ARTICLE

TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits

Sini Nagpal,[1,11] Xiaoran Meng,[2,3,11] Michael P. Epstein,[2,3] Lam C. Tsoi,[4] Matthew Patrick,[5] Greg Gibson,[1] Philip L. De Jager,[6] David A. Bennett,[7] Aliza P. Wingo,[8,9] Thomas S. Wingo,[3,10] and Jingjing Yang[3,*]

The transcriptome-wide association studies (TWASs) that test for association between the study trait and the imputed gene expression levels from *cis*-acting expression quantitative trait loci (*cis*-eQTL) genotypes have successfully enhanced the discovery of genetic risk loci for complex traits. By using the gene expression imputation models fitted from reference datasets that have both genetic and transcriptomic data, TWASs facilitate gene-based tests with GWAS data while accounting for the reference transcriptomic data. The existing TWAS tools like PrediXcan and FUSION use parametric imputation models that have limitations for modeling the complex genetic architecture of transcriptomic data. Therefore, to improve on this, we employ a nonparametric Bayesian method that was originally proposed for genetic prediction of complex traits, which assumes a data-driven nonparametric prior for *cis*-eQTL effect sizes. The nonparametric Bayesian method is flexible and general because it includes both of the parametric imputation models used by PrediXcan and FUSION as special cases. Our simulation studies showed that the nonparametric Bayesian model improved both imputation R^2 for transcriptomic data and the TWAS power over PrediXcan when $\geq 1\%$ *cis*-SNPs co-regulate gene expression and gene expression heritability ≤ 0.2. In real applications, the nonparametric Bayesian method fitted transcriptomic imputation models for 57.8% more genes over PrediXcan, thus improving the power of follow-up TWASs. We implement both parametric PrediXcan and nonparametric Bayesian methods in a convenient software tool "TIGAR" (Transcriptome-Integrated Genetic Association Resource), which imputes transcriptomic data and performs subsequent TWASs using individual-level or summary-level GWAS data.

Introduction

Genome-wide association studies (GWASs) have successfully identified thousands of genetic risk loci for complex traits. However, the majority of these loci are located within noncoding regions whose molecular mechanisms remain unknown.[1–3] Recent studies have shown that these associated regions were enriched for regulatory elements such as enhancers (H3K27ac marks)[4,5] and expression of quantitative trait loci (eQTL),[6,7] suggesting that the genetically regulated gene expression might play a key role in explaining the etiology of complex traits. Multiple studies have recently generated rich transcriptomic datasets for diverse tissues of the human body (besides genotype data), e.g., the Genotype-Tissue Expression (GTEx) project for >44 human tissues,[6] Genetic European Variation in Health and Disease (GEUVADIS) for lymphoblastoid cell lines,[8] Depression Genes and Networks (DGN) for whole-blood samples,[9] and the North American Brain Expression Consortium (NABEC) for cortex tissues.[10] Previous studies[11–16] have also shown that integrating transcriptomic data in GWASs can help identify functional loci.

The majority of GWAS projects do not profile transcriptomic data and thus cannot enable direct integrative analysis. However, existing studies[11,12] have shown that one can impute the genetically regulated gene expression (GReX) within such GWAS projects by using reference datasets like GTEx[6] and GEUVADIS[8] to train gene expression imputation models, and then test for the association between imputed GReX for GWAS samples and the trait of interest—referred to as transcriptome-wide association studies (TWASs).[11,12] Specifically, the gene expression imputation models are fitted by regressing assayed gene-expression levels on *cis*-eQTL genotypes with reference dataset. For examples, the PrediXcan[11] method uses an Elastic-Net[17] variable selection model and the FUSION[12] tool implements a Bayesian sparse linear mixed model (BSLMM)[18] to estimate the *cis*-eQTL effect sizes with reference dataset. The estimated *cis*-eQTL effect sizes are then used to impute the GReX for GWAS samples.

In short, the Elastic-Net[17] model used by PrediXcan[11] assumes a combination of LASSO[19] (L_1) and Ridge[20] (L_2) penalties on the *cis*-eQTL effect sizes, which is equivalent to a Bayesian model with a mixture Gaussian and Laplace

[1]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA; [2]Department of Biostatistics and Bioinformatics, Emory University School of Public Health, Atlanta, GA 30322, USA; [3]Center for Computational and Quantitative Genetics, Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA; [4]Department of Dermatology; Department of Computational Medicine & Bioinformatics; Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; [5]Department of Dermatology, University of Michigan Medical School, Ann Arbor, MI 48109, USA; [6]Medical Center Neurological Institute, Columbia University, New York, NY 10032, USA; [7]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL 60612, USA; [8]Division of Mental Health, Atlanta VA Medical Center, Decatur, GA, USA; [9]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA 30322, USA; [10]Department of Neurology, Emory University School of Medicine, Atlanta, GA 30322, USA
[11]These authors contributed equally to this work
*Correspondence: jingjing.yang@emory.edu
https://doi.org/10.1016/j.ajhg.2019.05.018.

prior.[21] In contrast, the BSLMM[18] used by FUSION[12] is a combination of Bayesian variable selection model (BVSR)[22] and linear mixed model (LMM)[23] by assuming a normal mixture prior. Since a parametric prior is assumed for the *cis*-eQTL effect sizes by both Elastic-Net and BSLMM, it restricts the capability of PrediXcan and FUSION for handling the underlying complex genetic architecture of transcriptomes. Existing studies[11,12] have also shown that both PrediXcan[11] and FUSION[12] estimated the average regression R^2 (i.e., the percentage of gene expression variation that can be explained by *cis*-genotypes) as ~5% for human whole-blood transcriptome, while the average genome-wide heritability of gene expression in human whole-blood transcriptome is estimated to be more than double that quantity.[24,25]

Therefore, to flexibly model *cis*-eQTL distributions, we use a nonparametric Bayesian method that was originally proposed for genetic prediction of complex traits,[26] where the prior for effect sizes is nonparametric and can be estimated from the data by assuming a Dirichlet process prior on effect-size variance. This Bayesian model is also known as latent Dirichlet process regression (DPR) model,[26] which can flexibly model the underlying complex genetic architecture of transcriptomes. Thus, DPR is a more generalized model that includes Elastic-Net (implemented in PrediXcan[11]) and BSLMM (implemented in FUSION[12]) as special cases. Consequently, DPR can robustly estimate *cis*-eQTLs and then improve imputation R^2 (the squared Pearson correlation between the observed and imputed values on test samples). Moreover, a variational Bayesian algorithm[26–28] can be employed as an alternative of Monte Carlo Markov Chain (MCMC)[29] to efficiently fit the Bayesian model.

Similar to PrediXcan[11] and FUSION[12] methods, we employ DPR to estimate *cis*-eQTLs effect sizes from a reference dataset, which can then be used for downstream TWASs using either individual-level or summary-level GWAS data. In subsequent sections, we first describe the DPR[26] approach for estimating *cis*-eQTL effect sizes from a reference dataset and how we can then use these effect sizes for a downstream TWAS. We then compare the performance of DPR with PrediXcan using both simulated data and real GWAS and transcriptomic data from the Religious Orders Study and Rush Memory Aging Project (ROS/MAP)[30–33] for studying Alzheimer disease (AD).

Our in-depth simulation studies demonstrated that the DPR method obtained higher imputation R^2 on test samples, when ≥1% *cis*-SNPs are true causal and the true expression heritability is ≤0.2. Consequently, better imputation R^2 resulted in improved power for follow-up association studies. Meanwhile, application of DPR to the ROS/MAP study imputed GReX for 57.8% more genes than PrediXcan. Using DPR, we also found a potentially associated gene *TRAPPC6A* for AD pathology indices, which was missed by PrediXcan. Further, by using the transcriptomic imputation models fitted from ROS/MAP data and summary-level GWAS data generated from the International Genomics of Alzheimer's Project (IGAP),[34] we identified three known AD loci[34–38] that potentially affect the late-onsite AD risk through transcript abundance. We conclude with a discussion of future topics and further describe our software tool TIGAR (Transcriptome-Integrated Genetic Association Resource) implementing both parametric Elastic-Net and nonparametric Bayesian DPR methods for public use.

Material and Methods

Here, we briefly describe the underlying statistical model of gene-expression imputation. Consider the following linear regression model for estimating the *cis*-eQTL effect sizes from a reference study that has both genetic and transcriptomic data available,

$$\mathbf{E_g} = \mathbf{Xw} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \mathbf{N}(0, \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2\mathbf{I}) \qquad \text{(Equation 1)}$$

where $\boldsymbol{E_g}$ denotes the gene expression levels (after corrections for confounding covariates such as age, sex, and principal components) for gene g, $\boldsymbol{X}$ denotes the genotype matrix for all *cis*-genotypes (encoded as the number of minor alleles or genotype dosages), $\boldsymbol{w}$ denotes the corresponding *cis*-eQTL effect-size vector, and ε denotes the error term. The intercept term is dropped in Equation 1 for assuming both $\boldsymbol{E_g}$ and $\boldsymbol{X}$ are centered at 0. Generally, SNPs within 1 Mb of the flanking 5′ and 3′ ends (*cis*-SNPs) are included in this regression model and non-zero $\widehat{\boldsymbol{w}}$ will be used for follow-up analysis. The GReX will be imputed by

$$\widehat{\boldsymbol{GReX}} = \boldsymbol{X_{new}}\widehat{\boldsymbol{w}},$$

with *cis*-SNP data $\boldsymbol{X_{new}}$ for GWAS samples.

Nonparametric Bayesian Method

Following the nonparametric Bayesian DPR model proposed in previous studies for genetic prediction of complex traits,[26] a normal prior $N(0, \sigma_w^2)$ is assumed for the *cis*-eQTL effect sizes (w_i, $i = 1,\ldots, p$) and a Dirichlet process (DP) prior[39] is assumed for the effect-size variance σ_w^2 (as in Equation 1):

$$\mathbf{w_i} \sim \mathbf{N}(0, \boldsymbol{\sigma}_{\mathbf{w}}^2), \boldsymbol{\sigma}_{\mathbf{w}}^2 \sim \mathbf{D}, \mathbf{D} \sim \mathbf{DP}(\mathbf{IG}(\mathbf{a}, \mathbf{b}), \xi). \qquad \text{(Equation 2)}$$

The prior distribution D deviates from the DP with base distribution as an inverse gamma (IG) distribution and concentration parameter ξ. Note that σ_w^2 can be viewed as a latent variable and integrating out σ_w^2 will induce a nonparametric prior distribution for w_i, which is equivalent to a DP normal mixture model,[26–28]

$$\mathbf{w_i} \sim \sum_{\mathbf{k}=0}^{+\infty} \pi_{\mathbf{k}} \mathbf{N}(0, \boldsymbol{\sigma}_{\mathbf{k}}^2), \boldsymbol{\sigma}_{\mathbf{k}}^2 \sim \mathbf{IG}(\mathbf{a_k}, \mathbf{b_k}), \pi_{\mathbf{k}} = \nu_{\mathbf{k}} \prod_{\mathbf{l}=0}^{\mathbf{k}-1} (1 - \nu_{\mathbf{l}}), \nu_{\mathbf{k}} \sim \mathbf{Beta}(1, \xi). \qquad \text{(Equation 3)}$$

Here, the nonparametric prior distribution on w_i is equivalently represented by a mixture normal prior that is a weighted sum of an infinitely number of normal distributions ($N(0, \sigma_k^2)$, $k = 0, \ldots, +\infty$), corresponding weight π_k is determined by (v_l, $l = 0,\ldots, k$) with a Beta prior, and ξ in the Beta prior (the same concentration parameter as in Equation 2) determines the number of components with non-zero weights in the mixture normal prior. Conjugate hyper priors $\xi \sim Gamma(a_\xi, b_\xi)$ and $\sigma_\varepsilon^2 \sim IG(a_\varepsilon, b_\varepsilon)$ are assumed.

Generally, the hyper parameters $a_k, b_k, a_\varepsilon, b_\varepsilon$ in the inverse gamma distributions can be set as 0.1 and (a_ξ, b_ξ) in the gamma distribution can be set as (1, 0.1) to induce non-informative priors for $(\sigma_k^2, \sigma_\varepsilon^2, \xi)$. That is, the parameters $(\sigma_k^2, \sigma_\varepsilon^2, \xi)$ will be adaptively estimated from the data and the nonparametric prior on w_i will be data driven. The posterior estimates for $\boldsymbol{w}$ can be obtained by the MCMC[29] or variational Bayesian algorithm,[28,40] from the following joint conditional posterior distribution

$$P(\boldsymbol{w}, \boldsymbol{\pi}, \boldsymbol{\nu}, \xi, \sigma_\varepsilon^2 \mid \boldsymbol{E_g}, \boldsymbol{X}) \propto$$

$$P(\boldsymbol{E_g} \mid \boldsymbol{w}, \boldsymbol{X}, \sigma_\varepsilon^2) P(\boldsymbol{w} \mid \boldsymbol{\pi}, \sigma_1^2, \ldots, \sigma_k^2, \ldots) \left(\prod_{k=0}^{+\infty} P(\sigma_k^2 \mid a_k, b_k) \right) \times$$
$$P(\boldsymbol{\pi} \mid \boldsymbol{\nu}) P(\boldsymbol{\nu} \mid \xi) P(\xi \mid a_\xi, b_\xi) P(\sigma_\varepsilon^2 \mid a_\varepsilon, b_\varepsilon).$$

Particularly, the variational Bayesian algorithm[28,40] is an approximation for the MCMC[29] with greatly improved computational efficiency, which is also used in our tool. Please refer to the Supplemental Material and Methods for technical details of both MCMC sampling and variational inference algorithms for obtaining the Bayesian posterior estimates for the *cis*-eQTL effect sizes.

Elastic-Net and BSLMM Methods

The Elastic-Net model[17] (used by PrediXcan[11]) estimates the *cis*-eQTL effect sizes $\widehat{\boldsymbol{w}}$ in Equation 1 with a combination of L_1 (LASSO)[19] and L_2 (Ridge)[20] penalties by

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left(\| \boldsymbol{E_g} - \boldsymbol{X}\boldsymbol{w} \|_2^2 + \lambda \left(\alpha \| \boldsymbol{w} \|_1 + \frac{1}{2}(1-\alpha) \| \boldsymbol{w} \|_2^2 \right) \right),$$

where $\| \cdot \|_2$ denotes L_2 norm, $\| \cdot \|_1$ denotes L_1 norm, $\alpha \in [0, 1]$ denotes the proportion of L_1 penalty, and λ denotes the penalty parameter. Particularly, PrediXcan[11] takes $\alpha = 0.5$ and tunes the penalty parameter λ by a 5-fold cross validation.

As pointed out by previous studies,[17,21] the Elastic-Net model is equivalent to a Bayesian model with a mixture Gaussian and Laplace (mixture normal) prior for $\boldsymbol{w}$, that is, $p(\boldsymbol{w}) \propto \exp\left(- \lambda \left(\alpha \| \boldsymbol{w} \|_1 + \frac{1}{2}(1-\alpha) \| \boldsymbol{w} \|_2^2 \right) \right)$. In contrast, the BSLMM[18] assumes a mixture of two normal as the prior for *cis*-eQTL effect sizes, $w_i \sim \pi N(0, (\sigma_1^2 + \sigma_2^2)) + (1-\pi) N(0, \sigma_2^2)$. That is, the BSLMM[18] assumes all *cis*-SNPs have at least a small effect, which are normally distributed with variance σ_2^2, and some proportion (π) of *cis*-SNPs have an additional effect, normally distributed with variance σ_1^2. Particularly, with $\sigma_2^2 = 0$, BSLMM becomes BVSR,[22] and with $\pi = 0$, the BSLMM becomes the LMM.[23] Therefore, the DP normal mixture[26–28] as assumed by the DPR method includes the parametric (mixture normal) priors used by Bayesian Elastic-Net[21] and BSLMM[18] as special cases, which is the main reason why DPR is a more generalized model including Elastic-Net and BSLMM as special cases. This is also why the DPR method can robustly model complex genetic architecture and improve the imputation R^2.

Association Study with Univariate Phenotype

Given individual-level GWAS data (genotype data $\boldsymbol{X_{new}}$, phenotype $\boldsymbol{Y}$, covariant matrix $\boldsymbol{C}$) and *cis*-eQTL effect size estimates $\widehat{\boldsymbol{w}}$, the follow-up TWAS (using a burden type gene-based test[41]) is to test the association between $\widehat{\boldsymbol{GReX}} = \boldsymbol{X_{new}}\widehat{\boldsymbol{w}}$ and $\boldsymbol{Y}$ based on the following generalized linear regression model

$$\mathbf{f}(\mathbf{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{C}]) = \boldsymbol{\eta}\mathbf{C} + \boldsymbol{\beta}\widehat{\mathbf{GReX}}. \qquad \text{(Equation 4)}$$

Here, $f(\cdot)$ is a pre-specified link function, which can be set as identity function for quantitative phenotype or set as logit function for dichotomous phenotype. The gene-based association test is equivalent to test $H_0 : \beta = 0$ in Equation 4.

