Supplementary Materials for "Integrative cross-omics and cross-context analysis elucidates molecular links underlying genetic effects on complex traits"

Yihao Lu[1], Meritxell Oliva[1], Brandon L. Pierce[1], Jin Liu[2]*, and Lin S. Chen[1]*

[1]Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA

[2]Centre for Quantitative Medicine, Program in Health Services & Systems Research, Duke-NUS Medical School, Singapore

*Correspondence should be addressed to Jin Liu (jin.liu@duke-nus.edu.sg) and Lin S. Chen (lchen@health.bsd.uchicago.edu)

Contents

S1 Algorithms

In this section, we present the details about the variational EM algorithms for the starting model (4) and the X-ING model (6) in the main text.

S1.1 An EM algorithm with variational inference for the starting model

The starting model (4) does not model shared data patterns, and the prior for $\boldsymbol{\gamma}_{\cdot j,\ell}$ is the same for all tested units in the j-th tissue from data ℓ. Maximizing the complete-data likelihood with respect to all data types is equivalent to maximizing complete-data likelihood for each data type ℓ separately. For each data type ℓ, we derive a computationally efficient EM algorithm with variational inference.

Let $q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell)$ be an approximation of the posterior $\Pr(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell | \boldsymbol{z}_\ell; \Theta_1^{(t)})$. The marginal likelihood can be decomposed as

$$\begin{aligned} \log \Pr(\boldsymbol{z}_\ell; \Theta_1^{(t)}) &= \mathrm{E}_q \log \left(\frac{p(\boldsymbol{z}_\ell, \widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell; \Theta_1^{(t)})}{q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell)} \right) + \mathrm{E}_q \log \left(\frac{q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell)}{p(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell | \boldsymbol{z}_\ell; \Theta_1^{(t)})} \right) \\ &\geq \mathrm{E}_q \log p(\boldsymbol{z}_\ell, \widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell; \Theta_1^{(t)}) - \mathrm{E}_q q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell) \coloneqq \mathcal{L}_{q,\ell}^{(t)}, \end{aligned} \tag{S1}$$

where the superscript indicates the estimates being from the t-th step, $\mathcal{L}_{q,\ell}^{(t)}$ is the evidence lower bound (ELBO), and the inequality holds due to Jensen's inequality. To overcome the computational intractability of the ELBO, we use a mean-field variational family, assuming that $q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell)$ can be factorized as

$$q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell) = \prod_{i=1}^{M} \prod_{j=1}^{K_\ell} q(\widetilde{z}_{ij,\ell}) q(\gamma_{ij,\ell}). \tag{S2}$$

Given variational family of distributions (S2), the optimal variational distribution $q^*(\widetilde{z}_{ij,\ell}, \gamma_{ij,\ell})$ maximizing the ELBO $\mathcal{L}_{q,\ell}^{(t)}$ has the following form:

$$\log q^*(\widetilde{z}_{ij,\ell}, \gamma_{ij,\ell}) = \mathrm{E}_{(i',j') \neq (i,j)} \log \Pr(\boldsymbol{z}_\ell, \widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell) + \text{const}, \tag{S3}$$

where the expectation is taken with respect to variational q distributions related to all other latent variables $(i', j') \neq (i, j)$. Denote the inverse of context-context covariance matrix as $\mathbf{R}_\ell^{-1} = \mathbf{\Lambda}_\ell = \{\lambda_{ij,\ell}\}$. Some derivations show that this optimal q distribution can be written as

$$q^*(\widetilde{z}_{ij,\ell}, \gamma_{ij,\ell}) = \alpha_{ij,\ell}^{\gamma_{ij,\ell}}(1-\alpha_{ij,\ell})^{1-\gamma_{ij,\ell}}\mathcal{N}(\mu_{ij,\ell}, s_{ij,\ell}^2)^{\gamma_{ij,\ell}}\mathcal{N}(0, \sigma_{j,\ell}^2)^{1-\gamma_{ij,\ell}}, \tag{S4}$$

where $\alpha_{ij,\ell}$, $\mu_{ij,\ell}$, and $s_{ij,\ell}^2$ are variational parameters defined as follows

$$\begin{aligned} \alpha_{ij,\ell} &= \frac{1}{1+\exp(-v_{ij,\ell})}, v_{ij,\ell} = \log\frac{\pi_{j,\ell}}{1-\pi_{j,\ell}} + \frac{1}{2}\left(\log\frac{s_{ij,\ell}^2}{\sigma_{j,\ell}^2} + \frac{\mu_{ij,\ell}^2}{s_{ij,\ell}^2}\right), \\ \mu_{ij,\ell} &= \frac{\sum_{j'=1}^{K_\ell}\lambda_{jj',\ell}z_{ij',\ell} - \sum_{j'\neq j}\lambda_{jj',\ell}\alpha_{ij',\ell}\mu_{ij',\ell}}{\lambda_{jj,\ell}+\frac{1}{\sigma_{j,\ell}^2}},\ s_{ij,\ell}^2 = \frac{1}{\lambda_{jj,\ell}+\frac{1}{\sigma_{j,\ell}^2}}. \end{aligned} \tag{S5}$$

In this way, $q^*(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell)$ can be obtained analytically and the ELBO $\mathcal{L}_{q,\ell}^{(t)}$ can be evaluated under the variational distribution q^*. This step can be viewed as a generalized E-step within the variational family of distributions (S2).

In the M-step, by taking partial derivatives of the ELBO $\mathcal{L}_{q,\ell}^{(t)}$ with respect to the model parameters $\Theta_1^{(t)}$ and setting them to zero, we obtain

$$\begin{aligned} \sigma_{j,\ell}^2 &= \frac{\sum_{i=1}^M \alpha_{ij,\ell}(s_{ij,\ell}^2 + \mu_{ij,\ell}^2)}{\sum_{i=1}^M \alpha_{ij,\ell}}, \\ \pi_{j,\ell} &= \frac{\sum_{i=1}^M \alpha_{ij,\ell}}{M}. \end{aligned} \tag{S6}$$

S1.2 An EM algorithm with variational inference for X-ING

To account for the major omics-shared and context-shared patterns in the estimation of association probabilities, we impose a low-rank structure to the latent matrix $\boldsymbol{U}_\ell$ for each data type ℓ that modulates the prior probability of the latent status as function (2) in the main text. Denote $\Theta_2 = \{\boldsymbol{U}_\ell, \boldsymbol{u}_{0,\ell}, \mathbf{R}_\ell, \sigma_{j,\ell}^2, \pi_{ij,\ell}, \ell = 1, \ldots, L, i = 1, \ldots, M, j = 1, \ldots, K_\ell\}$ as the parameter space for the X-ING model (6). Similar to the starting model (4), $\mathbf{R}_\ell$ can be pre-estimated and taken as known. To reduce computational complexity, we pre-estimate the intercepts $\boldsymbol{u}_{0,\ell}$'s using the estimated

$\pi_{j,\ell}$'s from function (S6), in the starting model (4).

By fixing the intercepts $\boldsymbol{u}_{0,\ell}$'s, it is straightforward to show that solving for $\boldsymbol{U}_\ell$'s is equivalent to solving for $\pi_{ij,\ell}$'s. By taking partial derivatives of the ELBO $\mathcal{L}_{q,\ell}^{(t)}$ with respect to $\pi_{ij,\ell}$'s and setting them to zero, we have $\pi_{ij,\ell} = \alpha_{ij,\ell}$. Then, we have

$$U_{ij,\ell} = \log\left(\frac{\pi_{ij,\ell}}{1-\pi_{ij,\ell}}\right) - u_{0j,\ell}. \tag{S7}$$

If no constraints are imposed, there would be an issue of over-parameterization for $\boldsymbol{U}_\ell$'s. Here in the M-step, we apply canonical correlation analysis (CCA) or generalized canonical correlation analysis (GCCA) to the standardized $\boldsymbol{U}_\ell$'s estimated as in formula (S7), where $\boldsymbol{U}_\ell$'s are standardized with mean zero and unit variance. When $L = 2$, GCCA reduces to the problem of CCA between two latent matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$. CCA/GCCA aims to maximize the pair-wise correlation between linear combinations of $\boldsymbol{U}_\ell$'s. Suppose we have rank-p_ℓ ($\leq K_\ell, \forall \ell \in \{1,\ldots,L\}$) approximation for $\boldsymbol{U}_\ell$, the corresponding canonical weight matrices $\boldsymbol{A}_\ell = [\boldsymbol{a}_\ell^1, \ldots, \boldsymbol{a}_\ell^{p_\ell}]$ can be estimated. Then, for each data type ℓ, the estimated low-rank matrix $\boldsymbol{U}_{\ell L} = \boldsymbol{U}_\ell \boldsymbol{A}_\ell \boldsymbol{A}_\ell^\dagger$ has rank$(\boldsymbol{U}_{\ell L}) = p_\ell$, where † refers to the Moore-Penrose pseudo-inverse. When the dimension of multivariate cellular contexts K_ℓ in each omic data is large, one may also use the regularized GCCA (RGCCA). We use R packages *CCA* and *RGCCA*[1] to perform CCA/GCCA/RGCCA analyses.

