The American Journal of Human Genetics, Volume *109*

Supplemental information

Multi-ancestry fine-mapping improves precision to identify causal genes in transcriptome-wide association studies

Zeyun Lu, Shyamalika Gopalan, Dong Yuan, David V. Conti, Bogdan Pasaniuc, Alexander Gusev, and Nicholas Mancuso

Supplemental Materials

Supplemental Notes

Estimating TWAS causal effect prior variance

Here we describe an estimator for the prior causal effect-size variance (i.e., $\sigma_{c,i}^2$) similar to the HESS model for local heritability[1]. Our model assumes that marginal Transcriptome-wide Association Study (TWAS) $\boldsymbol{z}$ statistics for m genes have a sampling distribution given by

$$\boldsymbol{z}_{twas,i} \mid \boldsymbol{V}_i, \boldsymbol{\Omega}_i, \boldsymbol{\alpha} \sim N\left(\frac{\sqrt{n_i}}{\sigma_{e,i}}\boldsymbol{\Psi}_i\boldsymbol{\alpha}, \boldsymbol{\Psi}_i\right)$$

where $\boldsymbol{\Psi}_i = \boldsymbol{\Omega}_i^T\boldsymbol{V}_i\boldsymbol{\Omega}_i$. We would like to define an unbiased estimator for the variance explained by (fixed) causal effects $\boldsymbol{\alpha}$. Specifically, $\mathbb{V}(\boldsymbol{G}_i\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T\mathbb{V}(\boldsymbol{G}_i)\boldsymbol{\alpha} = \boldsymbol{\alpha}^T\boldsymbol{W}_i{}^T\boldsymbol{V}_i\boldsymbol{W}\boldsymbol{\alpha} = \boldsymbol{\alpha}^T\boldsymbol{\Psi}_i\boldsymbol{\alpha}$. As a result, an intuitive (but biased) estimator for $\sigma_{c,i}^2$ would be

$$\left(\frac{\sigma_{e,i}}{\sqrt{n_i}}\boldsymbol{\Psi}_i{}^{-1}\boldsymbol{z}_{twas,i}\right)^T\boldsymbol{\Psi}_i\frac{\sigma_{e,i}}{\sqrt{n_i}}\boldsymbol{\Psi}_i{}^{-1}\boldsymbol{z}_{twas,i} = \frac{\sigma_{e,i}^2}{n_i}\boldsymbol{z}_{twas,i}\boldsymbol{\Psi}_i{}^{-1}\boldsymbol{z}_{twas,i}$$

In practice, $\sigma_{e,i}^2$ is extremely close to 1; hence an unbiased estimator for the sample-size scaled causal effect prior variance $n_i\sigma_{c,i}^2$ is given by

$$\begin{aligned}\mathbb{E}\left(n_i\frac{1}{n_i}\boldsymbol{z}_{twas,i}\boldsymbol{\Psi}_i{}^{-1}\boldsymbol{z}_{twas,i}\right) &= \mathrm{tr}(\boldsymbol{\Psi}_i{}^{-1}\boldsymbol{\Psi}_i) + (\sqrt{n_i}\boldsymbol{\Psi}_i\boldsymbol{\alpha}^*)^T\boldsymbol{\Psi}_i{}^{-1}(\sqrt{n_i}\boldsymbol{\Psi}_i\boldsymbol{\alpha}^*)\\ &= m + n_i\boldsymbol{\alpha}^T\boldsymbol{\Psi}_i\boldsymbol{\alpha}\\ \mathbb{E}\left(n_i\frac{1}{n_i}\boldsymbol{z}_{twas,i}\boldsymbol{\Psi}_i{}^{-1}\boldsymbol{z}_{twas,i}\right) - m &= n_i\boldsymbol{\alpha}^T\boldsymbol{\Psi}_i\boldsymbol{\alpha} = n_i\sigma_{c,i}^2.\end{aligned}$$

In practice, when the estimator is negative (e.g., when little TWAS signal exists), we use the biased estimator to ensure positivity.

GENOA data estimates higher LCL heritability and outputs better prediction accuracy

Using paired genotype and lymphoblastoid cell line (LCL) gene expression data of two ancestries in GENOA[2] (European Americans and African Americans; EA and AA) and GEUVADIS[3] (Europeans and Yoruba in Ibadan; EUR and YRI), we estimated the LCL SNP heritability (*cis*-h_g^2) respectively using GCTA[4,5], and we also fit prediction weights using FUSION[6] (see **Methods**). We found lower estimates of *cis*-h_g^2 in GEUVADIS than in GENOA with estimates of 0.06 compared to 0.15 for EUR/EA and 0.08 compared to 0.18 for YRI/AA ($P < 1 \times 10^{-100}$ for all tests). We observed that GENOA-based weights produced higher prediction accuracy with an average

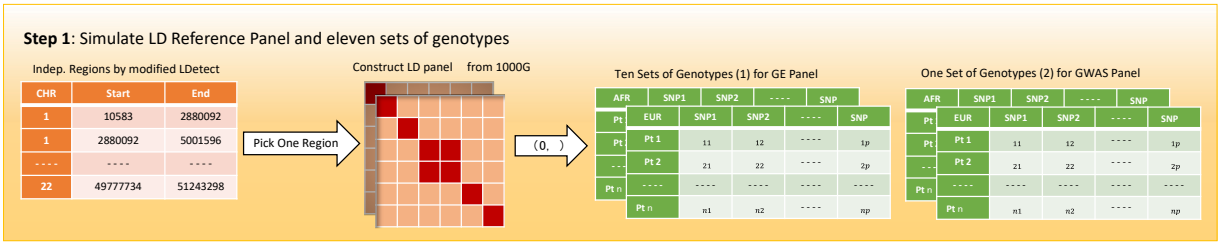
r^2 estimate of 0.04 compared to 0.03 for EA/EUR and of 0.05 compared to 0.02 for AA/YRI ($P = 6.33 \times 10^{-15}$ and $P = 1.36 \times 10^{-95}$; **Figure S14**).

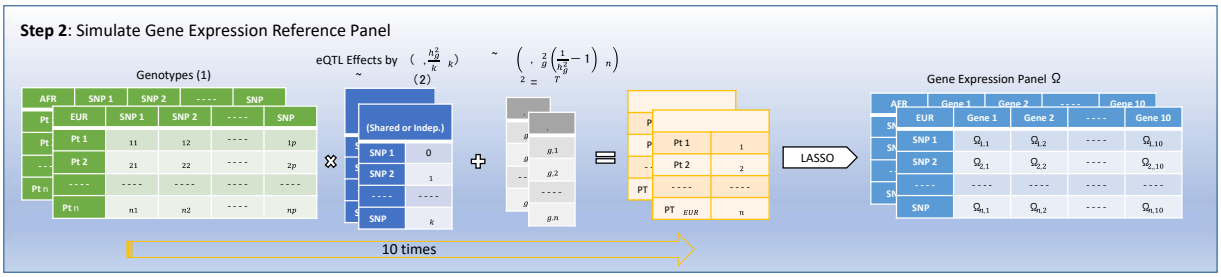
Comparison of different prior numbers of causal genes

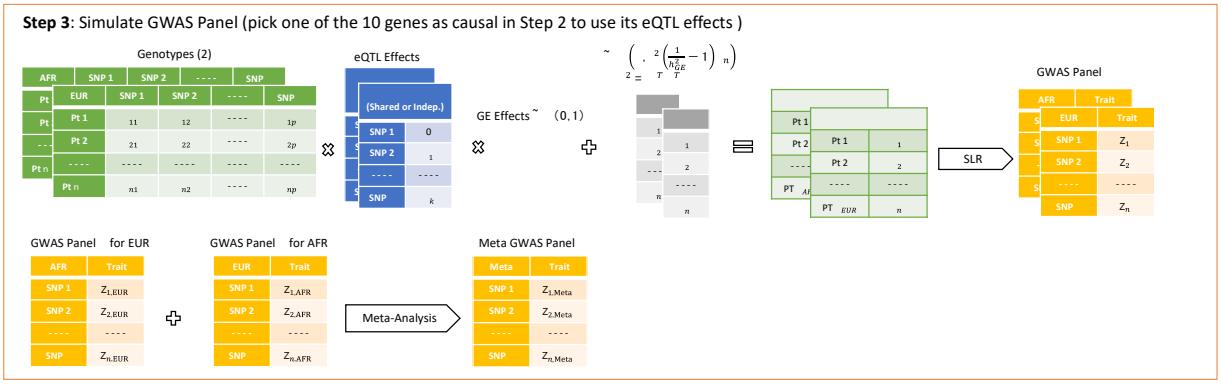
In practice, we recommend setting the maximum number of causal genes as three to avoid computation intractability. To explore the impact of this parameter, we re-ran MA-FOCUS using the same datasets but changing the maximum causal genes to one and five and analyzed the results by comparing the credible gene set size differences, computing the correlations in estimated PIPs, and counting how often the lead gene in the credible sets changed.

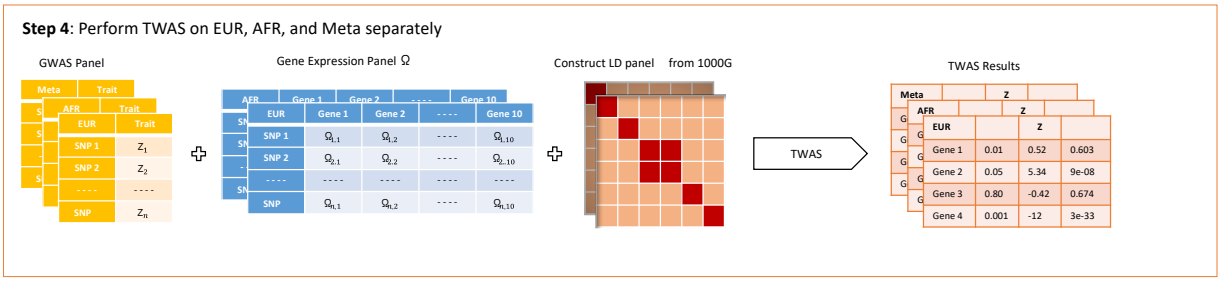
Focusing on the five-gene scenario first, we observed increased credible gene set sizes (4.83 compared to 3.17; $P = 7.79 \times 10^{-11}$), which are consistent with the updated prior. However, despite this increase in the credible set size, the inferred PIPs remained highly correlated with an estimate of 0.706 ($P = 5.07 \times 10^{-51}$; **Figure S20A**). In addition, out of 36 genes with the highest PIPs in the 23 credible sets of the three-gene scenario (to account for ties), 34 remained with the highest PIPs in the five-gene scenario, suggesting that leading genes remain robust, despite the overall increased credible set uncertainty. Next, focusing on the one-gene scenario, the average credible set size decreased from 3.17 to 1.09 ($P = 1.47 \times 10^{-16}$; **Figure S20B**), which is again consistent with the stricter prior penalty. Similarly, we found a significantly non-zero correlation across PIPs, however, to a lesser extent than the five-gene scenario (0.479; $P = 2.76 \times 10^{-20}$). Lastly, we found that 17 out of 36 genes retained their leading position in their respective credible sets. Of the remaining 19 genes, 12 genes dropped to either second or third place, and four genes dropped to more than the tenth rank. We also observed similar results for the baseline, AA FOCUS, and EA FOCUS. Overall, our results are largely stable to the prior number of causal genes but recommend our initial prior of three to strike a balance between uncertainty and computational simplicity.

