Supplement: Probabilistic fine-mapping of transcriptome-wide association studies

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

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

Supplementary Note

Efficient Bayes Factors via Eigen-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.[1] Namely,

$$\begin{aligned}
\mathrm{BF}_{\mathbf{c}} &= \frac{\mathcal{N}(\mathbf{z} \mid \mathbf{0}, \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{V}}\mathbf{D}_{\mathbf{c}}\boldsymbol{\mathcal{V}})}{\mathcal{N}(\mathbf{z} \mid \mathbf{0}, \boldsymbol{\mathcal{V}})} \\
&= \frac{\mathcal{N}(\mathbf{z}_{\mathbf{c}} \mid \mathbf{0}, \boldsymbol{\mathcal{V}}_{\mathbf{cc}} + n\sigma_c^2 \boldsymbol{\mathcal{V}}_{\mathbf{cc}}\boldsymbol{\mathcal{V}}_{\mathbf{cc}})}{\mathcal{N}(\mathbf{z}_{\mathbf{c}} \mid \mathbf{0}, \boldsymbol{\mathcal{V}}_{\mathbf{cc}})} \\
&= \frac{|\boldsymbol{\mathcal{V}}_{\mathbf{cc}} + n\sigma_c^2 \boldsymbol{\mathcal{V}}_{\mathbf{cc}}\boldsymbol{\mathcal{V}}_{\mathbf{cc}}|^{-1/2}}{|\boldsymbol{\mathcal{V}}_{\mathbf{cc}}|^{-1/2}} \cdot \frac{\exp[-\frac{1}{2}\mathbf{z}_{\mathbf{c}}^{\intercal}(\boldsymbol{\mathcal{V}}_{\mathbf{cc}} + n\sigma_c^2 \boldsymbol{\mathcal{V}}_{\mathbf{cc}}\boldsymbol{\mathcal{V}}_{\mathbf{cc}})^{-1}\mathbf{z}_{\mathbf{c}}]}{\exp[-\frac{1}{2}\mathbf{z}_{\mathbf{c}}^{\intercal}\boldsymbol{\mathcal{V}}_{\mathbf{cc}}^{-1}\mathbf{z}_{\mathbf{c}}]},
\end{aligned} \tag{1}$$

where $\boldsymbol{\mathcal{V}}_{\mathbf{cc}}$ is the $k \times k$ sub-matrix of $\boldsymbol{\mathcal{V}}$ restricted to k causal genes indicated by $\mathbf{c}$ (similarly for $\mathbf{z}_{\mathbf{c}}$). Evaluating (1) requires time $O(k^3)$ due to matrix multiplication and inversion. Computing the inverse of $\boldsymbol{\mathcal{V}}_{\mathbf{cc}}$ requires full-rank, which may not be the case in practice when predicted expression is highly correlated. We account for rank-deficiency in $\boldsymbol{\mathcal{V}}_{\mathbf{cc}}$ by performing a singular-value decomposition of $\boldsymbol{\mathcal{V}}_{\mathbf{cc}}$ and rotating $\mathbf{z}_{\mathbf{c}}$ to an independent basis and keeping positive singular values similar to that noted in ref.[2] 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[1] as,

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

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.

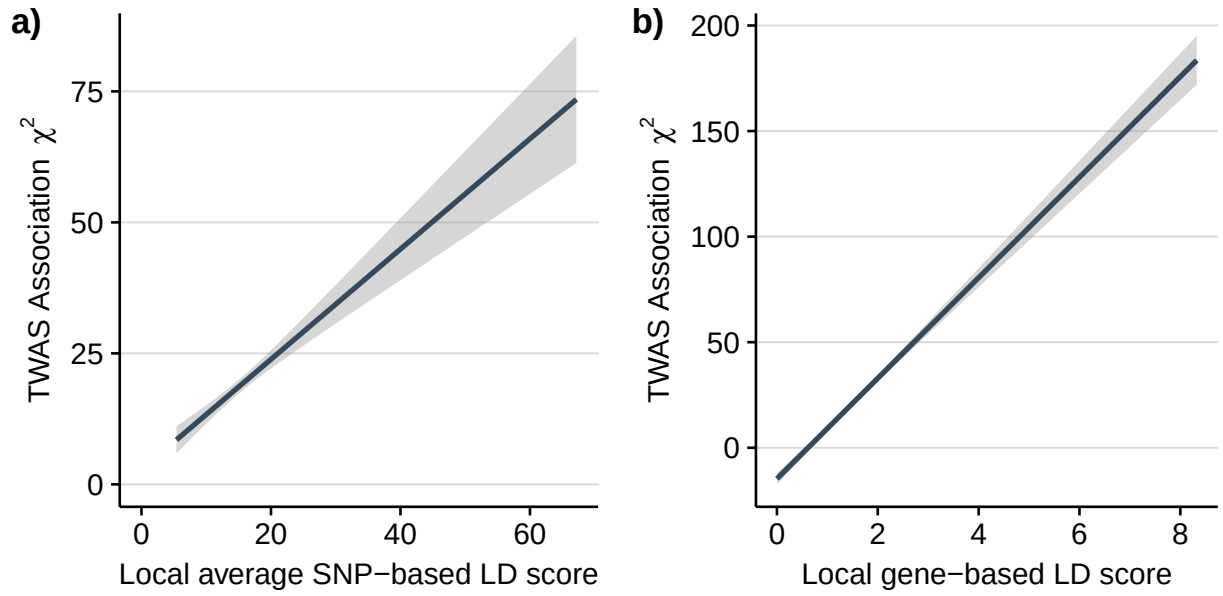
Supplementary Table 1: Summary of expression reference panels. See attached excel doc.

Method	No. Causals	Mean true-positive ratio	Standard deviation
FOCUS	Single	0.67	0.06
TWAS	Single	0.28	0.10
FOCUS	Multiple	0.57	0.06
TWAS	Multiple	0.18	0.11

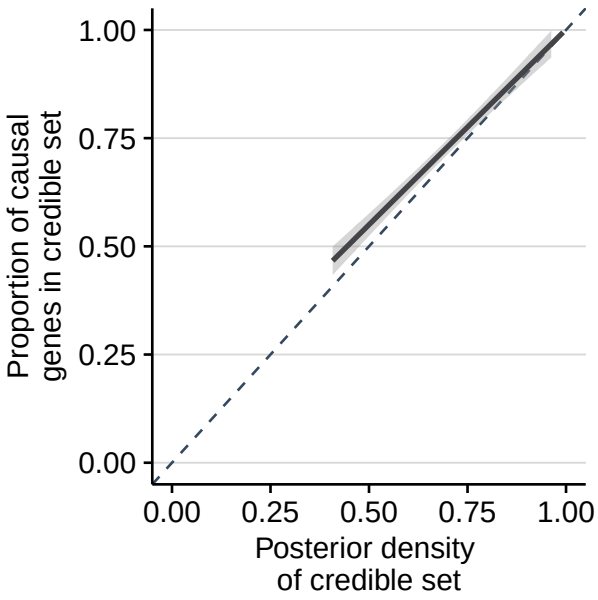
Supplementary Table 2: Simulation performance partitioned by number of causal genes in risk regions. We computed the average (standard deviation) true-positive ratio for results with at most a 5% false-positive ratio.

