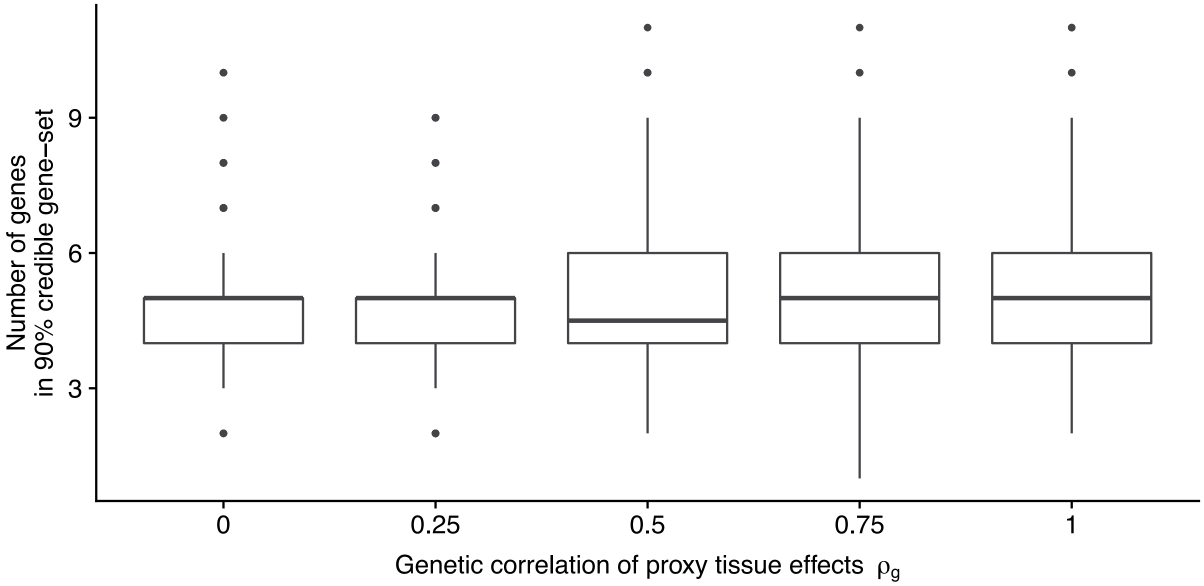
Supplementary Figure 12. Size of 90%-credible gene-sets as a function of heritability explained by causal gene expression. We varied h^2_{GE} in our simulation pipeline keeping other parameters fixed (i.e. $N_{GWAS} = 50{,}000$, $N_{eQTL} = 500$, gene expression $h^2_g = 0.2$). We ran FOCUS on the TWAS summary data using $n\sigma^2_c = 80$ and computed ρ-credible sets. Box-plots indicate the median, upper and lower quartiles, and 1.5 inter-quartile range of the number of causal genes captured in 90% credible gene-sets. Outliers are shown as points.



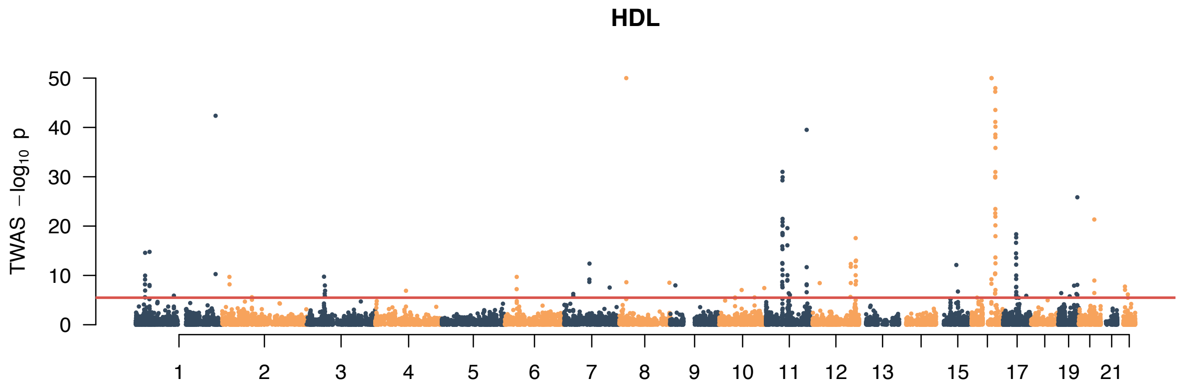
Supplementary Figure 13. Size of 90%-credible gene-sets as a function of causal gene effect-size variance parameter in posterior inference. We used our simulation pipeline keeping all parameters fixed (i.e. $N_{GWAS} = 50{,}000$, $N_{eQTL} = 500$, $h_{GE}^2 = 0.1$, gene expression $h_g^2 = 0.2$). We ran FOCUS on the TWAS summary data varying $n\sigma_c^2$ and computed 90%-credible gene sets. Box-plots indicate the median, upper and lower quartiles, and 1.5 inter-quartile range of the number of causal genes captured in 90% credible gene-sets. Outliers are shown as points.



