Isoform-level transcriptome-wide association uncovers extensive novel genetic risk mechanisms for neuropsychiatric disorders in the human brain

Supplemental Materials

Supplemental Methods

Here, we describe mathematical details of the predictive and hypothesis testing methods in the isoTWAS pipeline.

Predictive modeling

For a gene G with M isoforms across N samples, with expression measured across R inferential replicates, we consider the following multivariate linear model:

$$\mathbf{Y}_G^* = \mathbf{X}_G\mathbf{B}_G + \mathbf{E}_G, \tag{1}$$

where

- $\mathbf{Y}_G^*$ is the $N \times M$ matrix of isoform expression for gene G,
- $\mathbf{X}_G$ is the $N \times P$ matrix of genotype dosages (coded as 0,1, or 2 alternative alleles at a SNP) for SNPs within a *cis*-window of the body G,
- $\mathbf{B}_G$ is the $P \times M$ matrix of SNP effects on isoform expression, and
- $\mathbf{E}_G$ is a matrix of random errors, such that $\text{vec}(\mathbf{E}_G) \sim N_{NM}\left(0, \mathbf{\Sigma} = \mathbf{\Omega}^{-1} \otimes \mathbf{I}_N\right)$. Here, $\mathbf{\Sigma}$ is the variance-covariance matrix of the random errors, with $\mathbf{\Omega} = \mathbf{\Sigma}$ representing the precision matrix. The columns of $\mathbf{X}_G$ can be standardized to mean 0 and variance 1 to remove the intercept term from the model.

We implement 6 different methods to estimate $\hat{B}_G$.

Univariate modeling

The simplest method implemented is univariate predictive modelling, as implemented in Gusev et al's FUSION software[1]. We ignore the correlation structure between isoforms and train a univariate model. For the mth isoform, we fit:

$$y_{G,m}^* = \mathbf{X}_G\beta_{G,m} + \epsilon_{G,m} \tag{2}$$

We include three univariate methods:

1. **Elastic net regression with elastic net mixing parameter** $\alpha = 0.5$[2]. This procedure finds the $\hat{\beta}_{G,m}$ that minimizes

$$L(\beta_{G,m}) = \frac{1}{2N}\sum_{i=1}^{N}(y_{G,m,i} - x_{G,i}^T\beta_{G,m})^2 + \lambda[(1-\alpha)\|\beta_{G,m}\|_2^2/2 + \alpha\|\beta_{G,m}\|_1].$$

   We use the `glmnet` package in `R` for implementation with cross-validation.

2. **Best linear unbiased predictor (BLUP) using a linear mixed model**[3,4]. Here, we assume, in Equation 2, that $\beta_{G,m}$ are random SNP effects on the isoform m, such that $\beta_{G,m} \sim \mathbf{N}\left(\mathbf{0}, \frac{\sigma_m^2}{P} I_N\right)$. Here, σ_m^2 is a variance parameter for the SNP effects. We can calculate the BLUP of $\beta_{G,m}$ with the following solution of the Henderson mixed-model[3,4]:

$$\hat{\beta}_{G,m} = \frac{\hat{\sigma}_m^2}{M}\mathbf{X}_G^T\hat{\mathbf{V}}^{-1}y_{G,m}^*,$$

   where $\hat{\sigma}_m^2$ and $\mathbf{V} = \sigma_m^2\mathbf{X}_G\mathbf{X}_G^T/P + \sigma_\epsilon^2\mathbf{I}_N$ are estimated with restricted maximum likelihood estimation and subsequent matrix multiplication. We implement an estimation to this model using ridge regression with the `rrBLUP`package in `R`.

3. **Sum of Single Effects (SuSiE) regression.** Here, we assume that, in Equation 2, $\beta_{G,m} = \sum_{i=1}^{L}\beta_{l,G,m}$, where $\beta_{l,G,m}$ has exactly one non-zero element. SuSiE estimates the variance components using maximum likelihood prior to the estimating $\beta_{G,m}$ using an empirical Bayes approach. We implement this procedure using the `susieR` package in `R`[5].

