# Multi-omic strategies for transcriptome-wide prediction and association studies

Supplementary Materials

## 1 Supplementary methods

### 1.1 Asymptotic test of total mediation effect

In DePMA, a distal-eQTL $s$ is tested for its total mediation effect on gene $G$ through $m$ mediators that are local to $s$. Consider the following mediation model for $1 \leq j \leq m$:

$$\begin{aligned} Y_G &= X_s\beta_s + \mathbf{M}\beta_{\mathbf{M}} + \mathbf{X}_C\beta_C + \varepsilon_{Y_G} \\ M_j &= X_s\alpha_{M_j} + \mathbf{X}_C\alpha_{C,j} + \varepsilon_{M_j} \end{aligned} \tag{1}$$

.

Here, we construct the total mediation effect

$$\text{TME} = \alpha_{\mathbf{M}}^T\beta_{\mathbf{M}} = \sum_{i=1}^{m} \alpha_{M_i}\beta_{M_i}.$$

Note that TME is distributed as the product of two multivariate Normal distributions. By the multivariate Delta method[1], we can obtain the standard error for the estimated TME. Let $\boldsymbol{\theta} = (\alpha_{\mathbf{M}}, \beta_{\mathbf{M}})$ and define $f(\boldsymbol{\theta}) = TME = \sum_{i=1}^{m} \alpha_{M_i}\beta_{M_i}$.

The first order partial derivative of $f(\hat{\boldsymbol{\theta}})$ is

$$d_{\hat{\boldsymbol{\theta}}} = \frac{\partial(\sum_{i=1}^{m} \alpha_{M_i}\beta_{M_i})}{\partial\hat{\boldsymbol{\theta}}} = [\beta_{\mathbf{M}}\ \alpha_{\mathbf{M}}]^T.$$

We also obtain the estimated variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$ of $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{\alpha_{\mathbf{M}}} & \hat{\boldsymbol{\Sigma}}_{\alpha_{\mathbf{M}}\beta_{\mathbf{M}}} \\ \hat{\boldsymbol{\Sigma}}_{\alpha_{\mathbf{M}}\beta_{\mathbf{M}}} & \hat{\boldsymbol{\Sigma}}_{\beta_{\mathbf{M}}} \end{bmatrix},$$

where $\hat{\mathbf{\Sigma}}_{\alpha_\mathbf{M}}, \hat{\mathbf{\Sigma}}_{\beta_\mathbf{M}},$ and $\hat{\mathbf{\Sigma}}_{\alpha_\mathbf{M}\beta_\mathbf{M}}$ are the variances and covariance of $\hat{\alpha}_\mathbf{M}$, $\hat{\beta}_\mathbf{M}$, and between $\hat{\alpha}_\mathbf{M}$ and $\hat{\beta}_\mathbf{M}$, respectively. Sobel previously has shown, that with sufficient sample size, $\hat{\mathbf{\Sigma}}_{\alpha_\mathbf{M}\beta_\mathbf{M}} \approx 0$[2,3]. Thus, the standard error of $\hat{\boldsymbol{\theta}}$ is given by

$$\hat{\sigma}^2_{\hat{\boldsymbol{\theta}}} = d^T_{\hat{\boldsymbol{\theta}}} \hat{\mathbf{\Sigma}} d_{\hat{\boldsymbol{\theta}}}.$$

We can then test $H_0 : \text{TME} = 0$ against $H_1 : \text{TME} \neq 0$ with the two-sided Wald-type test with the test statistic $Z = \frac{\alpha^T_\mathbf{M}\beta_\mathbf{M}}{\sqrt{\hat{\sigma}^2_{\hat{\boldsymbol{\theta}}}}}$ and comparing to the null standard Normal distribution.

We illustrate the trade-off between power and computational speed using the asymptotic Sobel test and the permutation speed. Consider the following simulation framework with $m = 5$ mediators, 3 covariates and a sample size of $n \in \{200, 500, 700, 1000\}$ for the model in Equations **??**:

- an $n$-length genotype vector for SNP $s$ is drawn from Binomial$(2, MAF)$, where the minor allele frequency $MAF$ is set at 0.1 in **Supplemental Figure S1**below;
- Under the alternative, we simulated $\beta_X \sim N(0, 1)$, $\beta_\mathbf{M} \sim \mathbf{N}_5(\mathbf{0}, \mathbf{I}_5)$, $\beta_C \sim \mathbf{N}_3(\mathbf{0}, \mathbf{I}_3)$, $\alpha_{M_j}|^{m=5}_{j=1} \sim N(0, 1)$, $\alpha_C \sim \mathbf{N}_5(\mathbf{0}, \mathbf{I}_5)$.
- Under the null, all regression parameters were simulated as in the alternative case. However, we set $\alpha_{M_j} = 0|^m_{j=1}$ and $\beta_\mathbf{M} = \mathbf{0}$.
- Lastly, $\varepsilon_{Y_G} \sim N(0, 1 - h^2)$ and $\varepsilon_{M_j} \sim N(0, 1 - h^2_M)$, where $h^2 = h^2_M = 0.1$ in **Supplemental Figure S1** below.
- We then constructed $Y_G$ and $\mathbf{M}$ using Equations **??**.