If only summary-level GWAS data are available, we can take the same approach as implemented by the FUSION[12] method. Let $\boldsymbol{Z}$ denote the vector of Z-scores generated by single variant tests (Wald, likelihood ratio, score tests, etc.) for all *cis*-SNPs. The burden Z-score for gene-based association test is defined as

$$\tilde{\mathbf{Z}} = \frac{\mathbf{Z}\widehat{\mathbf{w}}}{\sqrt{\widehat{\mathbf{Z}\mathbf{w}}}} = \frac{\mathbf{Z}\widehat{\mathbf{w}}}{\sqrt{\widehat{\mathbf{w}}' \mathbf{V} \widehat{\mathbf{w}}}}, \qquad \text{(Equation 5)}$$

where $\boldsymbol{V}$ denotes the covariance matrix of analyzed SNPs that can be estimated from training data or reference panels such as 1000 Genomes Project[42] (of the same ethnicity).

Association Study with Multivariate Phenotype

To test the association between multivariate phenotypes and imputed GReX of the focal gene, we take a similar approach as the MultiPhen method.[43] For example, consider two phenotypes $(\boldsymbol{Y_1}, \boldsymbol{Y_2})$ and a covariate matrix $\boldsymbol{C}$, we first adjust for the covariates by taking the residuals $(\widetilde{\boldsymbol{Y}_1}, \widetilde{\boldsymbol{Y}_2})$ respectively from the linear regression models $\boldsymbol{Y_j} = \boldsymbol{\eta}\boldsymbol{C} + \varepsilon, j = 1, 2$. Then we test whether the regression R^2 is significantly greater than zero $(H_0 : R^2 = 0)$ for the following regression model

$$\widehat{\mathbf{GReX}}_\mathbf{g} = \boldsymbol{\beta}_1 \widetilde{\mathbf{Y}}_1 + \boldsymbol{\beta}_2 \widetilde{\mathbf{Y}}_2 + \boldsymbol{\varepsilon}. \qquad \text{(Equation 6)}$$

That is, we test whether the multivariate phenotypes can jointly explain a non-zero percentage of variance in the imputed GReX. The p value can be calculated by using the F-statistic for the regression R^2 in Equation 6.

Even when only summary-level GWAS data are available, we can first obtain a burden Z-score per phenotype from Equation 5, i.e., $\tilde{\boldsymbol{Z}} = (\widetilde{Z_1}, \widetilde{Z_2})$ with two phenotypes. Then, a similar burden approach can be used to obtain a joint Z-score for multi-phenotype test,

$$\tilde{Z}_{\text{joint}} = \frac{\tilde{Z}\boldsymbol{J}}{\sqrt{\tilde{Z}\boldsymbol{J}}} = \frac{\tilde{Z}\boldsymbol{J}}{\sqrt{\boldsymbol{J}' \boldsymbol{V}_\boldsymbol{Y} \boldsymbol{J}}}, \boldsymbol{J} = (1, \ldots, 1)',$$

where $\boldsymbol{V_Y}$ is the covariance matrix among multiple traits.

Simulation Study Design

We conducted in-depth simulation studies to compare the performance of both PrediXcan and DPR methods with respect to imputation R^2 in the test data and the power of TWASs. Specifically, we used data from 499 ROS/MAP participants[44] which contains both RNA-sequencing and genotype data as training data, and genotype data from an additional 1,200 ROS/MAP participants[44] as test data. The test sample size (1,200) was chosen arbitrarily (randomly selected from the ROS/MAP study) to be comparable with the sample size (1,164) in the real association study of AD pathology indices. The genotyped and imputed genetic data for 2,799 *cis*-SNPs (with minor allele frequency (MAF) > 5% and Hardy-Weinberg p value > 10^{-5}) of the arbitrarily chosen gene *ABCA7* (see Figure S1 for the LD block structure) were used to simulate gene expression levels.

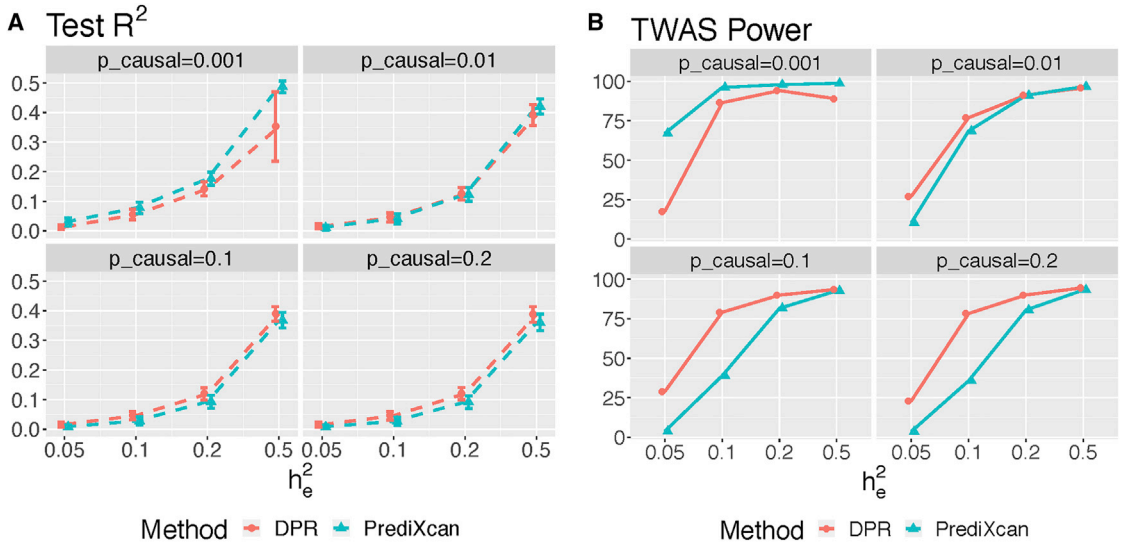


Figure 1. Performance Comparison of DPR versus PrediXcan
Plots of average imputation R^2 (A) and TWAS power (B) in test samples by DPR and PrediXcan, with various proportions of true causal SNPs $p_{causal} = (0.001, 0.01, 0.1, 0.2)$ and true expression heritability $h^2_e = (0.05, 0.1, 0.2, 0.5)$. TWAS power was evaluated with paired expression and phenotype heritability $(h^2_e, h^2_p) = ((0.05, 0.8), (0.1, 0.5), (0.2, 0.25), (0.5, 0.1))$.

We performed comprehensive scenarios that varied the proportion of causal SNPs (out of 2,799 SNPs, influenced gene expression) among values in the vector $p_{causal} = (0.001, 0.01, 0.1, 0.2)$. We varied the proportion of gene expression variance explained by causal SNPs (i.e., expression heritability), along with the proportion of phenotypic variance explained by simulated gene expression levels (i.e., phenotypic heritability), among values in the vector $(h^2_e, h^2_p) = ((0.05, 0.8), (0.1, 0.5), (0.2, 0.25), (0.5, 0.1))$. The phenotypic heritability was selected arbitrarily with respect to expression heritability such that the follow-up association study power fell within the range of (25%, 85%). We also considered various training sample sizes (100, 300, 499) for simulation scenario with $p_{causal} = 0.2$ and $(h^2_e, h^2_p) = (0.2, 0.25)$.

With genotype matrix $\boldsymbol{X_g}$ of the randomly selected causal SNPs (according to p_{causal}), we generated effect sizes w_i from $N(0,1)$ and then re-scaled the effect sizes to ensure the targeted h^2_e. Gene expression levels were generated by $\boldsymbol{E_g} = \boldsymbol{X_g w} + \varepsilon$, with $\varepsilon \sim N(0, (1 - h^2_e))$. Then the phenotype values were generated by $\boldsymbol{Y} = \beta \boldsymbol{E_g} + \varepsilon$, where β was selected with respect to h^2_p and $\varepsilon \sim N(0, (1 - h^2_p))$.

For each scenario, we repeated simulations for 1,000 times, where we applied both PrediXcan[11] and DPR methods to obtain imputation models with training samples, impute the GReX for test samples, and then conduct follow-up association studies using the imputed GReX. We did not compare with FUSION[12] using BSLMM because of the computational burden of estimating *cis*-eQTL effect sizes by MCMC (~2 h per gene). The association study power was calculated as the proportion of 1,000 repeated simulations with p value $< 2.5 \times 10^{-6}$ (genome-wide significance threshold adjusting for testing 20K independent genes).

ROS/MAP Data

Samples in the ROS/MAP data were collected from participants of the Religious Orders Study (ROS) and the Rush Memory and Aging Project (MAP), which are prospective cohort studies of studying aging and dementia.[30,31,33] The ROS/MAP study recruited senior adults without known dementia at enrollment who underwent annual clinical evaluation. Brain autopsy was done at the time of death for each participant. All participants signed an informed consent and Anatomic Gift Act, and the studies were approved by the Institutional Review Board of Rush University Medical Center, Chicago, IL. Specifically, microarray genotype data generated for 2,093 European-decent participants[44] were further imputed to the 1000 Genomes Project Phase 3[42] in our analysis. The post-mortem brain samples (gray matter of the dorsolateral prefrontal cortex) from ~30% these participants were profiled for transcriptomic data by next-generation RNA seqencing.[45] In this paper, we conducted TWASs for two important indices of AD pathology that were quantified with β-antibody specific immunostains:[30,31,33] neurofibrillary tangle density (tangles) with stereology and β-amyloid load (amyloid) with image analysis. The neurofibrillary tangle density quantifies the average Tau tangle density within two or more 20 μm sections from eight brain regions—hippocampus, entorhinal cortex, midfrontal cortex, inferior temporal, angular gyrus, calcarine cortex, anterior cingulate cortex, and superior frontal cortex. The β-amyloid load quantifies the average percent area of cortex occupied by β-amyloid protein in adjacent sections from the same eight brain regions.

Results

Simulation Studies

In the simulation studies, we observed that the DPR method performed robustly with respect to different causal proportions and gene expression heritability. Specifically, when $p_{causal} > 0.01$ DPR outperformed PrediXcan across all expression heritability values, giving higher imputation R^2 in test data (Figure 1A). For example, when $p_{causal} = 0.2$, the average imputation R^2 of 1,000 simulations was estimated as 4.55% by using DPR versus 2.64% by using PrediXcan with $h^2_e = 0.1$, while the average imputation R^2 was estimated as 12.02% by using DPR versus 9.13% by

Table 1. Simulation Prediction R^2 Comparison

	Causal Proportion 0.01		Causal Proportion 0.2	
h_e^2	DPR	PrediXcan	DPR	PrediXcan
0.05	1.60%*	1.12%	1.54%*	0.76%
0.1	4.54%*	4.13%	4.55%*	2.64%
0.2	12.54%*	12.29%	12.02%*	9.13%
0.5	39.31%	42.05%*	38.78%*	36.04%

Various simulation scenarios were considered, with the proportion of true causal SNPs p_{causal} = (0.01, 0.2) and expression heritability h_e^2 = (0.05, 0.1, 0.2, 0.5). The best prediction R^2 per scenario is indicated with asterisk (*).

using PrediXcan with $h_e^2 = 0.2$ (Table 1). When p_{causal} = 0.01, DPR performed slightly out-performed PrediXcan with $h_e^2 = (0.05, 0.1, 0.2)$ and PrediXcan outperformed DPR with $h_e^2 = 0.5$ (Table 1, Figure 1). On the other hand, under a sparse *cis*-eQTL causality model with p_{causal} = 0.001 (i.e., with 3 true causal *cis*-eQTL), the Elastic-Net method resulted in higher imputation R^2 and TWAS power on test data (Figure 1).

Consequently, when pcausal $\geq$ 0.01 and $h_e^2 \leq 0.2$, the power of association studies was higher by using DPR than using PrediXcan imputation models (Figure 1B). When $h_e^2 = 0.5$, using both imputation models led to comparable power for association studies (Figure 1B). Even though both methods had similar over-estimated training R^2 (Figure S2), the DPR method resulted in higher imputation R^2 for test data (Table 1; Figures 1A) and higher power for association studies under *cis*-eQTL causality models with $p_{causal} \geq 0.01$ and $h_e^2 \leq 0.2$ (Figure 1B). In addition, from the simulation studies with various training sample sizes (100, 300, 499), $p_{causal} = 0.2$, and $(h_e^2, h_p^2) = (0.2, 0.25)$, the imputation R^2 and TWAS power increases as sample size increases while the DPR method consistently outperforms PrediXcan (Figure 2). Overall, these results demonstrated the advantages of the DPR method for modeling the complex genetic architecture of transcriptomes, especially when the causal proportions $\geq$0.01 and the expression heritability $\leq$0.2.

Real Applications to ROS/MAP Data

To illustrate the performance of the DPR method in real studies, we applied both DPR and PrediXcan on the ROS/MAP data (see Material and Methods). We trained the gene expression imputation models using 499 samples that have both transcriptomic data for prefrontal cortex tissues and genotype data (imputed to 1000 Genomes Phase 3, with MAF > 5%, Hardy-Weinberg p value > 10^{-5}, and genotype imputation $R^2 > 0.3$). A total of 15,583 genes had gene expression levels after standard RNA-sequencing quality control. The gene expression levels were first adjusted for age at death, sex, postmortem interval, study (ROS or MAP), batch effects, RNA integrity number scores, and cell type proportions (with respect to oligodendrocytes, astrocytes, microglia, neurons) by linear regression models. For each gene, *cis*-SNPs within the 1 Mb of the flanking 5′ and 3′ ends were used in the imputation models as predictors.

First, we compared transcriptome-wide 5-fold cross validation (CV) regression R^2 estimated by using both DPR and PrediXcan methods. Specifically, we randomly split 499 training samples into 5 folds, where the imputation R^2 of each fold was calculated using the model trained with the other 4-fold samples. If the training model is null, we take the imputation R^2 as 0 and take the average imputation R^2 across all 5-fold test samples as 5-fold CV R^2. The transcriptome-wide median of 5-fold CV R^2 is 0.013 by DPR versus 0.005 by PrediXcan. The 5-fold CV R^2 was used as the criterion for selecting significant imputation models ($R^2 > 0.01$ as used by previous studies[11,46]). From Figure 3A, we can see that the DPR method obtained more imputation models and higher imputation R^2 when 5-fold CV R^2 is in the range of (0.01, 0.05), which is also consistent with our simulation studies. Overall, the DPR method obtained significant imputation models for 8,752 genes versus 5,547 genes by PrediXcan (with 57.8% increases). Thus, the DPR method featuring data-driven nonparametric prior for the *cis*-eQTL is preferred in real studies for identifying more genes with imputable expression levels.

Second, to investigate how both DPR and PrediXcan methods perform in real studies with independent prediction cohort, we used the ROS cohort (256 samples) to train gene expression imputation models and then used the MAP cohort (243 samples) as a test dataset. Specifically, we compared the median prediction R^2 by both DPR and PrediXcan with MAP test cohort. As shown in Table 2, the DPR method obtained higher median prediction R^2 than PrediXcan among 8,752 genes that have 5-fold CV $R^2 > 0.01$ by DPR (0.011 versus 0.003), performed similarly as PrediXcan among 5,547 genes that have 5-fold CV $R^2 >$ 0.01 by PrediXcan (0.026 versus 0.026), obtained slightly lower median predication R^2 among 4,819 genes that have 5-fold CV $R^2 > 0.01$ by both DPR and PrediXcan (0.033 versus 0.036). These results are also consistent with our simulation results and 5-fold cross validation results with ROS/MAP data. That is, PrediXcan method is preferred for genes with sparse causal eQTL that have relatively large effect sizes, whereas DPR is preferred for genes with less sparse causal eQTL that have minor effect sizes due to low expression heritability.

Third, we used all 499 training samples to fit imputation models for genes with respective 5-fold CV $R^2 > 0.01$ by both DPR and PrediXcan, and then used these models to impute the GReX for all GWAS samples. We conducted univariate phenotype association studies (Material and Methods) using all GWAS samples (n = 1,164) that have the AD pathology indices (neurofibrillary tangle density and β-amyloid load, with Pearson correlation 0.48) quantified. Possible confounding covariates including age at death, sex, study (ROS or MAP), smoking, education, and first three genotype principle components were adjusted

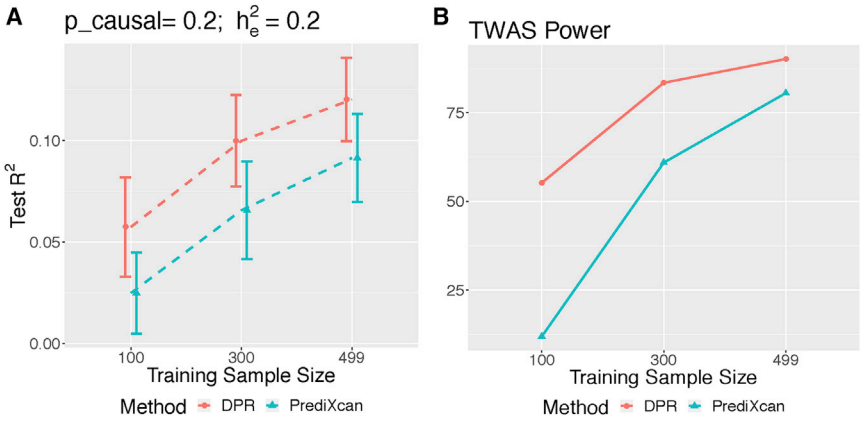


Figure 2. Performance of DPR and PrediXcan with Respect to Various Training Sample Sizes

Test R^2 (A) and TWAS power (B) from simulation studies with causal proportion $p_{causal} = 0.2$, expression heritability and phenotype heritability $(h_e^2, h_p^2) = (0.2, 0.25)$, and various training sample sizes (100, 300, 499).

in the association studies. Interestingly, the association studies for both AD pathology indices using the DPR imputation models identified the same top significant gene *TRAPPC6A* (within the 2 Mb region from the major risk gene *APOE*, encoding apolipoprotein E, but independent of *APOE*) with p values 1.64×10^{-5} and 5.35×10^{-5} (Figures S3A and S4A). Moreover, the multivariate phenotype association studies (Material and Methods) for both AD pathology indices identified *TRAPPC6A* as the most significant gene with p value 5.81×10^{-6} and FDR 0.08 (Figure 3C). On the other hand, the PrediXcan failed to obtain a transcriptomic imputation model for *TRAPPC6A* (Figures S3B, S4B, and S6). Quantile-quantile plots for these TWAS p values were presented in Figure S5.

