Probabilistic fine-mapping of transcriptome-wide association studies

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

1 Supplementary Note

GWAS & eQTL Simulation for COLOC comparison

COLOC[1] is a powerful statistical framework that computes the posterior probability that a single causal variant is shared between two traits. To compare FOCUS with COLOC, we simulated both eQTL and GWAS summary data. Each simulation sampled a random LD-block across the genome that overlapped between 10 and 25 gene transcription start sites as defined by RefSeq. For each gene definition, we sampled causal eQTLs preferentially within 100kb of the transcription start site. Expression was simulated as $\mathbf{G}_i = \mathbf{X}\mathbf{w}_i + \boldsymbol{\epsilon}_i$ where $\mathbf{X}$ is the $n \times p$ centered and standardized genotype matrix, $\mathbf{w}_i$ are the causal eQTL effects, and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, 1 - h^2_g)$ is random environmental noise. In practice we sampled $h^2_g \sim \text{Beta}(\theta_1, \theta_2)$ to match observations of SNP heritability in real data with $h^2_g = 0.2$ on average. After simulating gene expression levels, we generated a complex trait as $\mathbf{y} = \mathbf{G}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$ where $\mathbf{G}$ is the true genetic component of expression, $\boldsymbol{\alpha}$ are causal gene effects, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1 - h^2_{ge})$, with $h^2_{ge} = 0.25$ on average. To achieve statistical power similar to GWAS with our limited sample size, we generated multiple instances $\mathbf{y}_1, \ldots, \mathbf{y}_k$ keeping $\widehat{\mathbf{G}}\boldsymbol{\alpha}$ fixed. To generate SNP Z-scores $\mathbf{z}_{\text{gwas}}$, we performed a association scan using standard linear regression over $\mathbf{X}$ for each trait instance $\mathbf{y}_1, \ldots, \mathbf{y}_k$ and combined results using fixed effect meta analysis. We next performed a summary-based TWAS for each of the genes $\mathbf{G}_1, \ldots, \mathbf{G}_m$ defined by the region $\mathbf{X}$ given the meta-analysis trait association Z-scores $\mathbf{z}_{\text{gwas}}$ using GBLUP[2,3] estimated values $\widehat{\mathbf{W}}$.

Table S1: Summary of expression reference panels. See attached excel doc.

Table S2: TWAS results for lipids traits. See attached excel doc.

Table S3: Fine mapping for lipids traits. See attached excel doc.

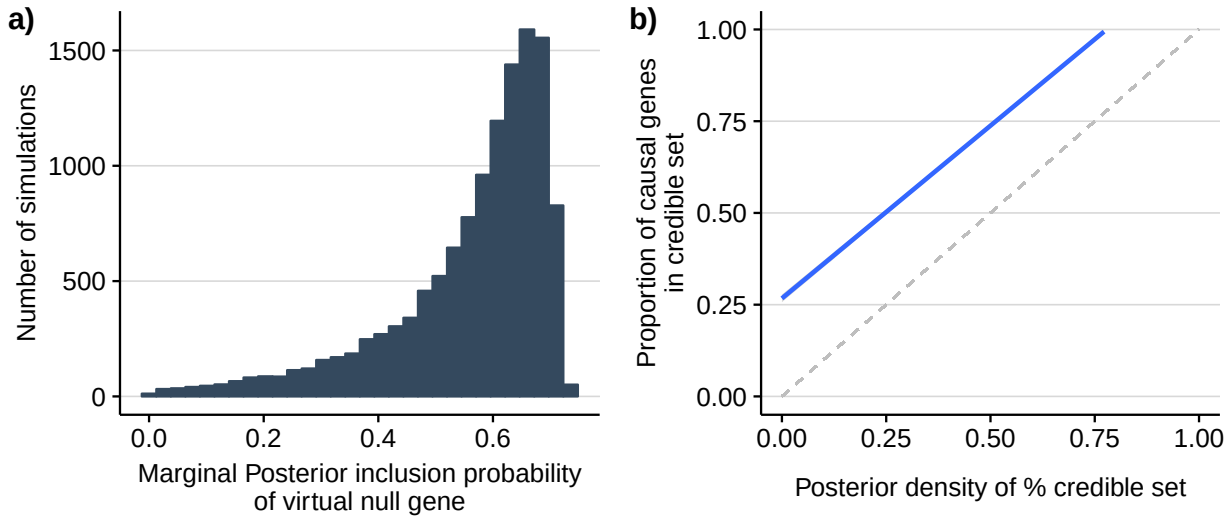

Figure S1: Simulation results where eQTL do not affect downstream trait. a) Histogram of the posterior inclusion probability for the null model across all simulated regions. The distribution skews upward with the null gene typically capturing most of the posterior. b) Credible sets and the proportion capturing the null gene. Computed credible sets tend to contain the null gene more than expected due to it typically having high posterior probability.