To further capture omic-specific patterns for each data type, we perform a PCA on the residual matrix from CCA/GCCA/RGCCA calculated as $\boldsymbol{U}_{\ell K} = \boldsymbol{U}_\ell - \boldsymbol{U}_{\ell L}$. Specifically, we first standardize the $\boldsymbol{U}_{\ell K}$'s with mean zero and unit variance. We calculate a truncated Singular Value Decomposition (SVD) and keep the top q_ℓ ($\leq K_\ell, \forall \ell \in \{1,\ldots,L\}$) largest singular values to approximate $\boldsymbol{U}_{\ell K}$. Those approximated matrices $\boldsymbol{U}_{\ell K}$'s capture the association patterns shared among and specific to different cellular contexts (tissues).

Combining omic-shared ($\boldsymbol{U}_{\ell L}$) and omic-specific ($\boldsymbol{U}_{\ell K}$) patterns across multivariate contexts/tissues, we can update the prior specification $\pi_{ij,\ell}$ as function (2). The algorithm for estimating X-ING model is given in Algorithm S1.

Algorithm S1 An EM algorithm with variational inference for the X-ING method

Input data: for $\ell = 1, \ldots, L$, $\boldsymbol{z}_\ell \in \mathbb{R}^{M \times K_\ell}$, $\mathbf{R}_\ell \in \mathbb{R}^{K_\ell \times K_\ell}$, $p_\ell \in \mathbb{Z}^+$, $q_\ell \in \mathbb{Z}^+$

Initialize parameters: $\alpha_{ij,\ell}, \mu_{ij,\ell}, s^2_{ij,\ell}, \sigma^2_{j,\ell}, U_{ij,\ell}$, $\pi_{ij,\ell} = \alpha_{ij,\ell}$, and specify $u_{0j,\ell}$, for $\ell = 1, \ldots, L$, $j = 1, \ldots, M$, $j = 1, \ldots, K_\ell$. This can be either user-specified or obtained by running the EM algorithm for the starting model (4) (by skipping Step 14-20 and setting $\boldsymbol{U}_{\ell L} + \boldsymbol{U}_{\ell K}$ as $\mathbf{0}$ in Step 24-25).

Initialize $\mathcal{L}^{(1)}_{q,\ell} = -\infty$

repeat $t = 2, 3, \ldots$

  E-step:

  for $\ell = 1, \ldots, L$ **do**

    for $i = 1, \ldots, M$ **do**

      for $j = 1, \ldots, K_\ell$ **do**

        $\mu_{ij,\ell} = \frac{\sum_{j'=1}^{K_\ell} \lambda_{jj',\ell} z_{ij',\ell} - \sum_{j' \neq j} \lambda_{jj',\ell} \alpha_{ij',\ell} \mu_{ij',\ell}}{\lambda_{jj,\ell} + \frac{1}{\sigma^2_{j,\ell}}}$,

        $s^2_{ij,\ell} = \frac{1}{\lambda_{jj,\ell} + \frac{1}{\sigma^2_{j,\ell}}}$,

        $v_{ij,\ell} = \log \frac{\pi_{ij,\ell}}{1 - \pi_{ij,\ell}} + \frac{1}{2}\Big(\log \frac{s^2_{ij,\ell}}{\sigma^2_{j,\ell}} + \frac{\mu^2_{ij,\ell}}{s^2_{ij,\ell}} \Big)$,

        $\alpha_{ij,\ell} = \frac{1}{1 + \exp(-v_{ij,\ell})}$.

  M-step:

  for $\ell = 1, \ldots, L$ **do**

    $\boldsymbol{U}_\ell = \Big\{ \log \frac{\alpha_{ij,\ell}}{1 - \alpha_{ij,\ell}} - u_{0j,\ell}, 1 \leq i \leq M, 1 \leq j \leq K_\ell \Big\}$,

  Perform a CCA/GCCA/RGCCA on the L standardized $\boldsymbol{U}_\ell$'s. Note that CCA applies when $L = 2$, GCCA applies when $L > 2$, and RGCCA is recommended when K_ℓ is large.

  for $\ell = 1, \ldots, L$ **do**

    Get the coefficient matrices $\boldsymbol{A}_\ell$'s. Using the top p_ℓ canonical coefficients to get $\boldsymbol{U}_{\ell L}$: $\boldsymbol{U}_{\ell L} = \boldsymbol{U}_\ell \boldsymbol{A}_\ell \boldsymbol{A}_\ell^\dagger$.

    Calculate the residual matrix: $\boldsymbol{U}_{\text{res},\ell} = \boldsymbol{U}_\ell - \boldsymbol{U}_{\ell L}$.

    Perform a PCA on each $\boldsymbol{U}_{\text{res},\ell}$ using SVD and keep the top q_ℓ singular values to get the low-rank approximation matrix $\boldsymbol{U}_{\ell K}$ for each omics data type.

  for $\ell = 1, \ldots, L$ **do**

    for $j = 1, \ldots, K_\ell$ **do**

      $\sigma^2_{j,\ell} = \frac{\sum_{i=1}^M \alpha_{ij,\ell} (s^2_{ij,\ell} + \mu^2_{ij,\ell})}{\sum_{i=1}^M \alpha_{ij,\ell}}$,

      for $i = 1, \ldots, M$ **do**

        $\pi_{ij,\ell} = \frac{1}{1 + \exp\left(-U_{ij,\ell L} - U_{ij,\ell K} - u_{0j,\ell}\right)}$.

until $\mathcal{L}^{(t)}_{q,\ell} - \mathcal{L}^{(t-1)}_{q,\ell} < \varepsilon$, where ε is a user determined threshold.

S2 Simulation Studies

S2.1 Comparisons with competing methods

In this section, we provided additional simulation results to evaluate the selection performance using the receiver operating characteristic (ROC), area under the ROC curve (AUC), and root-mean-square error (RMSE). We first considered sparse data with $\tau_\ell = 0.95$, proportion of phenotypic variation explained for Data 1 and Data 2 as $\theta_1 = \theta_2 = 0.5$, sample size $N_1 = N_2 = 1200$, within-data across-context correlation $\rho_1 = \rho_2 = 0.7$ and between-data correlation $r = 0.2$. X-ING exhibited higher AUC compared with other methods (Supplementary Fig. S1). Supplementary Fig. S2 shows the ROC curves of X-ING and competing methods. With $\rho_1 = \rho_2 = r = 0$, summary statistics across contexts were uncorrelated and there was no shared information across contexts. In this setting, all methods achieved similar performance.

We then compared X-ING with other competing methods using AUC. With the fixed number of contexts in omics Data 1, X-ING gained the improved AUC in Data 1 when the number of contexts in Data 2 increased (Supplementary Fig. S3). We also compared all methods using data with correlated predictors. With different levels of the proportion of phenotypic variation explained, θ_1, X-ING gained the improved AUC in data with unstructured effects and structured effects (Supplementary Fig. S4b). Here, the structured effects refer to non-null true effects correlated across contexts. Finally, we used different proportions of variance explained by simulated predictors for the two groups of contexts. We fixed the proportion as 0.3 for the first seven contexts and varied the proportion from 0.1 to 0.6 for the left three contexts (Supplementary Fig. S4c). The AUC of X-ING was consistently higher than competing methods.