Supplemental Figures









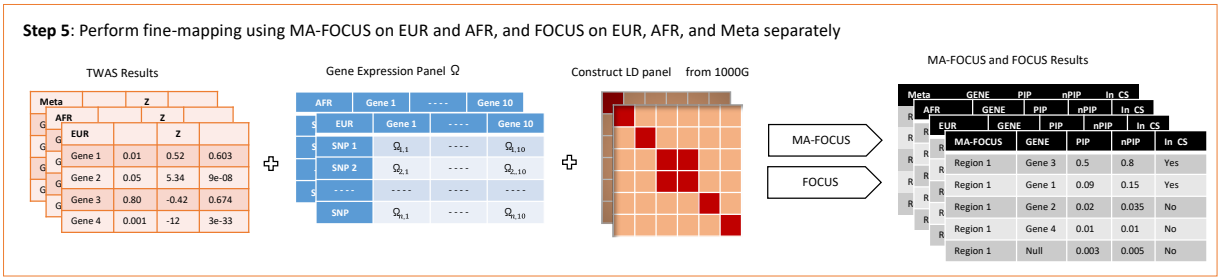


Figure S1. Simulation Process Illustration.

The general procedure of the simulation consists of 5 steps. First, we computed approximately independent Linkage Disequilibrium (LD) blocks for two ancestries. Second, we constructed the expression quantitative trait loci (eQTL) reference panels by sampling genotypes from the LD blocks, simulating eQTL effects, and computing gene expression at causal and non-causal genes. Third, we calculated the Genome-wide Association Study (GWAS) summary statistics by sampling genotypes and simulating a complex trait as a function of eQTL effects of the causal gene from the second step. Fourth, we performed a Transcriptome-wide Association Study (TWAS) using penalized models fitted in the eQTL reference panels. Fifth, we performed fine-mapping using different approaches, including our new method, MA-FOCUS.

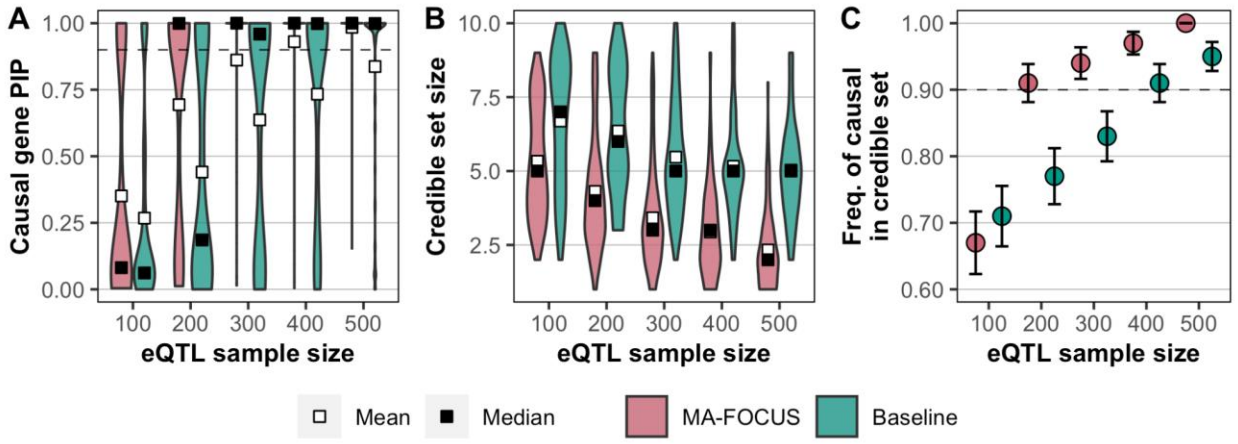


Figure S2. MA-FOCUS outperforms the baseline approach in all three metrics as eQTL sample sizes vary when eQTLs are independent across ancestries.

Posterior Inclusion Probabilities (PIPs) for 100 simulated causal genes (A), the distribution of 90% credible gene set sizes for 100 simulated gene regions (B), and the sensitivity (C) from MA-FOCUS and the baseline approach, varying expression quantitative trait loci (eQTL) panel size for both ancestries. The black dashed lines indicate 90%. See **Methods** section for default parameters. Error bars are constructed using a 95% confidence interval.

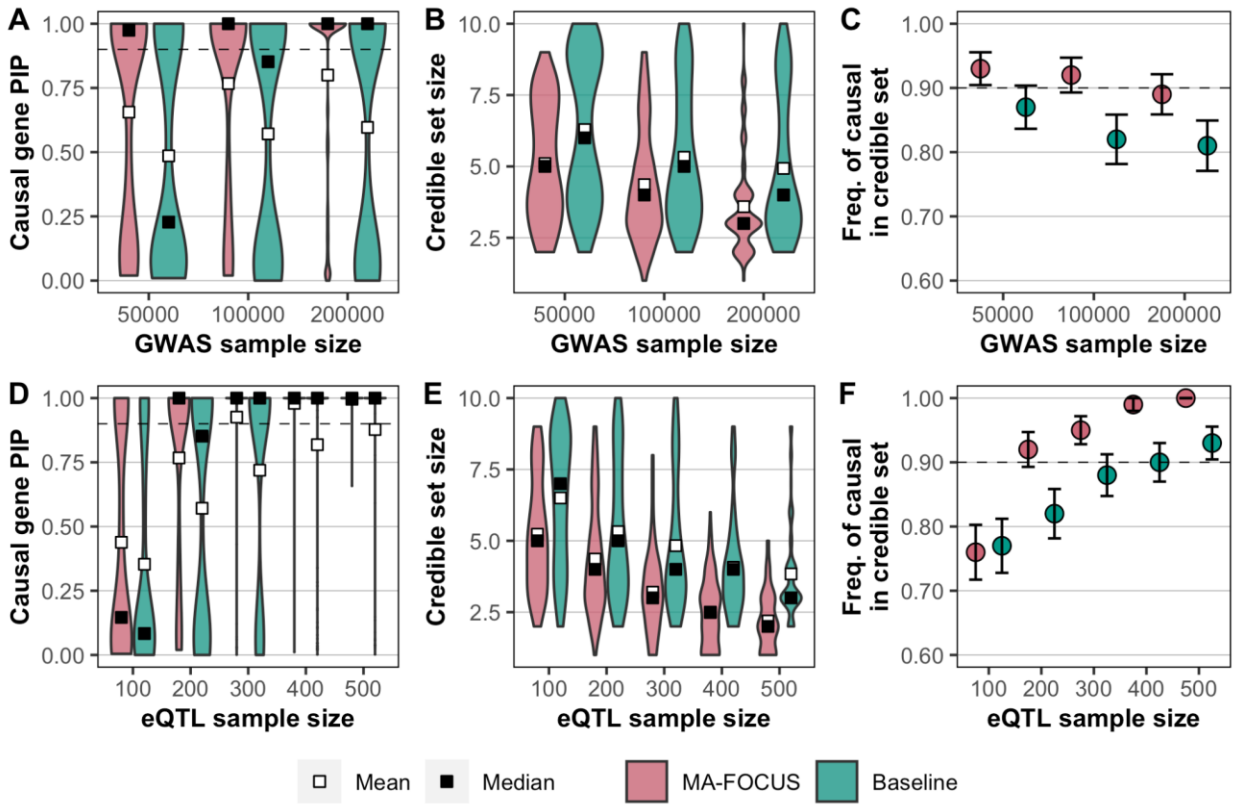


Figure S3. MA-FOCUS outperforms the baseline approach in all three metrics as GWAS and eQTL sample sizes vary when eQTLs are shared across ancestries.

Posterior Inclusion Probabilities (PIPs) for 100 simulated causal genes (A, D), the distribution of 90% credible gene set sizes for 100 simulated gene regions (B, E), and the sensitivity (C, F) from MA-FOCUS, and baseline approach, varying Genome-wide Association Study (GWAS; A, B, C) and expression quantitative trait loci (eQTL; D, E, F) sample sizes. The GWAS sample size was fixed at 100,000 when varying eQTL sample size, and the eQTL sample size was fixed at 200 when varying GWAS sample size. See **Methods** section for default parameters. The black dashed lines indicate 90%. Error bars are constructed using a 95% confidence interval.

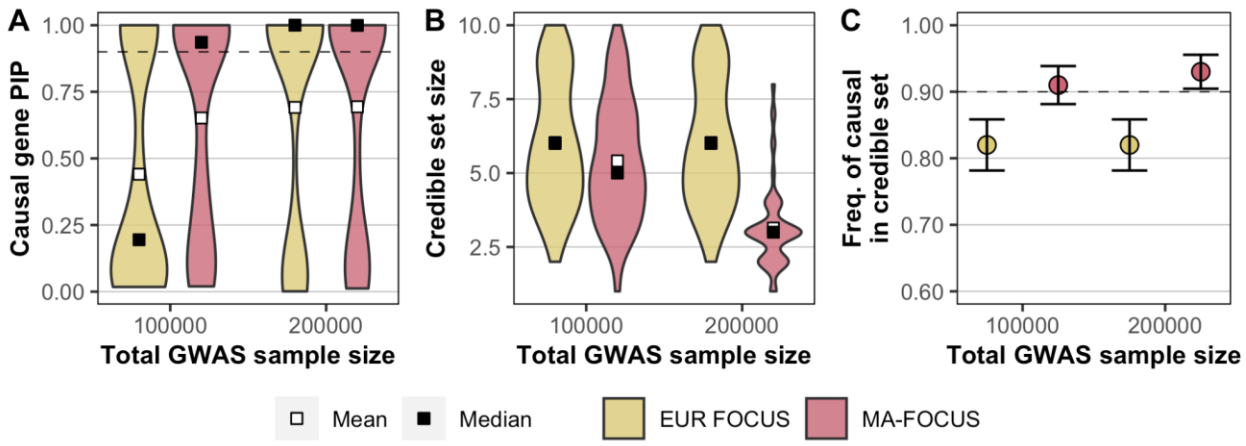


Figure S4. MA-FOCUS outperforms single-ancestry FOCUS in all three metrics when total GWAS samples are fixed, and eQTLs are independent across ancestries.