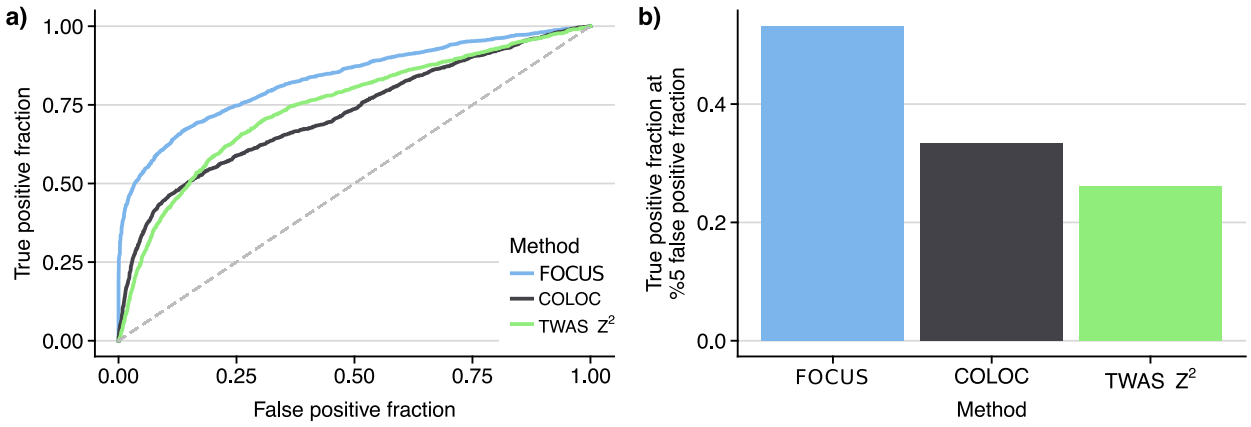

Figure S2: Simulation results using functional eQTL model. We simulated eQTL and GWAS summary data with complex trait being a linear function of gene expression. a) ROC plot for FOCUS, COLOC, and TWAS ranking. b) Snapshot at 5% false discovery rate.

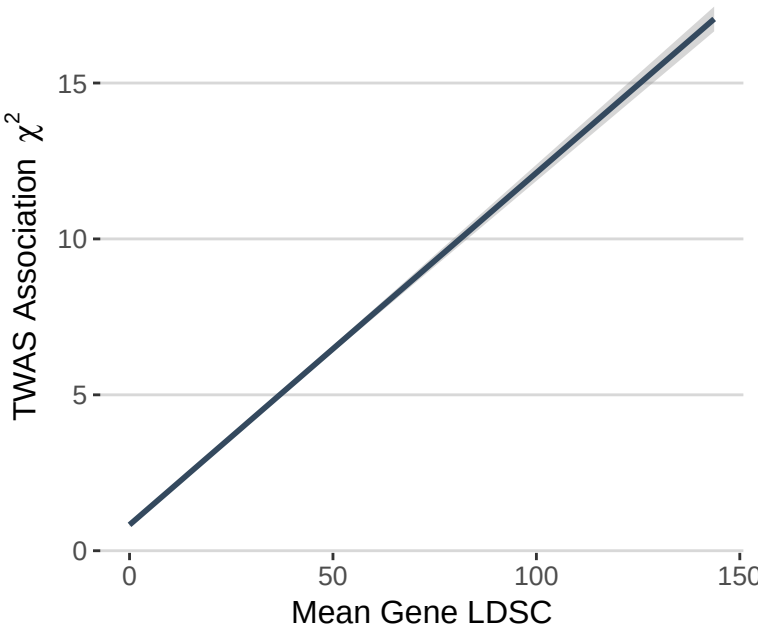

Figure S3: Genes with large average LD tag causal gene TWAS associations. We simulated TWAS association statistics using our model and then computed the association between TWAS association strength (χ^2) and the average LD-score for SNPs with non-zero eQTL weights. The dark line represents the best-fit regression line and the gray area represents the 95% predictive interval.