We evaluated the effect estimation of X-ING and mash using RMSE and compared the estimated posterior means of truly non-null effects of the two methods. We varied N_1 from 50 to 1200. With larger sample size, the RMSEs of both X-ING and mash decreased while X-ING had smaller RMSEs (Supplementary Fig. S5).

We compared the computational efficiency of X-ING with other competing meth-

ods. We varied the number of tested units from 1,500 to 25,000, with the number of contexts fixed at 10 in both Data 1 and Data 2. X-ING, MT-eQTL and Metasoft had similar computation time, all less than that of PAINTOR and mash (Supplementary Fig. S6). Moreover, we also varied the number of contexts from 3 to 50, with the number of tested units fixed at 2,500 in both Data 1 and Data 2. The computation time of X-ING was the least among all methods and increased linearly with the number of contexts (Supplementary Fig. S7).

S2.2 Sensitivity Analysis

We performed sensitivity analyses to evaluate the robustness of X-ING. We applied X-ING to data with correlated predictors and studied the impact of prior selection, the choice of the number of principal components (PCs) and canonical coefficients (CCs) on the performance of X-ING. First, we evaluated the performance of X-ING on data with correlated predictors. We divided the predictors into two groups. Predictors in each group were correlated with autoregressive coefficients ϕ_1 and ϕ_2. We varied ϕ_1 and ϕ_2 from 0.7 to 0.8 and from 0.1 to 0.3, respectively. We changed the proportion of predictors in the first group from 0 to 0.3. When the correlation between predictors increased, the performance of X-ING only decreased slightly (Supplementary Fig. S8). Second, we varied the choice of prior for fitting X-ING from 10^{-4} to 0.4. We considered two types of effects: unstructured effects with the true effects generated independently across contexts; structured effects with the non-null true effects correlated across contexts. The performance of X-ING was robust to different choices of priors for both structured and unstructured effects (Supplementary Fig. S9). Third, we evaluated the performance of X-ING in Data 1 generated with different ρ_1's and the number of PCs and CCs used for fitting X-ING ranging from 2 to 8. X-ING was robust to the number of PCs and CCs for both structured and unstructured effects (Supplementary Fig. S10-S11).

S3 Additional results

As shown in Supplementary Fig. S12, only a small proportion of risk SNPs used in multi-omics analysis were in linkage disequilibrium (LD). In the analysis of cis-eQTLs and cis-mQTLs, we observed that many cis-eQTLs and cis-mQTLs had effects shared across tissue types (Supplementary Fig. S13). For the 80 selected diseases/traits, at the 80% posterior probability cutoff, there were 644 to 15,490 SNP-gene-CpG site trios out of the examined disease/trait-specific trios with identified trans-eQTL effects in at least one out of the 28 examined eQTL tissues, or having non-zero trans-mQTL effects in at least one out of the nine examined mQTL tissues

S3.1 Disease-specific trans-e/mQTL hotspots explain more phenotypic variation than trait-associated ones

We further studied trans-QTLs with regulatory/association effects on multiple (≥ 5) genes/CpG sites, i.e., trans-e/mQTL hotspots, to examine their association patterns and contributions to disease/trait heritability. For each disease/trait, we first estimated the SNP-based heritability based on all SNPs, denoted as h^2, using LD score regression[2]. We used genotype data from Caucasian samples in the 1000 Genomes Project as the reference data. Similarly, we re-evaluated the SNP-based heritability, h^2_{r}, after removing T identified trans-QTL hotspots and their neighboring SNPs (within ± 1 MB). Then the percentage of change in heritability per hotspot region was evaluated as

$$\frac{h^2 - h^2_{\mathrm{r}}}{h^2} \cdot \frac{1}{T} \times 100\%.$$

Supplementary Fig. S14 shows the violin plots for the percentage of change in heritability per hotspot region. The average percentage of change in heritability per trans-eQTL hotspot region was 1.82% for the 31 examined diseases and 0.84% for the 21 examined traits, and the corresponding average percentage of change attributed to trans-mQTL hotspot regions was 0.73% and 0.36% for diseases and traits, respec-

tively. Disease-associated hotspot regions explained more phenotypic variation comparing with trait-associated ones, consistent with their relative higher contributions to expression difference and their higher fitness burdens[3].

S3.2 Tissue-sharing patterns of trans-association and cis-mediated trans-association effects

We selected trans-eQTLs that were also cis-eQTLs in at least one tissue to form 7,479 trios of SNP, cis-gene and trans-gene. Similarly, we formed 13,952 trios of SNP, cis-CpG site and trans-CpG site. We then quantified the indirect trans-eQTL effect through cis-genes and trans-mQTL effect through cis-CpG sites. For each trio of SNP, cis-gene and trans-gene, we estimated and tested the indirect mediation effect by regressing the trans-gene expression levels on the cis-gene expression levels adjusting for the cis-eQTL and other covariates. The indirect trans-mQTL effect through cis-CpG site can be estimated similarly. We then calculated the percentage of reduction in trans-effects by

$$\frac{\beta_{\text{total}} - \beta_{\text{direct}}}{\beta_{\text{total}}} \times 100\%,$$

where β_{total} is the marginal trans-effect of the e/mQTL on trans-gene/CpG, respectively, and β_{direct} is the trans-effect after adjusting for putative cis-mediator.

The above percentage of reduction in trans-effect could also be considered as the ratio of indirect effect (mediated via cis) to total effect. For trans-associations that are truly mediated, we expect positive reductions in trans-effects. A negative percentage of reduction in the trans-effect for a trio with a significant mediation P-value may imply a false discovery. For the 149 trios identified as having single-tissue trans-eQTL effects with significant mediation effect (FDR<0.05), we observed 23 (15.4%) trios had negative percentage of reduction in trans-effects (Supplementary Fig. S15). For the 226 trios identified as having trans-eQTL effects in at least two tissues, only 5 (2.2%) trios showed negative reductions. We observed a similar pattern for trans-mQTLs

(Supplementary Fig. S16). After accounting for cis-CpG site, 41 (24.8%) significant trios identified as having single-tissue trans-mQTL effects while 4 (3.1%) significant trios identified as having multi-tissue trans-mQTL effects showed negative reductions. This implies that mediation effects identified in trios from multi-tissue trans-e/mQTLs are less likely to be false discoveries.

S3.3 Integrating spatial transcriptomic data with multi-tissue eQTLs reveals spatially-defined molecular links underlying SCZ genetics

To study the molecular mechanisms underlying SCZ genetics, the CommonMind Consortium (CMC)[4] examined the bulk tissue gene expression variations implicated by SCZ risk loci, and found only a small number of SCZ-risk associated genes showed differential total expression levels between SCZ cases and controls. Besides the limiting power issue, the inadequate mechanistic findings could also be attributed to the heterogeneous nature of diverse cell types from bulk brain tissue data collected by CMC. A study of single-nuclei RNA sequencing (snRNA-seq) data[5] supported the biologically distinct roles of different cell types in SCZ etiology. Recently, a study of spatial transcriptomic data[6] examined the spatial expression variation from six-layered (and white matter) dorsolateral prefrontal cortex (DLPFC) of 12 adult human brain tissue slices, and reported enrichment of spatially differential expression variation for pre-defined gene sets proximal to SCZ GWAS loci.