In addition, for 14 known common and rare loci of late-onset AD[34–38] with significant imputation models, we conducted association studies using transcriptomic imputation models (DPR and PrediXcan) fitted from ROS/MAP data and summary-level GWAS data from IGAP.[34] Using the imputation models fit by DPR, we identified three significant loci with FDR < 0.05 (Figure 3B)—*ADAM10*, *CD2AP*, and *TREM2*—that potentially affect late-onset AD risk through transcriptomic changes. Here, *TREM2* was also identified by using the PrediXcan imputation model (Figure 3B). Particularly, the PrediXcan method imputed GReX for only 5 out of these 14 loci. In summary, these results show that the DPR method has superior power for follow-up TWASs.

Discussion

In this paper, by both in-depth simulations and real applications using individual-level ROS/MAP[30–33] and summary-level IGAP[34] GWAS data, we demonstrated that the nonparametric Bayesian DPR method is preferred for imputing gene expression when the proportion of causal *cis*-eQTL ≥ 0.01 and the true gene expression heritability ≤ 0.2. The advantage of DPR model is due to the flexible nonparametric modeling of *cis*-eQTL effect sizes that results in improved imputation R^2 for gene expression levels and higher power for TWASs. Here, we provide an integrated tool (freely available on GITHUB), referred as Transcriptome-Integrated Genetic Association Resource (TIGAR), which integrates both parametric Elastic-Net and nonparametric Bayesian DPR models as two options for transcriptomic data imputation, along with TWAS options using individual-level and summary-level GWAS data for univariate and multi-variate phenotypes. TIGAR also conducts 5-fold cross validation by default and output significant imputation models with CV $R^2 > 0.01$.

With respect to user-friendly interface and computational efficiency, TIGAR can (1) take standard input files such as genotype files in VCF and dosage formats, phenotype files in PED format, and a combined text file for gene annotations and expression levels; (2) load input data per gene by TABIX for memory efficiency; (3) filter SNPs based on input thresholds of MAF and Hardy-Weinberg p value; (4) provide options of training both Elastic-Net (use Python3 scripts) and DPR (generate input files and call the executable tool developed with C++[26]) imputation models with unified output format; and (5) implement multi-threaded computation to take full advantage of multi-core clusters. These features make TIGAR a preferred tool for saving tedious data preparation and computation time for users. For example, TIGAR can complete training imputation models for ~20K genes and ~1K samples within ~20 h and TWAS within ~1 h with a 2.4 GHz 16-core CPU.

It is important to notice that imputing GReX with *cis*-eQTL effect sizes estimated from a training dataset is analogous to the idea of estimating polygenic risk scores (PRSs).[47] Even though studies of population heterogeneity are lacked for imputing GReX, the same philosophy of estimating PRSs still applies because of the same underlying statistical models. That is, given both genetic and transcriptomic heterogeneities across different populations, one needs to be cautious not using training dataset of a different ethnicity for a TWAS.[47]

As observed in the real ROS/MAP studies, there remains a large gap between the 5-fold CV R^2 using *cis*-eQTL predictors (~5%) and the average genome-wide heritability of gene expression levels (21.8% estimated by GCTA[48] based on a LMM). This is likely due to the large *trans*-acting contribution to transcript abundance documented for most genes. Thus, we hypothesize that it is promising to further improve the imputation R^2 by fitting

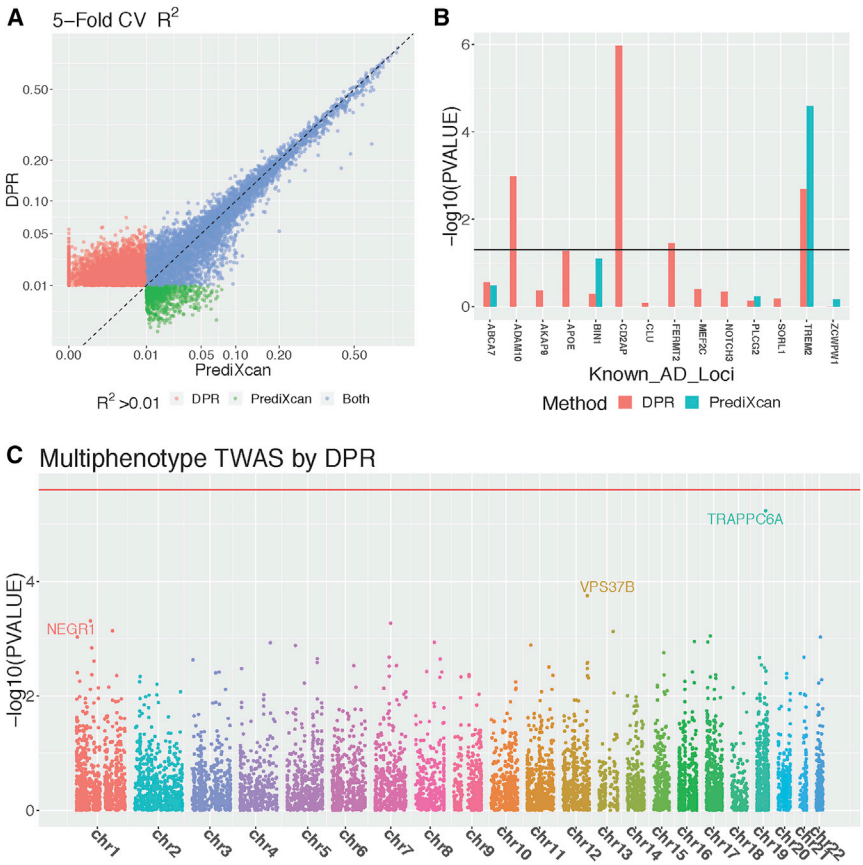


Figure 3. TWAS Results of Studying Alzheimer's Disease

Transcriptome-wide 5-fold cross validation R^2 (A) by PrediXcan and DPR with 499 ROS/MAP training samples, with different colors denoting whether the imputation $R^2 > 0.01$ by DPR, PrediXcan, or both methods (genes with $R^2 > 0.01$ by both DPR and PrediXcan were excluded from the plot). TWAS results (B) at known AD loci using GWAS summary-level statistics from IGAP and imputation models fitted from ROS/MAP data, where missing values are due to NULL imputation models by PrediXcan. Manhattan plot (C) for the multiphenotype TWAS (with neurofibrillary tangle density and β-amyloid load), using individual-level ROS/MAP data.

transcriptomic imputation models with genome-wide variants as predictors. Scalable Bayesian inference techniques such as the Expectation Maximization MCMC (EM-MCMC) algorithm[49] are required for incorporating genome-wide variants.

Another limitation of existing TWAS methods is that the uncertainty of *cis*-eQTL effect-size estimates has not been taken into accounted in the association studies. A Bayesian framework can also be derived by taking the standard errors of these *cis*-eQTL effect-size estimates as prior standard deviations, which is part of our continuing research.

Besides the follow-up gene-based association studies (i.e., TWASs) described in this paper, the transcriptomic imputation models can be further extended by incorporating environmental contributions. The imputed transcript abundance levels can then be used for gene network analysis, differential gene expression analysis, and transcriptome mediation analysis with GWAS data. Validation of transcriptomic prediction accuracy in independent datasets will be critical in this regard, but unfortunately multiple large and similar datasets are not yet generally available for tissues other than peripheral blood.

In conclusion, we expect our work will provide a convenient and improved tool for transcriptomic imputation using the currently available rich reference datasets, as well as enhanced gene mapping for better understanding the genetic etiology of complex traits.

Table 2. Real Study Prediction R^2 Comparison

Number of Genes	DPR	PrediXcan
8,752[a]	0.011	0.003
5,547[b]	0.026	0.026
4,819[c]	0.033	0.036

Median prediction R^2 in MAP test cohort by using imputation models trained with ROS cohort with both DPR and PrediXcan methods.
[a]Genes that have 5-fold CV $R^2 > 0.01$ by DPR.
[b]Genes that have 5-fold CV $R^2 > 0.01$ by PrediXcan.
[c]Genes that have 5-fold CV $R^2 > 0.01$ by both DPR and PrediXcan.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2019.05.018.

Acknowledgments

J.Y. was supported by the startup funding from Department of Human Genetics at Emory University School of Medicine. A.P.W. and T.S.W. were supported by National Institutes of Health (NIH) R01AG056533. M.P.E. was supported by NIH R01GM11796. L.C.T. was supported by the Dermatology Foundation, the Arthritis National Research Foundation, the National Psoriasis Foundation, and NIH K01AR072129. ROS/MAP study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, and U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute. In addition, we thank Thanneer Perumal and Benjamin Logsdon for performing quality control of the ROS/MAP RNA-sequencing data and for creating the brain cell type proportions.

Declaration of Interests

The authors declare no competing interests.

Received: December 27, 2018
Accepted: May 23, 2019
Published: June 20, 2019

Web Resources

FUSION, http://gusevlab.org/projects/fusion/
IGAP data, http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php
PrediXcan, https://github.com/hakyim/PrediXcan
RADC Research Resource Sharing Hub, http://www.radc.rush.edu/
ROS/MAP data, https://www.synapse.org/#!Synapse:syn3219045
TIGAR, https://github.com/yanglab-emory/TIGAR

References

1. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.
2. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356–369.
3. Huang, Q. (2015). Genetic study of complex diseases in the post-GWAS era. J. Genet. Genomics *42*, 87–98.
4. Farh, K.K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.
5. Tsoi, L.C., Stuart, P.E., Tian, C., Gudjonsson, J.E., Das, S., Zawistowski, M., Ellinghaus, E., Barker, J.N., Chandran, V., Dand, N., et al. (2017). Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants. Nat. Commun. *8*, 15382.
6. Battle, A., Brown, C.D., Engelhardt, B.E., Montgomery, S.B.; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis &Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; and eQTL manuscript working group (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.
7. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. *6*, e1000888.
8. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.
9. Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. *24*, 14–24.
10. Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J., et al. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. *6*, e1000952.
11. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., Im, H.K.; and GTEx Consortium (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.
12. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.
13. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., and Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*, 481–487.
14. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. Am. J. Hum. Genet. *100*, 473–487.
15. Su, Y.R., Di, C., Bien, S., Huang, L., Dong, X., Abecasis, G., Berndt, S., Bezieau, S., Brenner, H., Caan, B., et al. (2018). A Mixed-Effects Model for Powerful Association Tests in Integrative Functional Genomics. Am. J. Hum. Genet. *102*, 904–919.
16. Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S.M., Yu, Z., Li, B., Gu, J., Muchnik, S., et al.; Alzheimer's Disease Genetics Consortium (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. Nat. Genet. *51*, 568–576.
17. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol. *67*, 301–320.
18. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264.
19. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. B *58*, 267–288.
20. Hoerl, A.E., and Kennard, R.W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics *42*, 80–86.
21. Li, Q., and Lin, N. (2010). The Bayesian elastic net. Bayesian Anal. *5*, 151–170.
22. Guan, Y.T., and Stephens, M. (2011). Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems. Ann. Appl. Stat. *5*, 1780–1815.
23. Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. *38*, 203–208.
24. Huan, T., Liu, C., Joehanes, R., Zhang, X., Chen, B.H., Johnson, A.D., Yao, C., Courchesne, P., O'Donnell, C.J., Munson, P.J., and

Levy, D. (2015). A systematic heritability analysis of the human whole blood transcriptome. Hum. Genet. *134*, 343–358.
25. Lloyd-Jones, L.R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A., Metspalu, A., Dermitzakis, M., et al. (2017). The Genetic Architecture of Gene Expression in Peripheral Blood. Am. J. Hum. Genet. *100*, 371.
26. Zeng, P., and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat. Commun. *8*, 456.
27. Blei, D.M., and Jordan, M.I. (2006). Variational inference for Dirichlet process mixtures. Bayesian Anal. *1*, 121–143.
28. Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational Inference: A Review for Statisticians. J. Am. Stat. Assoc. *112*, 859–877.
29. Casella, G. (2001). Empirical Bayes Gibbs sampling. Biostatistics *2*, 485–500.
30. Bennett, D.A., Schneider, J.A., Arvanitakis, Z., and Wilson, R.S. (2012). Overview and findings from the religious orders study. Curr. Alzheimer Res. *9*, 628–645.
31. Bennett, D.A., Schneider, J.A., Buchman, A.S., Barnes, L.L., Boyle, P.A., and Wilson, R.S. (2012). Overview and findings from the rush Memory and Aging Project. Curr. Alzheimer Res. *9*, 646–663.
32. Ng, B., White, C.C., Klein, H.U., Sieberts, S.K., McCabe, C., Patrick, E., Xu, J., Yu, L., Gaiteri, C., Bennett, D.A., et al. (2017). An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat. Neurosci. *20*, 1418–1426.
33. Bennett, D.A., Buchman, A.S., Boyle, P.A., Barnes, L.L., Wilson, R.S., and Schneider, J.A. (2018). Religious Orders Study and Rush Memory and Aging Project. J. Alzheimers Dis. *64* (s1), S161–S189.
34. Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al.; European Alzheimer's Disease Initiative (EADI); Genetic and Environmental Risk in Alzheimer's Disease; Alzheimer's Disease Genetic Consortium; and Cohorts for Heart and Aging Research in Genomic Epidemiology (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. *45*, 1452–1458.
35. Reitz, C. (2014). Genetic loci associated with Alzheimer's disease. Future Neurol. *9*, 119–122.
36. Reitz, C. (2015). Novel susceptibility loci for Alzheimer's disease. Future Neurol. *10*, 547–558.
37. Sims, R., van der Lee, S.J., Naj, A.C., Bellenguez, C., Badarinarayan, N., Jakobsdottir, J., Kunkle, B.W., Boland, A., Raybould, R., Bis, J.C., et al.; ARUK Consortium; and GERAD/PERADES, CHARGE, ADGC, EADI (2017). Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. Nat. Genet. *49*, 1373–1384.
38. Yuan, X.Z., Sun, S., Tan, C.C., Yu, J.T., and Tan, L. (2017). The Role of ADAM10 in Alzheimer's Disease. J. Alzheimers Dis. *58*, 303–322.
39. Müller, P., and Mitra, R. (2013). Bayesian Nonparametric Inference - Why and How. Bayesian Anal. *8*, 8.
40. Carbonetto, P., and Stephens, M. (2012). Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. Bayesian Anal. *7*, 73–107.
41. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.
42. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.; and 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
43. O'Reilly, P.F., Hoggart, C.J., Pomyen, Y., Calboli, F.C., Elliott, P., Jarvelin, M.R., and Coin, L.J. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS ONE *7*, e34861.
44. De Jager, P.L., Shulman, J.M., Chibnik, L.B., Keenan, B.T., Raj, T., Wilson, R.S., Yu, L., Leurgans, S.E., Tran, D., Aubin, C., et al.; Alzheimer's Disease Neuroimaging Initiative (2012). A genome-wide scan for common variants affecting the rate of age-related cognitive decline. Neurobiol. Aging *33*, 1017.e1–1017.e15.
45. De Jager, P.L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L.C., Yu, L., Eaton, M.L., Keenan, B.T., Ernst, J., McCabe, C., et al. (2014). Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat. Neurosci. *17*, 1156–1163.
46. Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., Bolla, M.K., Shu, X.O., Lu, Y., Cai, Q., et al.; NBCS Collaborators; and kConFab/AOCS Investigators (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat. Genet. *50*, 968–978.
47. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. *100*, 635–649.
48. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
49. Yang, J., Fritsche, L.G., Zhou, X., Abecasis, G.; and International Age-Related Macular Degeneration Genomics Consortium (2017). A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies. Am. J. Hum. Genet. *101*, 404–416.

