bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Reversing Transcriptome-Wide Association Studies to improve expression Quantitative Trait Loci associations

Jie Yuan, Ben Lai, Itsik Pe'er

December 27, 2020

1 Abstract

Transcriptome-Wide Association Studies discover SNP effects mediated by gene expression through a two-stage process: a typically small reference panel is used to infer SNP-expression effects, and then these are applied to discover associations between imputed expression and phenotypes. We investigate whether the accuracy of SNP-expression and expression-phenotype associations can be increased by performing inference on both the reference panel and independent GWAS cohorts simultaneously. We develop EMBER (Estimation of Mediated Binary Effects in Regression) to re-estimate these effects using a liability threshold model with an adjustment to variance components accounting for imputed expression from GWAS data. In simulated data with only gene-mediated effects, EMBER more than doubles the performance of SNP-expression linear regression, increasing mean r^2 from 0.3 to 0.65 with a gene-mediated variance explained of 0.01. EMBER also improves estimation accuracy when the fraction of cis-SNP variance mediated by genes is as low as 30%. We apply EMBER to genotype and gene expression data in schizophrenia by combining 512 samples from the CommonMind Consortium and 56,081 samples from the Psychiatric Genomic Consortium. We evaluate performance of EMBER in 36 genes suggested by TWAS by concordance of inferred effects with effects reported independently for frontal cortex expression. Applying the EMBER framework to a baseline linear regression model increases performance in 26 out of 36 genes (sign test p-value .0020) with an increase in mean r^2 from 0.200 to 0.235.

2 Introduction

Genome-wide association methods have to date discovered thousands of associations with disease risks and other phenotypes [1]. However, functional studies to determine the causal genes or mechanisms behind these

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

associations have yet to be fully explored [2]. One such strategy is to identify expression Quantitative Trait Loci (eQTLs), variants in the genome which are implicated in mRNA expression. eQTLs have been shown to be enriched in GWAS-significant SNPs and further enrichment is found over fine-mapped cis-eQTLs both across all tissues and for specific pairs of tissues and phenotypes [3, 4]. Large consortia such as GTEx [5, 6] have discovered numerous eQTLs by repeating single-variant regressions over cis-SNPs [7, 8], though challenges remain such as low power and winner's curse stemming from small sample sizes of current RNA-seq experiments [9].

Transcriptome-Wide Association Studies (TWAS) have played a promising role in discovering eQTL associations with disease risk [10, 11]. The method uses reference panels of known SNP-gene effect sizes obtained from smaller gene expression experiments to predict associations between expression and disease phenotypes. SNP-gene effects are trained using multi-SNP regression methods: Gusev et al. [11] use primarily the best linear unbiased predictor (BLUP) whereas the PrediXcan model by Gamazon et al. [10] uses elastic net. Associations between imputed gene expression and disease phenotypes are then assessed either by imputing expression for each individual in large GWAS cohorts or computed directly from summary statistic information. While TWAS performs only a Z-score to test imputed gene-trait association, other iterations of this framework include MR-Egger, a method for estimating of the imputed gene effect size through Mendelian randomization, which accounts for direct SNP-trait effects not mediated by genes, and further refinements of these methods to factor out correlations between SNPs due to linkage disequilibrium [12]. To date, TWAS has discovered a large number of imputed gene-phenotype associations in a wide variety of phenotypes including schizophrenia [13]. However, challenges to interpretation remain, as TWAS associations are not necessarily causal, and genes may be mistakenly tagged due to pleiotropy at the SNP level. Additional noise is provided by small sample sizes in SNP-expression reference panels [14].

We developed EMBER (Estimation of Mediated Binary Effects in Regression) to investigate whether SNP-gene and gene-trait effect sizes in TWAS can be improved by performing inference on both a reference panel and GWAS data simultaneously, rather than a two-step process in which eQTLs are first inferred, then assumed to be fixed and true. As this model must account for both observed gene expression from the reference panel and imputed gene expression from an estimate of eQTLs, EMBER applies the liability threshold model and its treatment of variance explained to represent differing levels of confidence in gene expression estimates [15]. The liability threshold model assumes an observed binary trait, such as disease status, is sampled by thresholding a latent normally distributed liability function. The variance explained by a set of predictors, such as SNPs or imputed gene expression, is then a fraction of the variance of the standard normal liability distribution.

We find in simulations that EMBER significantly improves estimation of SNP-expression effects when

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SNP effects are mediated by gene expression. Even when non-mediated effects are present, EMBER can still improve SNP-expression estimation when the variance explained of mediated effects is at least 30%. In practice, we obtained the largest boost in SNP-expression estimation performance from a variable reduction preprocessing step, far outperforming the BLUP and elastic net on schizoprenia cohorts from the CommonMind Consortium [16] and the Schizophrenia Working Group of the Psychiatric Genomics Consortium [17]. After the variable reduction step, we evaluate EMBER in 36 genes with significant TWAS effects as reported by Gusev et al. [11]. While most of these genes mediate far less than 30% of the variance explained of their cis-SNPs, we nevertheless find that for a subset of these genes, EMBER results in slight improvements in r^2 of eQTL estimates in comparison to independent estimates from GTEx v7 [6].

3 Methods

3.1 Simulation Model

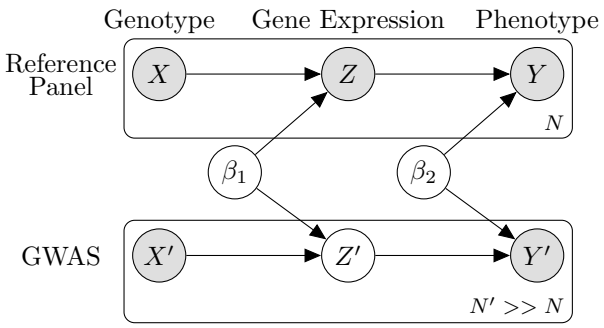


Figure 1: For a small reference panel cohort of size N, we observe SNPs X, transcripts Z, and binary trait Y. Assume we have a large GWAS cohort of size N' with only X' and Y' observed. EMBER attempts to improve inference of β_1 (the SNP-gene effects) given β_2 (the TWAS gene-trait effect).

We assume we have access to reference panel data comprising N individuals with genotype matrix X, gene expression matrix Z, and case/control phenotype vector Y. The conventional method for calculating eQTLs involves a linear model over the observed X and Z. We also assume that we have a large GWAS data set comprising $N' >> N$ individuals with genotype matrix X' and case/control labels for the same phenotype Z'. All SNPs and transcripts are normalized to mean 0 and standard deviation 1. We denote the set of SNP-gene linear effects β_1 and the gene-trait linear effects β_2. For our simulations, we oversimplify the genotypes to be sampled independently with an effect allele frequency set to 0.5 across all genes. We assume true gene expression quantities are distributed $Z = X\beta_1 + \epsilon$ with $\epsilon \sim N(0, 1 - \sigma_X^2)$, where σ_X^2 is the fraction of variance in Z explained by all predictors in X. The mediated SNP-trait effect is then their product $\beta_1\beta_2$, given all SNPs are sampled independently. For now, we assume there are no direct SNP-trait effects not

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mediated by gene expression. As Y is a binary variable, we opt to use a liability threshold model over the total SNP and gene effects $X\hat{\beta}_1\hat{\beta}_2$, described in the next section.