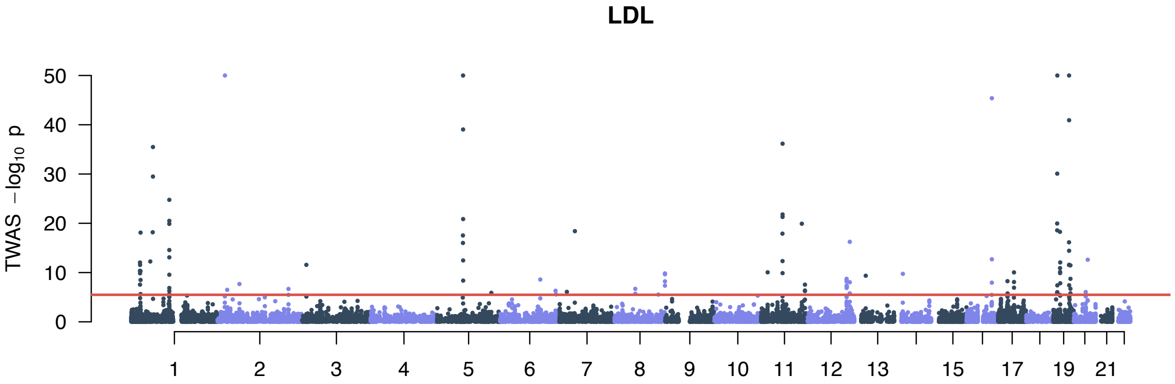
Supplementary Figure 14. Size of 90%-credible gene-sets as a function of heritability at proxy-tissue gene expression. We used our simulation pipeline keeping all parameters fixed (i.e. $N_{GWAS} = 50{,}000$, $N_{eQTL} = 500$, $h_{GE}^2 = 0.1$, gene expression $h_g^2 = 0.2$), but here we used proxy-tissue gene expression data with varying h_g^2 for the eQTL reference panel keeping $\rho_g = 0.9$. We ran FOCUS on the TWAS summary data using $n\sigma_c^2 = 80$ and computed 90%-credible gene sets. Box-plots indicate the median, upper and lower quartiles, and 1.5 inter-quartile range of the number of causal genes captured in 90% credible gene-sets. Outliers are shown as points.



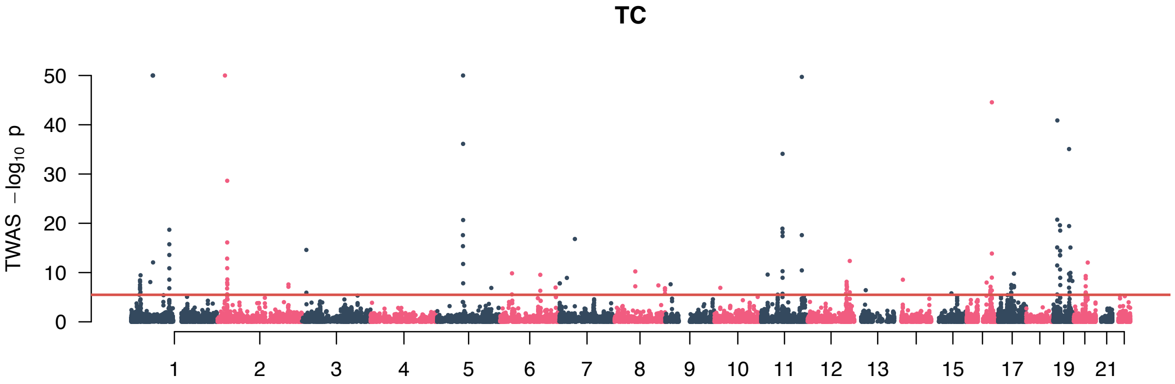
Supplementary Figure 15. Size of 90%-credible gene-sets as a function of genetic correlation for proxy-tissue expression. We used our simulation pipeline keeping all parameters fixed (i.e. $N_{GWAS} = 50{,}000$, $N_{eQTL} = 500$, $h^2_{GE} = 0.1$, gene expression $h^2_g = 0.2$), but here we used proxy-tissue gene expression data with varying ρ_g for the eQTL reference panel keeping $h^2_g = 0.2$. We ran FOCUS on the TWAS summary data using $n\sigma^2_c = 80$ and computed 90%-credible gene sets. Box-plots indicate the median, upper and lower quartiles, and 1.5 inter-quartile range of the number of causal genes captured in 90% credible gene-sets. Outliers are shown as points.