The American Journal of Human Genetics, Volume *105*

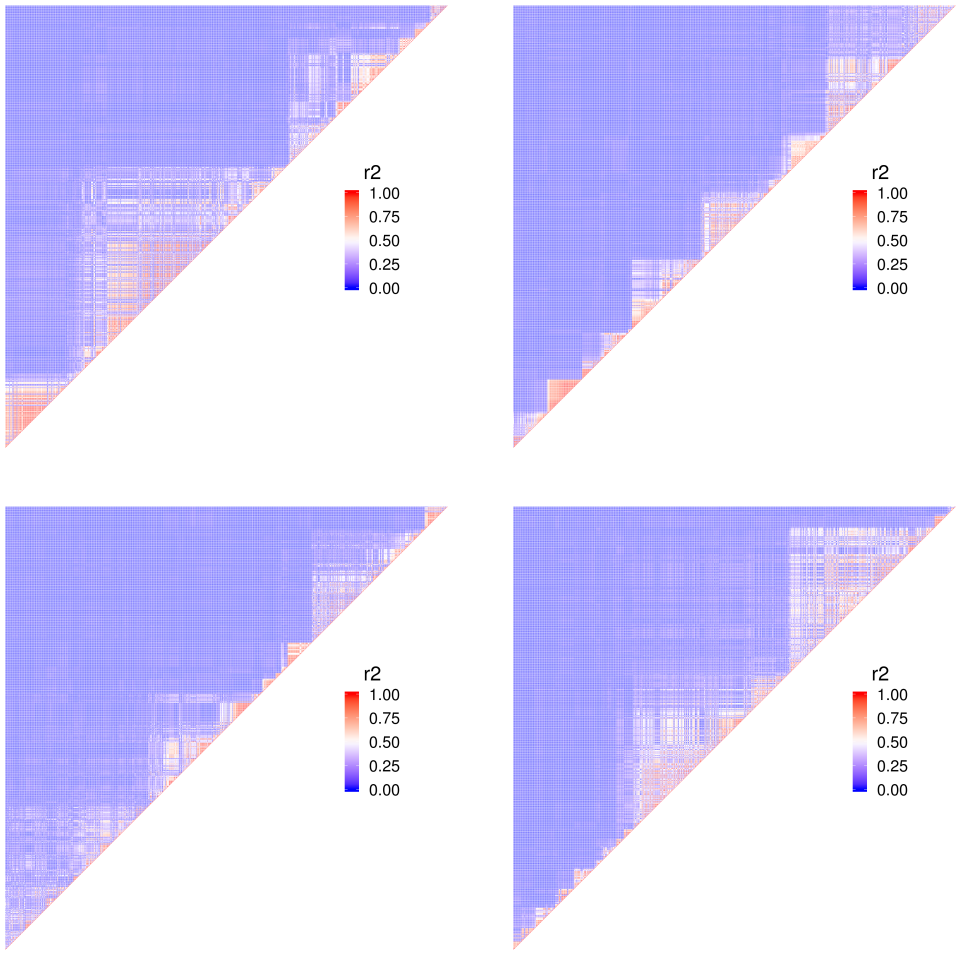
Supplemental Data

TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits

Sini Nagpal, Xiaoran Meng, Michael P. Epstein, Lam C. Tsoi, Matthew Patrick, Greg Gibson, Philip L. De Jager, David A. Bennett, Aliza P. Wingo, Thomas S. Wingo, and Jingjing Yang

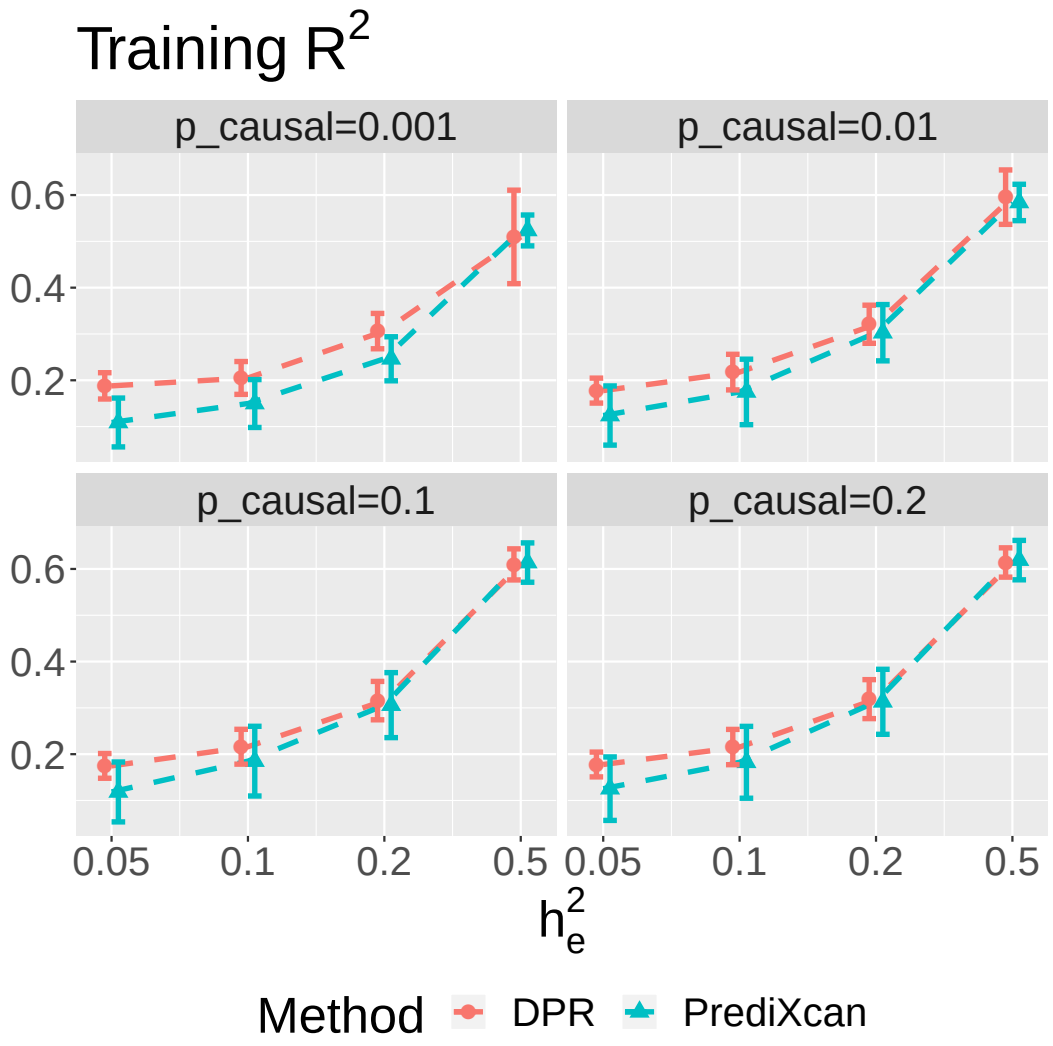
1 Supplemental Figures

Figure S 1: Linkage disequilibrium block structure for *ABCA7*.



Each plot represents a non-overlapped region of the genotype data used in our simulations studies.

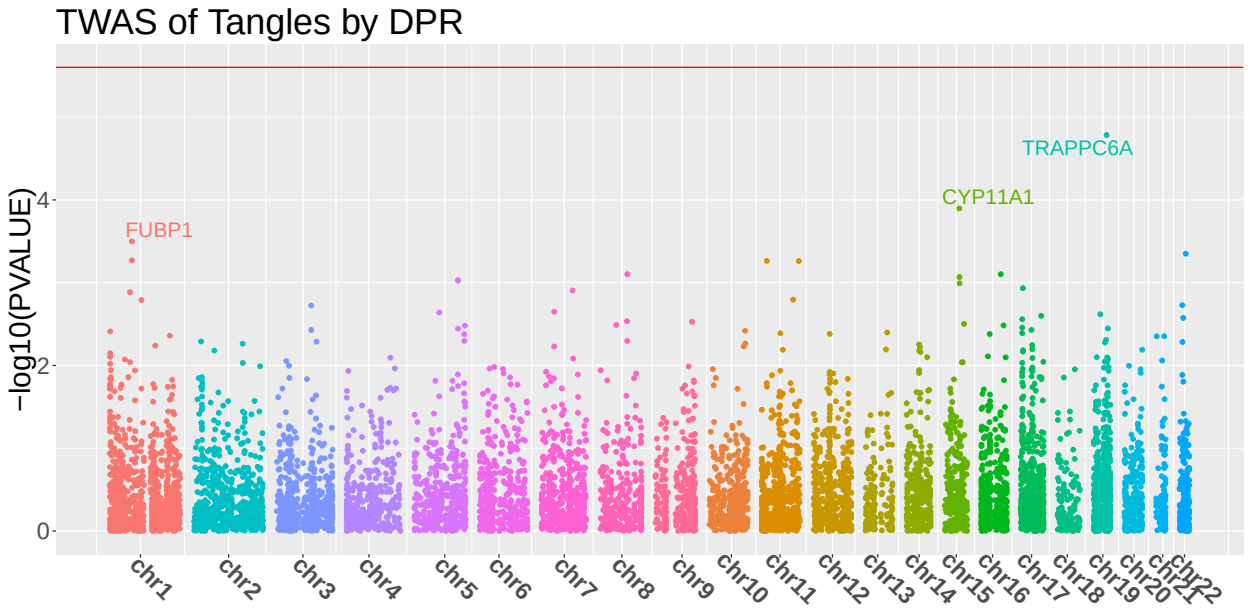
Figure S 2: Training R^2 by DPR and PrediXcan in simulation studies.



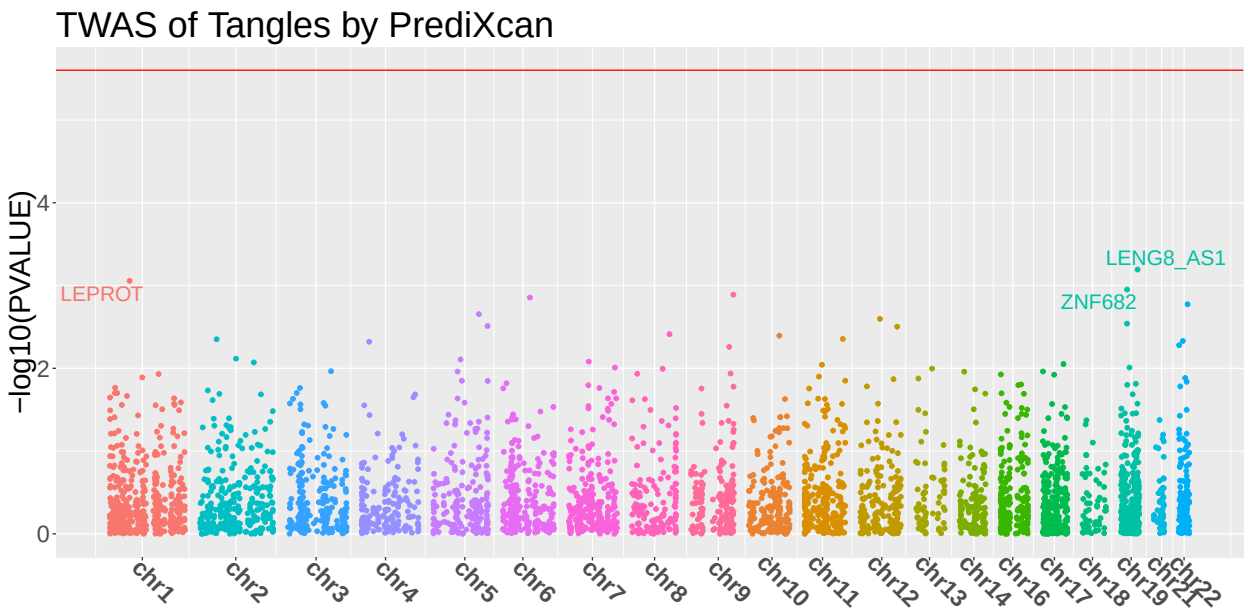
Under various simulation scenarios with the proportions of true causal SNPs $p_{causal} = (0.001, 0.01, 0.1, 0.2)$ and expression heritability $h^2_e = (0.05, 0.1, 0.2, 0.5)$.

Figure S 3: Manhattan plots for TWAS of neurofibrillary tangle density (Tangles).

(A)



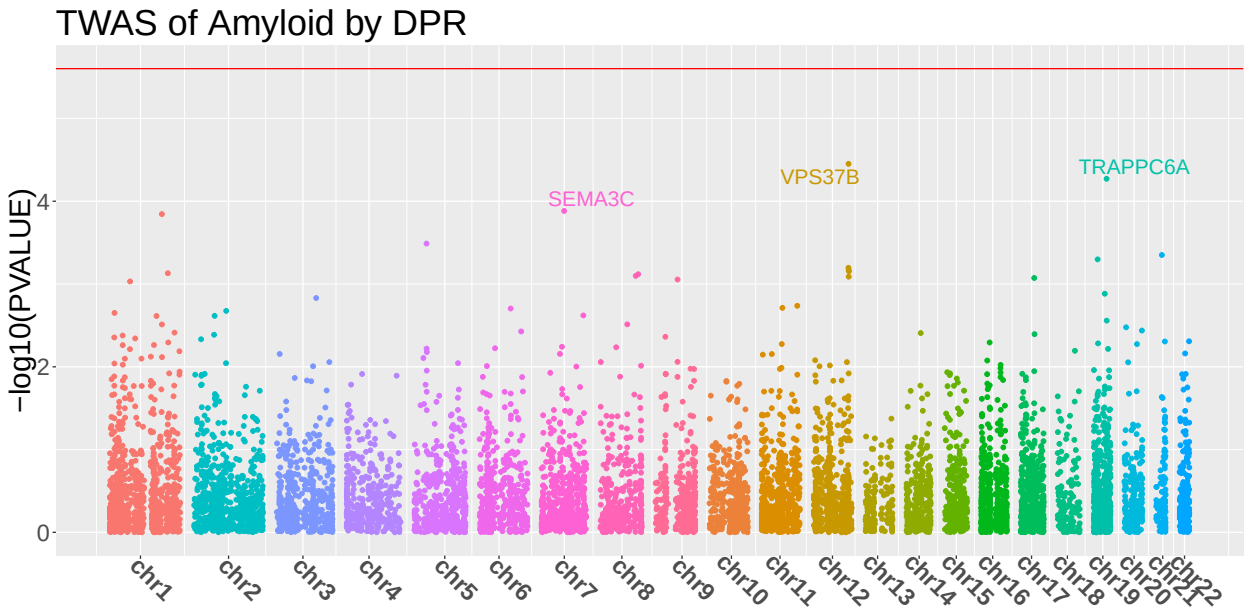
(B)



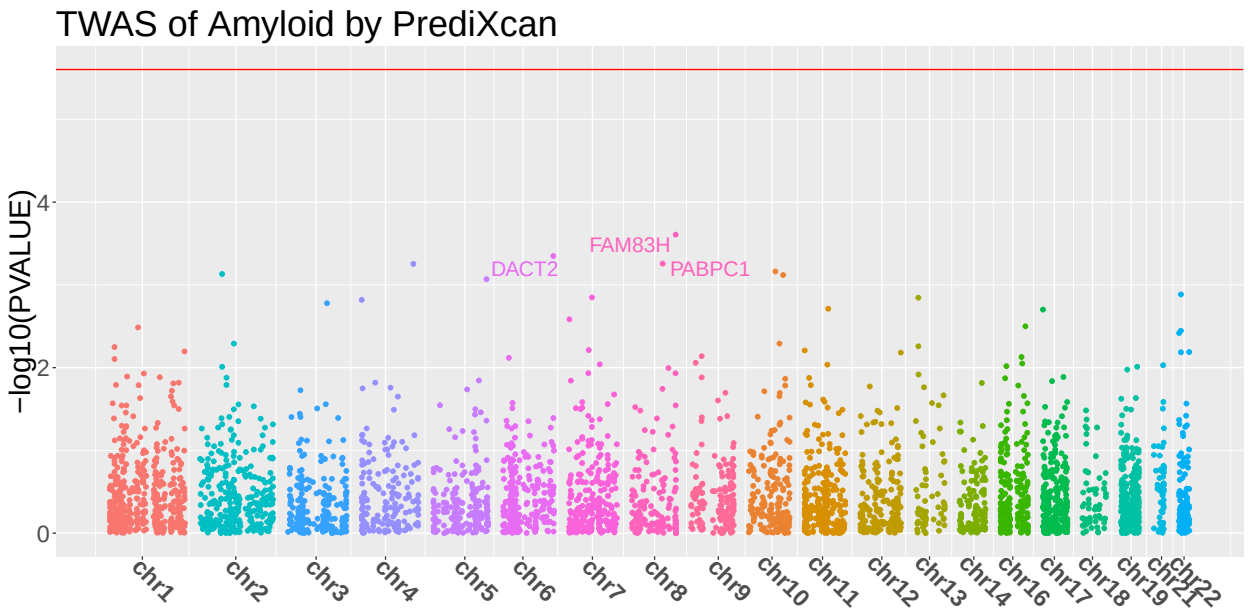
Using individual-level ROS/MAP data and imputation models by DPR (A) and PrediXcan (B).

Figure S 4: Manhattan plots for TWAS of β-amyloid load (Amyloid).