We jointly analyzed over 1.6 million SNP-gene pairs matched in the spatial data from LIBD and the GTEx data, with a focus on 8,962 SNP-gene pairs involving 527 SCZ risk loci and 3,184 genes in cis (1 MB) with a SCZ loci. At the 90% probability cutoff, 229 genes were associated with SCZ risk loci in at least two GTEx brain tissues, and those genes showed a significantly higher enrichment in layer 2 (L2; $P = 0.026$), layer 5 (L5; $P = 0.025$) and white matter (WM; $P = 0.070$) (Fig. 5) based on one-sided Student's t-test. To study L2- and L5-specific differentially expressed genes,

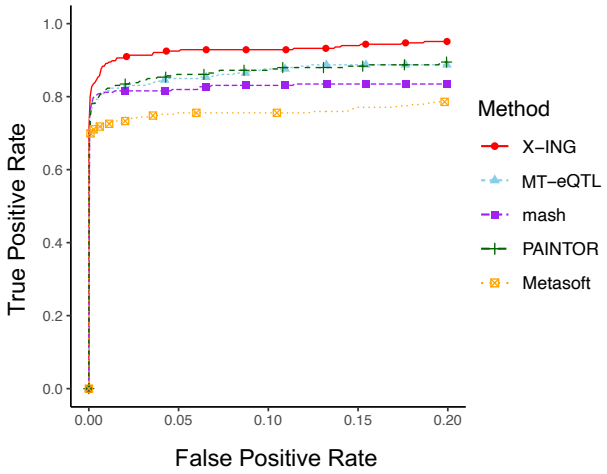
we examined their differential expression patterns in bulk tissues of SCZ cases versus controls from CMC. Restricting to Caucasian donors with SCZ ($N = 209$) and control subjects ($N = 206$), there were 1,679 genes showing significant case-versus-control differential total expression levels from bulk tissues (FDR<0.05) out of 15,437 genes (10.9%)[4]. We found a higher proportion of differentially expressed genes in CMC bulk tissues among the L2-specific differentially expressed genes (19.75%; $P = 0.006$) and the L5-specific differentially expressed genes (16.2%; $P = 0.089$). In other words, many spatially-expressed genes within cis regions of and associated with SCZ loci also showed differential expressions in bulk-tissue data of CMC study.

Among the 229 genes associated with SCZ loci in at least two GTEx brain tissues, X-ING identified a total of 100 genes with differential expression in L2, L5 or WM for at least one sample, and 55 of them were specific to only one layer. The L2-specific differentially expressed gene, *SNX19*, was previously identified as a putative causal gene for SCZ by a summary data-based Mendelian randomization analysis[7], and was reported to have higher expression levels in SCZ cases than controls[8]. X-ING also identified the gene, *NEK4*, as an L5-specific differentially expressed gene. The *NEK4* gene was reported to be associated with the risk of SCZ, intelligence and brain volume[9], and was involved in the formation of abnormal brain structure and cognitive performance that are related to SCZ. The gene, *MEF2C*, was identified as a differentially expressed gene in WM. Many mutations of SCZ cases occur in or near *MEF2C* region, which is active in excitatory and inhibitory neurons[10]. Abnormal function of *MEF2C* in brain may increase the risk of psychiatric disorders.

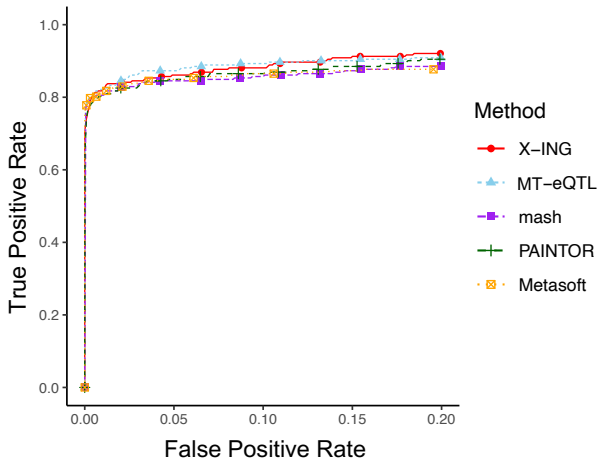
In addition to SCZ, in Supplementary Fig. S17, we also showed the heatmap of the enrichment of layer-specific differentially expressed genes among Autism spectrum disorder (ASD) risk-associated genes. Genes associated with ASD risk loci in at least two GTEx brain tissues showed a significantly higher enrichment in L2 ($P = 0.009$), L5 ($P = 0.030$), L6 ($P = 0.028$) and WM ($P = 0.020$). Our result is consistent with existing studies[6] and uniquely identified WM being enriched with differentially expressed genes. The tissue sample sizes for eQTL analyses are given in Supplementary Ta-

ble S1-S2. The list of genes associated with SCZ/ASD and layer-specific differentially expressed is given in Supplementary Table S4-S5.

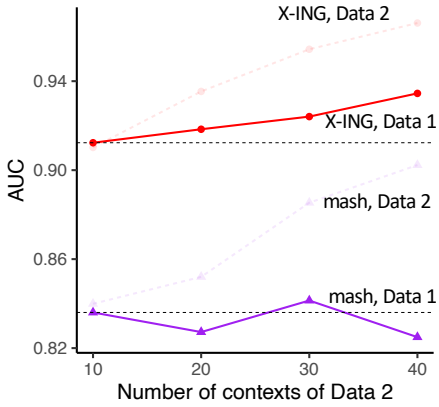
Our analysis highlights the importance of accounting for spatially defined expression variation when studying the mechanisms underlying complex diseases/traits with high context-specific expression. X-ING gains power in identifying individual genes with laminar variation and context-specific links to disease risk loci when integrating spatial transcriptomic data with multi-tissue eQTL data.



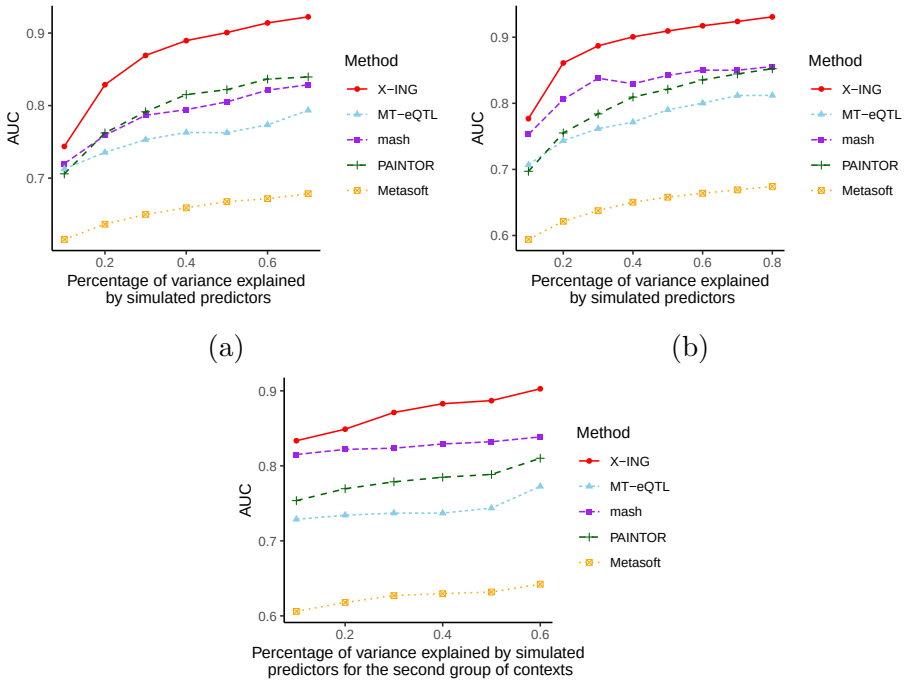
Supplementary Figure S1: ROC curves for detecting non-zero associations in sparse data, with $\tau_\ell = 0.95$. Here $N_1 = N_2 = 1200$, $\rho_1 = \rho_2 = 0.7$, $r = 0.2$ and the proportion of phenotypic variation explained by predictors for each data type was 0.5.



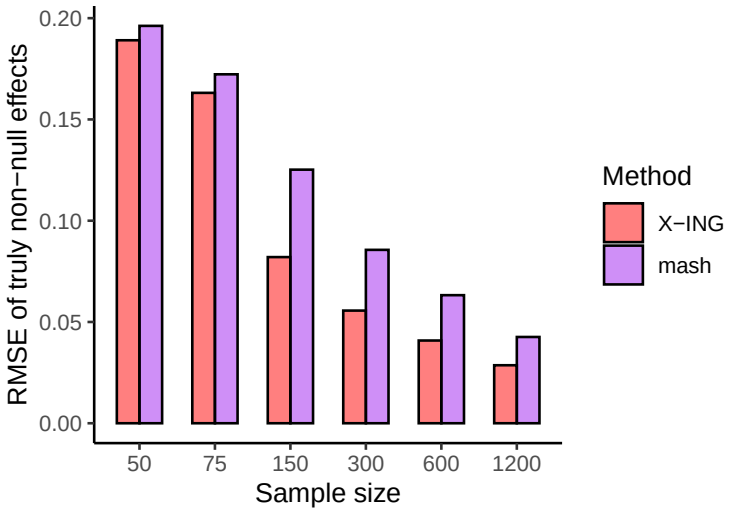
Supplementary Figure S2: ROC curves for detecting non-zero associations when input summary statistics are from independent contexts ($\rho_1 = \rho_2 = r = 0$). Here $N_1 = N_2 = 1200$ and the proportion of phenotypic variation explained by predictors is 0.5.