We found, that over 10,000 simulations, the permutation test was considerably more powerful, albeit considerably slower. However, in most cases of implementing DePMA, the number of tests of mediation are usually on the order of $10^1$ to $10^2$. We recommend the permutation test in most cases, unless gene $G$ has thousands of identified distal-eQTLs. Parallel implementations have been offered as options in the MOSTWAS package.

### 1.2 Added-last test of distal-SNPs

Here, we propose a test to assess the information added from distal-eSNPs in the weighted burden test beyond what we find from local SNPs. Let $\mathbf{Z}_l$ (an $n_l$-vector) and $\mathbf{Z}_d$ (an $n_d$-vector) be the $Z$-scores local and distal SNPs identified by a MOSTWAS model, with $\mathbf{Z} = [\mathbf{Z}_l\ \mathbf{Z}_d]^T$ (an $n$ vector). The local and distal SNP effects from the MOSTWAS model are represented in $\mathbf{w}_l$ (an $n_l$-vector) and $\mathbf{w}_d$ (an $n_d$-vector), with $\mathbf{w} = [\mathbf{w}_l\ \mathbf{w}_d]^T$ (an $n$ vector). Here, we are interested in testing

$$H_0 : \mathbf{w}_d^T \mathbf{Z}_d | \mathbf{w}_l^T \mathbf{Z}_l = \tilde{Z}_{l,\text{obs}} = 0,$$

where $\tilde{Z}_{l,\text{obs}}$ is the observed weighted $Z$-score from local SNPs.

Under the null distribution, as proposed by Pasaniuc et al and Gusev et al in the Imp-G framework[4,5], we assume that $Z \sim N_n(\mathbf{0}, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_l & \mathbf{\Sigma}_{l,d} \\ \mathbf{\Sigma}_{l,d}^T & \mathbf{\Sigma}_d \end{bmatrix}$$

is the LD matrix for the SNPs, as estimated from the reference panel. $\mathbf{\Sigma}_l$ and $\mathbf{\Sigma}_d$ represent the LD matrices for local and distal SNPs, respectively. The LD matrix between local and distal SNPs $\mathbf{\Sigma}_{l,d}$ can be assumed to be zero, though recent studies have showed long-range LD in the human genome[6,7]. We allow the user to set cross-chromosomal LD to 0, though by default, we estimate LD from the reference panel.

Now, we see that, under this null hypothesis, the joint distribution of $(\tilde{Z}_l, \tilde{Z}_d) = (w_l^T Z_l, w_d^T Z_d)$ is given by:

$$\begin{pmatrix} \tilde{Z}_l \\ \tilde{Z}_d \end{pmatrix} \sim N_2 \left( \mathbf{0}, \begin{bmatrix} w_l^T \mathbf{\Sigma}_l w_l & w_l^T \mathbf{\Sigma}_{l,d} w_d \\ w_d^T \mathbf{\Sigma}_{l,d}^T w_l & w_d^T \mathbf{\Sigma}_d w_d \end{bmatrix} \right).$$

It follows that, given $\tilde{Z}_l = \tilde{Z}_{l,\text{obs}}$,

$$\tilde{Z}_d | \tilde{Z}_l = \tilde{Z}_{l,\text{obs}} \sim N \left( \frac{w_l^T \mathbf{\Sigma}_{l,d} w_d}{w_l^T \mathbf{\Sigma}_l w_l} \tilde{Z}_{l,\text{obs}}, w_d^T \mathbf{\Sigma}_d w_d - \frac{[w_l^T \mathbf{\Sigma}_{l,d} w_d]^2}{w_l^T \mathbf{\Sigma}_l w_l} \right).$$

We can use this null distribution for the one-sided test of $H_0 : \mathbf{w}_d^T \mathbf{Z}_d | \mathbf{w}_l^T \mathbf{Z}_l = \tilde{Z}_{l,\text{obs}} = 0$ against $H_1 : \mathbf{w}_d^T \mathbf{Z}_d | \mathbf{w}_l^T \mathbf{Z}_l = \tilde{Z}_{l,\text{obs}} > 0$. These test is implemented in MOSTWAS as a follow-up to the weighted-burden test.

## 2 Supplemental Tables and Figures

| | TCGA-BRCA | ROS/MAP |
|---|---|---|
| **Local-only** | 0.037 (0.053) | 0.079 (0.119) |
| **MeTWAS** | 0.040 (0.066) | 0.135 (0.099) |
| **DePMA** | 0.383 (0.194) | 0.405 (0.118) |

Table S1: *Comparison of $h^2$ across local-only, MeTWAS, and DePMA predictive models.* The mean and standard deviation of $h^2$ across all genes that are significantly heritable with the genetic loci considered in the design matrix of each predictive model.