(A)

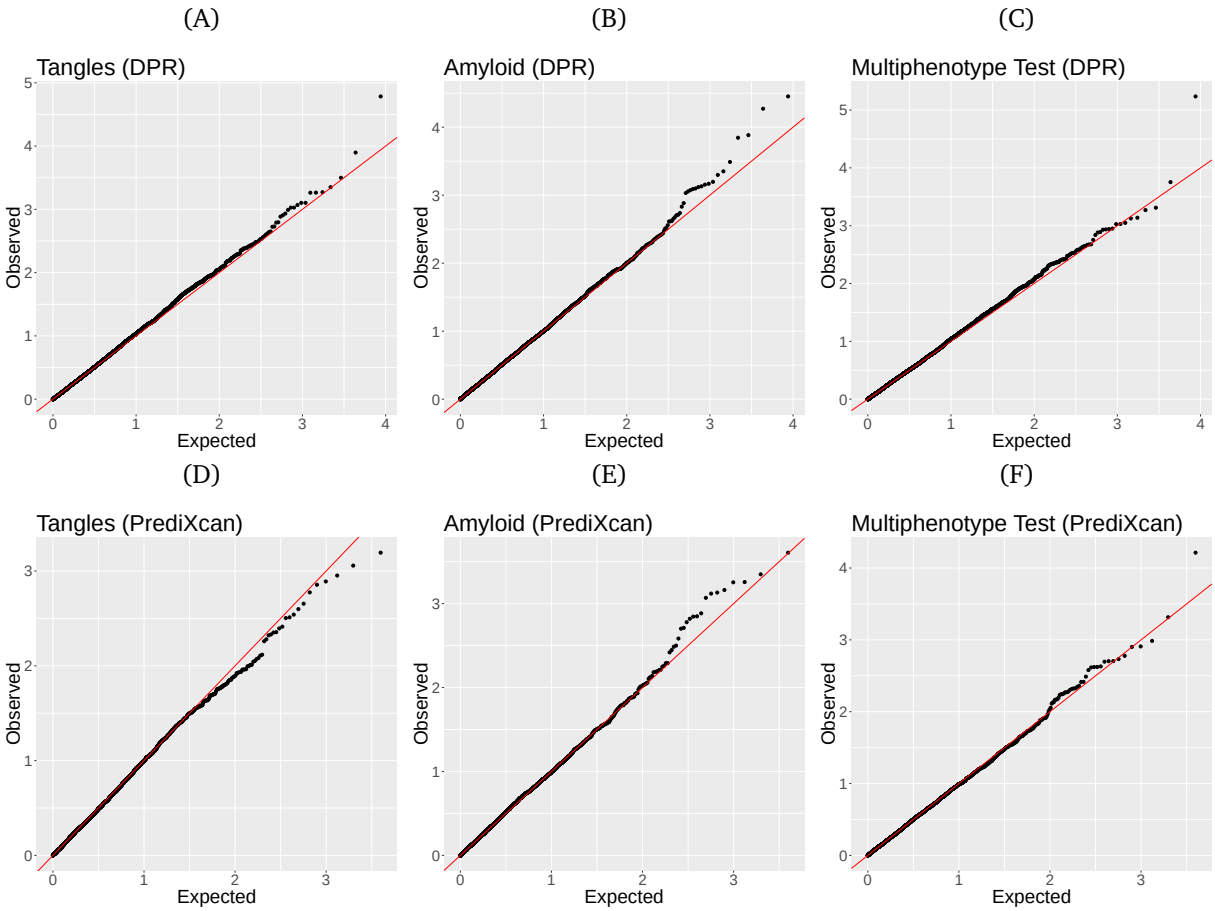


(B)



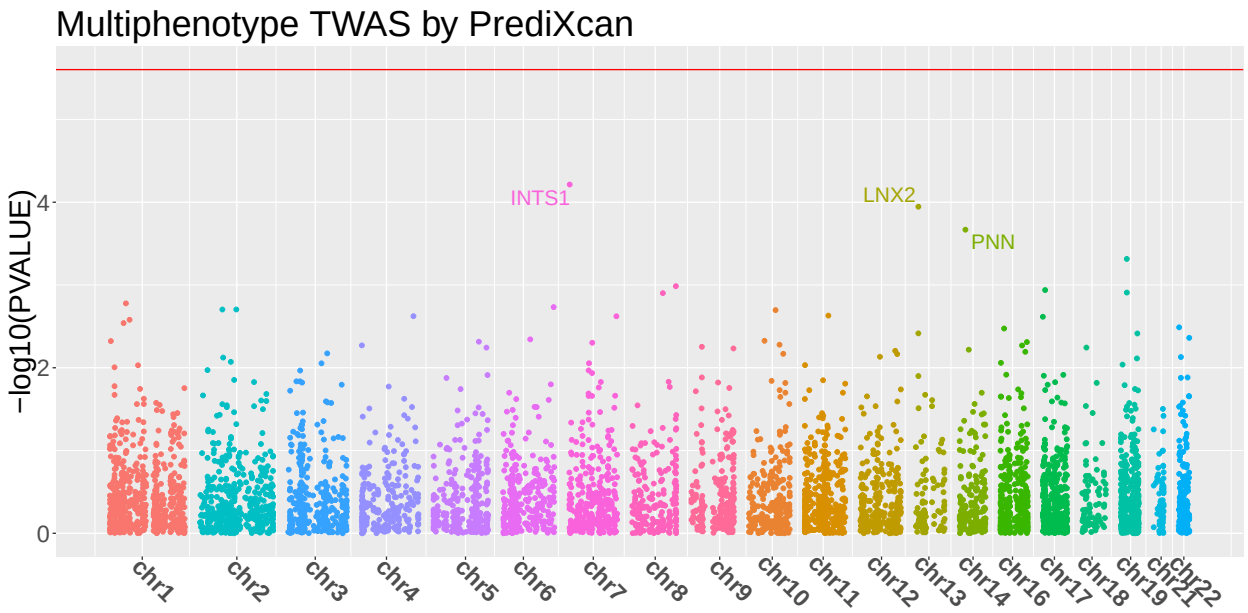
Using individual-level ROS/MAP data and imputation models by DPR (A) and PrediXcan (B).

Figure S 5: Quantile-Quantile (QQ) plots for univariate and multivariate TWAS.



With phenotypes of Tangles and Amyloid using individual-level ROS/MAP data and imputation models by DPR (A, B, C) and PrediXcan (D, E, F).

Figure S 6: Manhattan plot for multiphenotype TWAS of Tangles and Amyloid.



Using individual-level ROS/MAP data by PrediXcan.

2 Supplemental Method

2.1 Nonparametric Bayesian Dirichlet Process Regression Model

Consider the following additive linear regression model for estimating the cis-eQTL effect-sizes,

$$\mathbf{E_g} = \mathbf{X_{n\times p}} \mathbf{w_{p\times 1}} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma_\epsilon^{\mathbf{2}}\mathbf{I}),\ \sigma_\epsilon^{\mathbf{2}} \sim \mathbf{IG}(\mathbf{a}_\epsilon, \mathbf{b}_\epsilon), \tag{1}$$

where vector $\mathbf{E_g}$ denotes the gene expression levels for gene g, $\boldsymbol{X_{n\times p}}$ denotes the genotype matrix for n samples and p cis-SNPs (encoded as the number of minor alleles), $\boldsymbol{w}$ denotes the corresponding cis-eQTL effect-size vector, vector $\boldsymbol{\varepsilon}$ denotes the error term, and $\boldsymbol{I}$ denotes an identity matrix. The intercept term is dropped in model (1) for assuming both $\mathbf{E_g}$ and $\boldsymbol{X}$ are centered at 0. The error variance σ_ϵ^2 is assumed with an Inverse Gamma (IG) prior distribution.

Following the latent Dirichlet process regression (DPR) model [1], we assume a normal prior $N(0, \sigma_\epsilon^2\sigma_w^2)$ for the cis-eQTL effect-sizes and a Dirichlet process (DP) prior [2] for σ_w^2. That is,

$$w_i \sim N(0, \sigma_\epsilon^2\sigma_w^2),\ \sigma_w^2 \sim D,\ D \sim DP(IG(a,b), \xi),\ i = 1, \cdots, p, \tag{2}$$

where the prior distribution D deviates from the DP with base distribution as an Inverse Gamma distribution and concentration parameter ξ. In particular, different from the notation in the main text, the prior effect-size variance is assumed to be scaled by the inverse of the error variance $(\sigma_\epsilon^2)^{-1}$ for computational simplicity. Here, σ_w^2 can be viewed as a latent variable and integrating out σ_w^2 will induce a nonparametric prior distribution on w_i, which is equivalent to the following DP normal mixture model [3, 4],

$$\begin{aligned} w_i &\sim \pi_0 N(0, \sigma_\epsilon^2\sigma_0^2) + \sum_{k=1}^{+\infty} \pi_k N(0, \sigma_\epsilon^2(\sigma_k^2 + \sigma_0^2)),\ \pi_k = \nu_k \prod_{l=0}^{k-1}(1-\nu_l); \\ \nu_k &\sim Beta(1, \xi),\ \xi \sim Gamma(a_\xi, b_\xi),\ \sigma_k^2 \sim IG(a_k, b_k),\ k = 0, 1, \cdots, +\infty. \end{aligned} \tag{3}$$

Here, the nonparametric prior distribution on w_i is equivalently represented by a mixture normal prior that is a weighted sum of an infinitely number of zero-mean normal distributions. The variance terms of these normal distributions are assumed to be scaled by the inverse of the error variance $(\sigma_\epsilon^2)^{-1}$ and centered by $-\sigma_0^2$ for those corresponding to $k > 0$. Note that the σ_0^2 can also be viewed as the smallest variance component, such that other variance components can be written as $(\sigma_k^2 + \sigma_0^2)$. The weight parameter π_k is determined by $\{\nu_l, l = 0, \cdots, k\}$ with a Beta prior, where the parameter ξ determines the number of components with non-zero weights (ξ is the same concentration parameter as in (2)). Generally, the hyper parameters $\{a_k, b_k, a_\epsilon, b_\epsilon\}$ in the Inverse Gamma distributions can be set as 0.1 and (a_ξ, b_ξ) in the Gamma distribution can be set as $(1, 0.1)$ to induce non-informative priors for $(\sigma_k^2, \sigma_\epsilon^2, \xi)$.

For computational convenience, following previous Bayesian models [1, 5], we group the effect-sizes corresponding to the first normal component that has the smallest variance term, $N(0, \sigma_\epsilon^2\sigma_0^2)$,

as a random effect term $\boldsymbol{u}$. Then the above model (1, 3) is equivalent to

$$\begin{aligned}
\mathbf{E_g} &= \mathbf{X}\widetilde{\mathbf{w}} + \mathbf{u} + \boldsymbol{\varepsilon};\ \mathbf{u} \sim \mathbf{N}(\mathbf{0}, \sigma_\epsilon^{\mathbf{2}}\sigma_{\mathbf{0}}^{\mathbf{2}}\mathbf{K}),\ \boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma_\epsilon^{\mathbf{2}}\mathbf{I}),\ \sigma_\epsilon^{\mathbf{2}} \sim \mathbf{IG}(\mathbf{a}_\epsilon, \mathbf{b}_\epsilon); \\
\widetilde{w_i} &= \sum_{k=1}^{+\infty} \pi_k N(0, \sigma_\epsilon^2\sigma_k^2),\ \pi_k = \nu_k \prod_{l=0}^{k-1}(1-\nu_l),\ k = 1, \cdots, +\infty; \\
\nu_k &\sim Beta(1, \xi),\ \xi \sim Gamma(a_\xi, b_\xi);\ \sigma_k^2 \sim IG(a_k, b_k);\ k = 0, 1, , \cdots, +\infty;
\end{aligned} \tag{4}$$

where $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}'/p$ is the Genetic Relatedness Matrix (GRM) [1, 5]. Note that the random effect term can be written as

$$\boldsymbol{u} = \boldsymbol{X}\boldsymbol{\zeta};\ \zeta_i \sim N(0, \sigma_\epsilon^2\sigma_0^2/p),\ i = 1, ..., p; \tag{5}$$

where ζ_i denotes the random effect-size for SNP i. Our computational algorithms will be derived with respect to model (4) for estimating $(\widetilde{\boldsymbol{w}}, \boldsymbol{\zeta})$, which will then give estimates for the cis-eQTL effect-sizes $\boldsymbol{w} = \widetilde{\boldsymbol{w}} + \boldsymbol{\zeta}$ as in model (1).

Based on model (4), the joint conditional posterior function for all parameters in the model is given by

$$\begin{gathered}
P(\widetilde{\boldsymbol{w}}, \boldsymbol{u}, \boldsymbol{\nu}, \xi, \sigma_\epsilon^2, \sigma_0^2, \sigma_1^2, \cdots, \sigma_k^2, \cdots | \boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K}) \propto \\
P(\boldsymbol{E_g}|\boldsymbol{X}, \boldsymbol{K}, \widetilde{\boldsymbol{w}}, \boldsymbol{u}, \sigma_\epsilon^2) P(\boldsymbol{w}|\boldsymbol{\nu}, \sigma_1^2, \cdots, \sigma_k^2, \cdots) P(\boldsymbol{u}|\sigma_\epsilon^2, \sigma_0^2, \boldsymbol{K}) \\
\left(\prod_{k=0}^{+\infty} P(\sigma_k^2|a_k, b_k)\right) P(\boldsymbol{\nu}|\xi) P(\xi|a_\xi, b_\xi) P(\sigma_\epsilon^2|a_\epsilon, b_\epsilon).
\end{gathered} \tag{6}$$

One convenience for considering model (4) is that one can integrate out the random effect term $\boldsymbol{u}$ from (6) and then implement Gibbs sampling [6] for improved mixing in the Markov Chain Monte Carlo (MCMC) sampling. Specifically,

$$\begin{aligned}
P(\widetilde{\boldsymbol{w}}, \boldsymbol{\nu}, \xi, \sigma_\epsilon^2, \sigma_0^2, \sigma_1^2, \cdots, \sigma_k^2, \cdots | \boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K}) &= \int P(\widetilde{\boldsymbol{w}}, \boldsymbol{u}, \boldsymbol{\nu}, \xi, \sigma_\epsilon^2, \sigma_0^2, \sigma_1^2, \cdots, \sigma_k^2, \cdots | \boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K}) d\boldsymbol{u} \\
&= \int P(\boldsymbol{E_g}|\boldsymbol{X}, \boldsymbol{K}, \widetilde{\boldsymbol{w}}, \boldsymbol{u}, \sigma_\epsilon^2) P(\boldsymbol{u}|\sigma_\epsilon^2, \sigma_0^2, \boldsymbol{K}) d\boldsymbol{u} \times \\
&P(\boldsymbol{w}|\boldsymbol{\nu}, \sigma_1^2, \cdots, \sigma_k^2, \cdots) \left(\prod_{k=0}^{+\infty} P(\sigma_k^2|a_k, b_k)\right) P(\boldsymbol{\nu}|\xi) P(\xi|a_\xi, b_\xi) P(\sigma_\epsilon^2|a_\epsilon, b_\epsilon),
\end{aligned} \tag{7}$$

where

$$\begin{aligned}
\int P(\boldsymbol{E_g}|\boldsymbol{X}, \boldsymbol{K}, \widetilde{\boldsymbol{w}}, \boldsymbol{u}, \sigma_\epsilon^2) P(\boldsymbol{u}|\sigma_\epsilon^2, \sigma_0^2, \boldsymbol{K}) du &\propto |\sigma_\epsilon^2 \boldsymbol{H}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})' \boldsymbol{H}^{-1} (\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})\right\}, \\
\boldsymbol{H} &= \boldsymbol{I} + \sigma_0^2 \boldsymbol{K}.
\end{aligned}$$

The following MCMC Sampling and Variational Inference algorithms will be derived based on the joint conditional posterior density function (7).

2.2 MCMC Sampling

To facilitate MCMC sampling, for each SNP i, we assume a corresponding indicator vector $\boldsymbol{\gamma_i} = \{\gamma_{ik} \in \{0, 1\}, k = 1, ..., +\infty\}$ that indicates if the kth normal component contributes to the effect-size distribution. Consequently, γ_{ik} has a $Bernoulli(\pi_k)$ prior. Let $\boldsymbol{\gamma}$ denote the indicator vectors

$\{\boldsymbol{\gamma_i}, i = 1, \cdots, p\}$. Consequently, the joint conditional posterior density function (7) becomes

$$
\begin{gathered}
P(\widetilde{\boldsymbol{w}}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \xi, \sigma_\epsilon^2, \sigma_0^2, \sigma_1^2, \cdots, \sigma_k^2, \cdots | \boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K}) \propto \\
|\sigma_\epsilon^2 \boldsymbol{H}|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2\sigma_\epsilon^2}(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})' \boldsymbol{H}^{-1} (\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})\right\} \times \\
P(\boldsymbol{w}|\boldsymbol{\gamma}, \sigma_\epsilon^2, \sigma_1^2, \cdots, \sigma_k^2, \cdots) \left(\prod_{k=0}^{+\infty} P(\sigma_k^2 | a_k, b_k)\right) P(\boldsymbol{\gamma}|\boldsymbol{\nu}) P(\boldsymbol{\nu}|\xi) P(\xi|a_\xi, b_\xi) P(\sigma_\epsilon^2 | a_\epsilon, b_\epsilon) \propto \\
|\sigma_\epsilon^2 \boldsymbol{H}|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2\sigma_\epsilon^2}(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})' \boldsymbol{H}^{-1} (\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})\right\} \times \\
\left(\prod_{i=1}^{p} \prod_{k=1}^{+\infty} \gamma_{ik} N(\widetilde{w_{ik}}; 0, \sigma_\epsilon^2 \sigma_k^2)\right) \left(\prod_{k=0}^{+\infty} IG(\sigma_k^2; a_k, b_k)\right) \times \\
\left(\prod_{i=1}^{p} \prod_{k=1}^{+\infty} Bernoulli(\gamma_{ik}; \pi_k = \nu_k \prod_{l=0}^{k-1}(1 - \nu_l))\right) \left(\prod_{k=0}^{+\infty} Beta(\nu_k; 1, \xi)\right) \times \\
Gamma(\xi; a_\xi, b_\xi) IG(\sigma_\epsilon^2; a_\epsilon, b_\epsilon). \qquad (8)
\end{gathered}
$$

Denoting $\widetilde{w_i} = \sum_{k=1}^{+\infty} \gamma_{ik} \widetilde{w_{ik}}$, $\widetilde{w_{ik}} \sim N(0, \sigma_\epsilon^2 \sigma_k^2)$, Then the log joint conditional posterior density function is given by

$$
\begin{gathered}
log(P(\widetilde{\boldsymbol{w}}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \xi, \sigma_\epsilon^2, \sigma_0^2, \sigma_1^2, \cdots, \sigma_k^2, \cdots | \boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K})) = \\
C - \tfrac{1}{2} log|\sigma_\epsilon^2 \boldsymbol{H}| - \tfrac{1}{2\sigma_\epsilon^2}(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})' \boldsymbol{H}^{-1} (\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}}) + \\
\sum_{i=1}^{p} \sum_{k=1}^{+\infty} \gamma_{ik} \left[-\tfrac{1}{2} log(\sigma_\epsilon^2) - \tfrac{1}{2} log(\sigma_k^2) - \tfrac{\widetilde{w_{ik}}^2}{2\sigma_\epsilon^2 \sigma_k^2}\right] + \sum_{k=0}^{+\infty} \left[-(a_k + 1) log(\sigma_k^2) - \tfrac{b_k}{\sigma_k^2}\right] + \\
\sum_{i=1}^{p} \sum_{k=1}^{+\infty} \left[\gamma_{ik} \left(log(\nu_k) + \sum_{l=0}^{k-1} log(1 - \nu_l)\right) + (1 - \gamma_{ik}) log\left(1 - \nu_k \prod_{l=0}^{k-1}(1 - \nu_l)\right)\right] + \\
\sum_{k=0}^{+\infty} \left[(\xi - 1) log(1 - \nu_k) + log(\xi)\right] + (a_\xi - 1) log(\xi) - b_\xi \xi - (a_\epsilon + 1) log(\sigma_\epsilon^2) - \tfrac{b_\epsilon}{\sigma_\epsilon^2}, \qquad (9)
\end{gathered}
$$

where C denotes a normalization constant that is free of parameters. Based on the log conditional posterior density function (9), the following MCMC sampling scheme is derived for obtaining the posterior estimates for $\widetilde{\boldsymbol{w}}$.