The TWAS approach assumes first that an estimate $\hat{\beta}_1$ is calculated from the reference panel only, and then β_2 is obtained by regressing Y against $X\hat{\beta}_1$. Therefore the accuracy of $\hat{\beta}_1$ is one limitation of TWAS, as current sample sizes for gene expression data tend to be much smaller than for genotype data. Our goal is to perform the analysis in reverse: given the same data and an estimate of the TWAS effect size $\hat{\beta}_2$, we would like to improve estimates of $\hat{\beta}_1$. This relies on an assumption that $\hat{\beta}_2$, likely estimated initially from the same data, is robust to errors in $\hat{\beta}_1$ to a degree that allows its use to improve re-estimation of β_1, provided that the gene in question mediates a significant fraction of the observed SNP-trait effects. Improving the accuracy of SNP-gene effects has the potential to improve TWAS estimates further through more accurate gene imputation. Additionally, re-estimating significance of eQTLs may provide more insights into regulation of gene expression by SNPs.

3.2 Liability Threshold Model

The liability threshold model is an alternative to logistic regression for identifying predictors (SNPs or genes) that are highly associated with disease status. This model assumes that the variance of all modeled predictors is a fraction of a normally distributed underlying quantitative liability, which is thresholded to generate case/control status. This model provides an advantage over the more conventional logistic model by explicitly modeling the variance explained, rather than estimating it by pseudo-r^2 methods. It also provides a more natural interpretation of the Polygenic Risk Score (PRS).

Assume we are regressing measured or imputed quantitative gene expression values Z against case/control status Y. The liability threshold model assumes an underlying standard normal distribution representing the hidden liability. The PRS vector $Z^T\beta$ explains a fraction of the variance of this liability, represented by σ_Z^2. We describe in Section 3.3 a method to estimate σ_Z^2 independently of the inference process for effect sizes in β. The liability is dichotomized into a sampled phenotype by labeling the sampled individual a case if the total liability score passes a threshold T, set such that $1-\Phi(T)$ is equal to the population prevalence if $\Phi(\cdot)$ is the standard normal CDF. If an individual has polygenic risk score $Z_i^T\beta$, then the conditional mean of the liability is shifted by that value, and the variance of the liability is reduced by the variance of the PRS. Therefore, the probability of disease Y_i for an individual with predictors Z_i is given by Wray and Goddard [15]

$$P(Y_i = 1|Z_i, \beta) = \Phi\left(\frac{Z_i^T\beta - T}{\sqrt{1-\sigma_Z^2}}\right) \tag{1}$$

Our objective is to infer values of the coefficients β such that the probability of observed disease statuses

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

across all individuals is maximized given their PRS. We assume the prior for all β to be a normal distribution with standard deviation 1. Therefore, the joint log likelihood of the liability threshold model is

$$LL = \ln(p(Y|Z,\beta)p(\beta)) = -\frac{1}{2}\beta^T\beta + \sum_{i=1}^{N} Y_i \ln \Phi\left(\frac{Z_i^T\beta - T}{\sqrt{1-\sigma_Z^2}}\right) + \sum_{i=1}^{N}(1-Y_i)\ln\left(1-\Phi\left(\frac{Z_i^T\beta - T}{\sqrt{1-\sigma_Z^2}}\right)\right) \tag{2}$$

However, inference with respect to β is difficult because of the dependence of σ_Z^2 in the denominator on values of β.

3.3 Correction for effect size inflation

To resolve the dependence of σ_Z^2 on β during inference, we instead independently estimate σ_Z^2 using the sample data prior to inference of β itself. A separate step to estimate total variance explained of predictors serves two roles: first, it decouples the PRS variance explained from the predictors themselves, facilitating inference, and second, it allows for correction of inflated effect size magnitudes inherent to both liability threshold model and logistic regression. Even when the relative magnitudes of effect sizes are accurately estimated, the overall scale of the effects is poorly controlled in both liability threshold and logistic models (see Results for simulations in which inflated effects produce larger likelihoods). By estimating the true total variance explained σ_Z^2, we can correct for this inflation by re-scaling all effect sizes by a constant factor such that the theoretical total variance explained as a function of β equals the observed σ_Z^2.

When regressing predictors against a quantitative label, the total fraction of variance explained by the predictors can be obtained using Haseman-Elston regression [18, 19]. Golan et al. [20] demonstrate that a similar proportionality between labels and predictors applies to binary case/control labels in a liability threshold model. Assume the disease has a population prevalence K, study prevalence P, and liability threshold $T = 1 - \Phi(K)$ corresponding to the population prevalence. We can define a correlation between samples in Y by taking the outer product of all pairs (i, j) of standardized labels: a Bernoulli label Y_i is standardized as $\frac{Y_i - P}{\sqrt{P(1-P)}}$. As in quantitative traits, this label correlation can be regressed against the correlation between pairs of samples in Z. Specifically, Golan et al. [20] show that correlations between labels and between predictors are related by the following linear function

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$
\begin{aligned}
corr(Y_i, Y_j) &:= \frac{(Y_i - P)(Y_j - P)}{P(1-P)} \\
&\approx \frac{P(1-P)}{K^2(1-K)^2}\sigma_Z^2 \phi(T)^2 corr(Z_i, Z_j)
\end{aligned} \tag{3}
$$

where σ_Z^2 is the variance explained by Z. Therefore, we can regress the two sets of correlations over pairs of samples to obtain a slope, and then solve for σ_Z^2 given observed population parameters of P, K, and T.

3.4 Expectation Maximization of the liability threshold model

By estimating a fixed value for σ_Z^2, we have decoupled it from inference on β, resolving the aforementioned intractability. Therefore, we can perform inference by the conventional probit expectation maximization. First, we introduce a hidden variable l denoting the unobserved quantitative liability which determines case/control status $Y = P(l > T)$. This results in the following EM equation for the joint log likelihood:

$$
LL = \mathbb{E}_q\left(\ln \frac{p(Y, l, \beta | Z)}{q(l)}\right) + KL(q(l)||p(l|Y, Z, \beta)) \tag{4}
$$

We assume l follows a standard normal distribution, with a fraction of variance explained by SNP effects: therefore $l \sim N(Z\beta, 1 - \sigma_Z^2)$. Furthermore, the distribution of l_i is dependent on disease label Y_i, as l must fall on the correct side of the threshold T to impart the correct case/control status. Therefore, we have the following for the distribution $q(l)$

$$
\begin{aligned}
q(l) = p(l|Y, Z, \beta) &= \prod_{i=1}^{N} p(l_i|Y_i, Z_i, \beta) \\
&\propto \prod_{i=1}^{N} TN_{Y_i,T}(Z_i\beta, 1 - \sigma_Z^2)
\end{aligned} \tag{5}
$$

where $TN_Y(\cdot)$ is the truncated normal distribution over values greater than or less than T depending on whether Y_i indicates a case or control, respectively. We can then plug this distribution into the EM equation with joint log likelihood LL. By setting $q(l) = p(l|Y, Z, \beta)$, the KL divergence term cancels, allowing us to maximize β with respect to the expected log likelihood over $q(l)$. This term can be expanded as follows

$$
\sum_{i=1}^{N} \mathbb{E}_q \ln p(Y_i, l_i, \beta | Z_i) = \sum_{i=1}^{N} \mathbb{E}_q\left(\frac{1}{2(1-\sigma_Z^2)}(l - Z_i^T\beta)^2\right) - \frac{1}{2}\beta^T\beta + C \tag{6}
$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where C denotes terms which are constant with respect to β. Therefore, we have the following EM procedure, where in the E-step we update expected values of the hidden liability, and in the M-step we maximize β with respect to the likelihood informed by those expected liabilities.