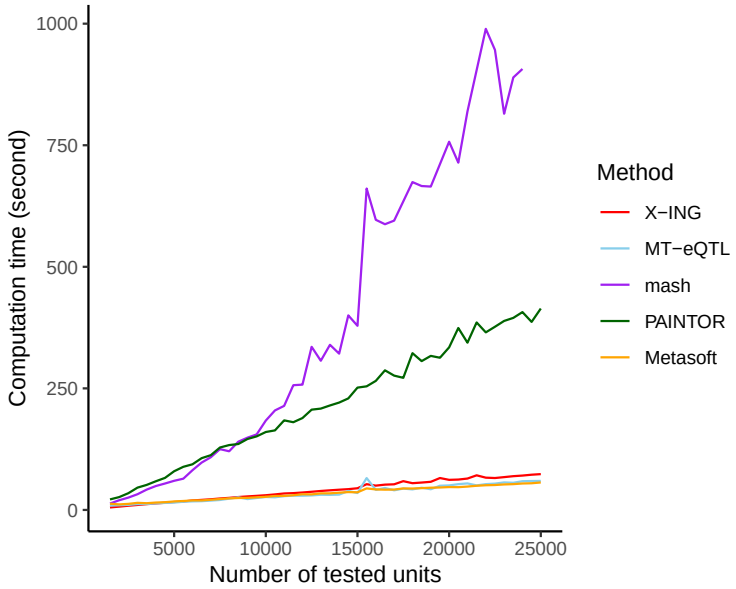
Supplementary Figure S3: Comparison of AUC between X-ING and mash on Data 1 and 2 with varying number of contexts for omics Data 2. K_2 varied from 10 to 40.



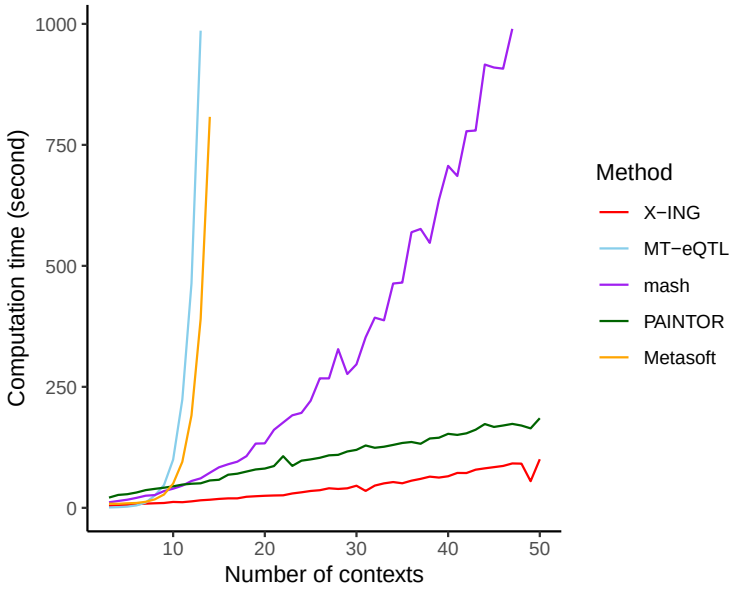
Supplementary Figure S4: Comparison of methods on simulated data. (a) AUC on Data 1 with unstructured effects. The proportion of variation explained by predictors θ_1 varied from 0.1 to 0.8. The simulated true effects were unstructured, i.e., true effects were independently generated. (b) AUC on Data 1 with structured effects. The proportion of phenotypic variation explained by predictors θ_1 varied from 0.1 to 0.8. The simulated true effects were structured, i.e., correlated for those with truely non-null association. (c) AUC on Data 1 with unstructured effects. The proportion of phenotypic variation explained by predictors was fixed as 0.3 for the first 7 contexts while that of the left 3 contexts ranged from 0.1 to 0.8.



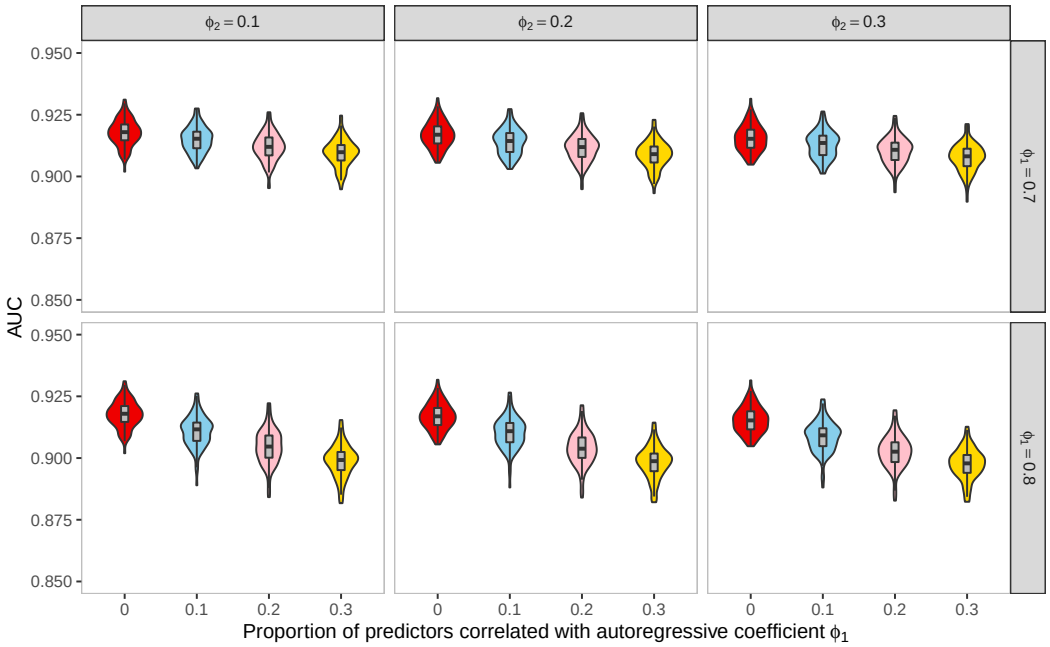
Supplementary Figure S5: Comparison of RMSEs for posterior means estimated by X-ING and mash on truly non-null effects. The sample size N_1 varied from 50 to 1200.



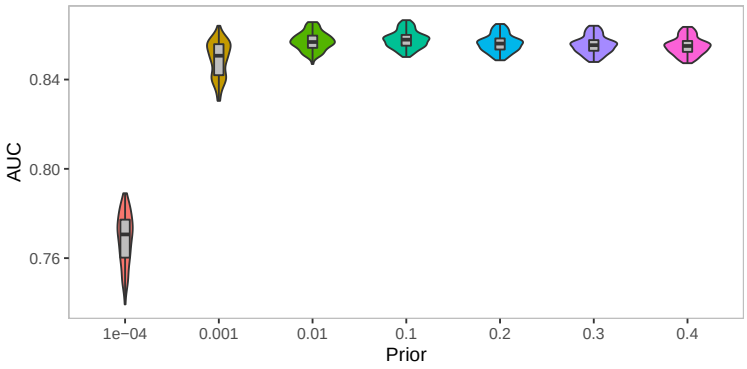
Supplementary Figure S6: Comparison of computation time of model fitting process of X-ING and other methods. The number of tested units of each data varied from 1,500 to 25,000.