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

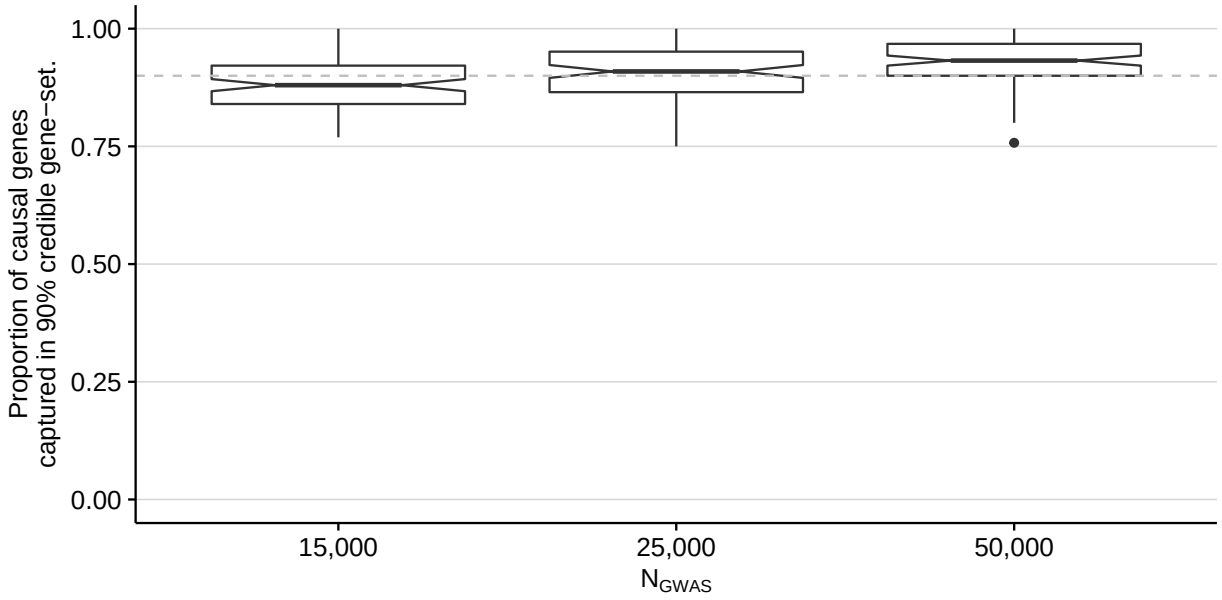
Supplementary Table 4: Fine mapping for lipids traits. See attached excel doc.



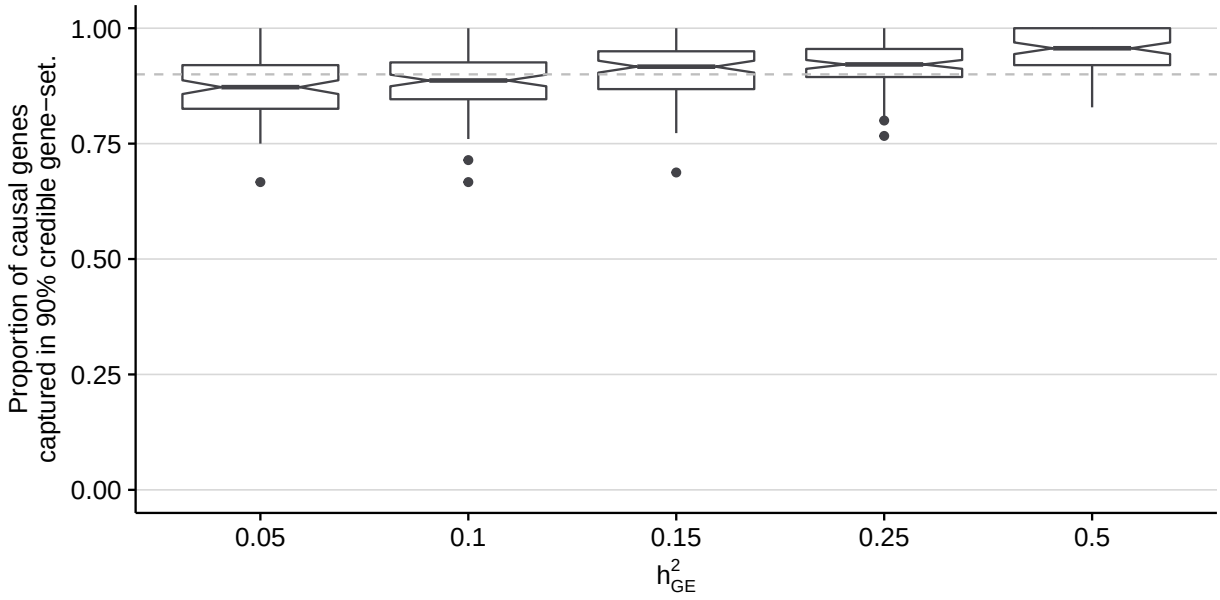
Supplementary Figure 1: Genes with large average LD tag causal gene TWAS associations. Using TWAS results from our simulation pipeline (i.e. $N_{\text{GWAS}} = 50,000$, $N_{\text{eQTL}} = 500$, $h^2_{\text{GE}} = 0.2$, gene expression $h^2_g = 0.2$), we computed the correlation between TWAS association 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. $p^{-1}\text{trace}(\mathbf{VV})$). b) The "gene-based" correlation score taking into account 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.