E-step

$$t = \frac{T - Z_i\beta}{\sqrt{1-\sigma_Z^2}} \tag{7}$$

$$\mathbb{E}_q[l_i] = \begin{cases} Z_i\beta + (\sqrt{1-\sigma_Z^2})\frac{\phi(t)}{(1-\Phi(t))} & Y_i = 0 \\ Z_i\beta - (\sqrt{1-\sigma_Z^2})\frac{\phi(t)}{\Phi(t)} & Y_i = 1 \end{cases} \tag{8}$$

M-step

$$\beta = \left(I + \sigma_Z^{-2}\sum_{i=1}^{N} Z_i Z_i^T\right)^{-1}\left(\sigma_Z^{-2}\sum_{i=1}^{N} Z_i\mathbb{E}_q[l_i]\right) \tag{9}$$

At each step the joint log likelihood is evaluated for convergence, after which the inferred β is returned.

3.4.1 Constrained EM incorporating h^2 correction

The method to estimate σ_Z^2 independently of β inference described by Golan et al. [20] allows us to correct for the aforementioned inflation inferred effect size magnitudes. As effect sizes tend to be scaled by an arbitrary constant factor, we can simply re-scale all effects such that their theoretical variance explained ($\beta^T\beta$ if predictors in Z are standardized) equals σ_Z^2 obtained by the Golan et al. [20] method.

However, we also investigated whether the re-scaling procedure can instead be incorporated directly into the EM algorithm as a constraint. Having obtained an estimate of σ_Z^2 from the sample data, we add a Lagrangian multiplier to the maximization step to enforce the inequality constraint $\sum_{j=1}^{M}\beta_j^2 \leq \sigma_Z^2$. We now maximize the modified log likelihood

$$LL \propto \sum_{i=1}^{N} -\frac{1}{2(1-\sigma_Z^2)}(\mathbb{E}_q[l_i] - Z_i\beta)^2 - \frac{1}{2}\beta^T\beta + \frac{\lambda}{2}(\beta^T\beta - \sigma_Z^2) \tag{10}$$

We notice that this modification conveniently adds only a constant term to the coefficient of $\beta^T\beta$. Therefore, the new update equation has the following form

$$\beta = \left(\left(1+\lambda\right)I + \frac{1}{(1-\sigma_Z^2)}\sum_{i=1}^{N} Z_i^T Z_i\right)^{-1}\left(\sum_{i=1}^{N} Z_i\frac{\mathbb{E}[l_i]}{(1-\sigma_Z^2)}\right) \tag{11}$$

Note that the addition of lambda effectively re-scales the variance of the prior distribution of β, which has until now been assumed to be 1. At every iteration, we would like to solve for lambda such that the constraint $\sum_{j=1}^{M}\beta_j^2 \leq \sigma_Z^2$ is satisfied. However, this operation does not appear to have an analytical solution,

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and we would like to limit the number of matrix inverse calculations required in a numeric solution. To accomplish this, we define the eigendecomposition for the covariance matrix $\frac{1}{(1-\sigma_Z^2)}\sum_{i=1}^{N} Z_i^T Z_i = Q\Lambda Q^{-1}$. We also substitute $\lambda' = \lambda + 1$. Therefore, the matrix inverse in equation 11 can be avoided by retaining the decomposition and only updating the diagonal matrix Λ, which requires only a vector division.

$$(Q\Lambda Q^{-1} + \lambda I)^{-1} = Q(\Lambda + \lambda I)^{-1}Q^{-1} \tag{12}$$

We can then express the predictors β in equation 11 in terms of their eigendecomposition, and plug in this term into the Lagrangian constraint.

$$H = Q^{-1}\left(\sum_{i=1}^{N} Z_i \frac{\mathbb{E}[l_i]}{(1-\sigma_Z^2)}\right) \tag{13}$$

$$\beta^T\beta = H^T(\Lambda + \lambda I)^{-2}H \leq \sigma_Z^2 \tag{14}$$

To our knowledge there is still no analytical solution for λ such that $\beta^T\beta$ is close to σ_Z^2. We instead take the numeric inverse to obtain $f^{-1}(\sigma_Z^2) = \lambda$. We perform this using the Python package pynverse, which performs a search over scalar values for λ. The operation in equation 13, though performed repeatedly, is relatively efficient as it can be performed as a dot product over 1-D vectors.

In the Supplementary Text, we have additionally derived an extension of the Lagrangian method for correlated predictors, in which case the variance of their weighted sum is dependent on the covariance between predictors.

3.5 Improving β_1 estimation

Here we describe the main contribution of EMBER: a method to improve SNP-gene associations (β_1) given an existing estimate of TWAS gene-trait effects (β_2). Instead of a linear regression over the reference panel only, we combine this stage with a liability threshold regression passing estimates of β_1 to scoring $X'\beta_1\beta_2$ against case/control status in a much larger GWAS cohort. We model one gene at a time, so β_2 is a scalar value. First, we describe an adjust to the liability threshold model which accounts for increased error in imputed gene samples.

3.5.1 Accounting for imputed gene expression in the liability model

To combine both observed expression Z and imputed expression $Z' = X'\beta_1$ into a single model, we introduce a hidden variable $\omega \sim N(X'\beta_1\beta_2, 1-\sigma_X^2\beta_2^2)$ representing the distribution of the imputed gene effect on disease

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

liability, where σ_X^2 is the total fraction of variance in Z explained by predictors in X. We can now redefine disease risk with respect to X' and β_1 by marginalizing over the distribution of ω.

$$P(Y_i = 1|X_i', \beta_1) = \int_{-\infty}^{\infty} \Phi\left(\frac{\omega - T}{\sqrt{1-\sigma_X^2\beta_2^2}}\right) N_\omega\left((X_i'\beta_1)\beta_2, 1-\sigma_X^2\beta_2^2\right) \quad d\omega \tag{15}$$

We can simplify this expression by applying the following lemma described by Owen [21]

$$\int_{-\infty}^{\infty} \Phi(a+bx)\phi(x)dx = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \tag{16}$$