Curds-and-whey (CW) procedure

We implement Brieman and Friedman's curds-and-whey (CW) procedure, a method that takes advantage of correlations between response variables to improve predictive accuracy[6]. The CW procedure follows these steps:

1. The columns of $\mathbf{Y}_G^*$ and $\mathbf{X}_G^*$ are standardized.
2. $\mathbf{Y}_G^*$ is transformed to the canonical coordinate system, $\mathbf{Y}'_G = \mathbf{T}\mathbf{Y}_G^*$, using the transformation matrix $\mathbf{T}$.
3. Separate univariate elastic net regressions are performed of each of the columns of $\mathbf{Y}'_G$, of the transformed isoform expression response variables. This leads to a new variable of fitted values for the response: $\hat{\mathbf{Y}}'_G$.
4. Each column of $\hat{\mathbf{Y}}'_G$, which we denote without bold face, is shrunk by the corresponding shrinkage factors $d_i = \frac{(1-r)(c_i^2-r)}{(1-r)^2c_i^2+r^2(1-c_i^2)}$ to form $\bar{Y}_G^*$, where c_i is the ith canonical coordinate in T and $r = p/n$.
5. Then, the responses are transformed back to the original coordinate system: $\bar{\mathbf{Y}}_G = \mathbf{T}^{-1}\bar{\mathbf{Y}}_G$. We then fit univariate elastic net models on each column of $\bar{\mathbf{Y}}_G$ to generate the final predictive model.

Curds-and-whey can be thought of as multivariate proportional shrinkage, which addresses sources of prediction error, namely the bias and variance in the model. Regularization tries to decrease the bias in the model by pulling model parameters to 0. Proportional shrinkage introduces a little bias to save a lot of variance by shrinking estimates and not sending them straight to 0. By decreasing variance in the model (and subsequently the prediction), we can get lower prediction error with this biased (but lower variance) model than with the unbiased model.

Multivariate elastic net

Multivariate elastic net is an extension of elastic net regression for a multivariate response variable. The optimization here, fit through coordinate descent, solves

$$\text{argmin}_{\mathbf{B}_G} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \|y_i - \mathbf{B}_G^T x_{G,i}\|_F^2 + \lambda \left[(1-\alpha)\|\mathbf{B}_G\|_F^2/2 + \alpha \sum_{j=1}^{P} \|\beta_{G,j}\|_2 \right] \right\}.$$

Here, $\beta_{G,j}$ is the jth row of the SNP effects matrix $\mathbf{B}_G$. There is a group-lasso penalty on each M-length vector of isoform effects for a single SNP. This penalty works on the whole group of coefficients for each response: either all coefficients are 0, or none are 0. All coefficients are shrunk by the λ penalty, optimally selected through cross-validation. Intuitively, multivariate elastic net should be optimal in settings where the causal isoQTLs are the same across different isoforms of the same gene. We fit this model using the `glmnet` package in `R`[7].

Multivariate Regression with LASSO with Covariance Estimation

From Equation 1, we jointly estimate $\mathbf{B}_G$ and $\mathbf{\Omega}$ by minimizing the following objective function:

$$\left(\hat{\mathbf{B}}_G, \hat{\mathbf{\Omega}}\right) = \text{argmin}_{\mathbf{B}_G, \mathbf{\Omega}} \left\{ g(\mathbf{B}_G, \mathbf{\Omega}) + \lambda_1 \sum_{j' \neq j} |\omega_{j',j}| + \lambda_2 \sum_{j=1}^{p} \sum_{k=1}^{q} |b_{jk}| \right\},$$

where

$$g(\mathbf{B}_G, \mathbf{\Omega}) = \text{tr}\left[n^-1(\mathbf{Y}_G^* - \mathbf{X}_G\mathbf{B}_G)^T(\mathbf{Y}_G^* - \mathbf{X}_G\mathbf{B}_G)\mathbf{\Omega}\right] - \log|\mathbf{\Omega}|.$$

This objective function can be iteratively minimized for both matrix parameters. In any given iteration, we first solve of $\hat{\mathbf{B}}_G$ with a fixed $\mathbf{\Omega}$ using coordinate descent. Then, we can solve for $\hat{\mathbf{\Omega}}$ with the fixed $\hat{\mathbf{B}}_G$ at the given iteration with graphical lasso. We iterate until the convergence tolerance parameter is met. Full details are outlined in Rothman et al[8].