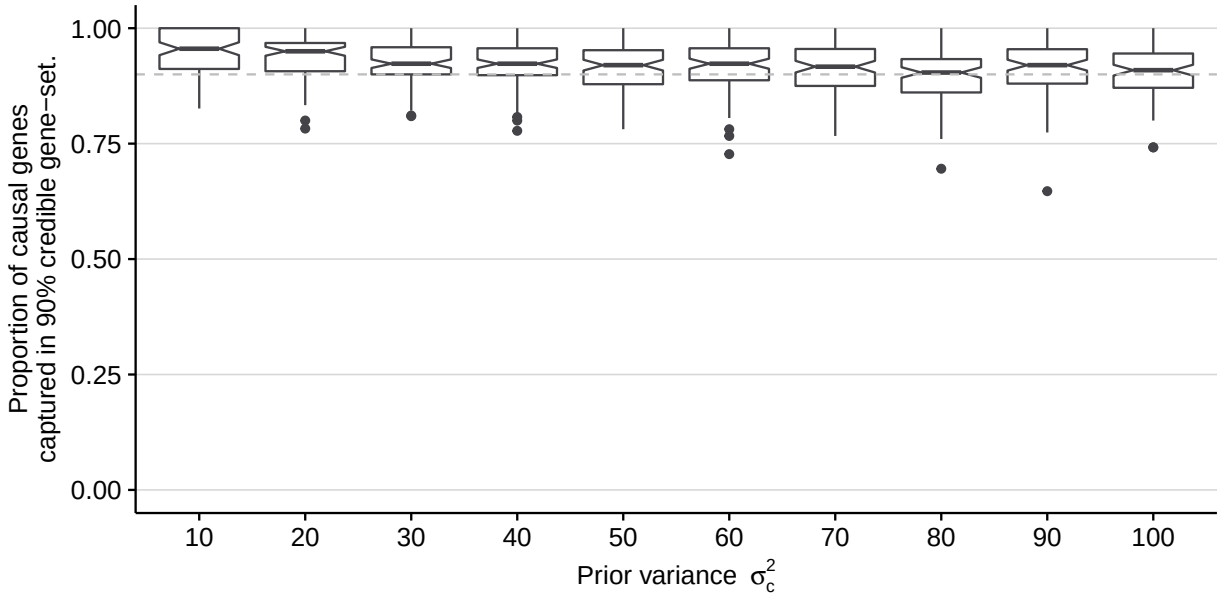
Supplementary Figure 2: Credible gene-sets are unbiased. We simulated 100 complex traits as a function of underlying gene expression (i.e. $N_{\text{GWAS}} = 50,000$, $N_{\text{eQTL}} = 500$, $h^2_{\text{GE}} = 0.2$, gene expression $h^2_g = 0.2$). We ran FOCUS on the TWAS summary data using $n\sigma^2_c = 80$ and computed ρ-credible sets. 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.



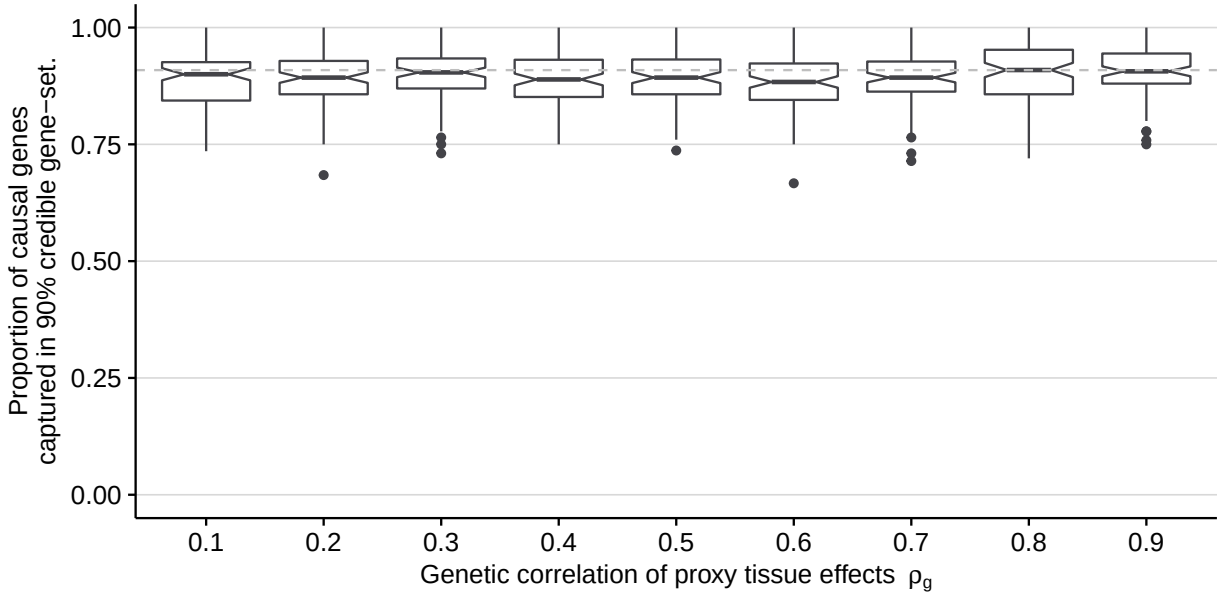
Supplementary Figure 3: Influence of GWAS sample size in prioritizing causal genes. We varied N_{GWAS} in our simulation pipeline keeping other parameters fixed (i.e. $N_{\text{eQTL}} = 500$, $h^2_{\text{GE}} = 0.2$, and gene expression $h^2_g = 0.2$). We ran FOCUS on the TWAS summary data using $n\sigma^2_c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