We define a standardized gene expression variable ω' and substitute this into equation 15

$$\omega' = \frac{\omega - X'\beta_1\beta_2}{\sqrt{1-\sigma_X^2\beta_2^2}} \tag{17}$$

$$\begin{aligned} P(Y_i = 1|X_i', \beta_1) &= \int_{-\infty}^{\infty} \Phi\left(\frac{[\omega'\sqrt{1-\sigma_X^2\beta_2^2} + (X_i'\beta_1)\beta_2] - T}{\sqrt{1-\sigma_X^2\beta_2^2}}\right)\phi(\omega') \quad d\omega \\ &= \int_{-\infty}^{\infty} \Phi\left(\omega' + \frac{(X_i'\beta_1)\beta_2 - T}{\sqrt{1-\sigma_X^2\beta_2^2}}\right)\phi(\omega') \quad d\omega \end{aligned} \tag{18}$$

We can plug this expression into equation 16 with $b = 1$ and $a = \frac{(X_i'\beta_1)\beta_2 - T}{\sqrt{1-\sigma_X^2\beta_2^2}}$, obtaining

$$P(Y_i = 1|X_i', \beta_1) = \Phi\left(\frac{(X_i'\beta_1)\beta_2 - T}{\sqrt{2(1-\sigma_X^2\beta_2^2)}}\right) \tag{19}$$

Therefore the only change from the liability model with observed Z instead of imputed $X'\beta_1$ is the non-explained liability variance is multiplied by a factor of 2. This distinction in variance explained allows us to perform regression on both observed gene expression in the reference panel and imputed gene expression from GWAS simultaneously.

3.5.2 Inference incorporating both observed and imputed gene expression

Our goal is to infer SNP-gene effects given access to both observed genotype and gene expression data from a small reference panel as well as genotype and trait data only from a large GWAS cohort. Therefore, the log joint likelihood now comprises a sum between the normal prior on β, a linear regression over observed gene expression in the reference panel, and liability threshold regression over imputed gene expression in a GWAS cohort. We first estimate σ_X^2, the variance explained of SNPs on gene expression, using either H-E

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

regression on the reference panel or the liability method by Golan et al. [20], making sure to divide by β_2^2 to remove the contribution of the gene-trait effect. The joint log likelihood can then be expressed as

$$\begin{aligned} L = &-\frac{1}{2}\beta_1^T\beta_1 - N\ln\sqrt{1-\sigma_X^2} \\ &- \frac{1}{2(1-\sigma_X^2)}\left[\sum_{i=1}^{N}(Z_i - X_i\beta_1)^2\right] \\ &+ \sum_{i=1}^{N'} Y_i' \ln\Phi\left(\frac{(X_i'\beta_2)^T\beta_1 - T}{\sqrt{2(1-\sigma_X^2)}}\right) \\ &+ \sum_{i=1}^{N'}(1-Y_i')\ln\left[1-\Phi\left(\frac{(X_i'\beta_2)^T\beta_1 - T}{\sqrt{2(1-\sigma_X^2)}}\right)\right] \end{aligned} \tag{20}$$

The inclusion of linear regression terms can be accommodated easily in the EM algorithm. The new updates are

E step: Calculate $\mathbb{E}_q[l_i]$ with $q(l_i) = TN((X_i^T\beta_2)\beta_1, 1-\sigma_X^2)$

$$t = \frac{T - (X_i'^T\beta_2)\beta_1}{\sqrt{1-\sigma_X^2\beta_2^2}} \tag{21}$$

$$\mathbb{E}_q[l_i] = \begin{cases} (X_i'^T\beta_2)\beta_1 + (\sqrt{1-\sigma_X^2\beta_2^2})\frac{\phi(t)}{(1-\Phi(t))} & Y_i = 0 \\ (X_i'^T\beta_2)\beta_1 - (\sqrt{1-\sigma_X^2\beta_2^2})\frac{\phi(t)}{\Phi(t)} & Y_i = 1 \end{cases} \tag{22}$$

M step:

$$\begin{aligned} \beta = &\left(I + \frac{1}{2(1-\sigma_X^2\beta_2^2)}\sum_{i=1}^{N} X_i'X_i'^T + \frac{1}{1-\sigma_X^2}\sum_{i=1}^{N} X_iX_i^T\right)^{-1} \cdot \\ &\left(\sum_{i=1}^{N} X_i'\frac{\mathbb{E}_q[l_i]}{2(1-\sigma_X^2\beta_2^2)} - \frac{1}{1-\sigma_X^2}X_i^TZ_i\right) \end{aligned} \tag{23}$$

The inferred β can be corrected for effect size inflation by either re-scaling according to σ_X^2 after EM inference, or since the form of the EM algorithm is largely unchanged, the Lagrangian can be incorporated as described in section 3.4.1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4 Results

4.1 Robustness of TWAS estimates (β_2) to sample size

Broadly, we are interested in whether simultaneous inference of SNP-gene (β_1) and gene-trait (β_2) effect sizes can improve accuracy over two-stage regression, given access to both a reference panel and an independent GWAS cohort. As GWAS cohorts are often orders of magnitude larger, hidden gene expression quantities must be inferred for a large majority of the data. The accuracy of these hidden values, which rely on estimates of β_1, may in turn improve estimation of β_2. But to improve β_1 estimation over an initial result from the observed reference panel data, we must incorporate the GWAS cohort and our best estimate of β_2. As the potential benefit of EMBER derives from providing additional samples with only partially observed data, we first investigate the robustness of TWAS associations to sample size. While Gusev et al. [11] report that increasing the size of reference panels past several hundred samples does not improve eQTL estimation, Wainberg et al. [14] nevertheless claim more recently that dependence on small eQTL reference panels is a potential limitation for TWAS.

We found in simulations that TWAS-imputed gene-trait effects are very robust to small sample size in the reference panel, corroborating Gusev et al. [11], and therefore we do not expect significant improvement in β_2 estimation from combining reference and GWAS data sets. We ran simulations comparing r^2 of inferred results from a probit regression to the true generating effects. These results are shown in Figure S1 as a function of sample size and varying expression-trait variance explained. We find across all conditions that the performance of TWAS plateaus at an r^2 of 0.8, and this value is reached at a reference panel size of roughly 500, similar to that of the CommonMind Consortium (CMC) data, for all values of gene-mediated variance explained except for the smallest of 0.01. These results indicate that TWAS is quite robust to sampling error due to small reference panel sizes, and there is little room for improvement on this front. This also suggests that an estimate of β_2 from data is sufficient to improve estimation of eQTLs (β_1) in one step, rather than requiring repeated updates of one set of effects given newer estimates of the other.

4.2 Constrained probit EM for correct effect sizes

We demonstrate that effect size inflation does occur in conventional applications of logistic and probit regression by simulating case/control data with a single predictor and known effect size. From generated case/control samples, we then calculate the log likelihood resulting from scaling the known effect size by a scalar factor (Fig S2A). We find that the scalar factor resulting in the maximum log likelihood does not occur at a value of 1.0, corresponding to the original value, but rather at roughly 1.4, indicating a significant

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

inflation from the inference.

By simulating case/control data from a probit/liability threshold model, we observe inflation in the inferred effects of both the logistic and probit models in Fig S2B. While the log odds ratios of the logistic model are not expected to map directly onto liability model used to simulate the data, we should expect to recover effects of the correct magnitude in the probit model. We rescale the effects by calculating the total variance explained of the predictors as described in the Methods and then divide each effect size by a constant such that their sum of squares equals the calculated value. Lastly, the faster constrained EM method produces nearly identical estimates to those of the slower rescaling method (Fig S2B).