Multivariate Sum of Single Effects

We employ a multivariate extension of Wang et al's Sum of Single Effects (SuSiE) method to address shared effects across isoforms of the same gene. Here, we assume that $\mathbf{E}_G \sim N_{N\times M}(\mathbf{0}, \mathbf{S} \otimes \mathbf{I}_N)$, where $\mathbf{S}$ is the estimated residual covariance. The main assumption of SuSiE is that $\mathbf{B}_G = \sum_{l=1}^{L} \mathbf{B}_{l,G}$, where $\mathbf{B}_{l,G}$ is a single effect matrix of isoQTLs[9]. We assume that $\mathbf{B}_{l,G} = \gamma_l \mathbf{b}_l^T$, where γ is the causal configuration of isoQTLs for the lth single effect. We draw $\gamma_l \sim \text{Mult}(1, \alpha)$ and $\mathbf{b}_l \sim \sum_k \pi_k N_M(\mathbf{0}, \mathbf{U}_k)$. The fitting procedure is as follows: first, the residual covariance matrix $\mathbf{S}$ is estimated. Then, all possible patterns of effect shared (coded in $\mathbf{U}_k$) are learned. These are used to estimate the mixture prior weights π_k. Using a mash prior[10], the multivariate SuSiE model is fit. This procedure is repeated until convergence.

Fine-mapping and ordinary least squares with clustered standard errors

We first conduct a feature selection step by estimated a 90% credible set of causal isoQTLs using `SuSiE`[11]. We restrict the number of isoQTLs in the 90% credible set to no more than 10% the sample size so as to not over-determine the eventual linear regression fitted by ordinary least squares. Call this reduced design matrix of estimated causal isoQTLs $\mathbf{X}_G^*$. We then fit the following linear model using ordinary least squares:

$$\mathbf{Y}_G^* = \mathbf{X}_G^* \mathbf{B}_G + \mathbf{E}_G,$$

where

$$\hat{\mathbf{B}}_G = [(\mathbf{X}_G^*)'\mathbf{X}_G^*)]^{-1}(\mathbf{X}_G^*)'\mathbf{Y}_G^*.$$

Accordingly, we can derive the classic sandwich estimator for the variance of $\hat{\mathbf{B}}_G$, with sandwich parameter $\mathbf{W}$, i.e.

$$\text{Var}(\hat{\mathbf{B}}_G) = (X'X)^{-1}X'\mathbf{W}X(X'X)^{-1}.$$

Instead of the classic OLS approach to set the estimated residuals $\hat{\mathbf{E}}_G = \mathbf{Y}_G^* - \mathbf{X}_G^*\hat{\mathbf{B}}_G$ and letting $\hat{\mathbf{W}} = \hat{\mathbf{E}_G}\hat{\mathbf{E}_G}'$. This flexible estimator does not converge to the variance of the estimated effects matrix as sample size goes to infinity. We employ a clustered standard error, where we assume $\mathbf{W}$ is block-diagonal according to the clusters or replicates in the sample. We thus define each replicated to have design matrix $\mathbf{X}_{c,G}^*$ and variance of estimated effects $\mathbf{W}_c$. Thus, we can estimate the "meat" of the sandwich estimator as

$$X_G^{*'}\mathbf{W}X_G^* = \sum_c X_{c,G}^{*'}\mathbf{W}_c X_{c,G}^*.$$

The $\hat{\mathbf{B}}_G$ alone can be used as weights in traditional summary-statistics based disease mapping in isoTWAS, but the variance in the estimated effects can give a prediction interval when using individual-level genotypes in the external GWAS panel.

Association testing procedure

We employ a stage-wise testing procedure, similar to the `stageR` method[12].