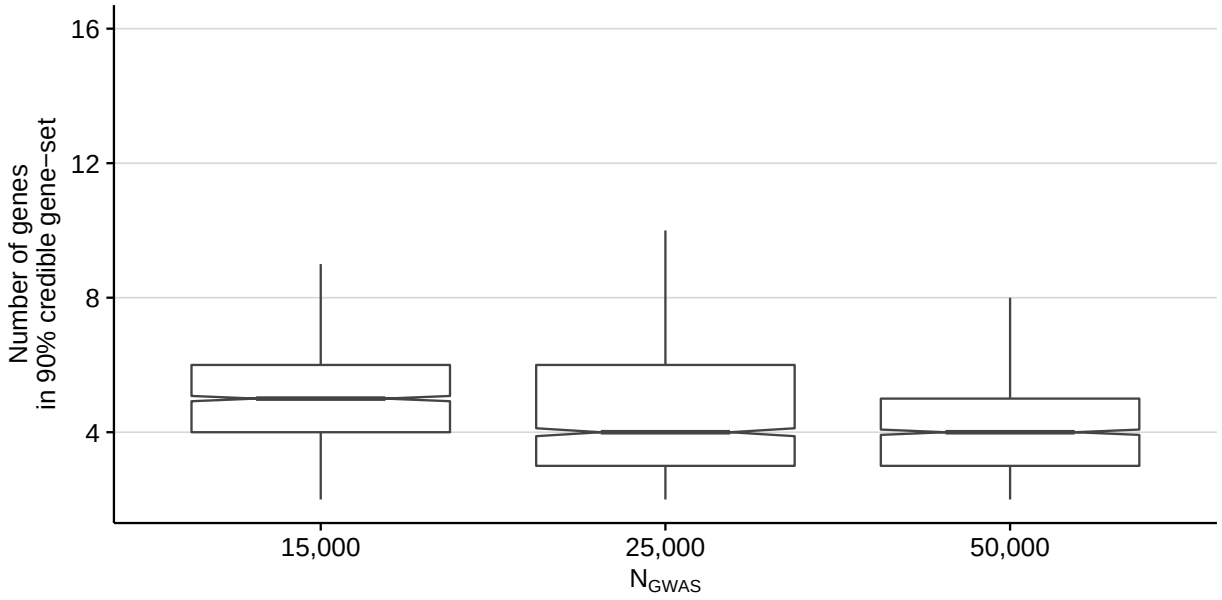
Supplementary Figure 4: Influence of trait-variance explained from predicted gene expression in prioritizing causal genes. We varied N_{eQTL} in our simulation pipeline keeping other parameters fixed (i.e. $N_{\text{GWAS}} = 50,000$, $h^2_{\text{GE}} = 0.2$, and gene expression $h^2_g = 0.2$). We ran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



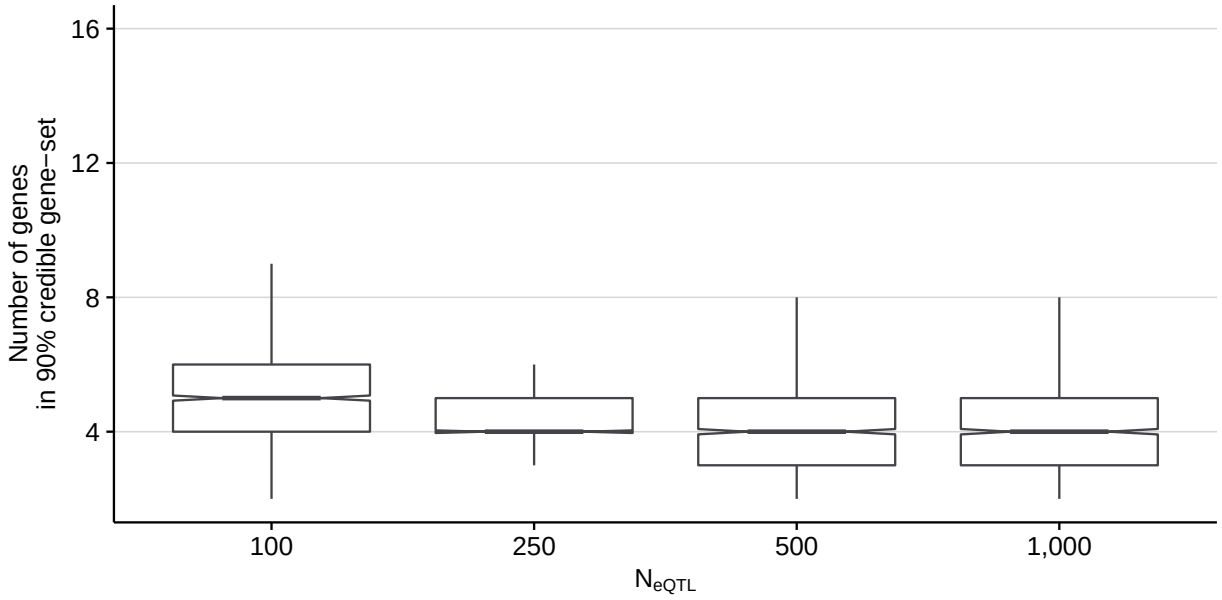
Supplementary Figure 5: FOCUS is stable to various settings of prior variance. We used our simulation pipeline keeping all parameters fixed (i.e. $N_{\text{GWAS}} = 50,000$, $N_{\text{eQTL}} = 500$, $h^2_{\text{GE}} = 0.2$, and gene expression $h^2_g = 0.2$). We ran FOCUS on the TWAS summary data varying $n\sigma^2_c$ and computed 90%-credible gene sets. Box-plots indicate the distribution of the proportion of causal genes captured in 90% credible gene-sets. The dashed line indicates 90%.



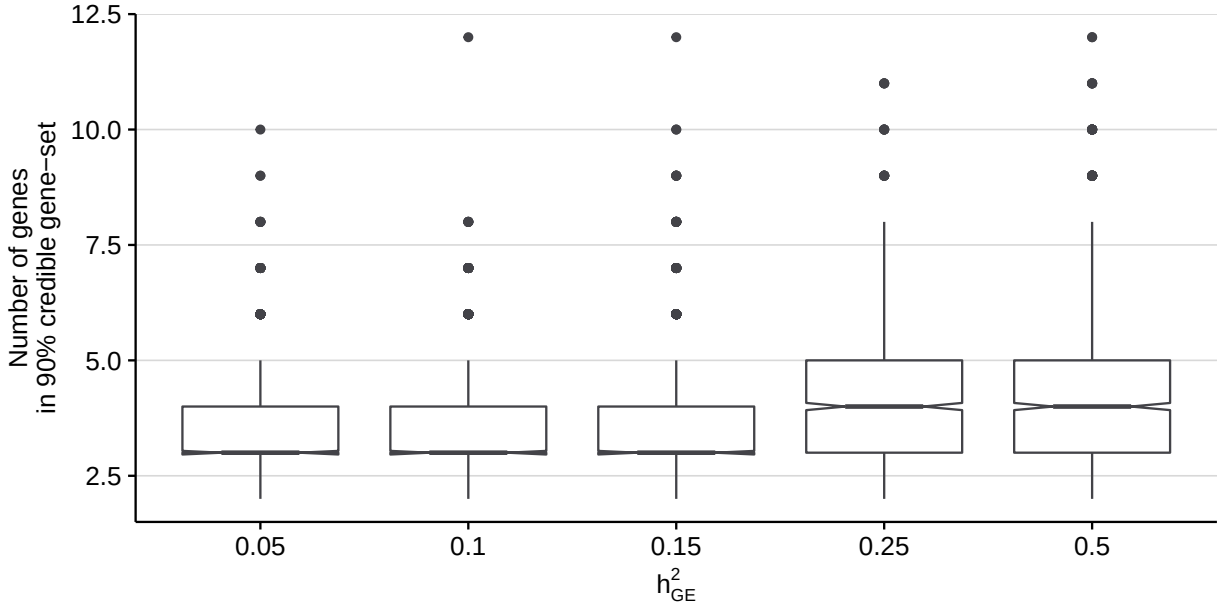
Supplementary Figure 6: FOCUS maintains performance when using expression in correlated proxy tissues. We used our simulation pipeline keeping all parameters fixed (i.e. $N_{\text{GWAS}} = 50,000$, $N_{\text{eQTL}} = 500$, $h^2_{\text{GE}} = 0.2$, and gene expression $h^2_g = 0.2$), but here we used proxy-tissue gene expression data (also $h^2_g = 0.2$) for the eQTL reference panel. We ran FOCUS on the TWAS summary data using $n\sigma^2_c = 80$ and computed 90%-credible gene sets. Box-plots indicate the distribution of the proportion of causal genes captured in 90% credible gene-sets. The dashed line indicates 90%.