2.2.1 Gibbs Sampling Scheme

From (9), for each parameter in the model, we derive the log conditional density function conditioning on other parameters as follows:

- $\widetilde{w_{ik}}$ **and** γ_{ik}
  The log joint conditional density function for $\widetilde{w_{ik}}, \gamma_{ik}$ is given by,

$$
\begin{gathered}
log(P(\widetilde{w_{ik}}, \gamma_{ik} | \cdot)) = C - \tfrac{\boldsymbol{x_i} \boldsymbol{H^{-1}} \boldsymbol{x_i}}{2\sigma_\epsilon^2} (\widetilde{w_{ik}}^2 + 2\widetilde{w_{ik}} \widetilde{w}_{i(-k)}) + \tfrac{1}{\sigma_\epsilon^2} \boldsymbol{x_i'} \boldsymbol{H}^{-1} (\boldsymbol{E_g} - \textstyle\sum_{j \neq i} \boldsymbol{x_j} \widetilde{w_j}) \widetilde{w_{ik}} + \\
\gamma_{ik} \left[-\tfrac{1}{2} log(\sigma_\epsilon^2) - \tfrac{1}{2} log(\sigma_k^2) - \tfrac{\widetilde{w_{ik}}^2}{2\sigma_\epsilon^2 \sigma_k^2}\right] + \\
\gamma_{ik} \left(log(\nu_k) + \textstyle\sum_{l=0}^{k-1} log(1 - \nu_l)\right) + (1 - \gamma_{ik}) log\left(1 - \nu_k \textstyle\prod_{l=0}^{k-1}(1 - \nu_l)\right), \qquad (10)
\end{gathered}
$$

  where $\widetilde{w}_{i(-k)} = \sum_{l \neq k} \widetilde{w}_{il}$.

Then the log conditional posterior distribution for $(\widetilde{w_{ik}}|\gamma_{ik}=1,\cdot)$ is given by

$$\begin{aligned}
log(P(\widetilde{w_{ik}}|\gamma_{ik}=1,\cdot)) &= C-\frac{\boldsymbol{x_i}\boldsymbol{H^{-1}}\boldsymbol{x_i}}{2\sigma_\epsilon^2}\widetilde{w_{ik}}^2-\frac{\widetilde{w_{ik}}^2}{2\sigma_\epsilon^2\sigma_k^2}+\frac{1}{\sigma_\epsilon^2}\boldsymbol{x_i'}\boldsymbol{H}^{-1}(\boldsymbol{E_g}-\sum_{j\neq i}\boldsymbol{x_j}\widetilde{w_j}-\boldsymbol{x_i}\widetilde{w}_{i(-k)})\widetilde{w_{ik}} \\
&= C-\frac{\boldsymbol{x_i}\boldsymbol{H^{-1}}\boldsymbol{x_i}+\sigma_k^{-2}}{2\sigma_\epsilon^2}\widetilde{w_{ik}}^2+\frac{1}{\sigma_\epsilon^2}\boldsymbol{x_i'}\boldsymbol{H}^{-1}(\boldsymbol{E_g}-\sum_{j\neq i}\boldsymbol{x_j}\widetilde{w_j}-\boldsymbol{x_i}\widetilde{w}_{i(-k)})\widetilde{w_{ik}}; \\
P(\widetilde{w_{ik}}|\gamma_{ik}=1,\cdot) &\sim N(m_{ik},s_{ik}^2), \\
m_{ik} &= \frac{\boldsymbol{x_i'}\boldsymbol{H}^{-1}(\boldsymbol{E_g}-\sum_{j\neq i}\boldsymbol{x_j}\widetilde{w_j}-\boldsymbol{x_i}\widetilde{w}_{i(-k)})}{\boldsymbol{x_i}\boldsymbol{H^{-1}}\boldsymbol{x_i}+\sigma_k^{-2}}, \\
s_{ik}^2 &= \frac{\sigma_\epsilon^2}{\boldsymbol{x_i}\boldsymbol{H^{-1}}\boldsymbol{x_i}+\sigma_k^{-2}}. \qquad (11)
\end{aligned}$$

After integrating out $\widetilde{w_{ik}}$ from (10), the conditional probability for $(\gamma_{ik}=1|\cdot)$ is given by

$$\begin{aligned}
P(\gamma_{ik}=1|\cdot) &= \pi_{ik}=\int P(\widetilde{w_{ik}},\gamma_{ik}=1|\cdot)d\widetilde{w_{ik}}\propto \\
&\exp\left\{-\frac{1}{2}log(\sigma_\epsilon^2)-\frac{1}{2}log(\sigma_k^2)+log(\nu_k)+\sum_{l=0}^{k-1}log(1-\nu_l)\right\}\times \\
&\int\exp\left\{-\frac{\boldsymbol{x_i}\boldsymbol{H^{-1}}\boldsymbol{x_i}+\sigma_k^{-2}}{2\sigma_\epsilon^2}\widetilde{w_{ik}}^2+\frac{1}{\sigma_\epsilon^2}\boldsymbol{x_i'}\boldsymbol{H}^{-1}(\boldsymbol{E_g}-\sum_{j\neq i}\boldsymbol{x_j}\widetilde{w_j}-\boldsymbol{x_i}\widetilde{w}_{i(-k)})\widetilde{w_{ik}}\right\}d\widetilde{w_{ik}}\propto \\
&\exp\left\{-\frac{1}{2}log(\sigma_\epsilon^2)-\frac{1}{2}log(\sigma_k^2)+log(\nu_k)+\sum_{l=0}^{k-1}log(1-\nu_l)\right\}\sqrt{2\pi s_k^2}\exp\{\frac{m_{ik}^2}{2s_k^2}\}\times \\
&\int\frac{1}{\sqrt{2\pi s_k^2}}\exp\left\{-\frac{1}{2s_{ik}^2}\left(\widetilde{w_{ik}}^2-2m_{ik}\widetilde{w_{ik}}+m_{ik}^2\right)\right\}d\widetilde{w_{ik}}\propto \\
&\sqrt{2\pi s_k^2}\exp\left\{\frac{m_{ik}^2}{2s_k^2}-\frac{1}{2}log(\sigma_\epsilon^2)-\frac{1}{2}log(\sigma_k^2)+log(\nu_k)+\sum_{l=0}^{k-1}log(1-\nu_l)\right\}\propto \\
&\exp\left\{\frac{m_{ik}^2}{2s_k^2}+log(s_k)-log(\sigma_\epsilon)-log(\sigma_k)+log(\nu_k)+\sum_{l=0}^{k-1}log(1-\nu_l)\right\}, \qquad (12)
\end{aligned}$$

where $\int\frac{1}{\sqrt{2\pi s_k^2}}\exp\left\{-\frac{1}{2s_{ik}^2}\left(\widetilde{w_{ik}}^2-2m_{ik}\widetilde{w_{ik}}+m_{ik}^2\right)\right\}d\widetilde{w_{ik}}=1$ because the integrand is just the density function of $N(\widetilde{w_{ik}};m_{ik},s_k^2)$.

- ν_k

The log conditional density function for ν_k is given by,

$$\begin{aligned}
log(P(\nu_k|\cdot)) &= \sum_{i=1}^{p}\sum_{k=1}^{+\infty}\left[\gamma_{ik}\left(log(\nu_k)+\sum_{l=0}^{k-1}log(1-\nu_l)\right)+(1-\gamma_{ik})log\left(1-\nu_k\prod_{l=0}^{k-1}(1-\nu_l)\right)\right]+ \\
&\quad \sum_{k=1}^{+\infty}(\xi-1)log(1-\nu_k)+C, \\
&= \sum_{i=1}^{p}\gamma_{ik}log(\nu_k)+\sum_{i=1}^{p}\sum_{l=k+1}^{+\infty}\gamma_{ik}log(1-\nu_k)+ \\
&\quad \sum_{i=1}^{p}\sum_{j=k}^{+\infty}(1-\gamma_{ij})log\left(1-\nu_j\prod_{l=0}^{j-1}(1-\nu_l)\right)+(\xi-1)log(1-\nu_k)+C, \\
&\approx \sum_{i=1}^{p}\gamma_{ik}log(\nu_k)+\sum_{i=1}^{p}\sum_{l=k+1}^{+\infty}\gamma_{ik}log(1-\nu_k)+(\xi-1)log(1-\nu_k)+C, \\
P(\nu_k|\cdot) &\sim Beta(\kappa_k,\xi_k),\ \kappa_k=\sum_{i=1}^{p}\gamma_{ik}+1,\ \xi_k=\sum_{i=1}^{p}\sum_{l=k+1}^{+\infty}\gamma_{il}+\xi. \qquad (13)
\end{aligned}$$

- $\sigma_k^2, k>0$
  The log conditional density function for $\sigma_k^2, k>0$ is given by,

$$\begin{aligned}
log(P(\sigma_k^2|\cdot)) &= C+\sum_{i=1}^{p}\sum_{k=1}^{+\infty}\gamma_{ik}\left[-\frac{1}{2}log(\sigma_k^2)-\frac{\widetilde{w_{ik}}^2}{2\sigma_\epsilon^2\sigma_k^2}\right]+\sum_{k=0}^{+\infty}\left[-(a_k+1)log(\sigma_k^2)-\frac{b_k}{\sigma_k^2}\right] \\
&= C+\sum_{i=1}^{p}\gamma_{ik}\left[-\frac{1}{2}log(\sigma_k^2)-\frac{\widetilde{w_{ik}}^2}{2\sigma_\epsilon^2\sigma_k^2}\right]-(a_k+1)log(\sigma_k^2)-\frac{b_k}{\sigma_k^2} \\
&= C-\left(\frac{1}{2}\sum_{i=1}^{p}\gamma_{ik}+a_k+1\right)log(\sigma_k^2)-\left(\frac{\sum_{i=1}^{p}(\gamma_{ik}\widetilde{w_{ik}}^2)}{2\sigma_\epsilon^2}+b_k\right)\frac{1}{\sigma_k^2}; \\
P(\sigma_k^2|\cdot) &\sim IG(\widetilde{a_k},\widetilde{b_k}),\ \widetilde{a_k}=\frac{1}{2}\sum_{i=1}^{p}\gamma_{ik}+a_k,\ \widetilde{b_k}=\frac{\sum_{i=1}^{p}(\gamma_{ik}\widetilde{w_{ik}}^2)}{2\sigma_\epsilon^2}+b_k. \qquad (14)
\end{aligned}$$

- ξ The log conditional density function for ξ is given by,

$$\begin{aligned}
log(P(\xi|\cdot)) &= C+\sum_{k=0}^{+\infty}[(\xi-1)log(1-\nu_k)+log(\xi)]+(a_\xi-1)log(\xi)-b_\xi\xi \\
&= C+\left(a_\xi+\sum_{k=0}^{+\infty}1_k-1\right)log(\xi)-\left(b_\xi-\sum_{k=0}^{+\infty}log(1-\nu_k)\right)\xi; \\
(P(\xi|\cdot)) &\sim Gamma(\widetilde{a_\xi},\widetilde{b_\xi}),\ \widetilde{a_\xi}=a_\xi+\sum_{k=0}^{+\infty}1_k,\ \widetilde{b_\xi}=b_\xi-\sum_{k=0}^{+\infty}log(1-\nu_k). \qquad (15)
\end{aligned}$$

- σ_ϵ^2

  The log conditional density function for σ_ϵ^2 is given by,

$$\begin{aligned} log(P(\sigma_\epsilon^2|\cdot)) &= C - \frac{1}{2}log|\sigma_\epsilon^2 \boldsymbol{H}| - \frac{1}{2\sigma_\epsilon^2}(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})'\boldsymbol{H}^{-1}(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}}) + \\ &\quad \sum_{i=1}^{p}\sum_{k=1}^{+\infty}\gamma_{ik}\left[-\frac{1}{2}log(\sigma_\epsilon^2) - \frac{\widetilde{w_{ik}}^2}{2\sigma_\epsilon^2\sigma_k^2}\right] - (a_\epsilon+1)log(\sigma_\epsilon^2) - \frac{b_\epsilon}{\sigma_\epsilon^2} \\ &= C - log(\sigma_\epsilon^2)\left(\frac{n}{2} + \frac{1}{2}\sum_{i=1}^{p}\sum_{k=1}^{+\infty}\gamma_{ik} + a_\epsilon + 1\right) - \\ &\quad \frac{1}{\sigma_\epsilon^2}\left(\frac{1}{2}(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})'\boldsymbol{H}^{-1}(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}}) + \sum_{i=1}^{p}\sum_{k=1}^{+\infty}\gamma_{ik}\frac{\widetilde{w_{ik}}^2}{2\sigma_k^2} + b_\epsilon\right); \\ P(\sigma_\epsilon^2|\cdot) &\sim IG(\widetilde{a_\epsilon}, \widetilde{b_\epsilon}), \qquad (16) \\ \widetilde{a_\epsilon} &= \frac{n}{2} + \frac{1}{2}\sum_{i=1}^{p}\sum_{k=1}^{+\infty}\gamma_{ik} + a_\epsilon, \\ \widetilde{b_\epsilon} &= \frac{1}{2}(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})'\boldsymbol{H}^{-1}(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}}) + \sum_{i=1}^{p}\sum_{k=1}^{+\infty}\gamma_{ik}\frac{\widetilde{w_{ik}}^2}{2\sigma_k^2} + b_\epsilon. \end{aligned}$$

- σ_0^2

  The log conditional density function for σ_0^2 is given by,

$$\begin{aligned} log(P(\sigma_0^2|\cdot)) &= C - \frac{1}{2}log|\boldsymbol{H}| - \frac{1}{2\sigma_\epsilon^2}(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})'\boldsymbol{H}^{-1}(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}}) - (a_0+1)log(\sigma_0^2) - \frac{b_0}{\sigma_0^2}, \\ \boldsymbol{H} &= \boldsymbol{I} + \sigma_0^2\boldsymbol{K}. \qquad (17) \end{aligned}$$

  Because the conditional density function (17) is of an unknown distribution, Metropolis-Hastings algorithm is needed to generate posterior samples for σ_0^2. For improved mixing property, we will re-parametrize σ_0^2 to $h^2 = \frac{\sigma_0^2}{\sigma_0^2+1}$, such that h^2 has domain $[0,1]$. Based on the probability density function for change of variables $P(h^2|\cdot) = \frac{d\sigma_0^2(h^2)}{dh^2}P(\sigma_0^2|\cdot)$, the log conditional density function for h^2 is given by

$$log(P(h^2|\cdot)) = log(\sigma_0^2(h^2)|\cdot) - 2log(1-h^2), \qquad (18)$$

  where $log(\sigma_0^2(h^2)|\cdot)$ is given by (17) with $\sigma_0^2(h^2) = \frac{h^2}{1-h^2}$. In the MCMC sampling, we will first sample h^2 based on (18) and then obtain the σ_0^2 sample from $\sigma_0^2 = \frac{h^2}{1-h^2}$.

- The above conditional posterior density functions will be used in the Gibbs sampling algorithm to generate posterior samples with respect to each parameter in turn. As a result, the average of the posterior samples will be taken as the posterior Bayesian estimate for the corresponding parameter.