Posterior Inclusion Probabilities (PIPs) for 100 simulated causal genes (A), the distribution of 90% credible gene set sizes for 100 simulated gene regions (B), and the sensitivity (C) from MA-FOCUS and the single-ancestry FOCUS for Europeans (EUR), varying total Genome-wide Association Study (GWAS) sample sizes. The total GWAS sample size is equally split for each ancestry for the MA-FOCUS approach. See **Methods** section for default parameters. The black dashed lines indicate 90%. Error bars are constructed using a 95% confidence interval.

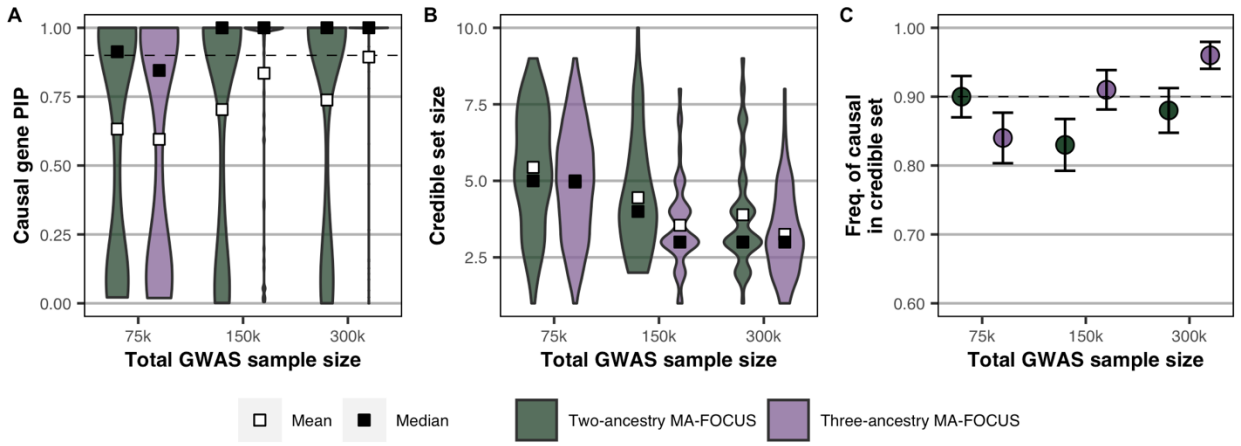


Figure S5. MA-FOCUS performs better using three ancestries compared to two when eQTLs are independent across ancestries.

Posterior Inclusion Probabilities (PIPs) for 100 simulated causal genes (A), the distribution of 90% credible gene set sizes for 100 simulated gene regions (B), and the sensitivity (C) from EUR-AFR (two-ancestry) MA-FOCUS and EUR-AFR-EAS (three-ancestry) MA-FOCUS, varying total Genome-wide Association Study (GWAS) sample sizes. EUR represents Europeans, AFR represents Africans, and EAS represents East Asians. The total GWAS sample size is equally split for each MA-FOCUS approach for each ancestry. See **Methods** section for default parameters. The black dashed lines indicate 90%. Error bars are constructed using a 95% confidence interval.

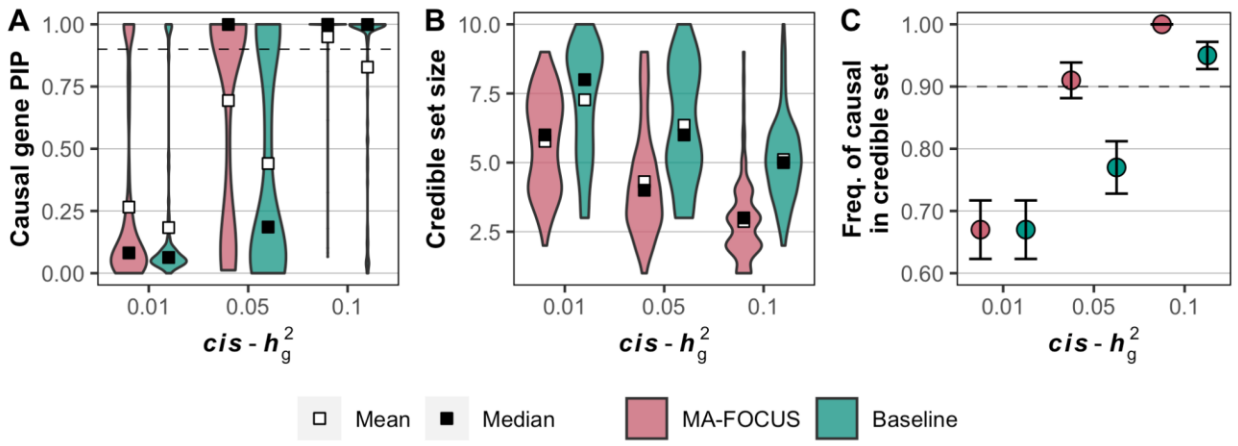


Figure S6. MA-FOCUS outperforms the baseline approach while varying *cis*-h_g^2 when eQTLs are independent across ancestries.

Posterior Inclusion Probabilities (PIPs) for 100 simulated causal genes (A), the distribution of 90% credible gene set sizes for 100 simulated gene regions (B), and the sensitivity (C) from MA-FOCUS and the baseline approach, varying SNP heritability (*cis*-h_g^2) for both ancestries. See **Methods** section for default parameters. The black dashed lines indicate 90%. Error bars are constructed using a 95% confidence interval.

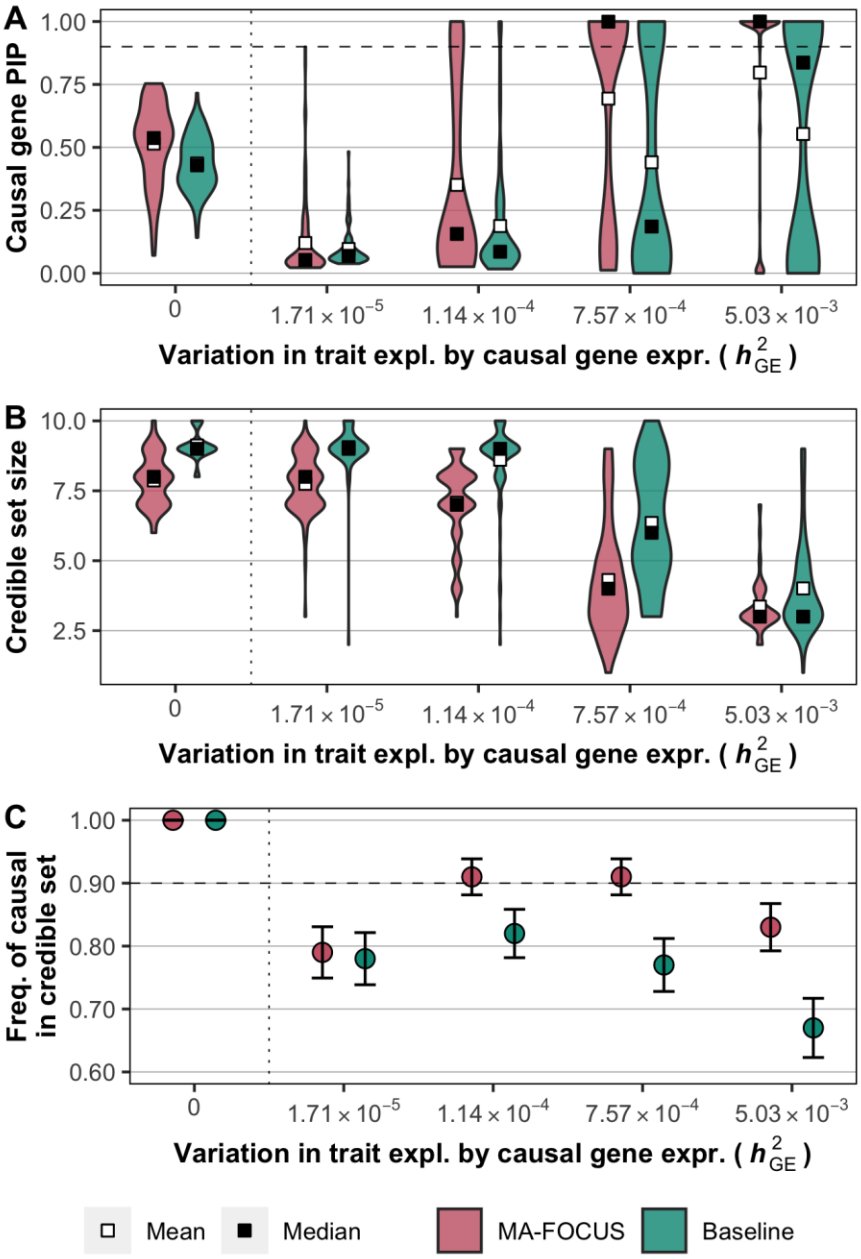


Figure S7. MA-FOCUS outperforms the baseline approach while varying h^2_{GE} when eQTLs are independent across ancestries.

Posterior Inclusion Probabilities (PIPs) for 100 simulated causal genes (A), the distribution of 90% credible gene set sizes for 100 simulated gene regions (B), and the sensitivity (C) from MA-FOCUS and the baseline approach, varying trait variation explained by causal gene expression (h^2_{GE}) for both ancestries. See **Methods** section for default parameters. The black dashed lines indicate 90%. Error bars are constructed using a 95% confidence interval.