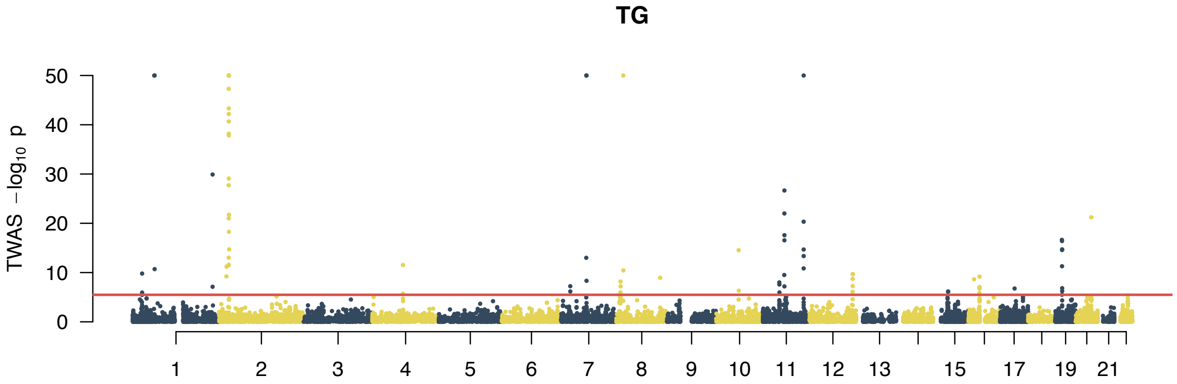
Supplementary Figure 16. Adipose-prioritized TWAS for high density lipoprotein (HDL) measurements. Manhattan plot of HDL TWAS results. Each point represents the association strength of each tested gene. We used a Bonferroni-adjusted transcriptome-wide significance of 0.05 / 15,277 (indicated by red line).



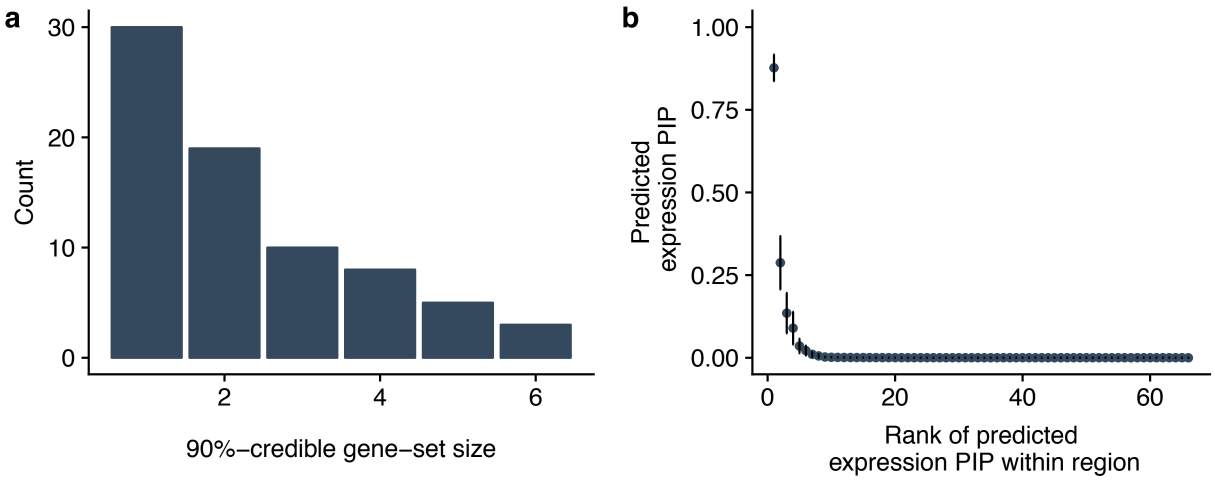
Supplementary Figure 17. Adipose-prioritized TWAS for low density lipoprotein (LDL) measurements. Manhattan plot of LDL TWAS results. Each point represents the association strength of each tested gene. results. We used a Bonferroni-adjusted transcriptome-wide significance of 0.05 / 15,277 (indicated by red line).