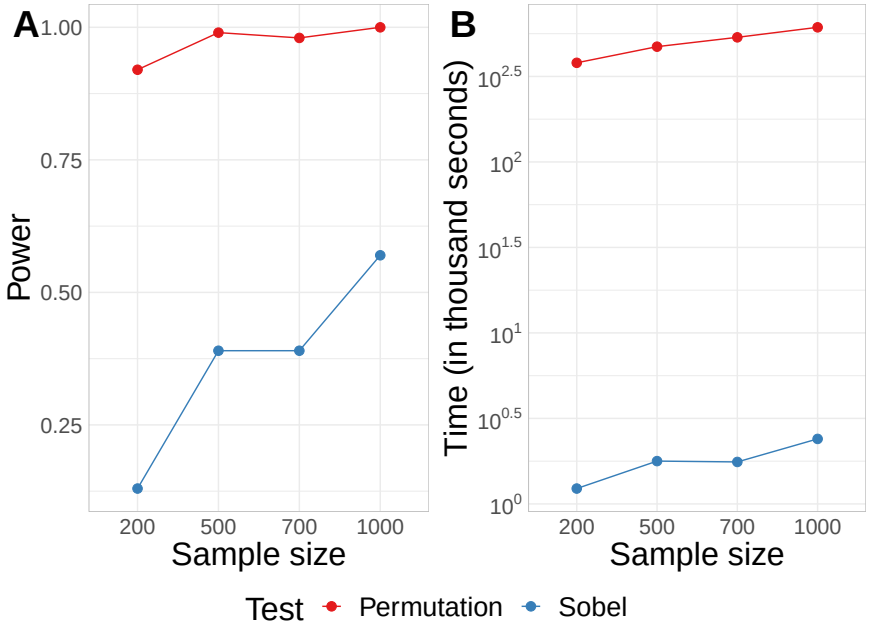


Figure S1: *Comparison of power and computational speed comparison of permutation and Sobel test.* Power (A) and computational speed (B) of permutation test (red) and asymptotic Sobel test (blue) in simulation framework

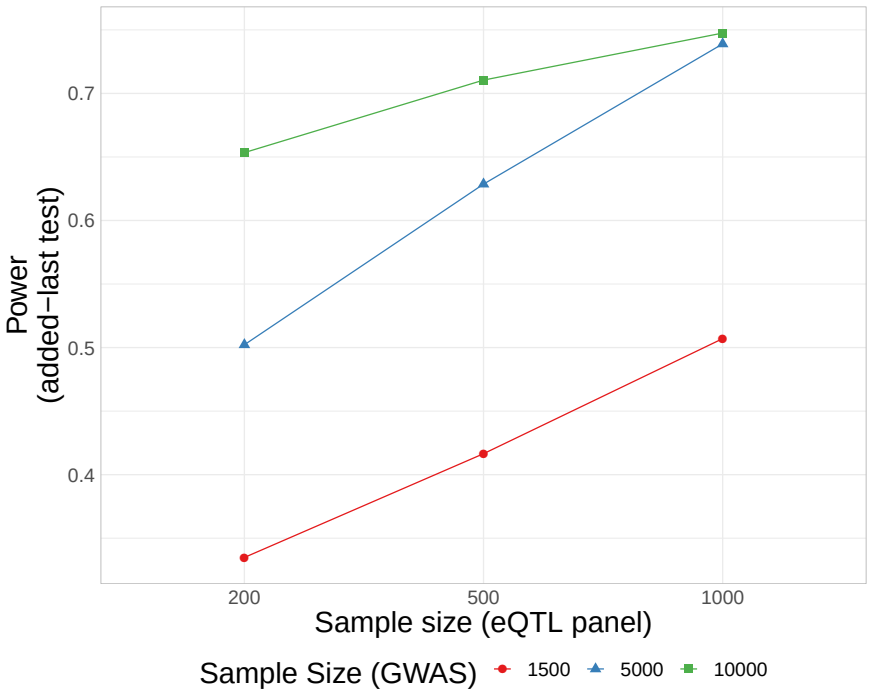


Figure S2: *Simulation analysis for the power of the distal variants added-last test.* Across various sample sizes for the eQTL reference ($X$-axis) panel and GWAS imputation panel (color), the power of the distal added-last test to detect a significant association with distal variants conditional on a significant local association at FDR-adjusted $P < 0.05$.