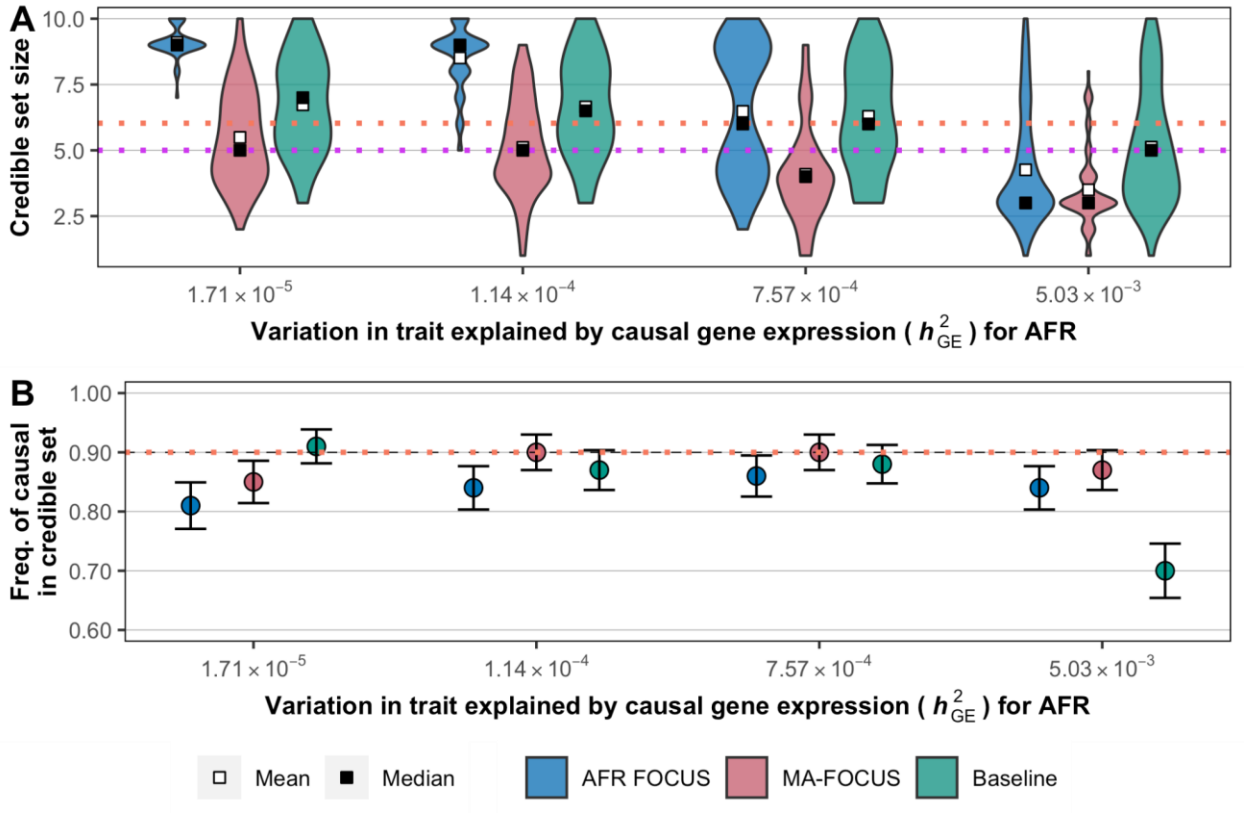


Figure S8. MA-FOCUS performs robustly when h^2_{GE} differs across ancestries.

The distribution of 90% credible gene set sizes for 100 simulated gene regions (A) and the sensitivity (B) from Africans (AFR) FOCUS, MA-FOCUS, and baseline approach, varying complex trait variation explained by causal gene expression (h^2_{GE}) for AFR and fixing for Europeans (EUR) whose $h^2_{GE} = 7.57 \times 10^{-4}$. See **Methods** section for default parameters. The orange and purple dotted lines in (A) indicate the mean and the median of credible gene set size using EUR FOCUS. The orange dotted line in (B) indicates the sensitivity using EUR FOCUS. Error bars are constructed using a 95% confidence interval.

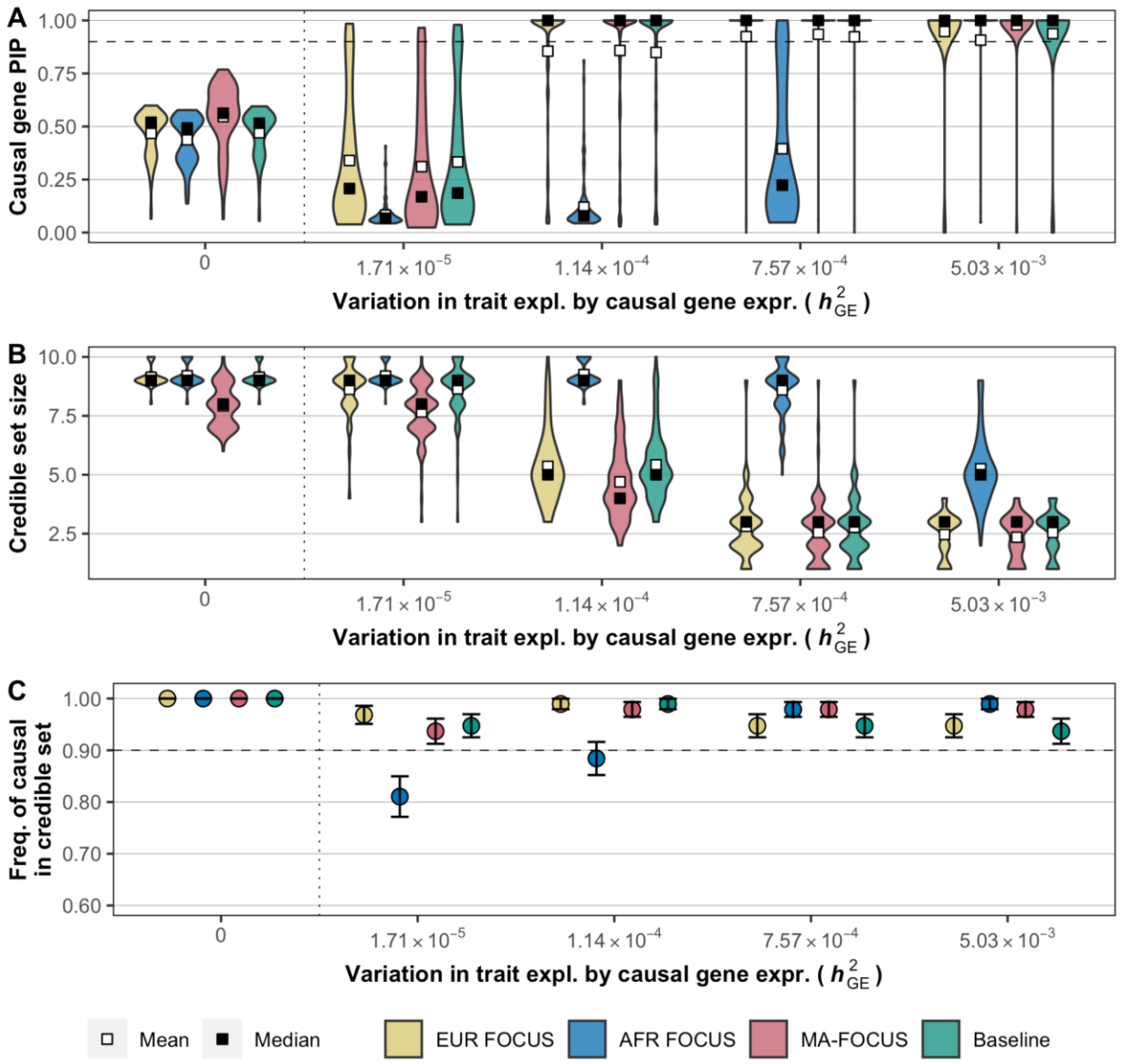


Figure S9. MA-FOCUS outputs smaller gene set sizes when a huge imbalance exists in GWAS sample sizes across ancestries.

Posterior Inclusion Probabilities (PIPs) for 100 simulated causal genes (A), the distribution of 90% credible gene set sizes for 100 simulated gene regions (B), and the sensitivity (C) from the single-ancestry FOCUS for Europeans (EUR) and Africans (AFR), MA-FOCUS, and baseline approach, varying complex trait variation explained by causal gene expression (h^2_{GE}) for each ancestry. The Genome-wide Association Study (GWAS) sample size was fixed at 503,717 and 13,313 for EUR and AFR. The expression quantitative trait loci (eQTL) panel sample size was fixed at 373 and 441 for EUR and AFR. See **Methods** section for default parameters. The black dashed lines indicate 90%. Error bars are constructed using a 95% confidence interval.