To demonstrate the performance advantages of the constrained EM algorithm in comparison to rescaling the conventional liability threshold model, we measured performance time and number of EM iterations as a function of increasing sample size with other parameters kept constant. Results are shown in Fig 2, where for both methods the number of iterations remains relatively constant as a function of sample size, and the total execution time increases linearly. By both metrics, the constrained probit method far outperforms the rescaling method by requiring fewer iterations despite performing more work per iteration.

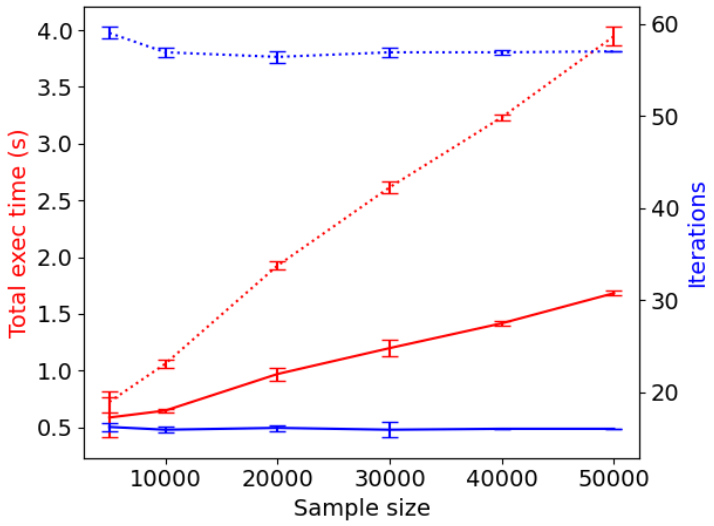


Figure 2: Performance of constrained probit model (solid line) versus conventional probit with rescaling (dotted line). Performance was measured as a function of sample size in terms of total execution time (red) and total number of iterations to convergence (blue). All simulations are run with 200 standard normal predictors with a total variance explained of 0.1. The mean and standard deviation of 10 trials is shown for each sample size.

4.3 Reversing TWAS to re-estimate eQTLs

To test whether we can improve estimation of SNP-expression effects (β_1) when a reference panel and GWAS cohorts are analyzed together, we ran the EMBER method for estimating eQTLs described in the Methods

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

on simulated data. We simulated sample sizes roughly corresponding to those of the CMC and PGC data: a reference panel size of 500 and GWAS size of 50,000. We also assumed for now that the variance explained of non-gene mediated SNP effects is zero. An example run is shown in Fig 3. In this single trial, the r^2 of EMBER more than doubles that of linear regression: 0.88863 versus 0.30665. Additionally, EMBER correctly re-scales all effects so that their magnitudes fall appropriately on the $y = x$ line.

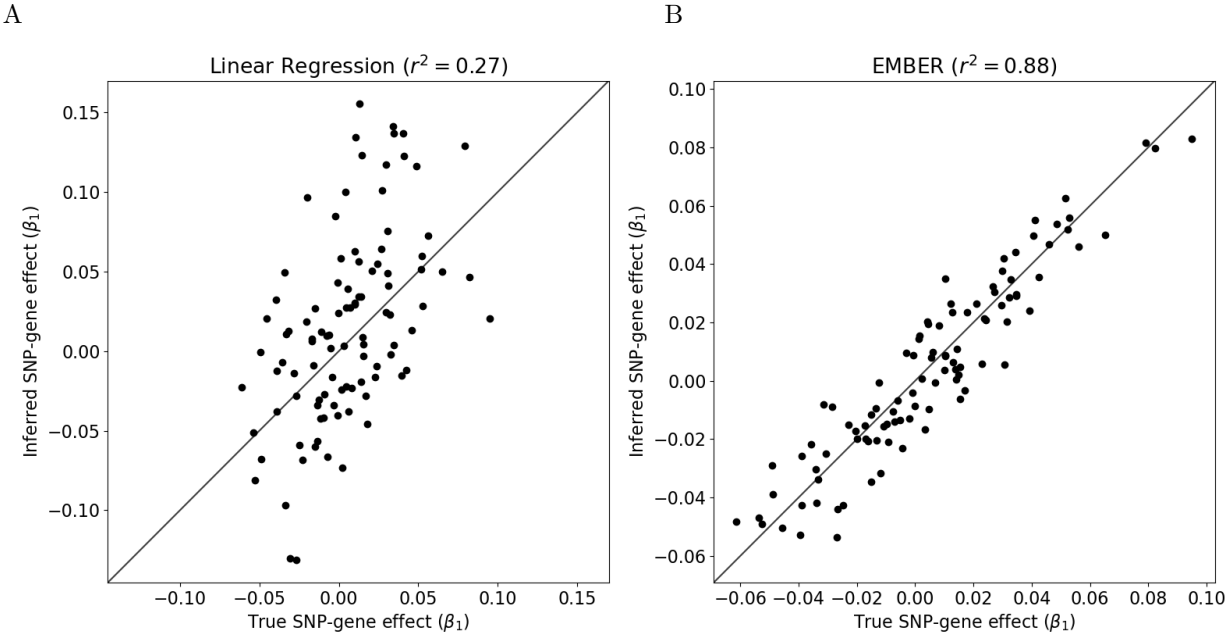


Figure 3: Inference of eQTLs on simulated data by (A) conventional linear regression on the reference panel only and (B) combining reference panel and GWAS data in EMBER. For simulation we assume a reference panel size of 500, GWAS sample size of 50,000, 100 SNPs and one gene with total SNP-gene variance explained of 0.1 and gene-trait variance explained of 0.01. A population prevalence of 0.01 was assumed and allele frequency of 0.5 for all SNPs.

EMBER assumes that there are no non-gene mediated SNP effects, but in practice, this assumption may be violated. To test this, we fixed the gene-mediated SNP variance explained to a constant, and varied the non-mediated variance as a percentage of that constant. These results are shown in Fig 4. We observe that when the percentage on the x-axis is zero, corresponding to results shown in Fig 3, EMBER significantly outperforms linear regression to a mediated variance explained of 0.005, after which the expression layer contributes very little signal to the inference. We also observe that even though the assumption of no non-mediated effects is violated, EMBER still outperforms linear regression up to a percentage of non-mediated variance between 0.7 and 0.8 across the tested conditions.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