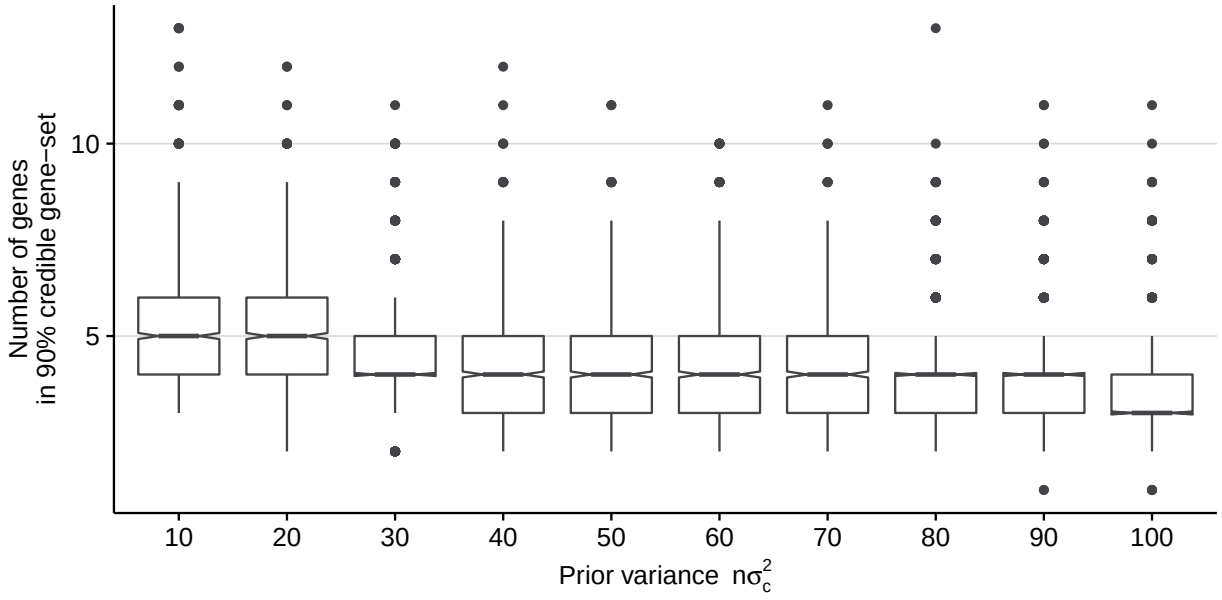
Supplementary Figure 7: Size of 90%-credible gene-sets as a function of GWAS size. We varied N_{GWAS} in our simulation pipeline keeping other parameters fixed (i.e. $N_{\text{eQTL}} = 500$, $h^2_{\text{GE}} = 0.2$, and gene expression $h^2_g = 0.2$). We ran FOCUS on the TWAS summary data using $n\sigma^2_c = 80$ and computed 90%-credible gene sets. Box-plots indicate the distribution of the total number of genes in 90% credible gene-sets.



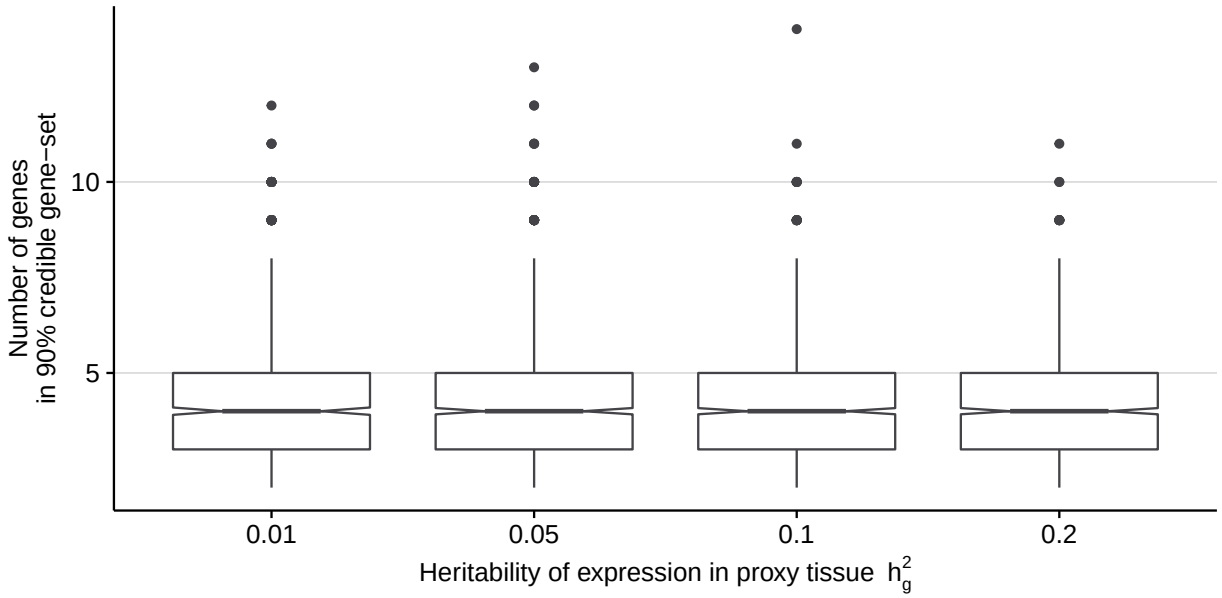
Supplementary Figure 8: Size of 90%-credible gene-sets as a function of eQTL reference panels size. We varied N_{eQTL} in our simulation pipeline keeping other parameters fixed (i.e. $N_{\text{GWAS}} = 50,000$, $h^2_{\text{GE}} = 0.2$, and gene expression $h^2_g = 0.2$). We ran FOCUS on the TWAS summary data using $n\sigma^2_c = 80$ and computed 90%-credible gene sets. Box-plots indicate the distribution of the total number of genes in 90% credible gene-sets.