1. We impute genetically-regulated expression of each isoform and estimate associations between each isoform using (1) the appropriate linear regression if we have access to individual-level genotypes in the GWAS and (2) the weighted burden test if we only have access to GWAS summary statistics[13]. We use an LD reference panel from the 1000 Genomes Project[14] that appropriately matches the ancestry of the GWAS sample and the eQTL sample the predictive models were trained with.

2. Given the Wald-type test statistics $Z_1, \ldots, Z_m$ for a given gene, we run an omnibus test to aggregate the test statistics of isoforms of the same gene. We employ either (1) minimum P-value aggregation (i.e. set the gene-level omnibus P-value to the minimum isoform-level P-value), (2) an aggregated Cauchy association test (ACAT)[15], or (3) Chi-square aggregation, where we define the gene-level test statistic $T_G = \sum_{i=1}^m Z_i^2$ and compare to the Chi-square distribution with m degrees of freedom. We correct for multiple comparisons using the Benjamini-Hochberg correction@benjamini1995.

3. We then run an isoform-level multiple testing procedure using the Shaffer MSRB method to assess all isoform-level associations[16]. This procedure controls the family-wide error rate when hypotheses are correlated within the family (i.e. isoforms of the same gene).

Given any overlapping isoforms (i.e. isoforms within 0.5 Megabases of one another), we use gene-level probabilistic fine-mapping[17] to generate a 90% credible set of associated isoforms.

Simulation framework and parameters

Here, we adopt techniques from Mancuso et al's `twas_sim` package[18] to simulate multivariate isoform expression. We consider the following model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},$$

where, for n total samples, $\mathbf{Y}$ is an $n \times m$ matrix of expression values for m isoforms, $\mathbf{X}$ is an $n \times p$ matrix of p SNPs within 1 Megabase of the isoforms in $\mathbf{Y}$, $\mathbf{B}$ is an $p \times m$ matrix of SNP-isoform effects, and $\mathbf{E}$ represents the random error. We first simulate the SNPs in $\mathbf{X}$ by selecting 1,107 SNPs within 1 Megabase of *CACNA1E* and *XRN2*, by using the linkage disequilibrium matrix from European samples of the 1000 Genomes Project and the framework outlined in `twas_sim`. We then simulate $\mathbf{B}$ by selecting p_c proportion of the SNPs in $\mathbf{X}$ as "causal" and generating a non-zero effect size for these SNPs. We allow for the SNPs to be "shared" or "different" across different isoforms. For example, a "shared" $\mathbf{B}$ matrix will have p_c proportion of its rows set to non-zero values and the rest all 0. Conversely, a "different" $\mathbf{B}$ matrix will have p_c proportion SNPs randomly selected to be non-zero for each column of $\mathbf{B}$. We then scale each column of $\mathbf{B}$ to ensure that the genetically-determined portion of each column of $\mathbf{Y}$ equals the isoform expression heritability parameter h_g^2.

To ensure the correlation between columns of $\mathbf{Y}$ reflects correlation matrices determined in simulation parameters, we match moments to generate a multivariate Normal random matrix for $\mathbf{E}$.

Let $\mathbf{C}$ be the desired correlation matrix of $\mathbf{Y}$. Let Y_i and Y_j be the vectors of expression for the ith and jth isoforms. We find

$$\begin{aligned}\text{Cov}(Y_i, Y_j) &= \text{Cov}(\mathbf{X}\beta_i + \epsilon_i, \mathbf{X}\beta_j + \epsilon_j) \\ &= \beta_i' \text{Var}(\mathbf{X})\beta_j + \text{Cov}(\epsilon_i, \epsilon_j)\end{aligned}$$

For $i \neq j$ and c_{ij} the desired correlation between Y_i and Y_j, we have

$$\text{Cov}(Y_i, Y_j) = c_{ij}\sigma_i\sigma_j = \beta_i' \text{Var}(\mathbf{X})\beta_j + \text{Cov}(\epsilon_i, \epsilon_j).$$