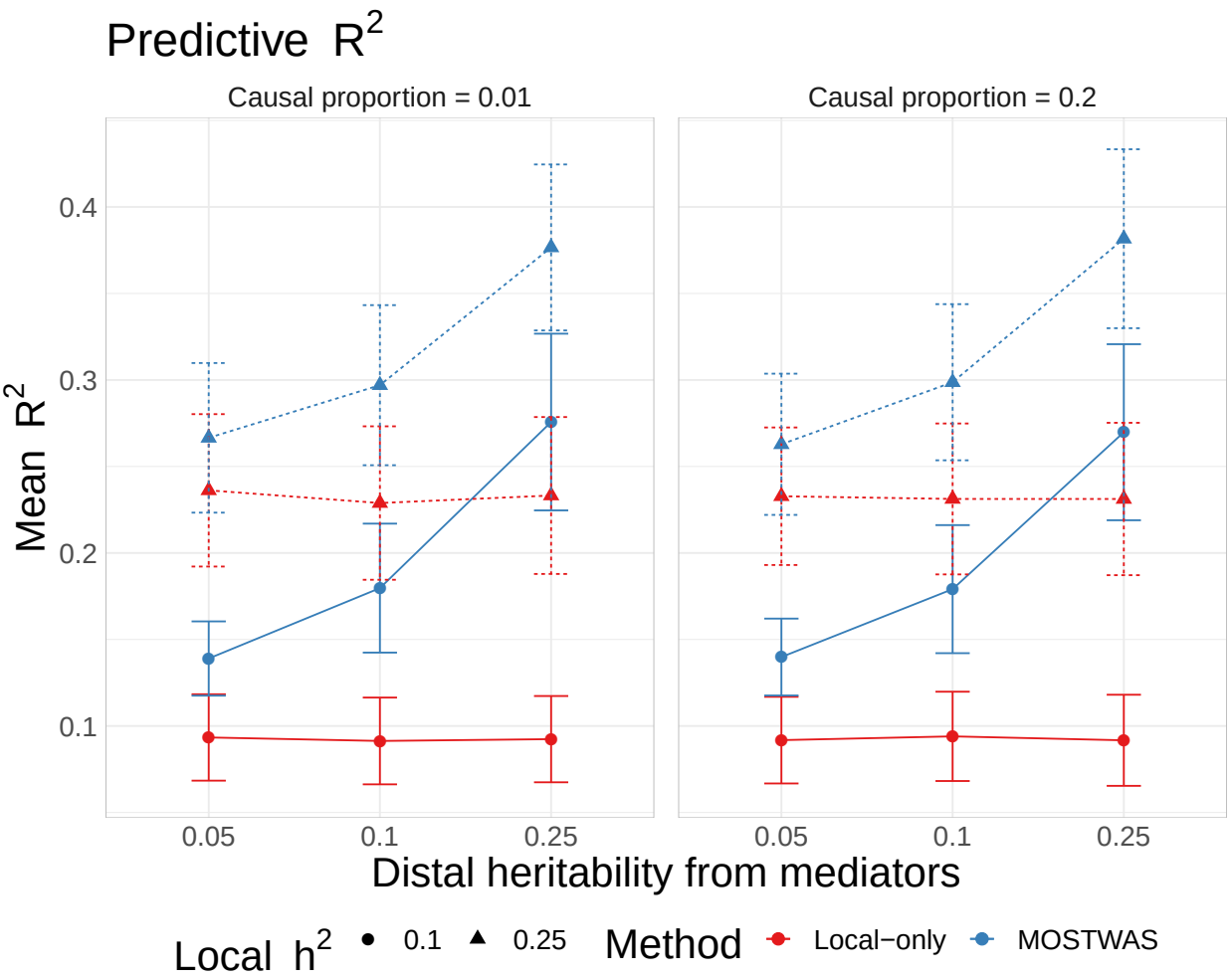


Figure S3: *Comparison of predictive $R^2$ in simulations.* Mean adjusted $R^2$ across various local and distal expression heritabilities, trait heritabilities, and causal proportions using local-only (red) and the best MOSTWAS (blue) models. The error bars reflect a width of 1 standard deviation of the 1,000 simulated adjusted $R^2$ values.