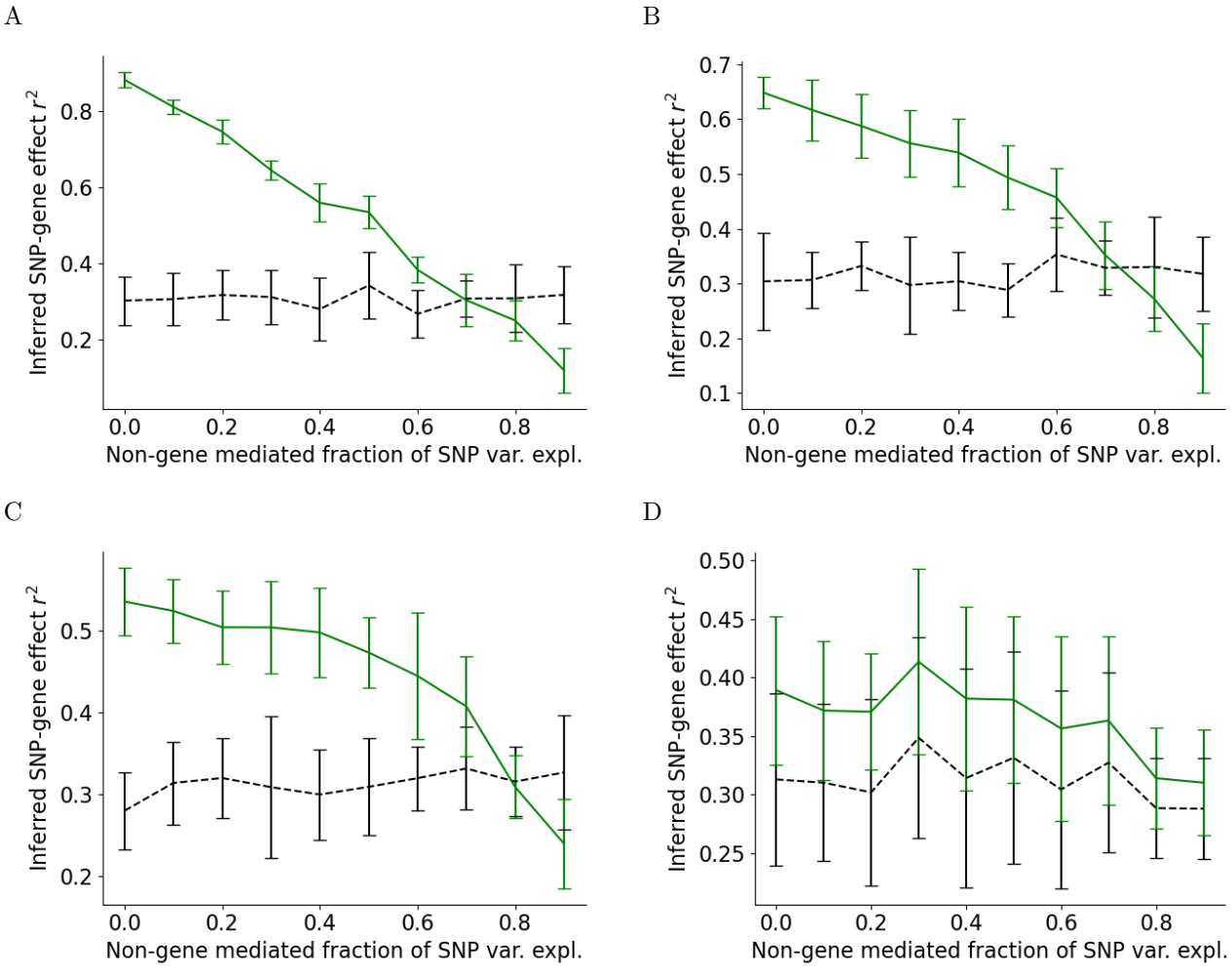


Figure 4: EQTL estimation (r^2) as an increasing fraction of non-mediated SNP variance explained for EMBER (green) and linear regression (black). The gene-mediated SNP variance explained is fixed to (A) 0.05, (B) 0.01, (C) 0.005, (D) 0.001. Percentages are specified as $\frac{\sigma^2_{res}}{\sigma^2_X \sigma^2_Y + \sigma^2_{res}}$ where σ^2_{res} denotes the non-gene mediated residual SNP variance explained, and σ^2_X and σ^2_Y denote the SNP-gene and gene-trait variances explained by β_1 and β_2, respectively.

4.4 Validation in schizophrenia

We applied EMBER to study eQTLs for genes relevant to schizophrenia. We obtained a gene expression reference panel from the CommonMind Consortium (CMC) and GWAS data from 34 cohorts of the Psychiatric Genomics Consortium (PGC). For quality control, we followed the suggestions in Marees et al. [22]: we removed individuals with genotype missingness rate greater than 0.05, and we removed SNPs with missingness rate in the data greater than 0.1, minor allele frequency less than 0.01, and Hardy-Weinberg equilibrium less than 1×10^{-10}. We additionally filtered for individuals in the CMC data with both genotype data and

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

expression data available. This preprocessing resulted in 512 individuals from the CMC data and 56,081 individuals from the PGC. We focused on genes which have been shown to be associated with schizophrenia in TWAS by Mancuso et al. [13], obtaining 36 genes with TWAS p-value $< 10^{-5}$ and which are also present in the CommonMind data. We imputed cis-SNPs using SHAPEIT [23] and Impute2 [24, 25] and obtained ancestry principal components from PLINK [26]. As EMBER is dependent on gene mediation to improve performance, we calculated the fraction of SNP variance mediated by expression by separately calculating total SNP variance, and SNP-expression variance multiplied by inferred TWAS effect squared. These results are shown in Fig S3A. To obtain the best possible TWAS effect sizes for our data, we re-calculated them in the usual manner by estimating eQTLs from the CMC reference data only, imputing gene expression for the PGC data, and then regressing imputed expression against case/control status. Covariates included for all analyses were sex and top 10 ancestry PCs.

To validate EMBER, we performed eQTL re-estimation in both EMBER and linear regression on the CMC and PGC data, and compared the level of concordance with reported eQTLs in GTEx v7 data for the frontal lobe. A high degree of concordance would suggest reproducibility of effects and a higher confidence that these are true eQTLs. We obtained GTEx v7 eQTLs effects from the TWAS/Fusion website (http://gusevlab.org/projects/fusion/), calculated using BLUP.

To sidestep challenges incorporating the model assumptions of EMBER into the BLUP method, we instead addressed significant LD in cis-SNP regions by performing dimensionality reduction over the input SNPs and covariates by performing a second PCA over all predictors, not to be confused with the initial PCA used to discover ancestry covariates. After inferring effects on the transformed PC variables, the predictor-level effects can be recovered by multiplying the PC loadings matrix by the inferred effects over PCs. As the number of components to include from the variable reduction PCA is an open question, we decided to test the performance of linear regression results compared to GTEx reported effects as a function of this parameter. The mean r^2 compared to GTEx across all 36 genes is shown in Fig S3B as a function of number of PCs of the variable reduction PCA to ensure independence of input predictors. We observe that the performance plateaus between 5 and 10 PCs and begins to decline afterward, which prompted us to select 10 PCs for this procedure.