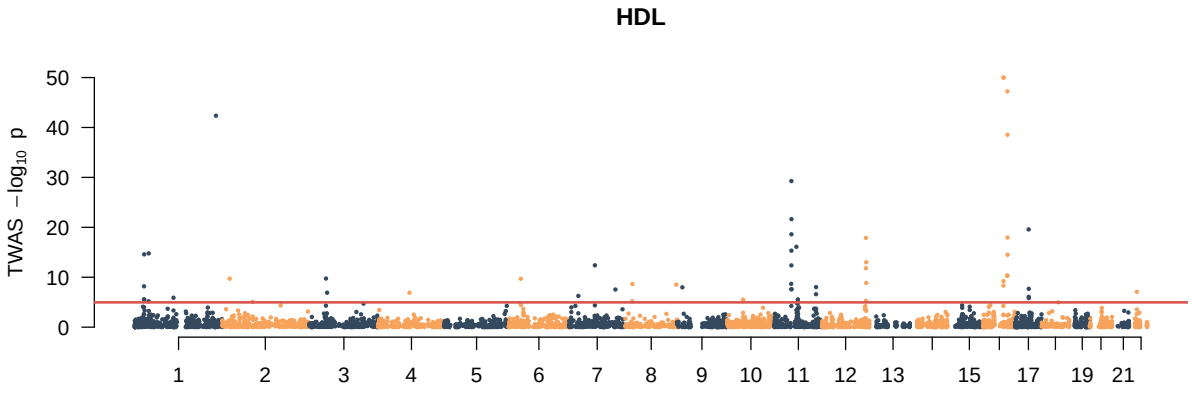

Figure S4: High density lipoprotein TWAS. Manhattan plot of HDL TWAS results. Each point represents the association strength of each tested gene.

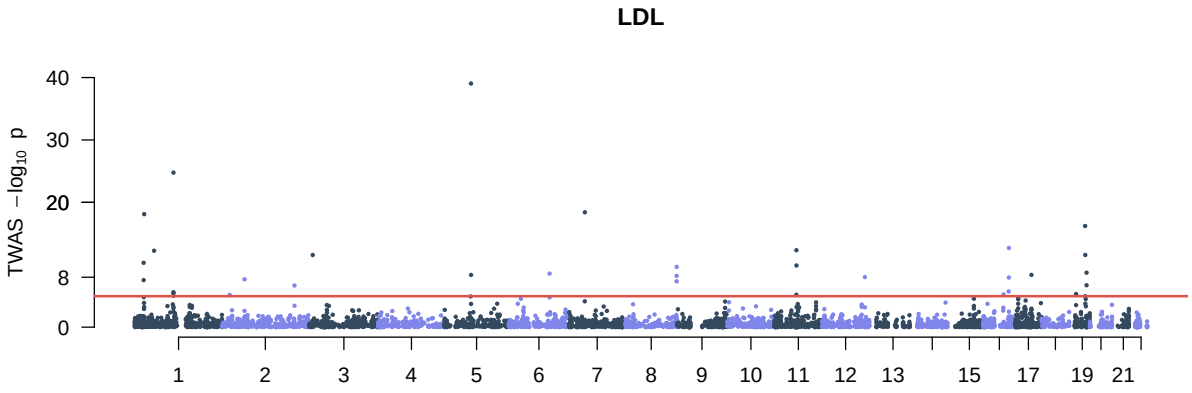

Figure S5: Low density lipoprotein TWAS. Manhattan plot of LDL TWAS results. Each point represents the association strength of each tested gene.