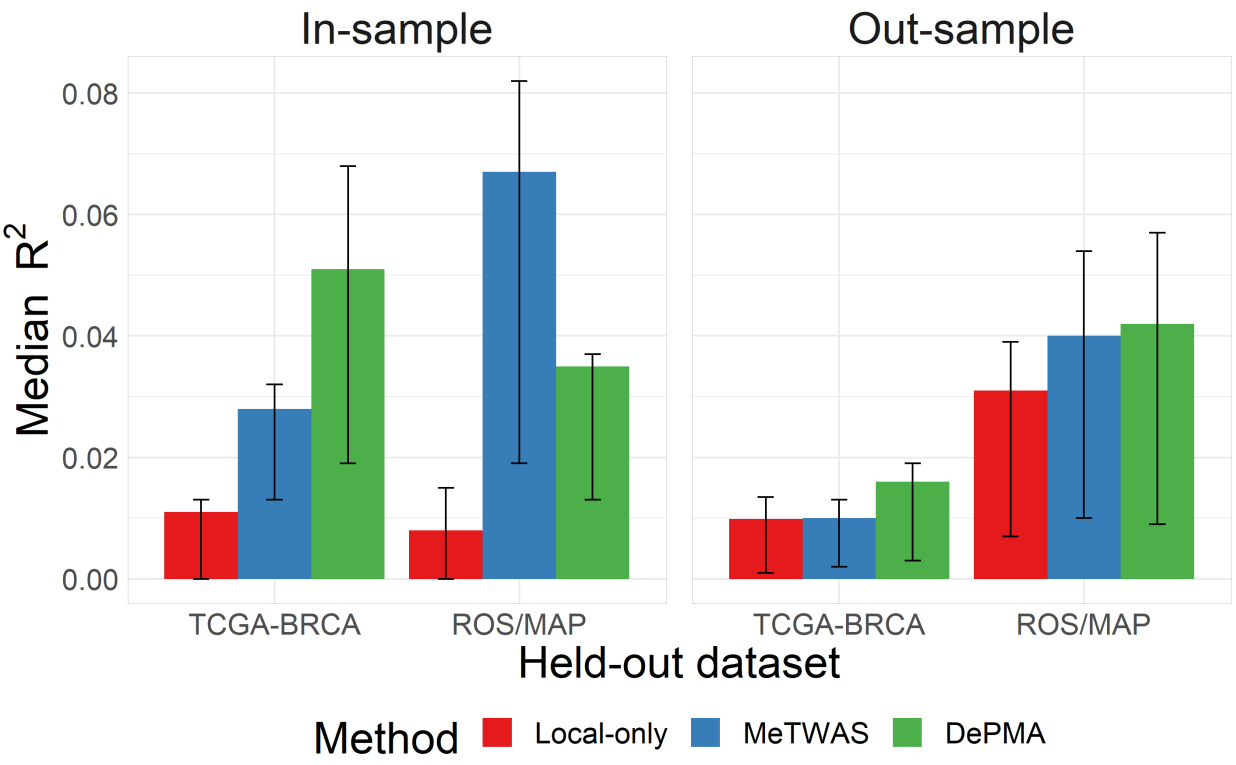


Figure S4: *Comparison of in- and out-sample predictive performance of local-only and MOSTWAS expression models.* Median predictive adjusted $R^2$ for in-sample (left) and out-sample (right) performance in TCGA-BRCA and ROS/MAP expression models using local-only (red), MeTWAS (blue), and DePMA (green) modelling. The interval provided shows the 25% and 75% quartiles. Only genes with significant $h^2$ at raw $P < 0.05$ are shown here.

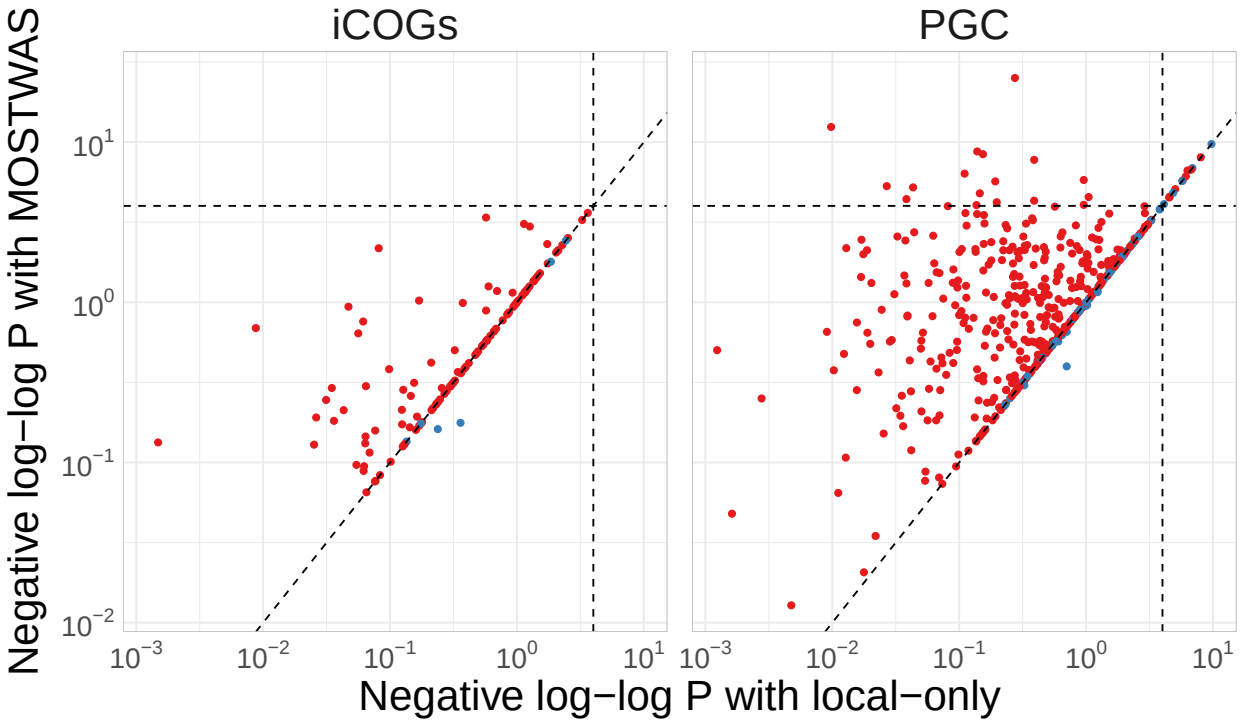


Figure S5: *Gene-trait associations in iCOGs and PGC using local-only and MOSTWAS models.* $-log_{10}P$-values of weighted burden gene-trait associations using iCOGs survival GWAS in European-ancestry women (left) and PGC MDD risk GWAS in predominantly European-ancestry patients (right) among genes that were predicted at cross-validation $R^2 \geq 0.01$ using both local-only and MOSTWAS models. The $X$- and $Y$-axes display the $-log_{10}P$-values for local-only and the best MOSTWAS model, respectively. Note that the scales of both axes are on a doubly logarithmic scale. Points are colored red if $P$-value of association is less than or equal using the MOSTWAS model. The horizontal and vertical reference lines indicate overall Bonferroni-corrected significance thresholds.