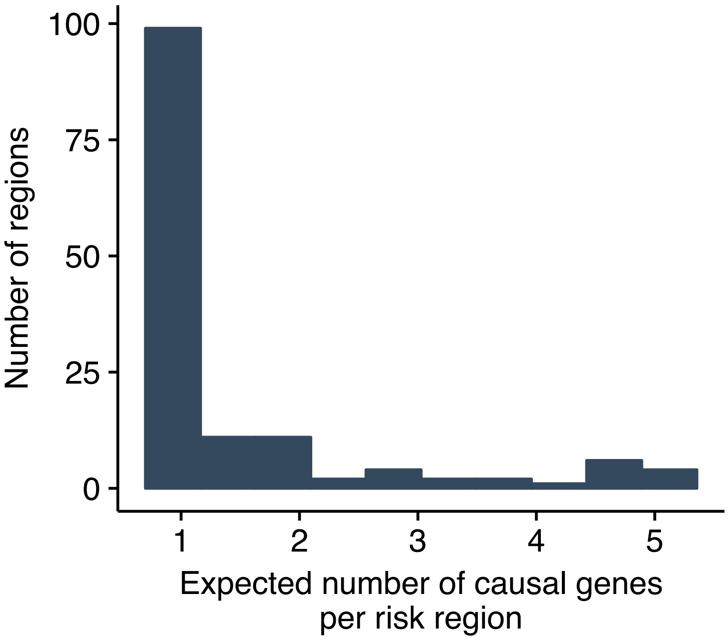
Supplementary Figure 18. Adipose-prioritized TWAS for total cholesterol (TC) measurements. Manhattan plot of TC TWAS results. Each point represents the association strength of each tested gene. results. We used a Bonferroni-adjusted transcriptome-wide significance of 0.05 / 15,277 (indicated by red line).



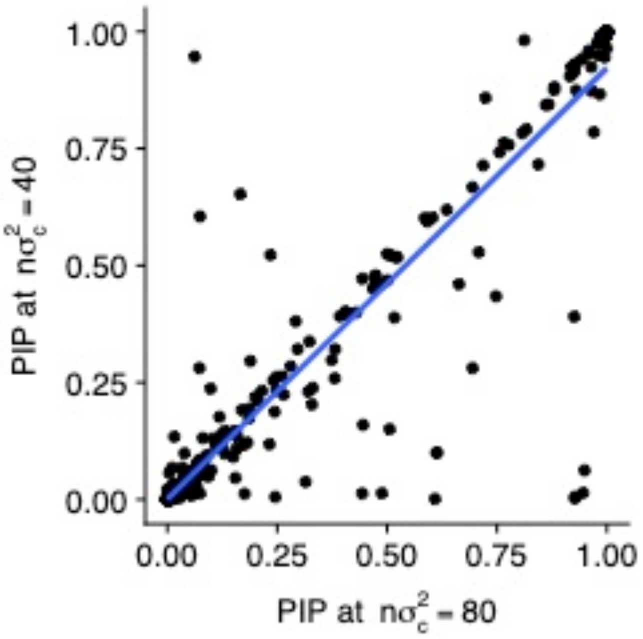
Supplementary Figure 19. Adipose-prioritized TWAS for triglyceride (TG) measurements. Manhattan plot of TG TWAS results. Each point represents the association strength of each tested gene. results. We used a Bonferroni-adjusted transcriptome-wide significance of 0.05 / 15,277 (indicated by red line).