2.2.2 Estimate $\boldsymbol{\zeta}$

Recall that the random effect term $\boldsymbol{u}$ can be represented by the random effect-size vector $\boldsymbol{\zeta}$ as in (5). The posterior conditional distribution for $\boldsymbol{\zeta}$ is given by

$$
\begin{aligned}
P(\boldsymbol{\zeta}|\cdot) &\propto P(\boldsymbol{E_g}|\boldsymbol{X}, \widetilde{\boldsymbol{w}}, \boldsymbol{\zeta}, \sigma_\epsilon^2) P(\boldsymbol{\zeta}|\sigma_\epsilon^2, \sigma_0^2) \\
&\propto MVN\left(\boldsymbol{E_g};\ \boldsymbol{X}\widetilde{\boldsymbol{w}} + \boldsymbol{X}\boldsymbol{\zeta},\ \sigma_\epsilon^2 \boldsymbol{I}\right) \times MVN\left(\boldsymbol{\zeta};\ 0,\ \frac{\sigma_\epsilon^2 \sigma_0^2}{p} \boldsymbol{I}\right) \\
&\propto \exp\left\{-\frac{1}{2}\left[-\frac{2}{\sigma_\epsilon^2}\boldsymbol{\zeta}'\boldsymbol{X}'(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}}) + \frac{1}{\sigma_\epsilon^2}\boldsymbol{\zeta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\zeta} + \frac{p}{\sigma_\epsilon^2\sigma_0^2}\boldsymbol{\zeta}'\boldsymbol{\zeta}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\zeta}'\left(\frac{1}{\sigma_\epsilon^2}\boldsymbol{X}'\boldsymbol{X} + \frac{p}{\sigma_\epsilon^2\sigma_0^2}\right)\boldsymbol{\zeta} - 2\boldsymbol{\zeta}'\frac{\boldsymbol{X}'(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})}{\sigma_\epsilon^2}\right]\right\}; \\
P(\boldsymbol{\zeta}|\cdot) &\sim MVN(\mu_\zeta, \Sigma_\zeta), \\
\mu_\zeta &= \frac{\boldsymbol{X}'(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})}{\sigma_\epsilon^2}\Sigma_\zeta = \frac{\sigma_0^2}{p}\boldsymbol{X}'(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}})\boldsymbol{H}^{-1}, \qquad (19) \\
\Sigma_\zeta &= \left(\frac{1}{\sigma_\epsilon^2}\boldsymbol{X}'\boldsymbol{X} + \frac{p}{\sigma_\epsilon^2\sigma_0^2}\right)^{-1} = \frac{\sigma_\epsilon^2\sigma_0^2}{p}\left(\sigma_0^2\boldsymbol{K} + \boldsymbol{I}\right)^{-1} = \frac{\sigma_\epsilon^2\sigma_0^2}{p}\boldsymbol{H}^{-1};\ \boldsymbol{K} = \frac{\boldsymbol{X}'\boldsymbol{X}}{p},\ \boldsymbol{H} = \sigma_0^2\boldsymbol{K} + \boldsymbol{I}.
\end{aligned}
$$

Instead of sampling $\boldsymbol{\zeta}$ in the MCMC algorithm, we take the Row-Blackwell approximation for the conditional posterior mean of $\boldsymbol{\zeta}$ as the posterior estimate. That is,

$$
\widehat{\boldsymbol{\zeta}} = \frac{1}{p}\boldsymbol{X}'\frac{1}{M}\sum_{m=1}^{M}(\sigma_0^2)^{(m)}(\boldsymbol{E_g} - \boldsymbol{X}\widetilde{\boldsymbol{w}}^{(\boldsymbol{m})})(\boldsymbol{H}^{(m)})^{-1}, \qquad (20)
$$

where M denotes the total number of MCMC iterations and (m) denotes the corresponding sample value for the mth MCMC iteration.

2.2.3 Estimate cis-eQTL Effect-sizes w for TWAS

Note that the random effect term u is generally specific for the training samples. That is, the random effect-size ζ is likely to be over-estimated for the training data. We investigated using either the fixed effect-sizes $\sum_{k=1}^{+\infty}\gamma_{ik}\widetilde{w_{ik}}$, or the additive effect-sizes $\sum_{k=1}^{+\infty}\gamma_{ik}\widetilde{w_{ik}} + \widehat{\zeta_i}$ as our cis-eQTL effect-size estimates $\widehat{\boldsymbol{w_i}}$. Our simulation studies show that using only fixed effect-sizes resulted slightly higher prediction R^2 and similar TWAS power in test data. Because our real TWAS analysis with ROS/MAP data includes training samples, we used the additive effect-size estimates to generate TWAS results.

2.2.4 Computation bottleneck for MCMC

To improve the computational efficiency of MCMC sampling, the previous DPR paper [1] utilized multiple techniques, e.g., approximating the infinite normal mixture by a truncated normal mixture

with K normal components as in [3], implementing an eigen decomposition for $\boldsymbol{K} = \boldsymbol{UDU'}$ and transfer the gene expression levels and genotype as $\boldsymbol{U'E_g}, \boldsymbol{U'X}$ such that the random effects become independent across samples, and taking the random scan Gibbs sampling algorithm [7, 8] to prioritize a set of 500 cis-eQTL that have the most significant marginal p-values for associated with $\boldsymbol{E_g}$. However, the MCMC sampling still takes $\sim$10 hours to run $50,000$ iterations per gene ($p \approx 10,000$) for the ROS/MAP data with $n = 499$ samples. Thus, we take the variational inference algorithm [1, 3, 9, 10] to fit the nonparametric Bayesian model (4), and then obtain the Bayesian estimates for cis-eQTL effect-sizes $\boldsymbol{w}$.

2.3 Variational Inference Algorithm

Besides the computational expensive MCMC algorithm, the variational inference algorithm [1, 3, 9, 10] provides an alternative, deterministic methodology for approximating likelihoods and posterior density functions for a Bayesian model such as (4). Following the derivations by [1, 3], we will use a particular class of variational methods known as mean-field methods, which is based on optimizing Kullback-Leibler (KL) divergence with respect to a variational distribution.

Let $\boldsymbol{\theta} = \{\widetilde{\boldsymbol{w}}, \boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{\nu}, \xi, \sigma_\epsilon^2, \sigma_0^2, \sigma_1^2, \cdots, \sigma_k^2, \cdots\}$ denote the parameters of interest from model (4). Let $q(\boldsymbol{\theta}) = \prod_j q(\theta_j)$ denote a variational distribution that assumes independence among parameters $\{\theta_j\}$ and approximates the joint posterior distribution for $\boldsymbol{\theta}$. We aim to identify a variational distribution such that the KL divergence between the variational distribution $q(\boldsymbol{\theta})$ and the joint posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K})$ is minimized. In particular,

$$\begin{aligned} KL(q(\boldsymbol{\theta})|P(\boldsymbol{\theta}|\boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K})) &= E_{q(\boldsymbol{\theta})}\left[log\left(\frac{q(\boldsymbol{\theta})}{P(\boldsymbol{\theta}|\boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K})}\right)\right] \\ &= E_{q(\boldsymbol{\theta})}\left[log(q(\boldsymbol{\theta}))\right] - E_{q(\boldsymbol{\theta})}\left[log(P(\boldsymbol{\theta}, \boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K}))\right] + log(P(\boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K})). \end{aligned} \tag{21}$$

Since $P(\boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K})$ in (21) is independent of $q(\boldsymbol{\theta})$, minimizing the KL divergence (21) is equivalent to maximizing the evidence lower bound (ELBO),

$$ELBO = E_{q(\boldsymbol{\theta})}\left[log(P(\boldsymbol{\theta}, \boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K}))\right] - E_{q(\boldsymbol{\theta})}\left[log(q(\boldsymbol{\theta}))\right]. \tag{22}$$

Because of the independence among $\{q(\theta_j)\}$, a gradient ascent algorithm [11] can be taken to maximize (22) by maximizing the ELBO with respect to each $q(\theta_j)$ in turn. We can derive the partial derivative of ELBO with respect to $q(\theta_j)$ as follows,

$$\begin{aligned} \frac{\partial ELBO}{\partial q(\theta_j)} &= \frac{\partial\left(\int q(\theta_j) E_{q(-\theta_j)}[log(P(\boldsymbol{\theta}, \boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K}))]d\theta_j - \int q(\theta_j) log(q(\theta_j))d\theta_j - \int q(\theta_j) C_{-j} d\theta_j\right)}{\partial q(\theta_j)} \\ &= E_{q(-\theta_j)}[log(P(\boldsymbol{\theta}, \boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K}))] - log(q(\theta_j)) - C_{-j}, \end{aligned} \tag{23}$$

where $E_{q(\boldsymbol{\theta})}\left[log(q(\boldsymbol{\theta}))\right] = E_{q(\boldsymbol{\theta})}\left[\sum_j log(q(\theta_j))\right] = E_{q(\boldsymbol{\theta})}\left[log(q(\theta_j))\right] + E_{q(\boldsymbol{\theta})}\left[\sum_{l \neq j} log(q(\theta_l))\right] = E_{q(\boldsymbol{\theta})}\left[log(q(\theta_j))\right] + \int q(\theta_j) C_{-j} d\theta_j$, and C_{-j} is a constant free of $q(\theta_j)$. By setting the partial

derivative (23) equal to zero, we obtain the following variational distribution for θ_j,

$$q(\theta_j) \propto \exp\left\{E_{q(\theta_{-j})}[log(P(\boldsymbol{\theta}, \boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K}))]\right\} \propto \exp\left\{E_{q(\theta_{-j})}[log(P(\theta_j|\theta_{-j}, \boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K}))]\right\}, \quad (24)$$

where θ_{-j} denotes other parameters excluding θ_j, and $q(\theta_{-j})$ denotes the variational distribution for θ_{-j}.

As a result, the optimal variational distributions per θ_j can be obtained by the gradient ascent algorithm with a partial derivative quantity (23). The ELBO quantity (22) will be used to check convergence of the gradient ascent algorithm. Then the posterior mean of the corresponding optimal variational distribution $q(\theta_j)$ from the last iteration will be taken as the Bayesian estimator for θ_j.

2.3.1 Joint Posterior Distribution for $\boldsymbol{\theta}$

Because the ELBO quantity (22) is difficult to compute if the variational distributions are of unknown distributions, we will derive the variational inference algorithm without integrating out the random effect term $\boldsymbol{u}$. For computational convenience, we transform the model (4) as follows,

$$\mathbf{U'E_g} \;=\; \mathbf{U'X\widetilde{w}} + \mathbf{U'u} + \boldsymbol{\varepsilon};\; \boldsymbol{\eta} = \mathbf{U'u} \sim \mathbf{N}(\mathbf{0}, \sigma_\epsilon^2\sigma_0^2\mathbf{D});\; \mathbf{K} = \frac{\boldsymbol{XX'}}{\boldsymbol{p}} = \mathbf{U'DU}, \quad (25)$$

where $\boldsymbol{D}$ is a diagonal matrix and prior distributions for $\{\widetilde{\boldsymbol{w}}, \sigma_0^2, \sigma_\epsilon^2\}$ are the same as in model (4).

With indicator vectors $\boldsymbol{\gamma}$, the joint conditional posterior function for $\boldsymbol{\theta}$ is given by,

$$\begin{gathered} P(\boldsymbol{\theta}|\boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K}) = P(\widetilde{\boldsymbol{w}}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\nu}, \xi, \sigma_\epsilon^2, \sigma_0^2, \sigma_1^2, \cdots, \sigma_k^2, \cdots|\boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K}) \propto \\ P(\boldsymbol{U'E_g}|\boldsymbol{U'X}, \boldsymbol{D}, \widetilde{\boldsymbol{w}}, \boldsymbol{\eta}, \sigma_\epsilon^2)P(\boldsymbol{w}|\boldsymbol{\gamma}, \boldsymbol{\nu}, \sigma_1^2, \cdots, \sigma_k^2, \cdots)P(\boldsymbol{\eta}|\sigma_\epsilon^2, \sigma_0^2, \boldsymbol{D}) \\ \left(\prod_{k=0}^{+\infty} P(\sigma_k^2|a_k, b_k)\right) P(\sigma_\epsilon^2|a_\epsilon, b_\epsilon)P(\boldsymbol{\gamma}|\boldsymbol{\nu})P(\boldsymbol{\nu}|\xi)P(\xi|a_\xi, b_\xi). \end{gathered} \quad (26)$$

Consequently, the log joint posterior density function is given by

$$\begin{gathered} log(P(\widetilde{\boldsymbol{w}}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\nu}, \xi, \sigma_\epsilon^2, \sigma_0^2, \sigma_1^2, \cdots, \sigma_k^2, \cdots|\boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K})) = \\ C - \tfrac{n}{2}log(\sigma_\epsilon^2) - \tfrac{1}{2\sigma_\epsilon^2}(\boldsymbol{U'E_g} - \boldsymbol{U'X}\widetilde{\boldsymbol{w}} - \boldsymbol{\eta})'(\boldsymbol{U'E_g} - \boldsymbol{U'X}\widetilde{\boldsymbol{w}} - \boldsymbol{\eta}) + \\ \sum_{i=1}^{p}\sum_{k=1}^{+\infty}\gamma_{ik}\left[-\tfrac{1}{2}log(\sigma_\epsilon^2) - \tfrac{1}{2}log(\sigma_k^2) - \tfrac{\widetilde{w_{ik}}^2}{2\sigma_\epsilon^2\sigma_k^2}\right] - \tfrac{n}{2}log(\sigma_\epsilon^2) - \tfrac{n}{2}log(\sigma_0^2) - \tfrac{1}{2}\boldsymbol{\eta'}(\sigma_\epsilon^2\sigma_0^2\boldsymbol{D})^{-1}\boldsymbol{\eta} + \\ \sum_{k=0}^{+\infty}\left[-(a_k+1)log(\sigma_k^2) - \tfrac{b_k}{\sigma_k^2}\right] - (a_\epsilon+1)log(\sigma_\epsilon^2) - \tfrac{b_\epsilon}{\sigma_\epsilon^2} + \\ \sum_{i=1}^{p}\sum_{k=1}^{+\infty}\left[\gamma_{ik}\left(log(\nu_k) + \sum_{l=0}^{k-1}log(1-\nu_l)\right) + (1-\gamma_{ik})log\left(1-\nu_k\prod_{l=0}^{k-1}(1-\nu_l)\right)\right] + \\ \sum_{k=0}^{+\infty}\left[(\xi-1)log(1-\nu_k) + log(\xi)\right] + (a_\xi-1)log(\xi) - b_\xi\xi, \end{gathered} \quad (27)$$

where C denotes a normalization constant that is free of parameters.

2.3.2 Variational Distribution for θ_j

Based on the above log joint posterior density function (27), the following variational distributions are derived with respect to each θ_j.

- $\widetilde{w_{ik}}$ **and** γ_{ik}

$$\begin{aligned} log(q(\widetilde{w_{ik}}, \gamma_{ik})) &= C - \frac{1}{2}E[\sigma_\epsilon^{-2}]\boldsymbol{x_i'UU'x_i}(\widetilde{w}_{ik}^2 + 2\widetilde{w}_{ik}E[\widetilde{w}_{i(-k)}]) + \\ & E[\sigma_\epsilon^{-2}]\boldsymbol{x_i'U}\left(\boldsymbol{U'E_g} - \sum_{l\neq i}\left(\boldsymbol{U'x_l}E[\widetilde{w_l}]\right) - E[\boldsymbol{\eta}]\right)\widetilde{w_{ik}} + \\ & \gamma_{ik}\left[-\frac{1}{2}E[log(\sigma_\epsilon^2)] - \frac{1}{2}E[log(\sigma_k^2)] - \frac{1}{2}E[\sigma_\epsilon^{-2}]E[\sigma_k^{-2}]\widetilde{w_{ik}}^2\right] + \\ & \gamma_{ik}\left(E[log(\nu_k)] + \sum_{l=0}^{k-1}E[log(1-\nu_l)]\right) + \\ & (1-\gamma_{ik})E\left[log\left(1-\nu_k\prod_{l=0}^{k-1}(1-\nu_l)\right)\right], \end{aligned} \tag{28}$$

where $\widetilde{w}_{i(-k)} = \sum_{j\neq k}\widetilde{w}_{ij}$.