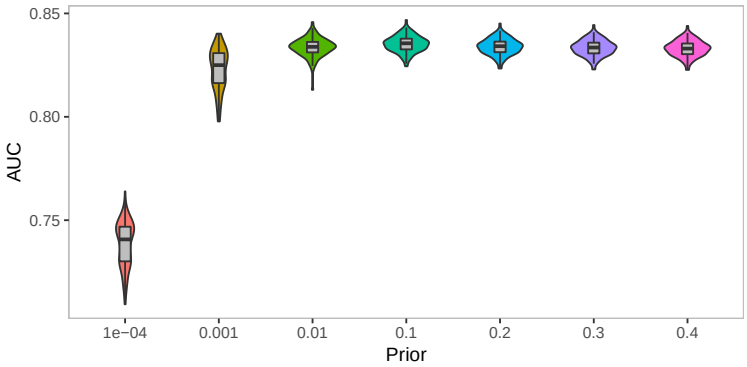
Supplementary Figure S7: Comparison of computation time of model fitting process of X-ING and other methods. The number of contexts of each data varied from 3 to 50.



Supplementary Figure S8: Comparison of AUC of X-ING using simulated Data 1. The proportion of highly correlated predictors varied from 0 to 0.3. For those highly correlated predictors, the autoregressive coefficient ϕ_1 varied from 0.7 to 0.8. All other predictors were correlated with lower autoregressive coefficient ϕ_2, which varied from 0.1 to 0.3.

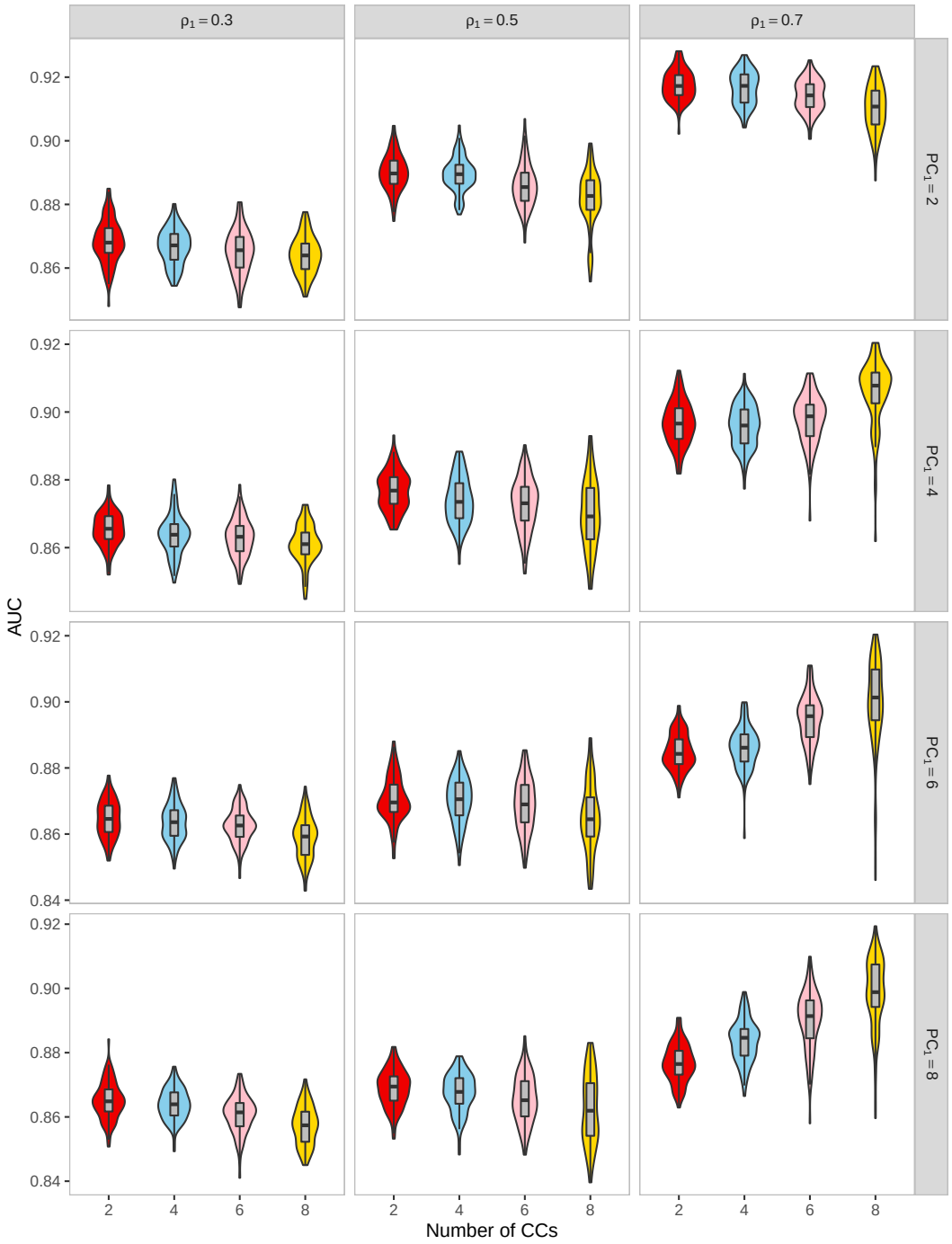


(a)

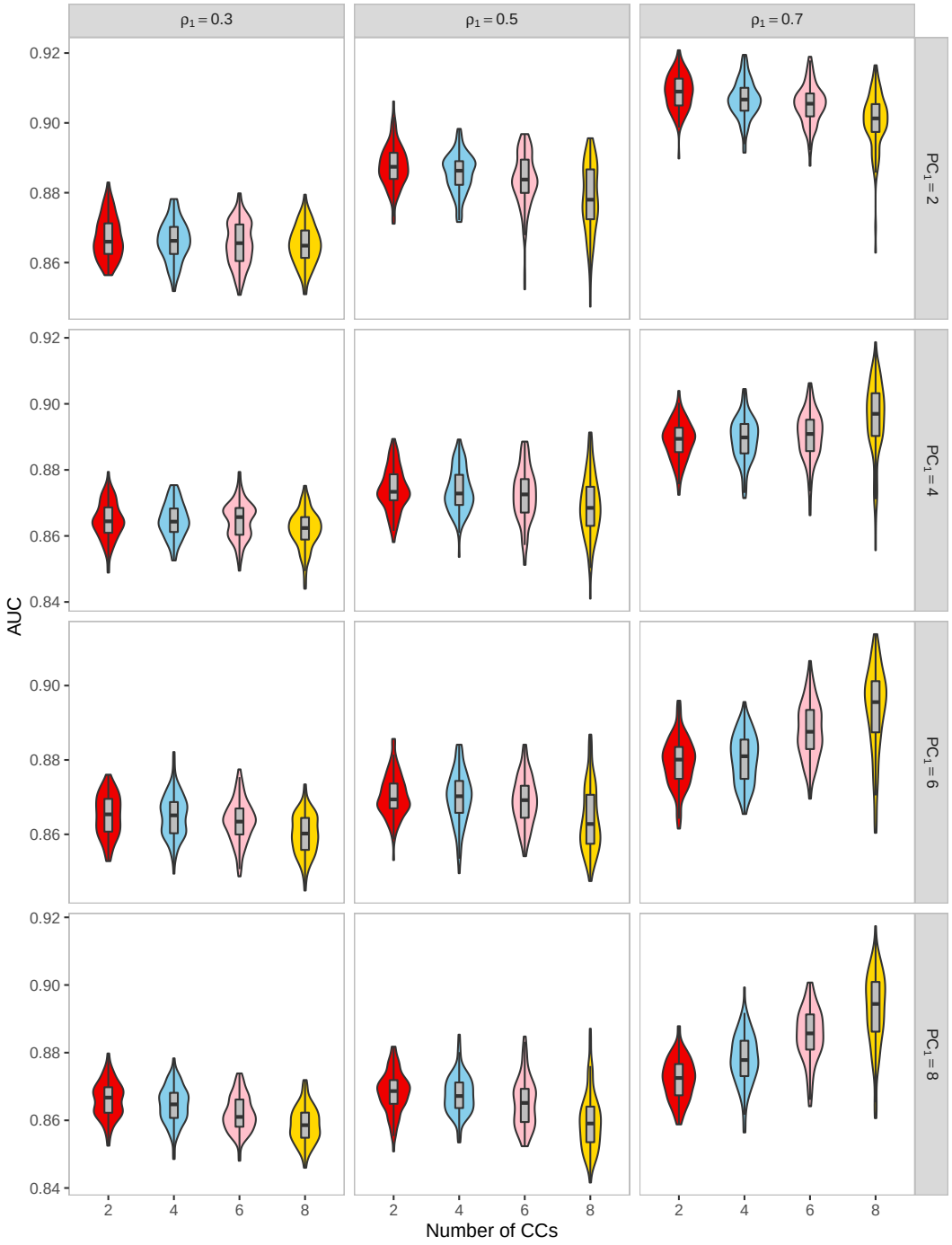


(b)