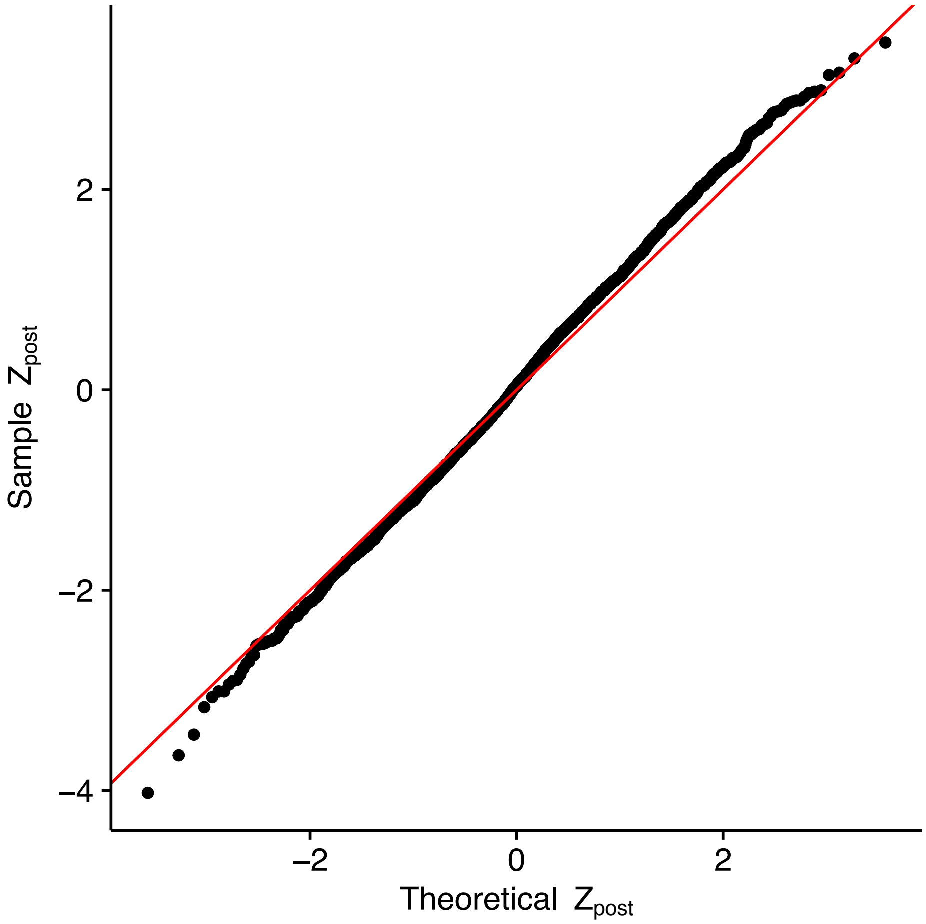
Supplementary Figure 20. Fine mapping of lipids TWAS risk regions. a) Histogram of 90%-credible gene-set sizes. **b)** Average PIP across risk regions according to ranking within each credible set. Points are the average PIP across a given rank and lines represent the 95% confident interval estimated across regions.