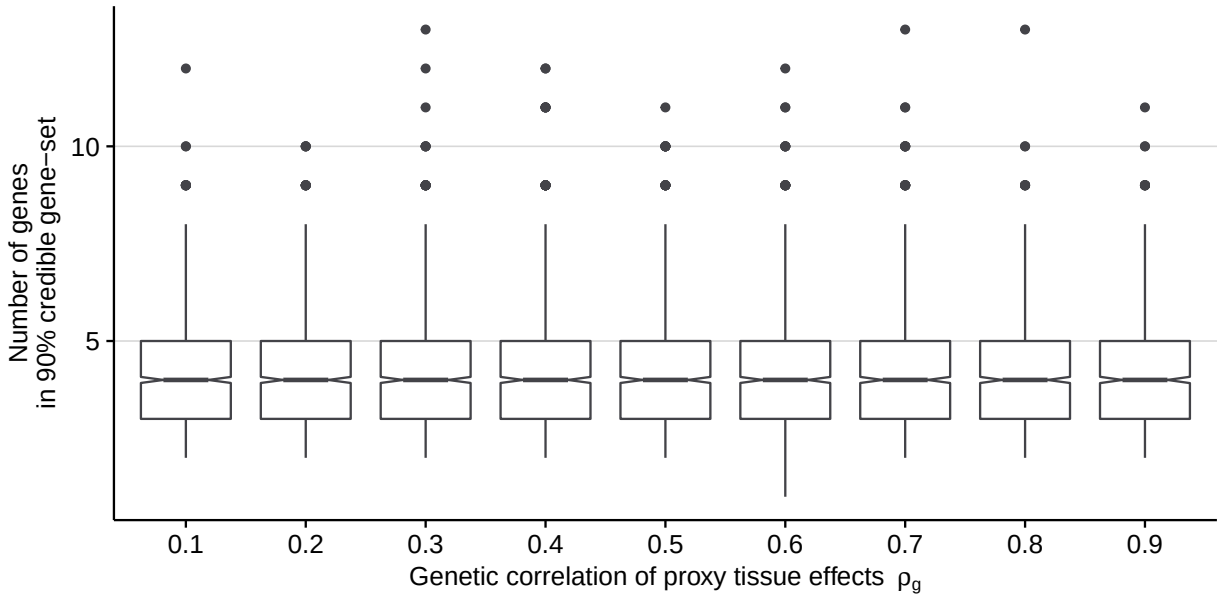
Supplementary Figure 9: Size of 90%-credible gene-sets as a function heritability at causal gene expression. We varied h^2_{GE} in our simulation pipeline keeping other parameters fixed (i.e. $N_{\text{GWAS}} = 50,000$, $N_{\text{eQTL}} = 500$, and gene expression $h^2_g = 0.2$). We ran FOCUS on the TWAS summary data using $n\sigma^2_c = 80$ and computed 90%-credible gene sets. Box-plots indicate the distribution of the total number of genes in 90% credible gene-sets.



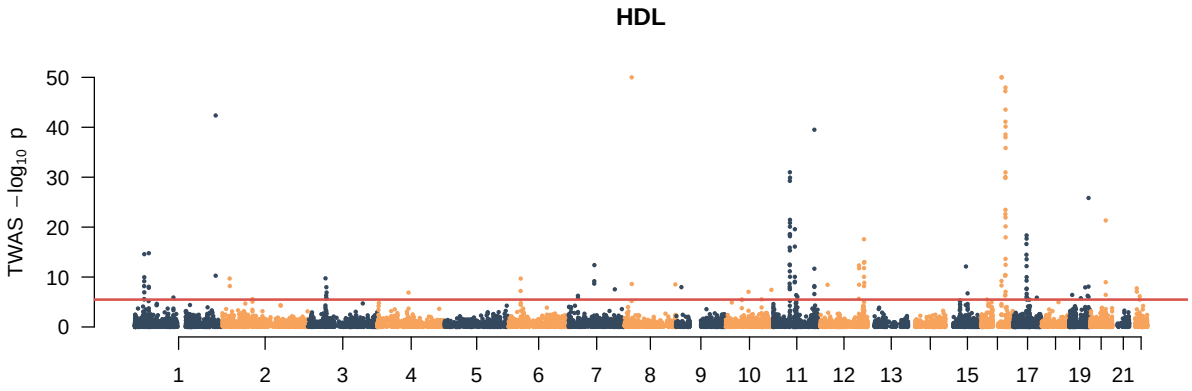
Supplementary Figure 10: Size of 90%-credible gene-sets as a function heritability at causal gene expression. We used our simulation pipeline keeping all parameters fixed (i.e. $N_{\text{GWAS}} = 50,000$, $N_{\text{eQTL}} = 500$, $h^2_{\text{GE}} = 0.2$, and gene expression $h^2_g = 0.2$). We ran FOCUS on the TWAS summary data varying $n\sigma^2_c$ and computed 90%-credible gene sets. Box-plots indicate the distribution of the total number of genes in 90% credible gene-sets.