| Gene | Cross-validation $R^2$ | iCOGs $Z$-statistic (Added-last $Z$) | Top GWAS SNP location ($P$) | Permutation FDR-adjusted $P$ |
|---|---|---|---|---|
| **C16orf13** | 0.019 | 4.51 (5.18) | Chr3:10720351 (0.13) | 0.03 |
| **C9orf169** | 0.011 | 4.18 (3.97) | Chr19:44949849 (0.04) | 0.03 |
| **CTRL** | 0.051 | 4.51 (3.88) | Chr10:2798136 (0.06) | 0.03 |
| **DNAL4** | 0.013 | -3.94 (-4.57) | Chr22:38681840 (0.01) | 0.04 |
| **LOC221710** | 0.014 | 5.34 (5.08) | Chr1:152983865 ($8.5 \times 10^{-4}$) | 0.03 |
| **MAP3K6** | 0.021 | -4.10 (-4.00) | Chr1:27686314 (0.01) | 0.05 |
| **MAP4K5** | 0.020 | 3.76 (1.26) | Chr14:50502944 ($1.3 \times 10^{-4}$) | 0.03 |
| **NPAT** | 0.115 | -3.92 (-3.72) | Chr20:4217738 (0.02) | 0.04 |
| **RPLP1** | 0.040 | -3.82 (-3.83) | Chr18:1592917 (0.18) | $1.4 \times 10^{-4}$ |
| **SPATA5L1** | 0.042 | 3.76 (No distal SNPs in model) | Chr15:45593323 (0.01) | $1.4 \times 10^{-4}$ |
| **TXNRD2** | 0.047 | 3.91 (4.64) | Chr22:19735425 ($3.5 \times 10^{-3}$) | 0.05 |

Table S2: *Summary statistics for 11 breast cancer-specific survival-associated loci identified by MOSTWAS models.* TWAS associations with breast cancer survival from GWAS statistics from iCOGs with permutation test results and added-last $Z$-statistics. The top iCOGs GWAS SNP in the identified loci with its location and $P$-value are provided.

| Gene | Z-statistic (FDR-adjusted $P$) | Cross-validation $R^2$ | TOP GWAS SNP location ($P$) | Permutation FDR-adjusted $P$ | Added last FDR-adjusted $P$ |
|---|---|---|---|---|---|
| **ABCA7** | -1.82 (0.09) | 0.011 | Chr19:553,066 (0.135) | NA | NA |
| **ADAM10** | -1.25 (0.23) | 0.014 | Chr15:59,052,072 ($1.68 \times 10^{-4}$) | NA | NA |
| **APOE** | 2.82 (0.02) | 0.119 | Chr19:45,545,562 ($3.0 \times 10^{-5}$) | $5.0 \times 10^{-3}$ | 0.03 |
| **BIN1** | 1.91 (0.08) | 0.010 | Chr22:24,199,787 ($8.53 \times 10^{-4}$) | NA | NA |
| **CD2AP** | 1.52 (0.15) | 0.011 | Chr6:47,432,637 ($1.23 \times 10^{-4}$) | NA | NA |
| **CLU** | -2.41 (0.04) | 0.012 | Chr8:27,465,312 ($1.33 \times 10^{-4}$) | 0.83 | 0.44 |
| **FERMT2** | 2.13 (0.06) | 0.017 | Chr14:53,305,626 ($1.38 \times 10^{-4}$) | NA | NA |
| **MEF2C** | 2.20 (0.06) | 0.016 | Chr5:88,359,039 (0.020) | NA | NA |
| **PLCG2** | -2.48 (0.04) | 0.010 | Chr16:81,879,218 (0.037) | 0.66 | 0.07 |
| **SORL1** | 2.91 (0.02) | 0.043 | Chr11:121,446,813 (0.032) | 0.04 | $4.5 \times 10^{-3}$ |
| **ZCWPW1** | -4.56 ($6.1 \times 10^{-5}$) | 0.018 | Chr7:100,435,157 (0.074) | 0.03 | $1.3 \times 10^{-5}$ |

Table S3: *Summary statistics for known Alzheimer's risk-associated loci identified by MOSTWAS models.* TWAS associations (weighted $Z$-score and FDR-adjusted[8] $P$-value) with late-onset Alzheimer's risk from GWAS statistics from IGAP[9]. The top IGAP GWAS SNP in the identified loci with its location and $P$-value are provided. For the 6 loci with significant TWAS associations, the FDR-adjusted $P$-value for the follow-up distal SNP added last test is provided.