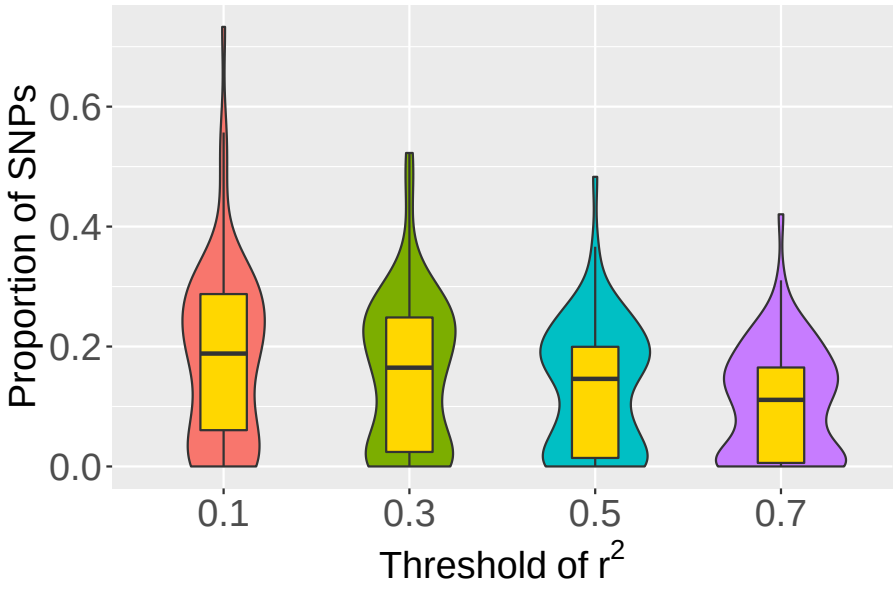
Supplementary Figure S9: Comparison of AUC of X-ING, varying the prior used for model fitting from 10^{-4} to 0.4. (a) The true effects were unstructured. (b) The true effects were structured.



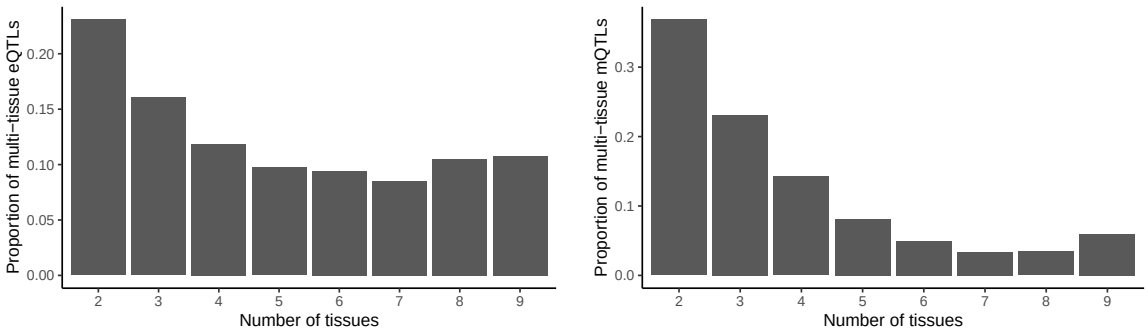
Supplementary Figure S10: Performance of X-ING on unstructured effects (true effects were generated independently across contexts) with different choices of the number of PCs and CCs for fitting X-ING. Data were generated with ρ_1 varying from 0.3 to 0.7. The number of PCs and CCs used for fitting X-ING ranged from 2 to 8.



Supplementary Figure S11: Performance of X-ING on structured effects with different choices of the number of PCs and CCs in data generated with different ρ_1's. The data were generated with ρ_1 varying from 0.3 to 0.7. The number of PCs and CCs used for fitting X-ING ranged from 2 to 8.

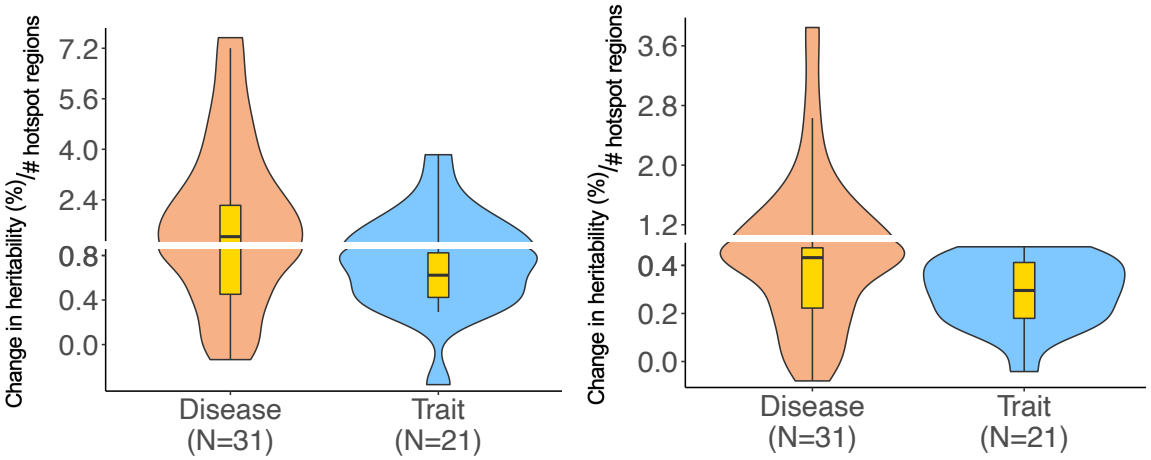


Supplementary Figure S12: Proportion of risk SNPs that are in LD for the 80 diseases/traits. Here the threshold of r^2 varied from 0.1 to 0.7 with fixed window size at 25KB.



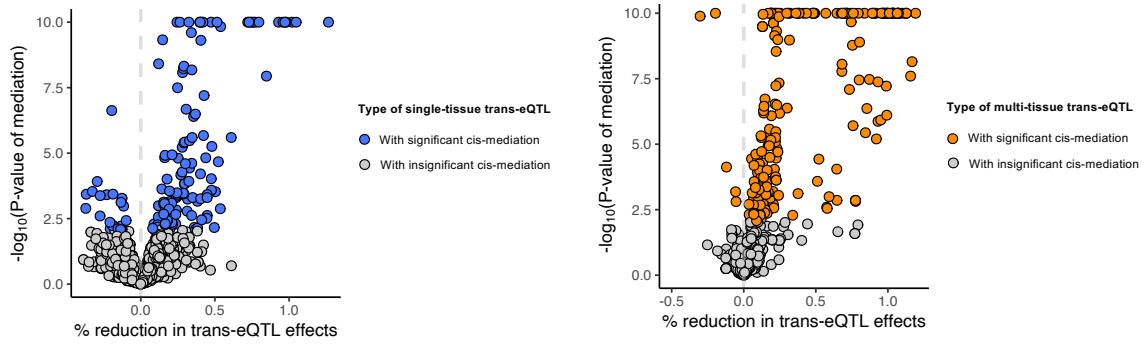
(a) (b)

Supplementary Figure S13: (a) Tissue-sharing patterns of disease/trait-associated cis-eQTLs. We focused on cis-eQTL effects that were non-zero in at least two tissues out of the nine tissues. (b) Tissue-sharing patterns of disease/trait-associated cis-mQTLs. We focus on cis-mQTL effects that were non-zero in at least two tissues out of the nine tissues.