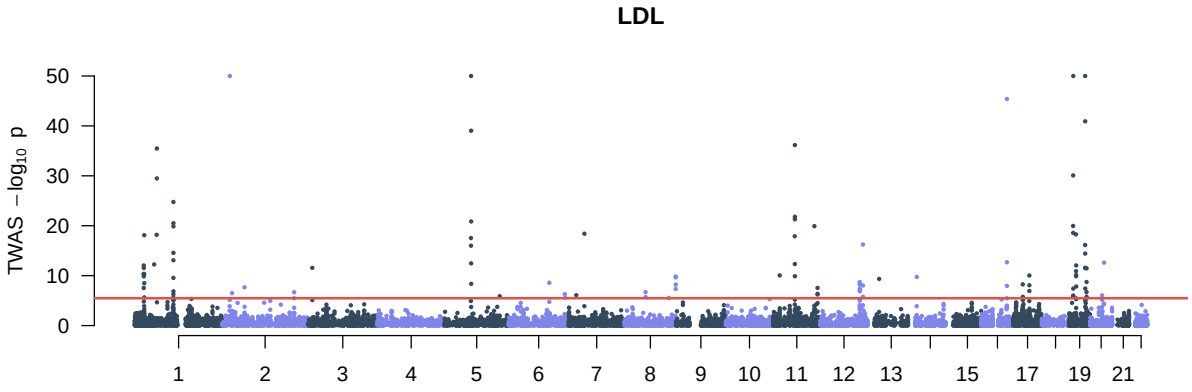
Supplementary Figure 11: Size of 90%-credible gene-sets as a function heritability at proxy-tissue gene expression. We used our simulation pipeline keeping all parameters fixed (i.e. $N_{\text{GWAS}} = 50,000$, $N_{\text{eQTL}} = 500$, $h^2_{\text{GE}} = 0.2$, and gene expression $h^2_g = 0.2$), but here we used proxy-tissue gene expression data with varying h^2_g for the eQTL reference panel keeping $\rho_g = 0.9$. We ran FOCUS on the TWAS summary data using $n\sigma^2_c = 80$ and computed 90%-credible gene sets. Box-plots indicate the distribution of the total number of genes in 90% credible gene-sets.



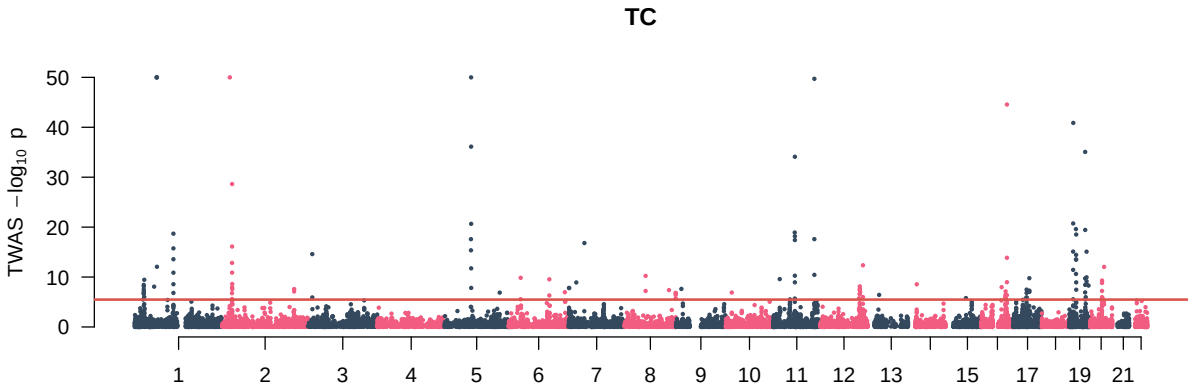
Supplementary Figure 12: Size of 90%-credible gene-sets as a function of genetic correlation for proxy-tissue expression. We used our simulation pipeline keeping all parameters fixed (i.e. $N_{\text{GWAS}} = 50,000$, $N_{\text{eQTL}} = 500$, $h^2_{\text{GE}} = 0.2$, and gene expression $h^2_g = 0.2$), 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 distribution of the total number of genes in 90% credible gene-sets.



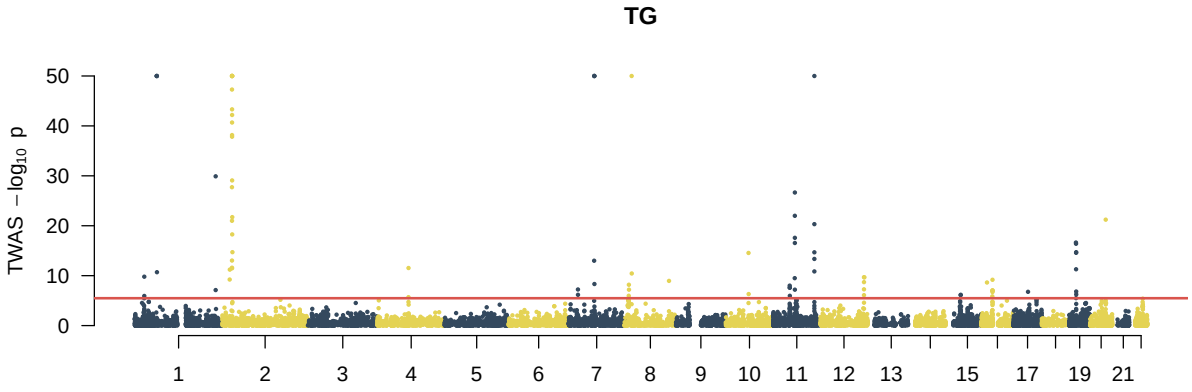
Supplementary Figure 13: Adipose-prioritized TWAS for high density lipoprotein (HDL) measurements. Manhattan plot of HDL TWAS results. Each point represents the association strength of each tested gene. We used a Bonferroni-adjusted transcriptome-wide significance of $0.05/15,277$ (indicated by red line).



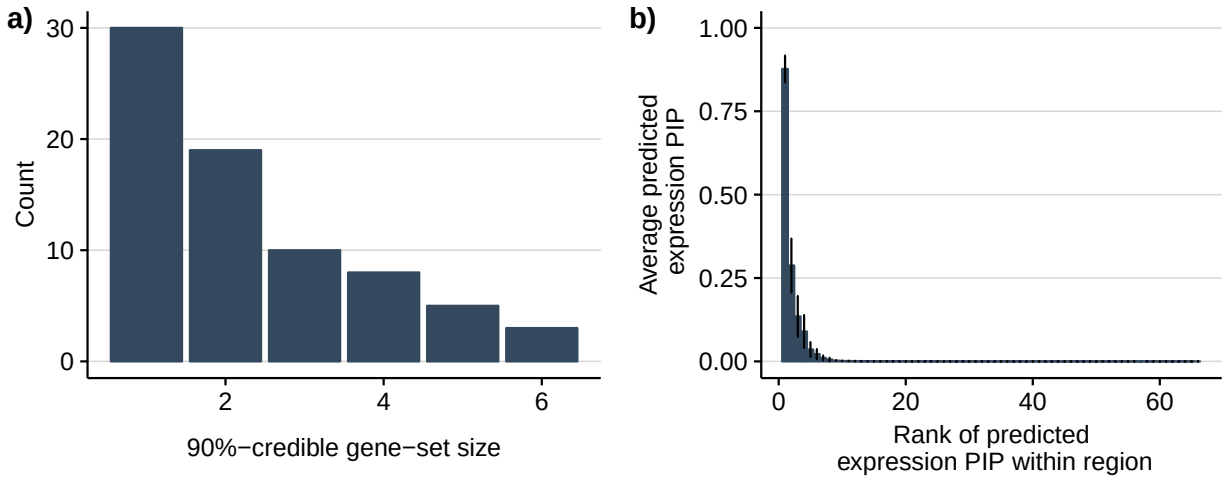
Supplementary Figure 14: Adipose-prioritized TWAS for low density lipoprotein measurements (LDL). Manhattan plot of LDL TWAS results. Each point represents the association strength of each tested gene. results. We used a Bonferroni-adjusted transcriptome-wide significance of $0.05/15,277$ (indicated by red line).