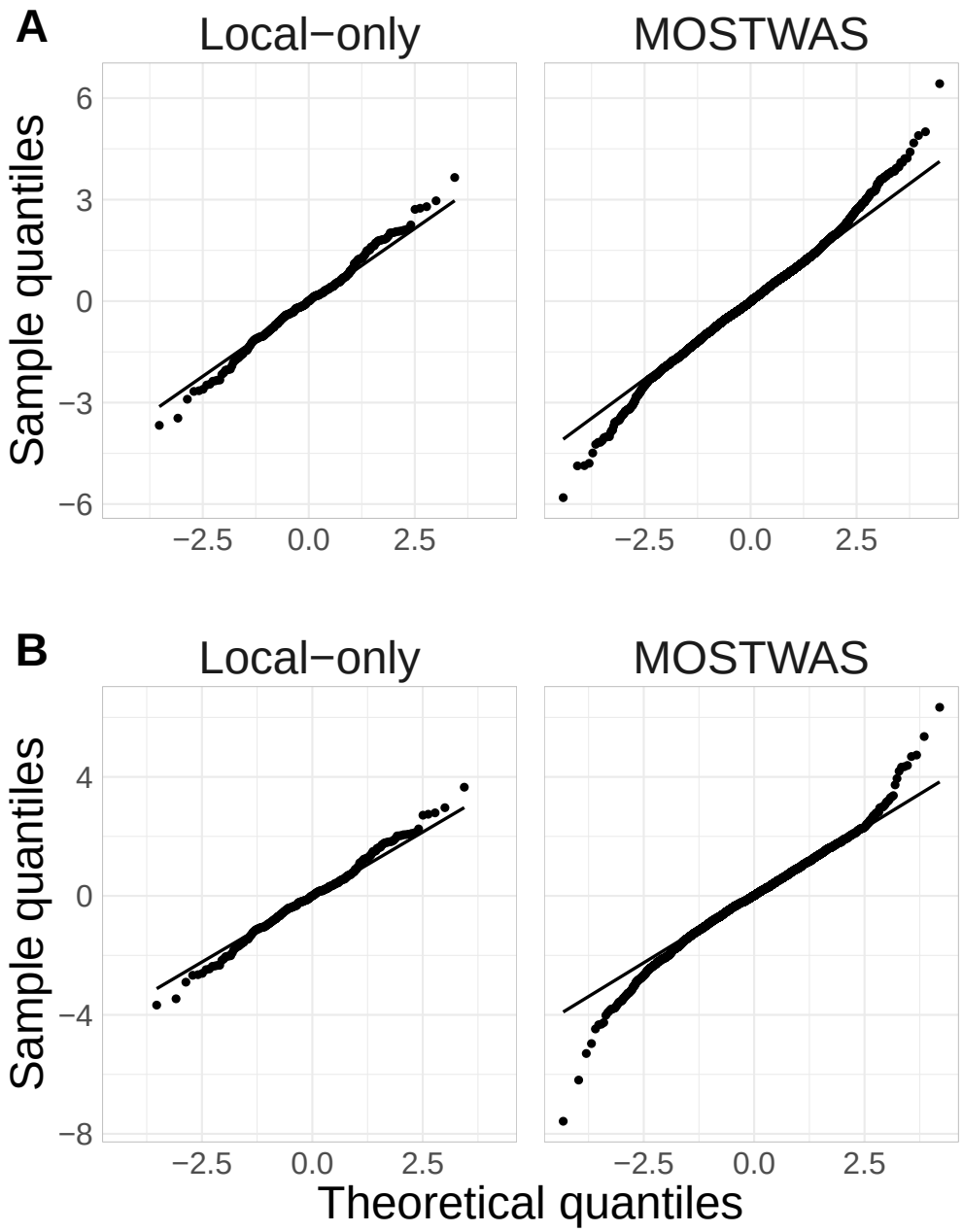


Figure S6: *Comparison of QQ-plots from TWAS associations.* QQ-plots from TWAS for breast cancer-specific survival in iCOGs (A) and MDD in PGC (B) with local-only models (left) and MOSTWAS (right)

| Gene | Cross-validation $R^2$ | PGC $Z$-statistic (UKBB $Z$) | Top GWAS SNP location ($P$) | Permutation FDR-adjusted $P$ |
|---|---|---|---|---|
| **ADAD2** | 0.050 | 5.89 (4.16) | Chr5:35,639,107 ($4.05 \times 10^{-3}$) | $3.5 \times 10^{-5}$ |
| **CACNA2D3** | 0.033 | 3.41 (2.88) | Chr7:12,268,243 ($1.27 \times 10^{-2}$) | 0.046 |
| **FAM43B** | 0.035 | -4.03 (-2.85) | Chr2:73,148,399 ($2.09 \times 10^{-2}$) | 0.028 |
| **MGC29506** | 0.022 | 3.51 (5.54) | Chr5:139,536,922 ($1.48 \times 10^{-3}$) | $3.5 \times 10^{-5}$ |
| **OR8U1** | 0.022 | -3.19 (-4.21) | Chr11:56,676,947 ($4.90 \times 10^{-5}$) | 0.049 |
| **SYT1** | 0.015 | -5.58 (-3.16) | Chr7:12,269,417 ($1.29 \times 10^{-2}$) | 0.040 |
| **YJEFN3** | 0.010 | 5.82 (7.22) | Chr7:12,276,011 ($1.35 \times 10^{-2}$) | 0.038 |

Table S4: *Summary statistics for 7 MDD risk-associated loci identified by MOSTWAS models.* TWAS associations with major depressive disorder from GWAS statistics from Psychiatric Genomics Consortium that were replicated with GWAX summary statistics in UK Biobank. The top PGC GWAS SNP in the identified loci with its location and $P$-value are provided.