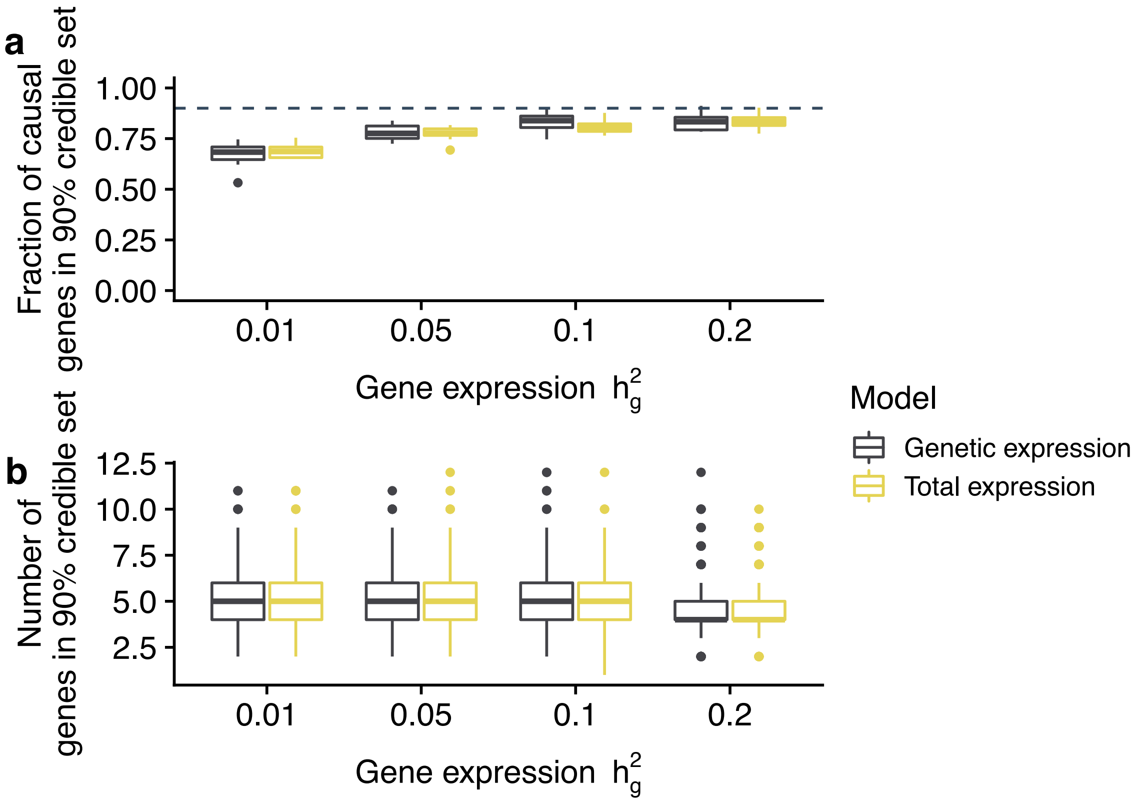
Supplementary Figure 21. Histogram of expected number of causal genes per risk region in lipids. Histogram of the expected number of causal genes in lipids TWAS under our model by summing across posterior inclusion probabilities per risk region. Specifically, we can compute this quantity at the jth risk region as $\sum_{i=1}^{m_j} \mathrm{PIP}_i$ where m_j is the number of genes at risk region j.



Supplementary Figure 22. Posterior inclusion probabilities for lipids fine-mapping results using different prior variance parameters. We ran FOCUS on the lipids TWAS data using $n\sigma_c^2 = 80$ and compared with our earlier results for $n\sigma_c^2 = 40$. Results were highly similar (linear regression $\beta = 0.92$; SE = 6.1×10^{-3}; two-sided t-test $P < 2 \times 10^{-16}$).



Supplementary Figure 23. FOCUS model is largely consistent with latent generative process. We performed a posterior predictive check for each gene (black point) using marginal PIPs to compute posterior p-value statistics. Here the null is that the average TWAS statistic for the ith gene is equal to its observed TWAS Z-score (i.e. $Z_{post,i} = \frac{\text{mean}(Z^*_{twas,i}) - Z_{twas,i}}{sd(Z^*_{twas,i})}$). The red line indicates the identity line.



Supplementary Figure 24. Comparison of simulations using total expression with genetically regulated expression. We compared performance of FOCUS under two simulation scenarios. A two-stage simulation where total expression has a direct effect on trait compared with a single-stage simulation where genetically regulated gene expression has a mediated effect such that h2GE is fixed to 0.1. We varied h_g^2 in our simulation pipeline keeping other parameters fixed (i.e. $N_{GWAS} = 50{,}000$, $N_{eQTL} = 500$, $h_{GE}^2 = 0.1$). We ran FOCUS on the TWAS summary data using $n\sigma_c^2 = 80$ and computed 90%-credible gene sets. **a)** Distribution for the fraction of causal genes captured in 90%-credible gene sets across simulations. **b)** Distribution of the number of genes in the 90%-credible gene sets. Box-plots indicate the median, upper and lower quartiles, and 1.5 inter-quartile range. Outliers are shown as points. The dashed line indicates 90%.

Supplementary References

1. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genet* **9**, e1003264 (2013).
2. Habier, D., Fernando, R. & Dekkers, J.C. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389-2397 (2007).
3. VanRaden, P.M. Efficient methods to compute genomic predictions. *Journal of dairy science* **91**, 4414-4423 (2008).
4. The Genomes Project, C. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
5. Bowden, J., Davey Smith, G., Haycock, P.C. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology* **40**, 304-314 (2016).
6. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44**, 512-525 (2015).
7. Barfield, R. *et al.* Transcriptome-wide association studies accounting for colocalization using Egger regression. *Genetic epidemiology* (2018).
8. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-1501 (2016).
9. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833 (2011).