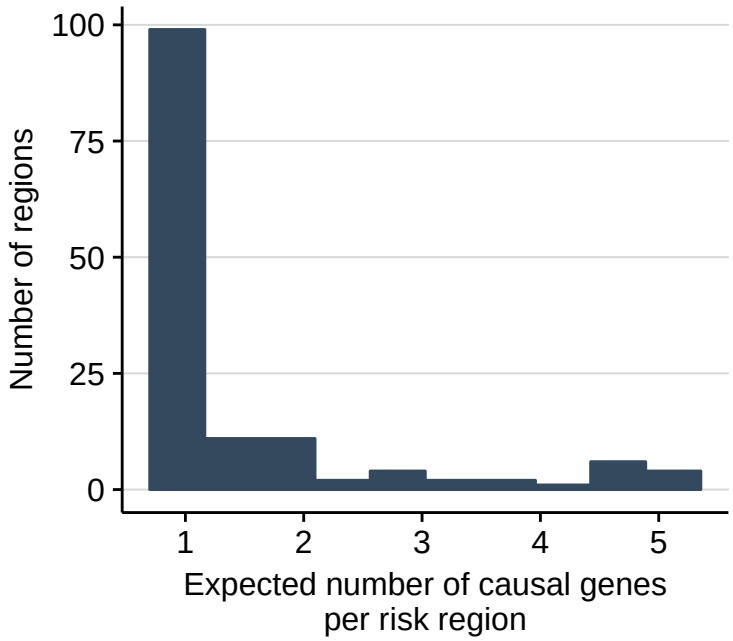
Supplementary Figure 15: Adipose-prioritized TWAS for total cholesterol measurements (TC). Manhattan plot of TC TWAS results. Each point represents the association strength of each tested gene. results. We used a Bonferroni-adjusted transcriptome-wide significance of $0.05/15,277$ (indicated by red line).



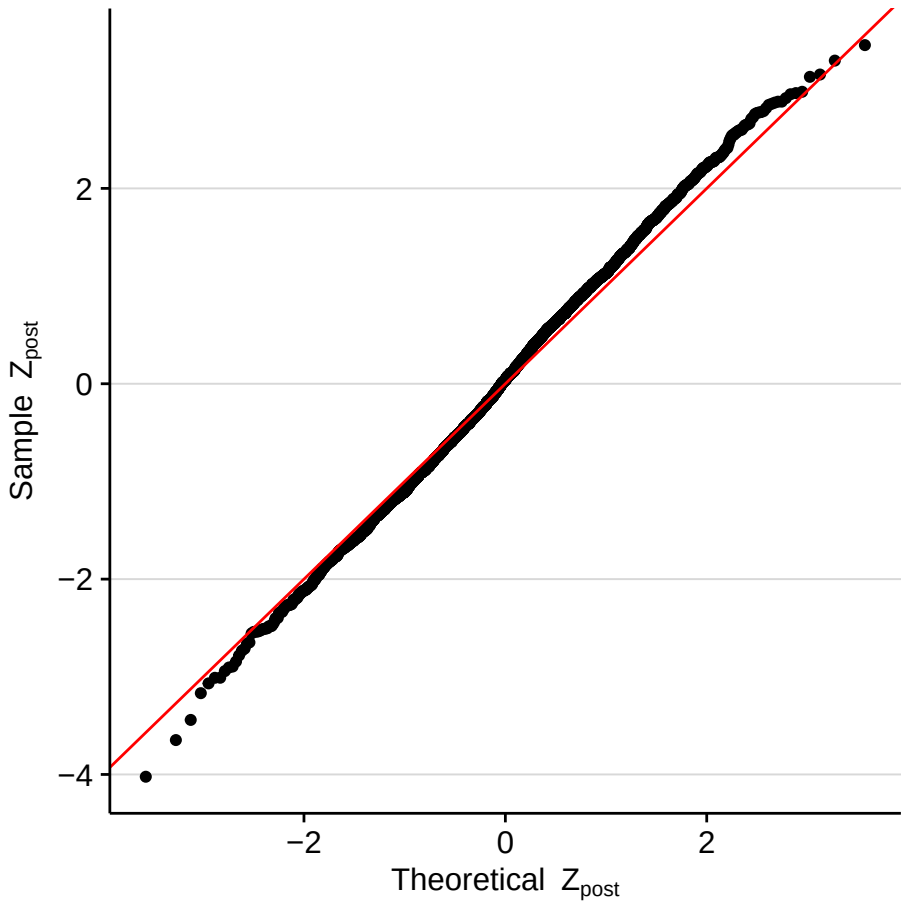
Supplementary Figure 16: Adipose-prioritized TWAS for triglyceride measurements (TG). Manhattan plot of TG TWAS results. Each point represents the association strength of each tested gene. results. We used a Bonferroni-adjusted transcriptome-wide significance of $0.05/15,277$ (indicated by red line).



Supplementary Figure 17: Fine mapping of lipids TWAS risk regions. a) Number of genes with predicted expression at TWAS risk regions for each lipids trait. b) The 90%-credible gene-set for most lipids TWAS risk regions contained a single gene. c) Average PIP across risk regions according to ranking within each credible set.



Supplementary Figure 18: Distribution 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.



Supplementary Figure 19: 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. $Z_{post,i} = \frac{\text{mean}(Z^*_{twas,i}) - Z_{twas,i}}{sd(Z_{twas,i})}$).

References

[1] Christian Benner, Chris CA Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, 2016.

[2] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833, 2011.