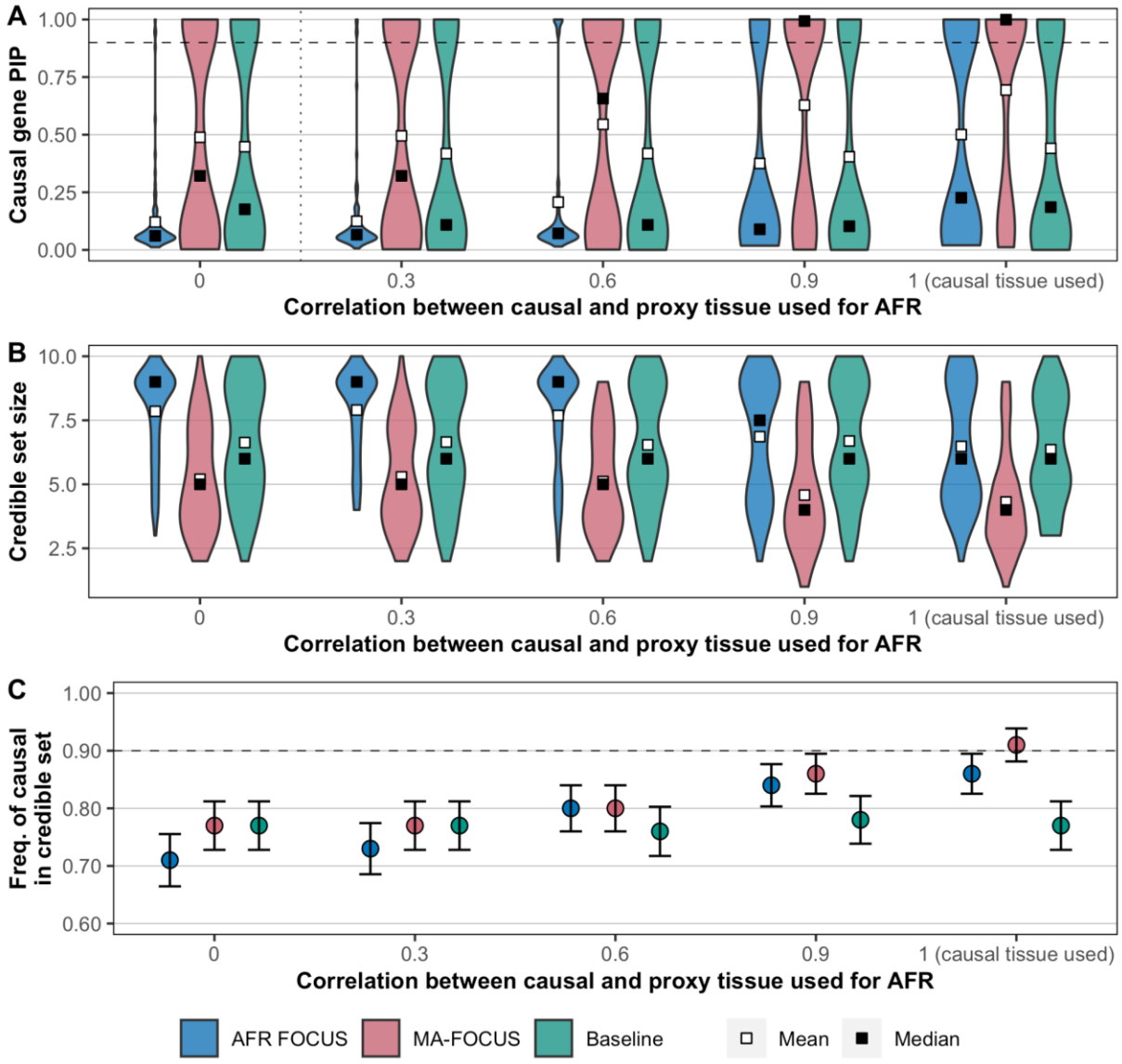


Figure S10. MA-FOCUS outperforms the baseline method when proxy tissue is used.

Posterior Inclusion Probabilities (PIPs) for 100 simulated causal genes (A), the distribution of 90% credible gene set sizes for 100 simulated gene regions (B), and the sensitivity (C) from the single-ancestry FOCUS for Africans (AFR), MA-FOCUS, and baseline approach, varying gene expression correlation between causal and proxy tissues. See **Methods** section for default parameters. The black dashed lines indicate 90%. Error bars are constructed using a 95% confidence interval.

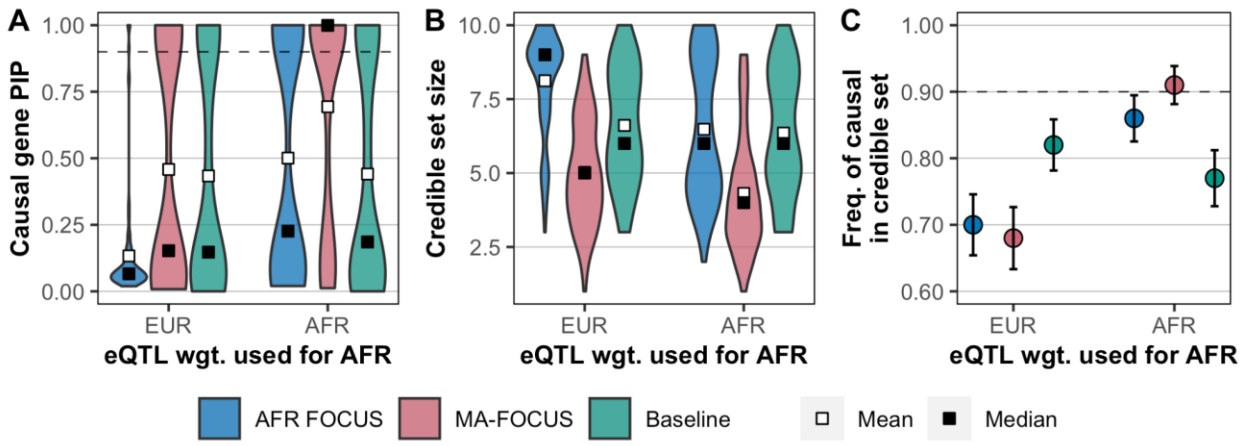


Figure S11. MA-FOCUS remains robust when the EUR eQTL panel is substituted for AFR.

Posterior Inclusion Probabilities (PIPs) for 100 simulated causal genes (A), the distribution of 90% credible gene set sizes for 100 simulated gene regions (B), and the sensitivity (C) from the single-ancestry FOCUS for Africans (AFR), MA-FOCUS, and baseline approach when Europeans (EUR) expression quantitative trait loci (eQTL) weights are substituted for AFR eQTL weights or still used original one. See **Methods** section for default parameters. The black dashed lines indicate 90%. Error bars are constructed using a 95% confidence interval.

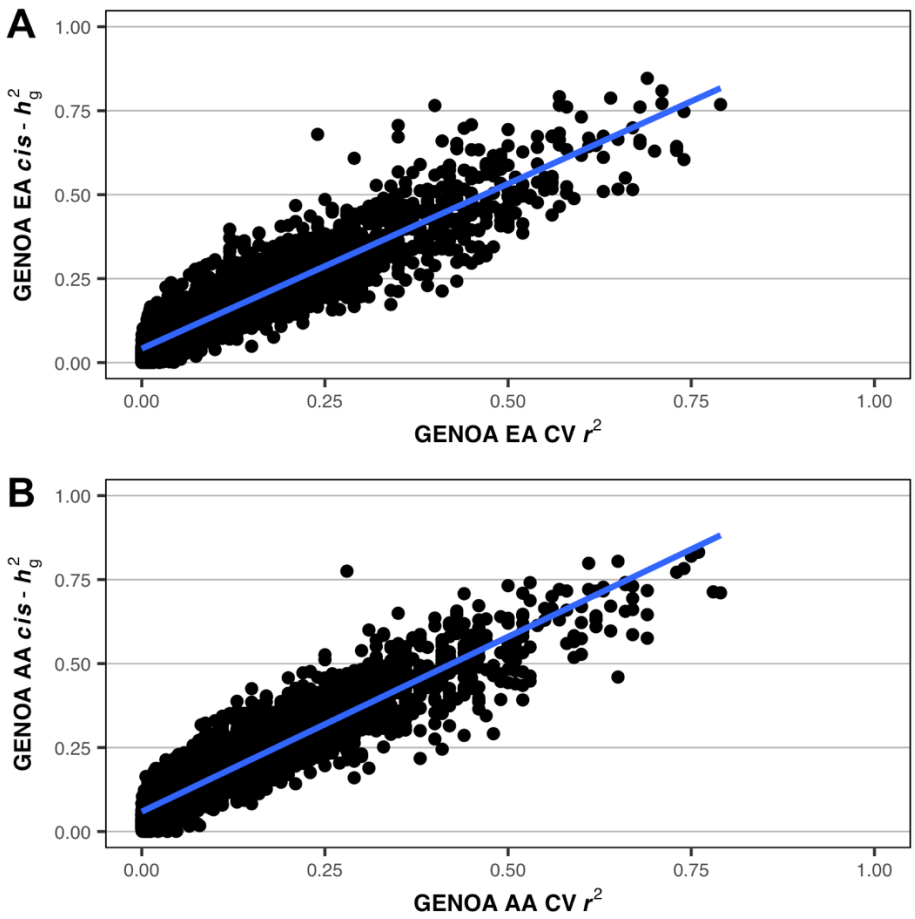


Figure S12. Significant correlation between cross-validation r^2 and heritability *cis*-h_g^2 indicates the better performance of the predictive models in genes with higher expression heritability.

(A) (B) Each point represents a gene. The y-axis is the SNP heritability (*cis*-h_g^2) estimated using GCTA (see **Methods**). The x-axis is the average cross-validation (CV) r^2 of the corresponding fitting model. The blue line is estimated using ordinary linear regression.

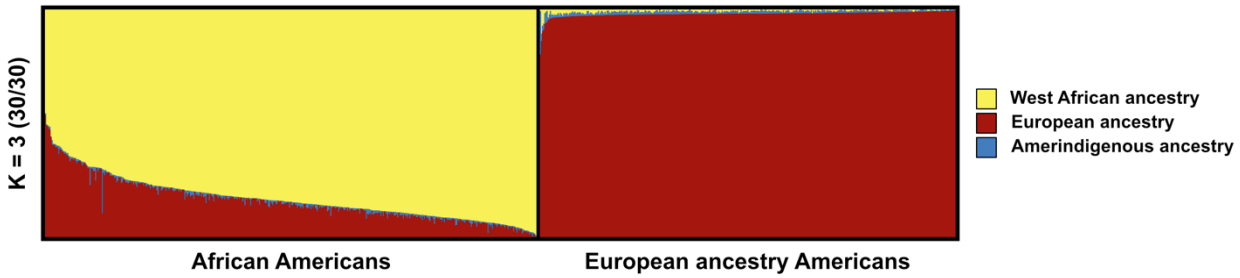


Figure S13. Genome-wide ancestry proportions of GENOA African American and European ancestry individuals in the gene expression panels.

We used 1000 Genomes[7] phase three data from ancestries sampled in Africa, Europe, and the Americas (Table **S3**) to estimate the global ancestry composition of the GENOA individuals comprising our gene expression panels[8] (see **Methods**). Each vertical bar represents one individual's genome modeled as a combination of three possible ancestries. $K = 3$ means individuals can be modeled as a combination of up to 3 ancestries. $30/30$ means 30 out of 30 replicates support this admixture model. We used pong[9] to draw this figure.