We evaluated the accuracy of baseline linear regression and EMBER, both using the variable reduction preprocessing step, by the concordance of their inferred SNP-expression effects with those reported by TWAS/FUSION in GTEx v7 data. As TWAS/FUSION also reports effect size calculations from the CMC data for 26 out of the 36 studied genes, we compare the r^2 of these results against the GTEx effects when possible, along with our own implementation of the BLUP method for all genes. We assume in both the EMBER and BLUP models a normal prior effect size distribution with a variance of 1×10^{-4}. Results for

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

these methods across all 36 genes are shown in Figure 5. Across all genes, EMBER achieves an average r^2 of 0.235 ± 0.145, improving over an r^2 of 0.200 ± 0.129 obtained by linear regression. EMBER also improves r^2 in 26 out of 36 genes (Figure 5B), with a one-sided sign test p-value of .00197. However, we do not observe a significant relationship between improvement in performance and the fraction of gene-mediated variance explained in Figure 5B. Given the generally low fraction of variance explained across all genes (see Figure 4), it is possible that cis-SNP effects not mediated by the studied genes contribute to a reduction in estimation accuracy. The average r^2 achieved by EMBER is slightly lower that of our implementation of the BLUP model at 0.240 ± 0.123 (Figure 5A), though both methods perform significantly better than the same BLUP model on CMC data as reported by TWAS/FUSION, with an r^2 of 0.201 ± 0.109. As EMBER is currently applied to a baseline linear regression model with a variable reduction step, further improvements in performance may be possible by accessing the full SNP covariance matrix as in the BLUP method. Figure 6 displays effect size comparisons for the gene GLT8D1, where we observed the largest increase in r^2 in EMBER compared to linear regression. Additional plots for all genes are shown in Figure S4.

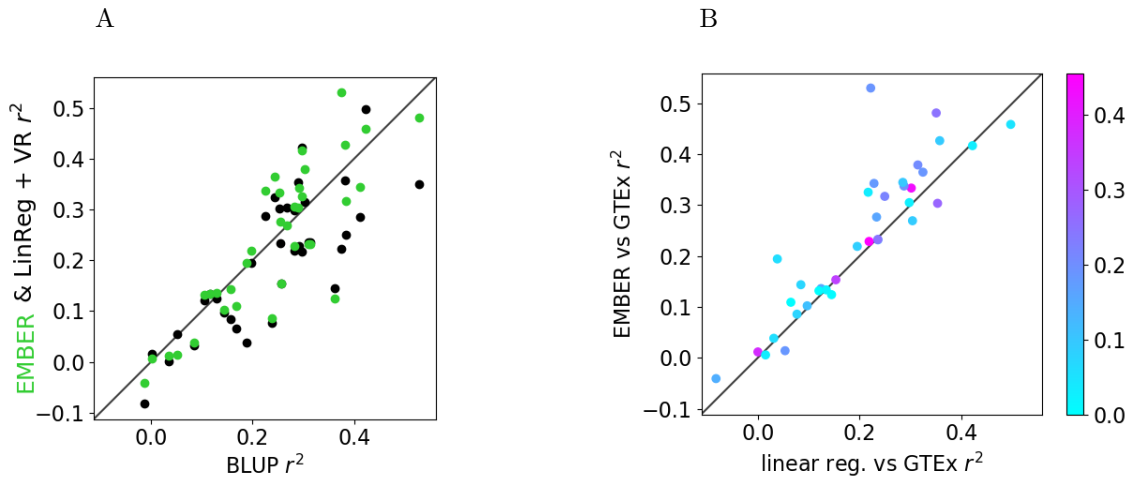


Figure 5: (A) For 36 genes, concordance of estimated SNP-expression effects with reported GTEx v7 effects using BLUP versus using EMBER (green) and linear regression (black) with a variable reduction step. (B) Concordance with reported GTEx v7 effects of linear regression versus EMBER. Colors indicate fraction of SNP variance explained that is mediated by gene expression.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

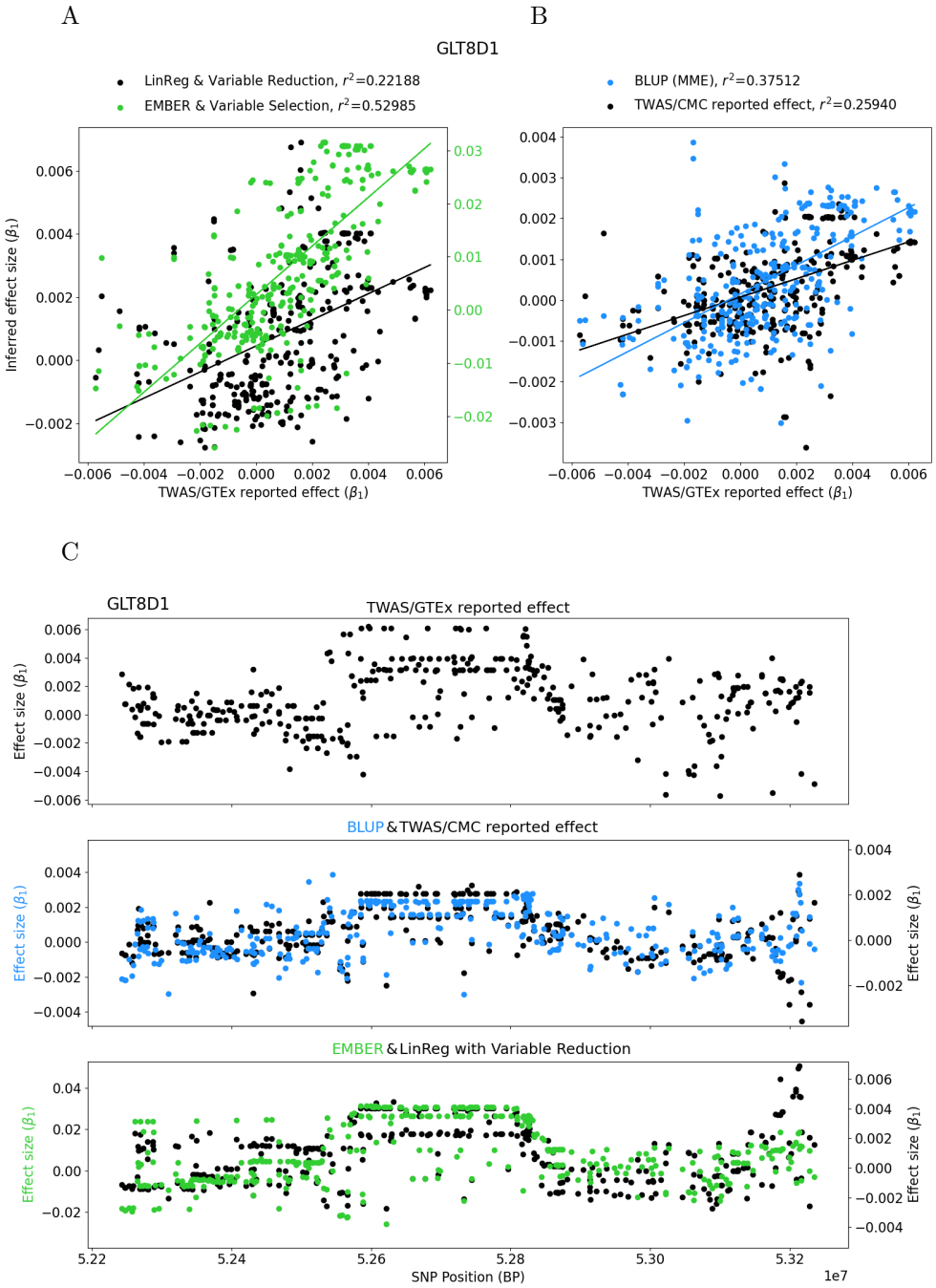


Figure 6: Example results for the gene GLT8D1, for which we observe the largest improvement in r^2 using EMBER. (A) GTEx v7 effects reported by TWAS/FUSION versus effects inferred from linear regression (black) and EMBER (green) with variable reduction. (B) GTEx v7 effects reported by TWAS/FUSION versus effects inferred from BLUP (blue), as well as BLUP effects reported by TWAS/FUSION on CMC data (black). (C) Effect sizes for all models arranged by SNP position in the locus. The top row displays TWAS/FUSION reported BLUP effects on GTEx v7 data; the middle row displays BLUP effects (blue) and TWAS/FUSION reported BLUP effects on CMC data (black); and the bottom row displays results of EMBER (green) and linear regression (black), both using variable reduction to resolve high LD across SNP predictors. For all tested genes, see Fig S4.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