We can solve for $\text{Cov}(\epsilon_i, \epsilon_j)$, as the other values are known, and use these values across $i, j \in \{1, \ldots, t\}$ to simulate $\mathbf{E}$. The ith diagonal entry for $\mathbf{E}$ (the variance of ϵ_i) can be taken from the equivalence $\sigma_i^2 = \beta_i^2 \text{Var}(\mathbf{X}) + \text{Var}(\epsilon_i)$. After scaling the genetic value and the error to ensure expression heritability is set of h_g^2, this gives us a simulated $\mathbf{Y}$ with a given correlation matrix. We conduct these simulations 10,000 times across the following set of parameters:

- $n \in \{200, 500, 1000\}$
- $p_c \in \{0.001, 0.01, 0.05\}$
- $h_g^2 \in \{0.05, 0.10, 0.25\}$
- SNP-isoform effect are either "shared" or "different"
- Correlation between isoforms is either "sparse" or "dense"

We generate "sparse" and "dense" correlation matrices using Joe's 2006 C-vine method[19], as implemented in the `clusterGeneration::genPositiveDefMat()` function[20] with $\eta = 1000$ for a sparse correlation matrix and $\eta = 1$ for a dense correlation matrix.

For simulations involving traits, we use the same framework to estimate a multivariate isoform expression matrix. We then estimate traits in 3 scenarios:

1. **Only gene-level expression has a non-zero effect on trait.** Here, we sum the isoform expression to generate a simulated gene expression. We randomly simulate the effect size and scale the error to ensure trait heritability $h_t^2 = 0.10$.

2. **Only 1 isoform has a non-zero effect on the trait.** Here, we generate a multivariate isoform expression matrix with 2 isoforms and scale the total gene expression value such that one isoform (called the effect isoform) makes up $p_g \in \{0.10, 0.30, 0.50, 0.70, 0.90\}$ proportion of total gene expression. We then generate effect size for one of the isoforms and scale the error to ensure trait heritability $h_t^2 = 0.10$.

3. **Two isoforms with different effects on traits**. Here, we generate a multivariate isoform expression matrix with 2 isoforms that make up equal portions of the total gene expression. We then generate an effect size of α for one isoform and $p_e\alpha$ for the other isoform, such that $p_e \in \{-1, -0.5, -0.2, 0.2, 0.5, 1\}$. We then scale the error to ensure trait heritability $h_t^2 = 0.10$.

References

1. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**, 245–252 (2016).
2. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22 (2010).
3. Liu, X. *et al.* GBAT: A gene-based association test for robust detection of trans-gene regulation. *Genome Biology* **21**, 211–211 (2020).
4. Endelman, J. B. Ridge regression and other kernels for genomic selection with r package rrBLUP. *The Plant Genome* **4**, 250–255 (2011).
5. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B* **82**, 1273–1300 (2020).
6. Breiman, L. & Friedman, J. H. No title. **59**, 3–54 (1997).
7. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22 (2010).
8. Rothman, A. J., Levina, E. & Zhu, J. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics* **19**, 947–962 (2010).
9. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B* **82**, 1273–1300 (2020).
10. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics* **51**, 187–195 (2019).
11. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B* **82**, 1273–1300 (2020).
12. Van den Berge, K., Soneson, C., Robinson, M. D. & Clement, L. stageR: A general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome Biology 2017 18:1* **18**, 1–14 (2017).
13. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**, 245–252 (2016).
14. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
15. Liu, Y. *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *The American Journal of Human Genetics* **104**, 410–421 (2019).

16. Shaffer, J. P. Modified Sequentially Rejective Multiple Test Procedures. *Journal of the American Statistical Association* **81**, 826–831 (1986).

17. Mancuso, N. *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. *Nature Genetics* **51**, 675–682 (2019).

18. Mancuso, N. *Twas_sim*. (MancusoLab, 2021).

19. Joe, H. Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis* **97**, 2177–2189 (2006).

20. Qiu, W. & Joe, H. *clusterGeneration: Random cluster generation (with specified degree of separation)*. (2020).