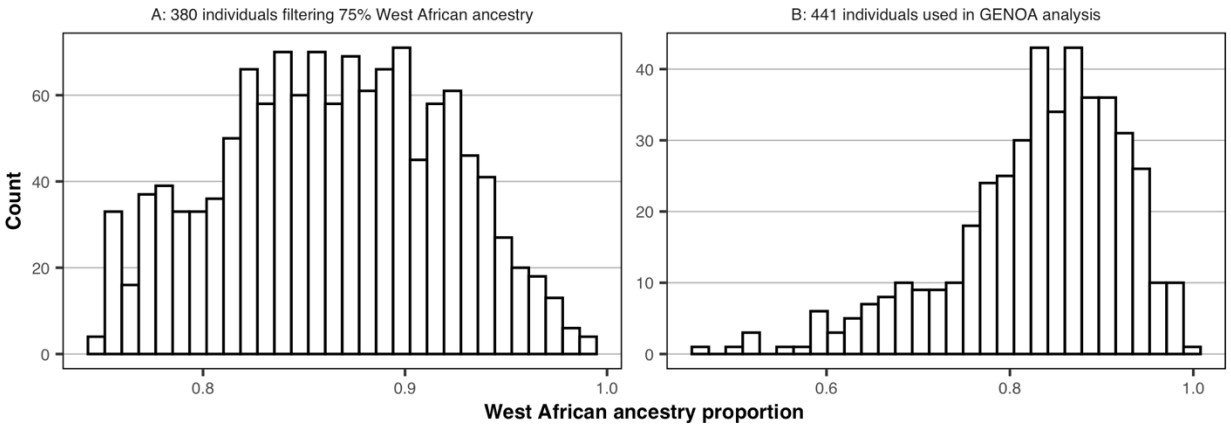


Figure S14. 441 GENOA AA individuals used in LCL prediction models had an average of 83% West African ancestry while first filtering on over 75% West African ancestry resulted in 380 individuals.

(A) (B) Each bar represents the count of individuals with the corresponding proportion of West African ancestry on the x-axis. For (A), the total number of individuals is 380 after first filtering on over 75% of West African ancestry. (B), the total number of individuals is 441 without first filtering on over 75% of West African ancestry, which we used to fit lymphoblastoid cell line (LCL) prediction models.

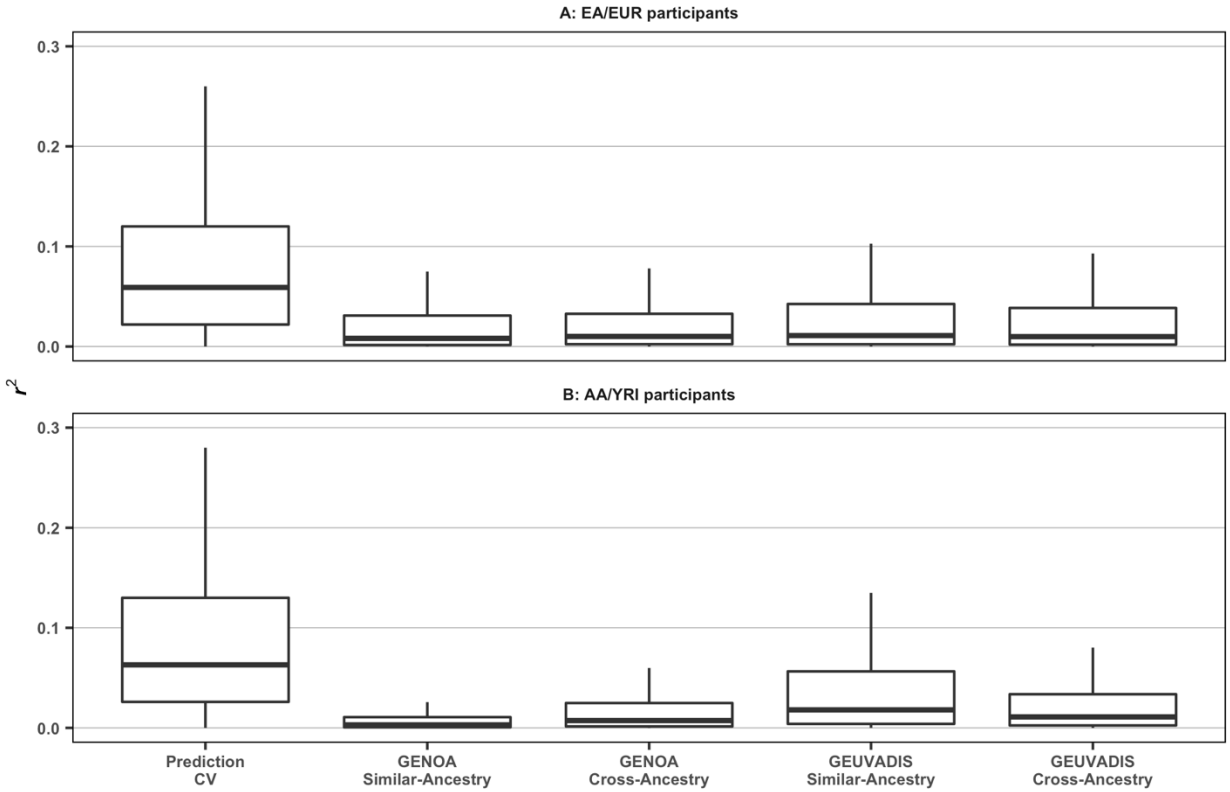


Figure S15. GENOA-based prediction models replicate well in GEUVADIS individuals, and GENOA prediction models top GEUVADIS prediction models in accuracy.

We compared the transportability of genetically predicted gene expression across various settings. The left-most boxplot (Prediction CV) is the in-sample cross-validation r^2 when fitting prediction models in GENOA for European Americans (EA; Subfigure **A**) and African Americans (AA; Subfigure **B**). The next four reflect the distribution of r^2 computed using measured and predicted gene expression levels of EA/EUR (Subfigure **A**) and AA/YRI (Subfigure **B**) when the source of the prediction model changes: [J] [K] -> [L] [M] means using prediction weights of study J ancestry K to predict expression levels of individuals of study L ancestry M. The second is GEUVADIS EUR -> GENOA EA and GEUVADIS YRI -> GENOA AA. The third is GENOA EA -> GENOA AA and GENOA AA -> GENOA EA. The fourth is GENOA EA -> GEUVADIS EUR and GENOA AA -> GEUVADIS YRI. The final is GENOA EA -> GEUVADIS YRI and GENOA AA -> GEUVADIS EUR). EUR represents Europeans, and YRI represents Yoruba in Ibadan.

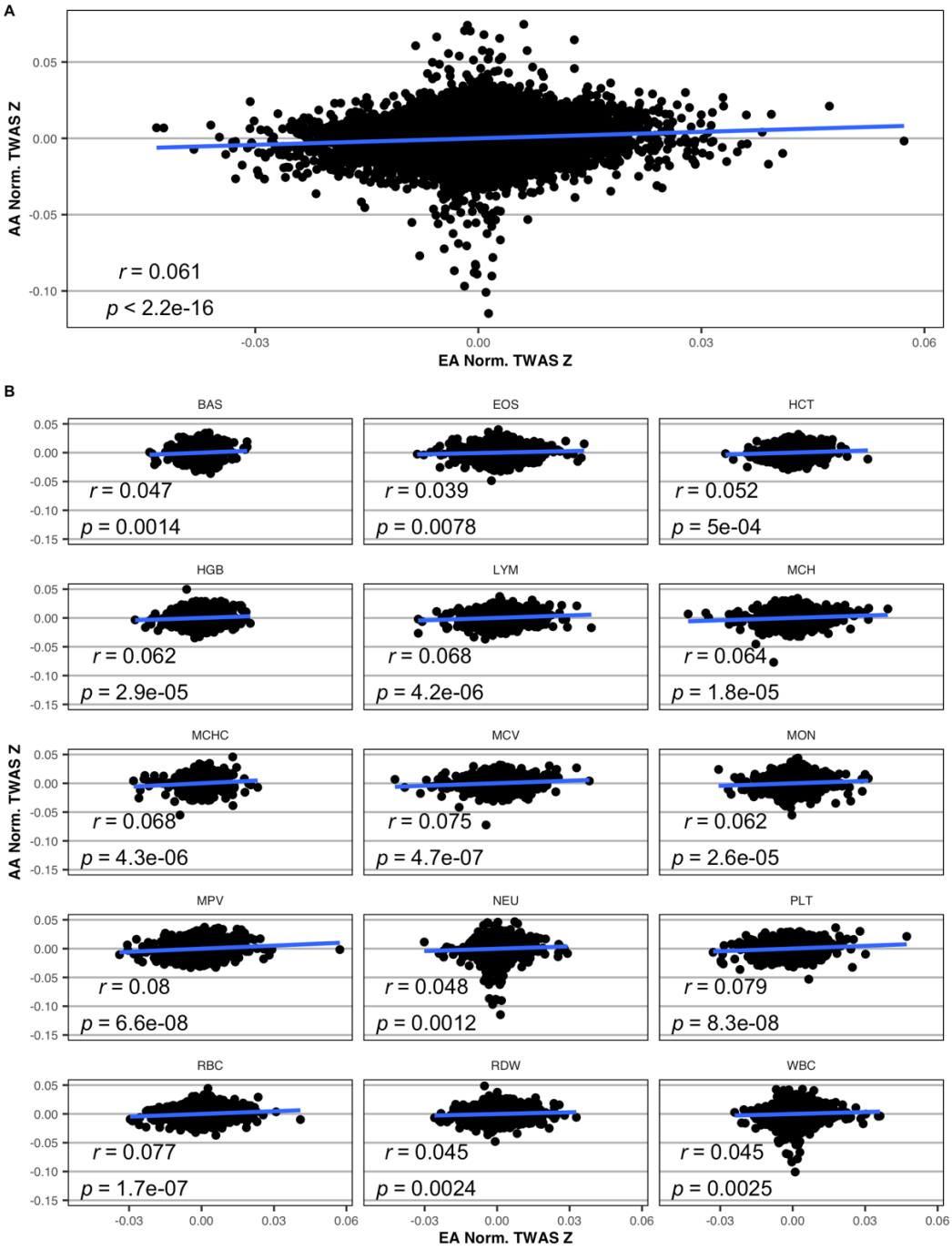


Figure S16. Scatterplot of sample-size normalized TWAS z scores between EA and AA.

(A) (B) Each point represents a gene where the x-axis is European American (EA) normalized Transcriptome-wide Association Study (TWAS) Z-scores, and the y-axis is African American (AA) normalized TWAS Z-scores (see **Methods**). The blue line is a regression slope estimated using ordinary linear regression. Pearson's method is used for the correlation r. (A) is the integration across all blood traits. (B) is trait-specific break-down[10]. See **Table S3** for each trait's full name.