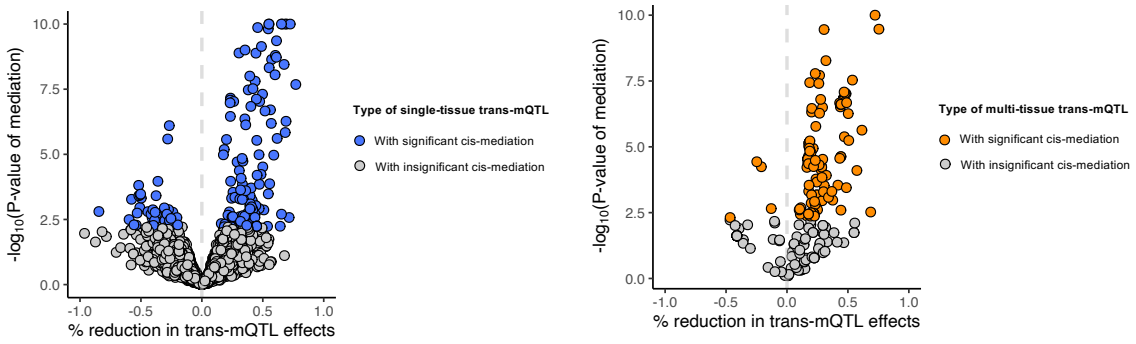
(a) (b)

Supplementary Figure S14: (a) The changes in estimated SNP-based heritability per hotspot region for each disease/trait after removing the trans-eQTL hotspot regions. The average change for the 31 examined diseases (orange) was 1.82%. The average change for the 21 examined traits (blue) was 0.84%. The maximum change was 7.54% for diseases and was 3.83% for traits. (b) The changes in heritability attributed to trans-mQTL hotspot regions. The average changes were 0.72% and 0.35% for diseases (orange) and traits (blue), respectively. The maximum change was 3.85% for diseases and was 0.87% for traits.



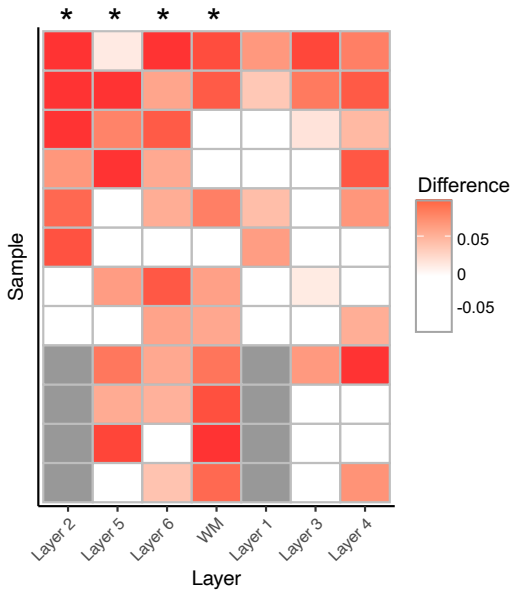
(a) (b)

Supplementary Figure S15: Trans-eQTL effect reduction (x-axis) after accounting for cis-gene versus mediation P-values (y-axis; in negative log base of 10). The mediation P-values were calculated for (a) SNP-cis-trans trios identified as having single-tissue trans-eQTL effects, and (b) trios identified as having multi-tissue trans-eQTL effects. Trios identified as having single-tissue trans-eQTL effects with significant mediation effects (FDR<0.05) are in blue. Trios identified as having multi-tissue trans-eQTL effects with significant mediation effects (FDR<0.05) are in orange. Trios with insignificant mediation effects are in grey. Here P-values are truncated at 10^{-10}. Reduction percentage in trans-eQTL effects is given by $(\beta_{\text{total}} - \beta_{\text{direct}})/\beta_{\text{total}} \times 100\%$, where β_{total} is the total trans-eQTL effect, and β_{direct} is the direct trans-eQTL effect after adjusting for cis-gene. For trios identified as having single-tissue trans-eQTL effects, 23 out of 149 (15.4%) with mediation effect (FDR< 0.05) showed negative reductions, while for trios identified as having multi-tissue trans-eQTL effects, 5 out of 226 trios (2.2%) showed negative reductions.



(a) (b)

Supplementary Figure S16: Trans-mQTL effect reduction (x-axis) after accounting for cis-CpG site versus mediation P-values (y-axis; in negative log base of 10). The mediation P-values were calculated for (a) SNP-cis-trans trios identified as having tissue-specific trans-mQTL effects, and (b) trios identified as having multi-tissue trans-mQTLs. Trios identified as having single-tissue trans-mQTL effects and having significant mediation effects (FDR<0.05) are in blue. Trios identified as having multi-tissue trans-mQTLs and having significant mediation effects (FDR<0.05) are in orange. Trios with insignificant mediation effects are in grey. Here P-values are truncated at 10^{-10}. For trios identified as having tissue-specific trans-mQTL effects in (a), 41 out of 165 trios (24.8%) with significant mediation effect (FDR< 0.05) showed negative reductions, while for trios identified as having multi-tissue trans-mQTL effects, 4 out of 127 trios (3.1%) showed negative reductions.



Supplementary Figure S17: Enrichment heatmap of layer-specific differentially expressed genes among ASD risk-associated genes with the proportions of layer-specific differentially expressed genes across the genome. Color in each cell indicates the difference between the two proportions for each sample and layer. Red represents an enrichment of differentially expressed genes in specific sample and layer, and white represents a depletion of differentially expressed genes. Grey cells indicate missing values (no distinct layer information). There is an enrichment of differentially expressed genes in layer 2 ($P = 0.009$), layer 5 ($P = 0.030$), layer 6 ($P = 0.028$) and white matter ($P = 0.020$) for cis-genes associated with ASD risk loci.

	Expression	Methylation
Breast Mammary Tissue	396	49
Colon Transverse	368	189
Kidney Cortex	73	47
Lung	515	190
Muscle Skeletal	706	42
Ovary	167	140
Prostate	221	105
Testis	322	47
Whole Blood	670	47

Supplementary Table S1: Tissues and e/mQTL analysis sample sizes of the nine tissues with both DNA methylation and expression data from GTEx.

Tissue	Number of samples with expression data
Artery Aorta	387
Artery Coronary	213
Artery Tibial	584
Brain Amygdala	129
Brain Anterior cingulate cortex (BA24)	147
Brain Caudate (basal ganglia)	194
Brain Cerebellar Hemisphere	175
Brain Cerebellum	209
Brain Cortex	205
Brain Frontal Cortex (BA9)	175
Brain Hippocampus	165
Brain Hypothalamus	170
Brain Nucleus accumbens (basal ganglia)	202
Brain Putamen (basal ganglia)	170
Brain Spinal cord (cervical c-1)	126
Brain Substantia nigra	114
Colon Sigmoid	318
Heart Atrial Appendage	372
Heart Left Ventricle	386

Supplementary Table S2: Tissues and tissue sample sizes of the other 19 tissues with only express data used in trans-eQTL analyses of GTEx data. Note that trans-eQTL analysis used a total of 28 tissue types.

References

1. Tenenhaus, A. & Tenenhaus, M. Regularized generalized canonical correlation analysis. *Psychometrika* **76**, 257 (2011).

2. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

3. Kirsten, H. et al. Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci. *Hum. Mol. Genet.* **24**, 4746–4763 (2015).

4. Hoffman, G. E. et al. CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. *Sci. Data* **6**, 1–14 (2019).

5. Skene, N. G. et al. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).

6. Maynard, K. R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).

7. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

8. Ma, L. et al. Schizophrenia risk variants influence multiple classes of transcripts of sorting nexin 19 (SNX19). *Mol. Psychiatry* **25**, 831–843 (2020).

9. Smeland, O. B. et al. Genome-wide analysis reveals extensive genetic overlap between schizophrenia, bipolar disorder, and intelligence. *Mol. Psychiatry* **25**, 844–853 (2020).

10. Harrington, A. J. et al. MEF2C regulates cortical inhibitory and excitatory synapses and behaviors relevant to neurodevelopmental disorders. *eLife* **5**, e20059 (2016).