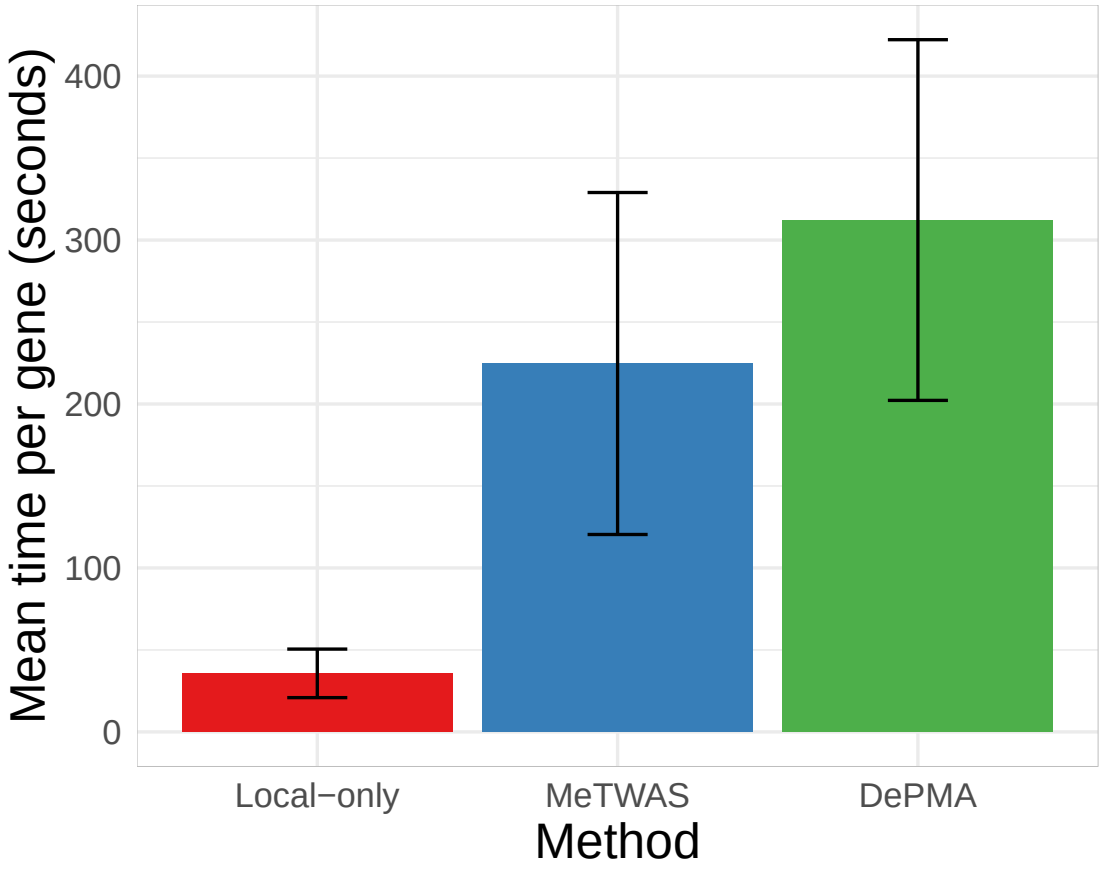


Figure S7: *Comparison of computation times between local-only and MOSTWAS modelling.* Mean and standard deviation of per-gene computation time across 50 randomly selected genes in TCGA-BRCA. Computations here were done with a 24-core, 3.0 GHz processor.

## References


[1] Y. M. Bishop et al. *Discrete Multivariate Analysis: Theory and Practice*. Springer, New York, 1975.

[2] M. E. Sobel. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology*, 13:290, 1982.

[3] M. E. Sobel. Direct and Indirect Effects in Linear Structural Equation Models. *Sociological Methods & Research*, 16(1):155–176, 8 1987.

[4] B. Pasaniuc et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20):2906–2914, 2014.

[5] A. Gusev et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, 3 2016.

[6] E. Koch et al. Long Range Linkage Disequilibrium across the Human Genome. *PLoS ONE*, 8(12):e80754, 12 2013.

[7] L. Park. Population-specific long-range linkage disequilibrium in the human genome and its influence on identifying common disease variants. *Scientific Reports*, 9(1):1–13, 12 2019.

[8] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple. Technical Report 1, 1995.

[9] J. C. Lambert et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45(12):1452–1458, 12 2013.