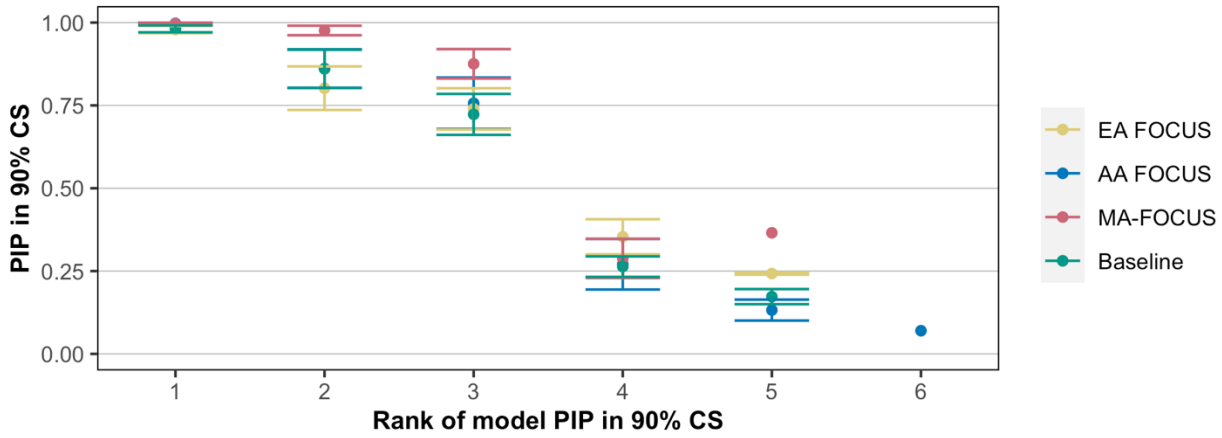
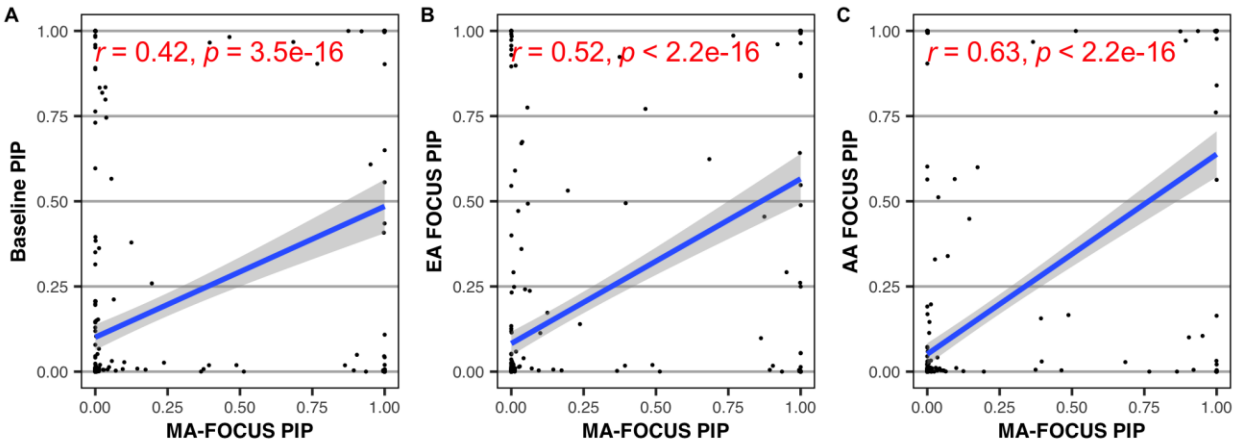


Figure S17. Mean PIPs within ranked genes in credible sets by each method across all traits.

The y-axis is the mean Posterior Inclusion Probability (PIP) of all models (including genes and null model) with the corresponding rank in their respective credible sets (CS) by each method. The x-axis is the rank of the model. Error bars are constructed using a 95% confidence interval. EA represents European Americans, and AA represents African Americans.



S18. PIPs are highly correlated between MA-FOCUS and other approaches.

Each point represents a gene fine-mapped by each method, where the x-axis and y-axis are Posterior Inclusion Probability (PIP) for the corresponding method. The blue line is estimated using ordinary linear regression. The grey band represents the 95% confidence interval. Pearson's method is used for the correlation. EA represents European Americans, and AA represents African Americans.

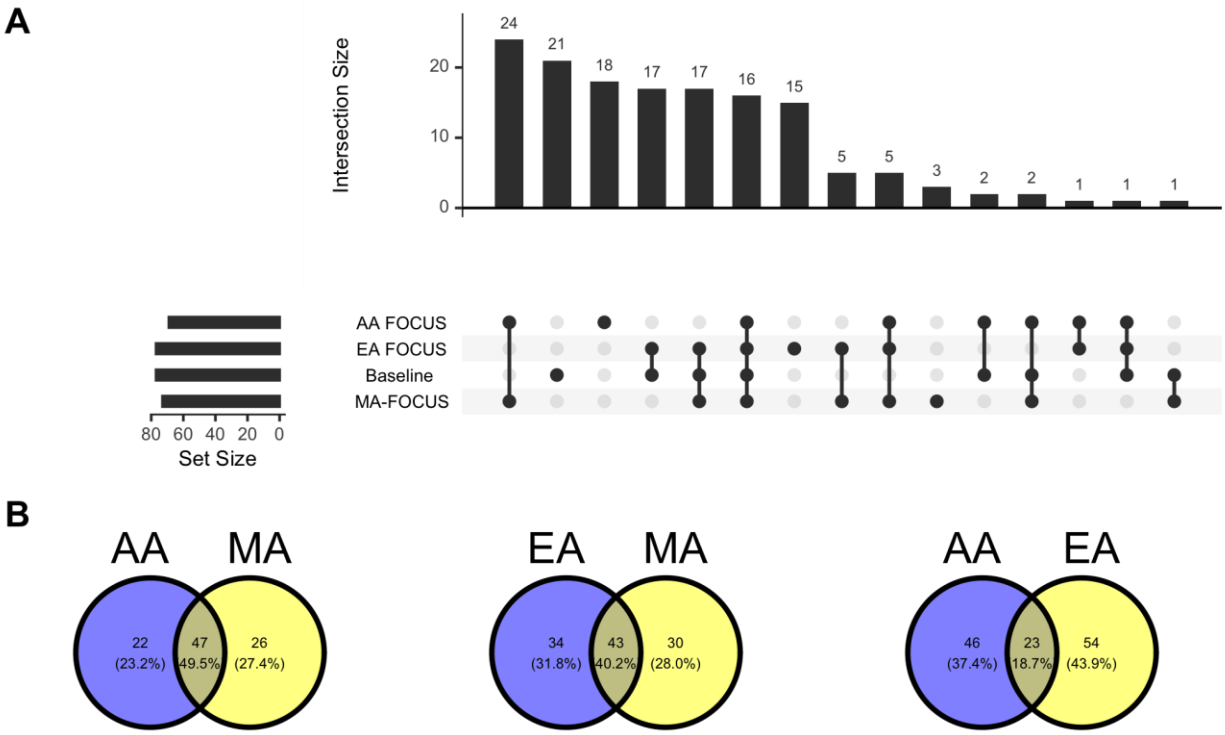


Figure S19. Intersections of credible sets computed by MA-FOCUS, Baseline, and EA FOCUS.

(A) The upset plot is created by the UpsetR package (see **Web Resources**). The upper bar plot indicates the count of genes identified in the lower method combinations. For instance, the first column, where the bar plot shows 24 and the dot plot shows African Americans (AA) FOCUS and MA-FOCUS represents 24 genes are exclusively identified in AA FOCUS and MA-FOCUS only. The left bar plot indicates the aggregation count of the genes identified in the corresponding method in the right. (B) The Venn diagram is created by the ggvenn package (see **Web Resources**). For instance, AA FOCUS identified $22 + 47 = 69$ genes in its credible sets, and 47 are also identified by MA-FOCUS. Both (A) and (B) results are aggregated from 23 fine-mapped regions that contain both European Americans (EA) and AA Transcriptome-wide Association Study (TWAS) signals across 11 blood traits.

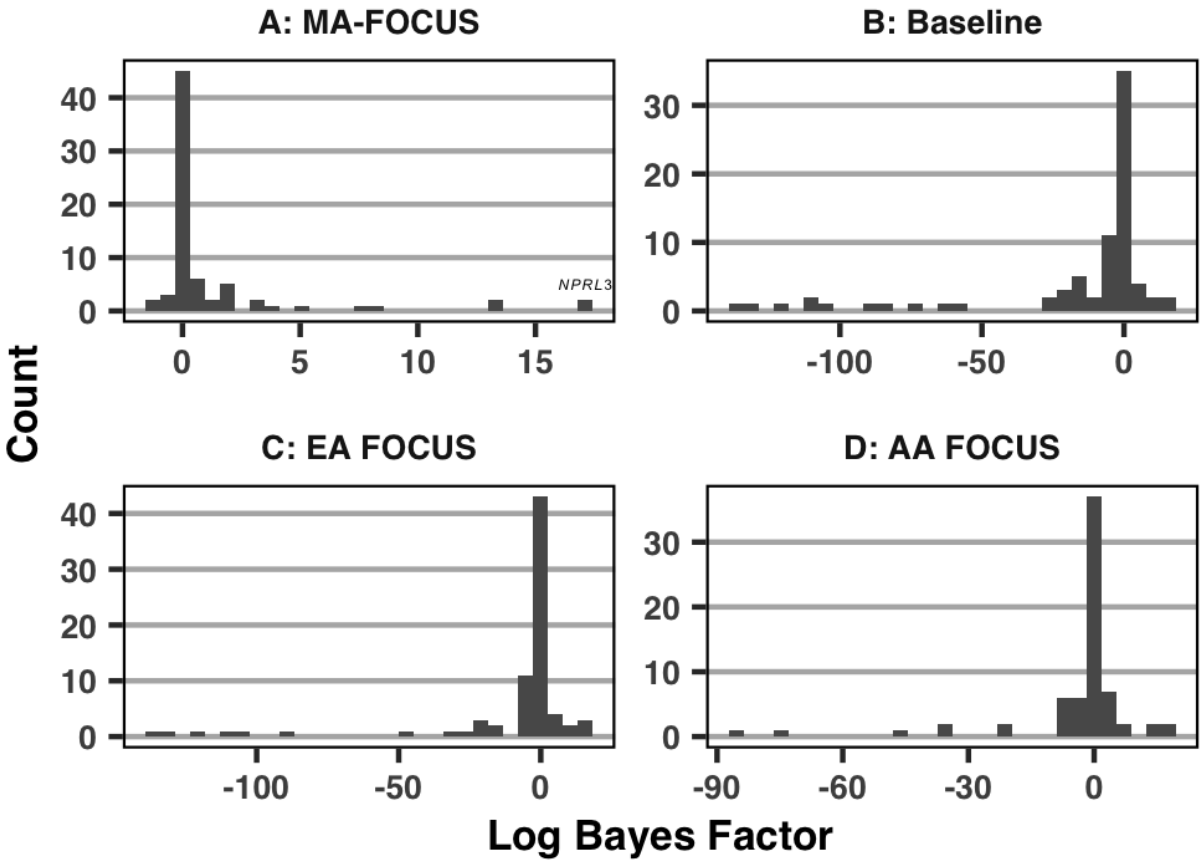


Figure S20. Positive log-scale Bayes factors by MA-FOCUS indicate evidence for shared causal genes across ancestries.