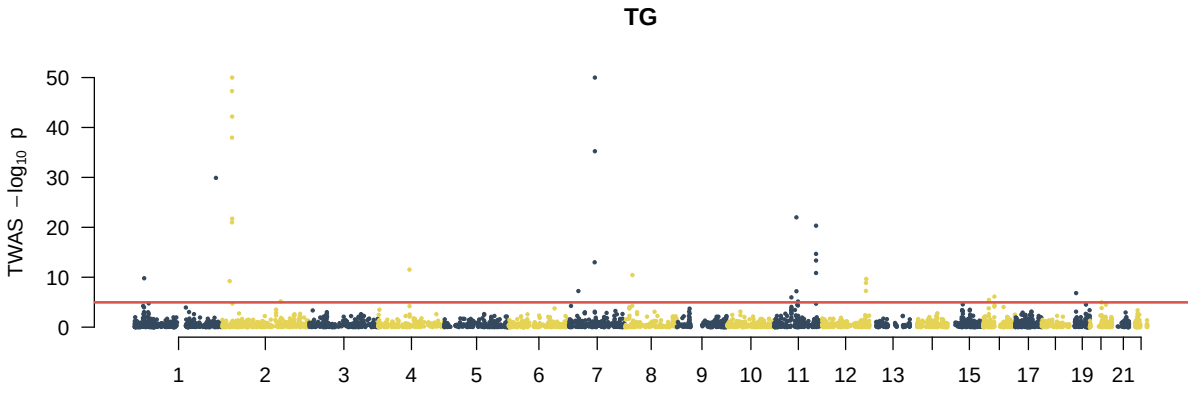

Figure S6: Triglyceride TWAS. Manhattan plot of TG TWAS results. Each point represents the association strength of each tested gene.

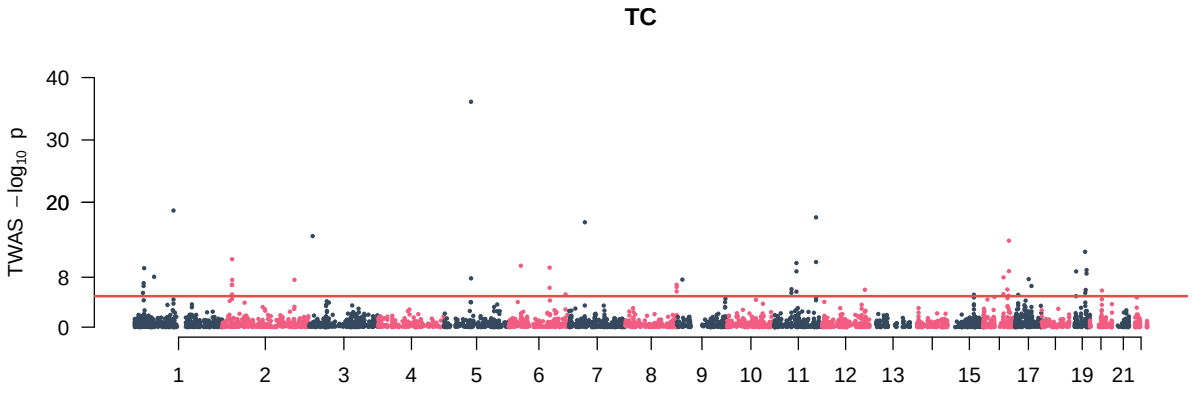

Figure S7: Total cholesterol TWAS. Manhattan plot of TC TWAS results. Each point represents the association strength of each tested gene.

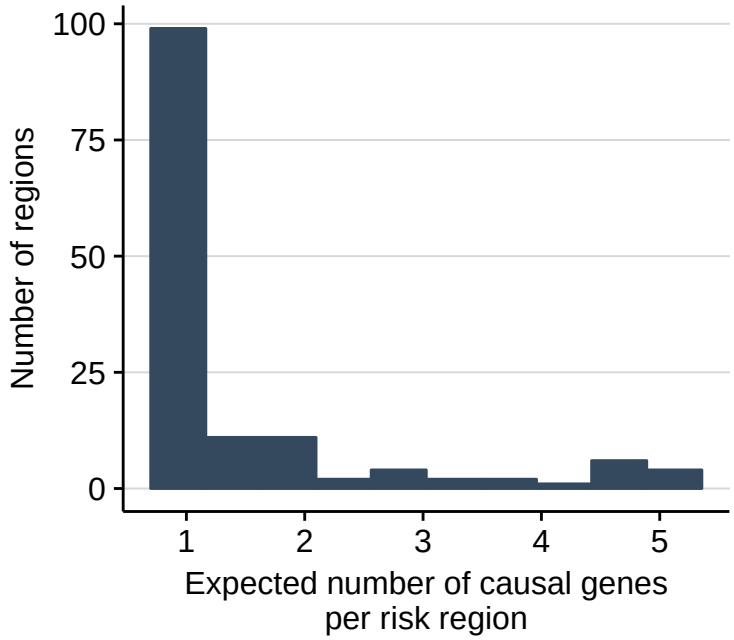

Figure S8: Expected number of causal genes per risk region. We computed the expected number of causal genes under our model by summing across posterior inclusion probabilities per risk region.

References

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

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

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