5 Discussion

We presented EMBER, a method to perform inference over gene expression reference panel data alongside GWAS data simultaneously to improve estimation of effect sizes over a two-stage regression approach. In developing EMBER, we introduced also a variance correction to incorporate both measured and imputed gene expression as predictors into a single regression, and a constrained EM method to rapidly re-scale effect sizes based on expected variance explained.

We mapped the parameter space where EMBER is effective. Through simulations, we found that imputed gene-trait associations are quite robust to low reference panel sample sizes, but that significant improvements in eQTL estimates are possible by including GWAS data with no gene expression data observed. Additionally, while EMBER focuses on gene mediated effects, we found that accuracy of inferred effects still improved when the fraction of gene-mediated variance explained was as low as 0.3.

We observed significant improvement in concordance of inferred eQTL effects when we performed a variable reduction step involving transformation of predictors to lower dimensions by PCA. This allows us to account for significant multicollinearity resulting from LD. With this step, ordinary least squares regression achieves an r^2 performance nearly equivalent to BLUP effects reported by TWAS/FUSION on CMC data. This allowed us to build EMBER from a linear regression model, although further improvements may be possible if more sophisticated multi-SNP models such as BLUP can be incorporated. These designs may present additional challenges: for example, a model combining EMBER and BLUP would require features from both the probit and linear mixed models, and while such algorithm have been proposed [27, 28], these typically require performing operations on the full covariance matrix of random effects, which is increasingly impractical for large GWAS.

EMBER is not without its limitations. After variable reduction, for several genes the improvement observed in CMC and PGC data is quite small. This is likely because the fraction of variance explained for these genes is lower than the threshold of 0.2-0.3 for gene-mediated fraction of SNP variance, and EMBER is not benchmarked to improve results in this range. Even so, we do not observe significant reductions in performance in EMBER relative to linear regression, with the largest drop in r^2 being -0.05 in the gene MAPK3.

Going forward, EMBER demonstrates a more general approach to the investigation of genotype-to-phenotype association. Instead of relying on a single study with its traded-off weaknesses of sample size vs. extensiveness of phenotyping, we show that one can leverage modern inference tools in order to get the best of both worlds across multiple studies of different designs. In the era of ubiquitous availability of genotype data, this can greatly facilitate the discovery of new signals.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

6 Web Resources

A Github page for EMBER, including code to reproduce figures is available at:

`https://github.com/jyuan1322/EMBER`

References

[1] Melinda C Mills and Charles Rahal. A scientometric review of genome-wide association studies. *Communications biology*, 2(1):1–11, 2019.

[2] Michael D Gallagher and Alice S Chen-Plotkin. The post-gwas era: from association to function. *The American Journal of Human Genetics*, 102(5):717–730, 2018.

[3] Jason M Torres, Eric R Gamazon, Esteban J Parra, Jennifer E Below, Adan Valladares-Salgado, Niels Wacher, Miguel Cruz, Craig L Hanis, and Nancy J Cox. Cross-tissue and tissue-specific eqtls: partitioning the heritability of a complex trait. *The American Journal of Human Genetics*, 95(5):521–534, 2014.

[4] Farhad Hormozdiari, Steven Gazal, Bryce van de Geijn, Hilary K Finucane, Chelsea J-T Ju, Po-Ru Loh, Armin Schoech, Yakir Reshef, Xuanyao Liu, Luke O'Connor, et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nature genetics*, 50(7): 1041–1047, 2018.

[5] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[6] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675): 204–213, 2017.

[7] Andrey A Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[8] Halit Ongen, Alfonso Buil, Andrew Anand Brown, Emmanouil T Dermitzakis, and Olivier Delaneau. Fast and efficient qtl mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, 2016.

[9] Qin Qin Huang, Scott C Ritchie, Marta Brozynska, and Michael Inouye. Power, false discovery rate and winner's curse in eqtl studies. *Nucleic acids research*, 46(22):e133–e133, 2018.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[10] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091, 2015.

[11] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–252, 2016.

[12] Richard Barfield, Helian Feng, Alexander Gusev, Lang Wu, Wei Zheng, Bogdan Pasaniuc, and Peter Kraft. Transcriptome-wide association studies accounting for colocalization using egger regression. *Genetic epidemiology*, 42(5):418–433, 2018.

[13] Nicholas Mancuso, Huwenbo Shi, Pagé Goddard, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *The American Journal of Human Genetics*, 100(3):473–487, 2017.

[14] Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N Barbeira, David A Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, et al. Opportunities and challenges for transcriptome-wide association studies. *Nature genetics*, 51(4):592–599, 2019.

[15] Naomi R Wray and Michael E Goddard. Multi-locus models of genetic risk of disease. *Genome Medicine*, 2(2):10, 2010.

[16] Menachem Fromer, Panos Roussos, Solveig K Sieberts, Jessica S Johnson, David H Kavanagh, Thanneer M Perumal, Douglas M Ruderfer, Edwin C Oh, Aaron Topol, Hardik R Shah, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature neuroscience*, 19(11): 1442–1453, 2016.

[17] Stephan Ripke, Benjamin M Neale, Aiden Corvin, James TR Walters, Kai-How Farh, Peter A Holmans, Phil Lee, Brendan Bulik-Sullivan, David A Collier, Hailiang Huang, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014.

[18] JK Haseman and RC Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior genetics*, 2(1):3–19, 1972.

[19] Robert C Elston, Sarah Buxbaum, Kevin B Jacobs, and Jane M Olson. Haseman and elston revisited. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 19 (1):1–17, 2000.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.27.424460; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[20] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281, 2014.

[21] Donald Bruce Owen. A table of normal integrals: A table. *Communications in Statistics-Simulation and Computation*, 9(4):389–419, 1980.

[22] Andries T Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, 27(2):e1608, 2018.

[23] Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–181, 2012.

[24] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529, 2009.

[25] Bryan Howie, Jonathan Marchini, and Matthew Stephens. Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*, 1(6):457–470, 2011.

[26] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[27] Meli Baragatti et al. Bayesian variable selection for probit mixed models applied to gene selection. *Bayesian Analysis*, 6(2):209–229, 2011.

[28] Elisa Tosetti and Veronica Vinciotti. A computationally efficient correlated mixed probit model for credit risk inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(4): 1183–1204, 2019.