(A) (B) (C) (D) Bar plots for log Bayes factor filtering on genes prioritized by each method. Results are calculated based on 23 regions that contain both European Americans (EA) and African Americans (AA) Transcriptome-wide Association Study (TWAS) signals across all 11 traits (see **Data Availability**). See **Methods** section for how to compute the log-scale Bayes factors.

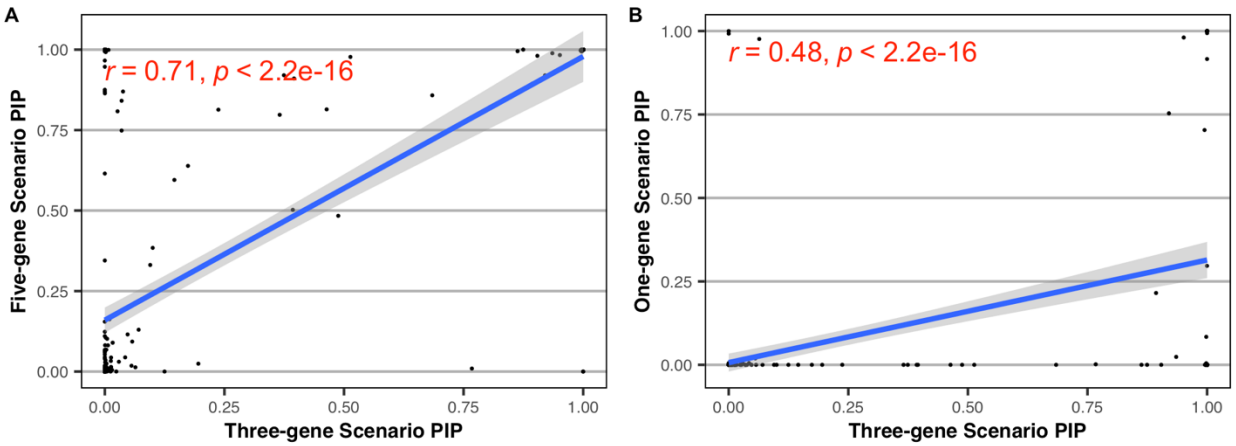


Figure S21. The five-gene scenario correlates more to the three-gene scenario than the one-gene scenario.

(A) (B) Scatter plots for Posterior Inclusion Probabilities (PIPs) between different maximum causal gene scenarios. Each point stands for each gene fine-mapped. The grey band represents the 95% confidence interval. The blue line is estimated using ordinary linear regression. Pearson's method is used for the correlation r.

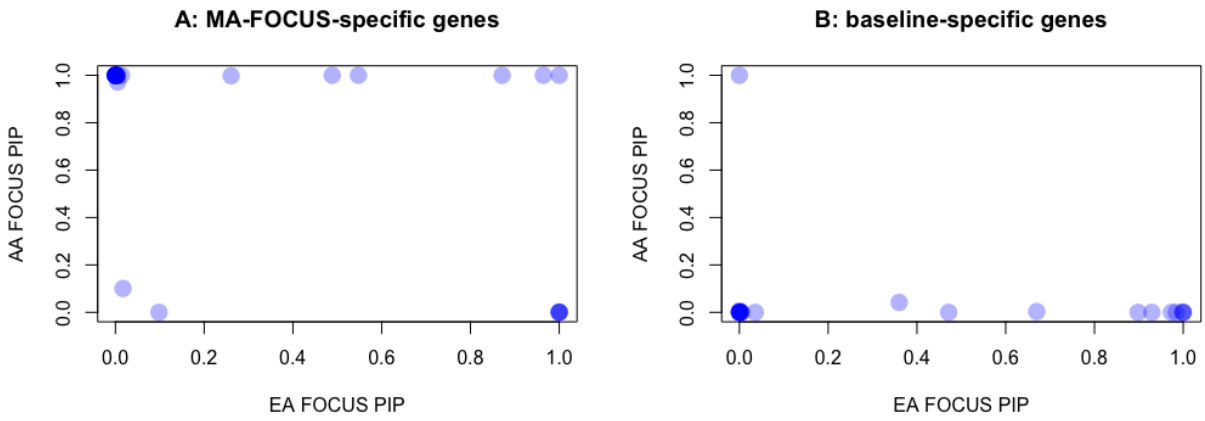


Figure S22. MA-FOCUS-specific genes tend to have higher PIPs in AA- or EA-specific FOCUS fine-mapping than baseline-specific genes.

(A) (B) Scatter plots for MA-FOCUS-specific genes and baseline-specific genes. The x-axis is the Posterior Inclusion Probabilities (PIPs) in European Americans (EA) FOCUS credible sets. The y-axis is the PIPs in African Americans (AA) FOCUS credible sets. The darker blue indicates the overlapping of the genes with similar PIPs.

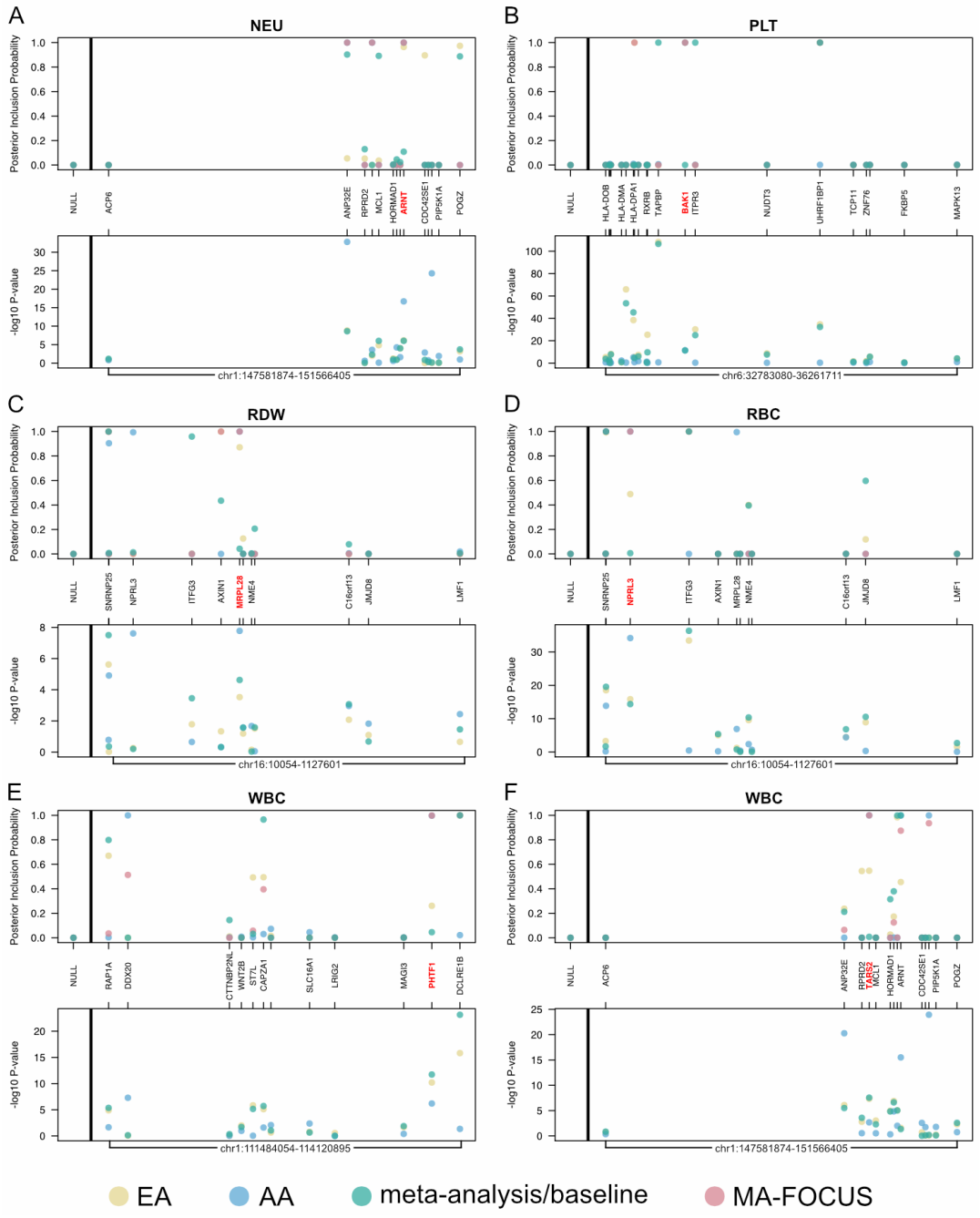


Figure S23. PIPs and TWAS P-values at independent genomic regions with six MA-FOCUS-specific genes.

Each panel highlights a particular gene (red) in its genomic context. For each subfigure, the y-axis represents the Posterior Inclusion Probabilities (PIPs) in the top and the Transcriptome-wide Association Study (TWAS) $-\log_{10} P$ in the bottom, and the x-axis represents the corresponding genomic position. The leftmost section of each top panel shows the PIP for the null model under each method. EA represents European Americans, and AA represents African Americans. See **Table S3** for each trait's full name.

Supplemental References

1. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).
2. Shang, L. *et al.* Genetic Architecture of Gene Expression in European and African Americans: An eQTL Mapping Study in GENOA. *Am. J. Hum. Genet.* **106**, 496–512 (2020).
3. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
4. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
5. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
6. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
7. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
8. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
9. Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P. & Ramachandran, S. Pong: Fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* **32**, 2817–2823 (2016).
10. Chen, M.-H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182**, 1198-1213.e14 (2020).
11. Piñero, J., Ramírez-Anguita, J.M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L.I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. *48*, D845–D855.
12. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics *14*, 128.
13. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90-7.