Then $q(\widetilde{w_{ik}}|\gamma_{ik}=1)$ is given by

$$\begin{aligned} q(\widetilde{w_{ik}}|\gamma_{ik}=1) &= C - \frac{1}{2}E[\sigma_\epsilon^{-2}]\boldsymbol{x_i'UU'x_i}\widetilde{w_{ik}}^2 - \frac{1}{2}E[\sigma_\epsilon^{-2}]E[\sigma_k^{-2}]\widetilde{w_{ik}}^2 + \\ & E[\sigma_\epsilon^{-2}]\boldsymbol{x_i'U}\left(\boldsymbol{U'E_g} - \sum_{l\neq i}(\boldsymbol{U'x_l}E[\widetilde{w_l}]) - E[\boldsymbol{\eta}] - \boldsymbol{U'x_i}E[\widetilde{w}_{i(-k)}]\right)\widetilde{w_{ik}} \\ &= C - \frac{E[\sigma_\epsilon^{-2}]}{2}\left(\boldsymbol{x_i'U'Ux_i} - E[\sigma_k^{-2}]\right)\widetilde{w_{ik}}^2 + \\ & E[\sigma_\epsilon^{-2}]\boldsymbol{x_i'U'}\left(\boldsymbol{UE_g} - \sum_{l\neq i}(\boldsymbol{U'x_l}E[\widetilde{w_l}]) - E[\boldsymbol{\eta}] - \boldsymbol{U'x_i}E[\widetilde{w}_{i(-k)}]\right)\widetilde{w_{ik}} \\ q(\widetilde{w_{ik}}|\gamma_{ik}=1) &\sim N(m_{ik}, s_{ik}^2), \\ m_{ik} &= \frac{\boldsymbol{x_i'U}(\boldsymbol{U'E_g} - \sum_{l\neq i}\boldsymbol{U'x_l}E[\widetilde{w_l}] - E[\boldsymbol{\eta}] - \boldsymbol{U'x_i}E[\widetilde{w}_{i(-k)}])}{\boldsymbol{x_i'U'Ux_i} + E[\sigma_k^{-2}]}, \\ s_{ik}^2 &= \frac{1}{E[\sigma_\epsilon^{-2}](\boldsymbol{x_i'UU'x_i} + E[\sigma_k^{-2}])}. \end{aligned} \tag{29}$$

Similar to the derivation for (12), after integrating out $\widetilde{w_{ik}}$ from (28), the variational

probability for $(\gamma_{ik} = 1)$ is given by

$$q(\gamma_{ik} = 1) = \phi_{ik} = \int q(\widetilde{w_{ik}}, \gamma_{ik} = 1) d\widetilde{w_{ik}} \propto$$
$$\exp\left\{ \frac{m_{ik}^2}{2s_k^2} + log(s_k) - E[log(\sigma_\epsilon)] - E[log(\sigma_k)] + E[log(\nu_k)] + \sum_{l=0}^{k-1} E[log(1-\nu_l)] \right\}. \tag{30}$$

- ν_k

$$\begin{aligned} log(q(\nu_k)) &= C + \sum_{i=1}^{p} E[\gamma_{ik}] log(\nu_k) + \sum_{i=1}^{p} \sum_{l=k+1}^{+\infty} E[\gamma_{ik}] log(1-\nu_k) + (E[\xi] - 1) log(1-\nu_k), \\ q(\nu_k) &\sim Beta(\kappa_k, \xi_k),\ \kappa_k = \sum_{i=1}^{p} E[\gamma_{ik}] + 1,\ \xi_k = \sum_{i=1}^{p} \sum_{l=k+1}^{+\infty} E[\gamma_{il}] + E[\xi]. \end{aligned} \tag{31}$$

- $\sigma_k^2, k > 0$

$$\begin{aligned} log(q(\sigma_k^2)) &= C - \left(\frac{1}{2} \sum_{i=1}^{p} E[\gamma_{ik}] + a_k + 1 \right) log(\sigma_k^2) - \left(\frac{\sum_{i=1}^{p} E[\gamma_{ik} \widetilde{w_{ik}}^2] E[\sigma_\epsilon^{-2}]}{2} + b_k \right) \frac{1}{\sigma_k^2}; \\ q(\sigma_k^2) &\sim IG(\widetilde{a_k}, \widetilde{b_k}),\ \widetilde{a_k} = \frac{1}{2} \sum_{i=1}^{p} E[\gamma_{ik}] + a_k,\ \widetilde{b_k} = \frac{1}{2} \sum_{i=1}^{p} E[\gamma_{ik} \widetilde{w_{ik}}^2] E[\sigma_\epsilon^{-2}] + b_k. \end{aligned} \tag{32}$$

- ξ

$$\begin{aligned} log(q(\xi)) &= C + \left(a_\xi + \sum_{k=0}^{+\infty} 1_k - 1 \right) log(\xi) - \left(b_\xi - \sum_{k=0}^{+\infty} E[log(1-\nu_k)] \right) \xi; \\ q(\xi) &\sim Gamma(\widetilde{a_\xi}, \widetilde{b_\xi}),\ \widetilde{a_\xi} = a_\xi + \sum_{k=0}^{+\infty} 1_k,\ \widetilde{b_\xi} = b_\xi - \sum_{k=0}^{+\infty} E[log(1-\nu_k)]. \end{aligned} \tag{33}$$

- σ_ϵ^2

$$
\begin{aligned}
log(q(\sigma_\epsilon^2)) &= C - log(\sigma_\epsilon^2)\left(\frac{n}{2} + \frac{1}{2}\sum_{i=1}^{p}\sum_{k=1}^{+\infty} E[\gamma_{ik}] + a_\epsilon + 1\right) - \\
&\quad \frac{1}{2\sigma_\epsilon^2}(\boldsymbol{U'E_g} - \boldsymbol{U'XE}[\widetilde{\boldsymbol{w}}] - E[\boldsymbol{\eta}])'(\boldsymbol{U'E_g} - \boldsymbol{U'XE}[\widetilde{\boldsymbol{w}}] - E[\boldsymbol{\eta}]) - \\
&\quad \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{p}\sum_{k=1}^{+\infty} E[\gamma_{ik}\widetilde{w_{ik}}^2]E[\sigma_k^{-2}] - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{p}\frac{E[\eta_i^2]E[\sigma_0^{-2}]}{D_{ii}} - \frac{1}{\sigma_\epsilon^2}b_\epsilon; \qquad (34) \\
q(\sigma_\epsilon^2) &\sim IG(\widetilde{a_\epsilon}, \widetilde{b_\epsilon}), \qquad (35) \\
\widetilde{a_\epsilon} &= \frac{n}{2} + \frac{1}{2}\sum_{i=1}^{p}\sum_{k=1}^{+\infty} E[\gamma_{ik}] + a_\epsilon, \\
\widetilde{b_\epsilon} &= \frac{1}{2}(\boldsymbol{U'E_g} - \boldsymbol{U'XE}[\widetilde{\boldsymbol{w}}] - E[\boldsymbol{\eta}])'(\boldsymbol{U'E_g} - \boldsymbol{U'XE}[\widetilde{\boldsymbol{w}}] - E[\boldsymbol{\eta}]) + \\
&\quad \frac{1}{2}\left(\sum_{i=1}^{p}\sum_{k=1}^{+\infty} E[\gamma_{ik}\widetilde{w_{ik}}^2]E[\sigma_k^{-2}] + \sum_{i=1}^{p}\frac{E[\eta_i^2]E[\sigma_0^{-2}]}{D_{ii}}\right) + b_\epsilon.
\end{aligned}
$$

- σ_0^2

$$
\begin{aligned}
log(q(\sigma_0^2)) &= C - \frac{n}{2}log(\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{p}\frac{E[\eta_i]^2E[\sigma_\epsilon^{-2}]}{D_{ii}} - (a_0+1)log(\sigma_0^2) - \frac{b_0}{\sigma_0^2}, \\
q(\sigma_0^2) &\sim IG(\widetilde{a_0},\ \widetilde{b_0}),\ \widetilde{a_0} = \frac{n}{2} + a_0,\ \widetilde{b_0} = \frac{1}{2}\sum_{i=1}^{p}\frac{E[\eta_i]^2E[\sigma_\epsilon^{-2}]}{D_{ii}} + b_0 \qquad (36)
\end{aligned}
$$

- $\boldsymbol{\eta}$

$$
\begin{aligned}
log(q(\boldsymbol{\eta})) &= C - \frac{E[\sigma_\epsilon^{-2}]}{2}(\boldsymbol{U'E_g} - \boldsymbol{U'XE}[\widetilde{\boldsymbol{w}}] - \boldsymbol{\eta})'(\boldsymbol{U'E_g} - \boldsymbol{U'XE}[\widetilde{\boldsymbol{w}}] - \boldsymbol{\eta}) - \\
&\quad \frac{1}{2}\boldsymbol{\eta}'E[(\sigma_\epsilon^2\sigma_0^2\boldsymbol{D})^{-1}]\boldsymbol{\eta}; \\
q(\boldsymbol{\eta}) &\sim MVN(\boldsymbol{\mu_\eta},\ \boldsymbol{\Sigma_\eta}), \\
\boldsymbol{\mu_\eta} &= (E[\sigma_0^{-2}]\boldsymbol{D}^{-1} + \boldsymbol{I})^{-1}(\boldsymbol{U'E_g} - \boldsymbol{U'XE}[\widetilde{\boldsymbol{w}}]), \\
\boldsymbol{\Sigma_\eta} &= \frac{(E[\sigma_0^{-2}]\boldsymbol{D}^{-1} + \boldsymbol{I})^{-1}}{E[\sigma_\epsilon^{-2}]}; \qquad (37)
\end{aligned}
$$

where $\boldsymbol{\Sigma_\eta}$ is a diagonal matrix. That is, $\{q(\eta_i); i = 1, \cdots, p\}$ are independent of each other.

- Evaluations of the expectations in the above variational distributions:

$$
\begin{aligned}
E[\boldsymbol{\eta}] &= \boldsymbol{\mu}_{\boldsymbol{\eta}}; \\
E[\gamma_{ik}] &= \phi_{ik},\ i = 1, \cdots, p,\ k = 1, 2, \cdots; \\
E[\gamma_i \widetilde{w_i}^2] &= \sum_k \phi_{ik}(m_{ik}^2 + s_{ik}^2);\ E[\widetilde{w_i}] = \sum_k \phi_{ik} m_{ik};\ i = 1, \cdots, p \\
E[log(\nu_k)] &= \psi(\kappa_k) - \psi(\kappa_k + \xi_k);\ E[log(1 - \nu_k)] = \psi(\xi_k) - \psi(\kappa_k + \xi_k);\ k = 0, 1, \cdots \\
E[log(\sigma_k)] &= \frac{1}{2}(log(\widetilde{b_k}) - \psi(\widetilde{a_k}));\ E[\sigma_k^{-2}] = \frac{\widetilde{a_k}}{\widetilde{b_k}};\ k = 0, 1, \cdots \\
E[log(\sigma_\epsilon)] &= \frac{1}{2}(log(\widetilde{b_\epsilon}) - \psi(\widetilde{a_\epsilon}));\ E[\sigma_\epsilon^{-2}] = \frac{\widetilde{a_\epsilon}}{\widetilde{b_\epsilon}}; \\
E[\log(\xi)] &= \psi(\widetilde{a_\xi}) - log(\widetilde{b_\xi});\ E[\xi] = \frac{\widetilde{a_\xi}}{\widetilde{b_\xi}};
\end{aligned}
$$

  where $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function and the expectations are taken with respect to each variational distribution $q(\theta_j)$.

2.3.3 Evaluate the ELBO

Recall the formula (22) for ELBO, we will need to calculate the expectation quantities $E_{q(\boldsymbol{\theta})}\left[log(q(\boldsymbol{\theta}))\right]$ and $E_{q(\boldsymbol{\theta})}\left[log(P(\boldsymbol{\theta}, \boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{K}))\right]$. In particular,

$$
\begin{aligned}
E_{q(\boldsymbol{\theta})}\left[log(q(\boldsymbol{\theta}))\right] &= \sum_j E_{q(\boldsymbol{\theta})}\left[log(q(\theta_j))\right] \\
&= \sum_{i=1}^{p}\sum_{k=1}^{+\infty} E_{q(\widetilde{w_i}, \gamma_i))}[log(q(\widetilde{w_i}, \gamma_i))] + \sum_{i=1}^{p} E_{q(\eta_i)}[log(q(\eta_i))] + \\
&\quad \sum_{k=1}^{+\infty} E_{q(\nu_k)}[log(q(\nu_k))] + \sum_{k=1}^{+\infty} E_{q(\sigma_k^2)}[log(q(\sigma_k^2))] + \\
&\quad E_{q(\sigma_\epsilon^2)}[log(q(\sigma_\epsilon^2))] + E_{q(\xi)}[log(q(\xi))],
\end{aligned} \tag{38}
$$

where

$$
\begin{aligned}
E_{q(\widetilde{w_i}, \gamma_i))}[log(q(\widetilde{w_i}, \gamma_i))] &= \sum_{k=1}^{+\infty} \phi_{ik} E[log(q(\widetilde{w_i}, \gamma_i = 1))] = \sum_{k=1}^{+\infty} \phi_{ik}(E[log(q(\widetilde{w_i}|\gamma_i = 1)q(\gamma_i = 1))]) \\
&= \sum_{k=1}^{+\infty} \phi_{ik}\left(log(\phi_{ik}) + E[log(q(\widetilde{w_i}|\gamma_i = 1))]\right) \\
&= \sum_{k=1}^{+\infty} \phi_{ik}\left(log(\phi_{ik}) - \frac{1}{2}log(2\pi s_{ik}^2) - \frac{1}{2}\right); \\
E_{q(\eta_i)}[log(q(\eta_i))] &= -\frac{1}{2}log(2\pi\Sigma_{\eta_{ii}}) - \frac{1}{2},\ i = 1, \cdots, p;
\end{aligned}
$$

$$
\begin{aligned}
E_{q(\nu_k)}[log(q(\nu_k))] &= log(\Gamma(\kappa_k+\xi_k)) - log(\Gamma(\kappa_k)) - log(\Gamma(\xi_k)) + \\
&\quad (\kappa-1)E[log(\nu_k)] + (\xi_k-1)E[log(1-\nu_k)] \\
&= log(\Gamma(\kappa_k+\xi_k)) - log(\Gamma(\kappa_k)) - log(\Gamma(\xi_k)) + \\
&\quad (\kappa-1)(\psi(\kappa_k)-\psi(\kappa_k+\xi_k)) - (\xi_k-1)(\psi(\kappa_k+\xi_k)-\psi(\xi_k)); \\
E_{q(\sigma_k^2)}[log(q(\sigma_k^2))] &= \widetilde{a_k}log(\widetilde{b_k}) - log(\Gamma(\widetilde{a_k})) - 2(\widetilde{a_k}+1)E[log(\sigma_k)] - \widetilde{b_k}E[\sigma_k^{-2}] \\
&= \widetilde{a_k}log(\widetilde{b_k}) - log(\Gamma(\widetilde{a_k})) - (\widetilde{a_k}+1)(log(\widetilde{b_k}) - \psi(\widetilde{a_k})) - \widetilde{a_k},\ k=1,2,\cdots; \\
E_{q(\sigma_\epsilon^2)}[log(q(\sigma_\epsilon^2))] &= \widetilde{a_\epsilon}log(\widetilde{b_\epsilon}) - log(\Gamma(\widetilde{a_\epsilon})) - (\widetilde{a_\epsilon}+1)(log(\widetilde{b_\epsilon}) - \psi(\widetilde{a_\epsilon})) - \widetilde{a_\epsilon}; \\
E_{q(\xi)}[log(q(\xi))] &= \widetilde{a_\xi}log(\widetilde{b_\xi}) - log(\Gamma(\widetilde{a_\xi})) + (\widetilde{a_\xi}-1)E[log(\xi)] - \widetilde{b_\xi}E[\xi] \\
&= log(\widetilde{b_\xi}) - log(\Gamma(\widetilde{a_\xi})) + (\widetilde{a_\xi}-1)\psi(\widetilde{a_\xi}) - \widetilde{a_\xi}.
\end{aligned}
$$

And

$$
E_{q(\boldsymbol{\theta})}\left[log(P(\boldsymbol{\theta}, \boldsymbol{E_g}, \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{D}))\right] \tag{39}
$$

$$
\begin{aligned}
&= E_{q(\boldsymbol{\theta})}\left[log(P(\boldsymbol{E_g}|\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{U})P(\widetilde{w}|\boldsymbol{\gamma}, \sigma_\epsilon^2, \sigma_k^2, k=1,\cdots)P(\boldsymbol{\eta}|\boldsymbol{D}, \sigma_\epsilon^2, \sigma_0^2)P(\boldsymbol{\gamma}|\boldsymbol{\nu})P(\boldsymbol{\nu}|\xi)P(\xi))\right] \\
&= E_{q(\boldsymbol{\theta})}[log(P(\boldsymbol{E_g}|\boldsymbol{\theta}, \boldsymbol{X})] + E_{q(\boldsymbol{\theta})}[log(P(\widetilde{w}|\boldsymbol{\gamma}, \sigma_\epsilon^2, \sigma_k^2, k=1,\cdots))] + E_{q(\boldsymbol{\theta})}[log(P(\boldsymbol{\eta}|\boldsymbol{D}, \sigma_\epsilon^2, \sigma_0^2))] + \\
&\quad E_{q(\boldsymbol{\theta})}[log(P(\boldsymbol{\gamma}|\boldsymbol{\nu}))] + E_{q(\boldsymbol{\theta})}[log(P(\boldsymbol{\nu}|\xi))] + E_{q(\boldsymbol{\theta})}[log(P(\xi))].
\end{aligned} \tag{40}
$$

2.3.4 Gradient Ascent Algorithm

Starting with initial parameter values $\boldsymbol{\theta}_0$, we will iteratively evaluate the above variational distributions and take the variational means as the corresponding parameter values in next iteration. Iterations will stop when the ELBO quantity converges.

The variational means from the last iteration will be taken as the Bayesian estimates for corresponding parameters, e.g., $\widehat{\widetilde{\boldsymbol{w}}}, \widehat{\boldsymbol{\eta}}$. Recall that $\boldsymbol{\eta} = \boldsymbol{U'X\zeta}$, then the Bayesian estimate for $\boldsymbol{\zeta}$ is given by $\widehat{\boldsymbol{\zeta}} = (\boldsymbol{U'X})^{-1}\widehat{\boldsymbol{\eta}}$. Thus, the Bayesian estimate for cis-eQTL effect-sizes is given by $\widehat{\boldsymbol{w}} = \widehat{\widetilde{\boldsymbol{w}}} + \widehat{\boldsymbol{\zeta}}$.

Supplemental References

[1] P. Zeng and X. Zhou. Non-parametric genetic prediction of complex traits with latent dirichlet process regression models. *Nat Commun*, 8(1):456, 2017.

[2] P. Muller and R. Mitra. Bayesian nonparametric inference - why and how. *Bayesian Anal*, 8(2), 2013.

[3] David M. Blei and Michael I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian Anal.*, 1(1):121–143, 2006.

[4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[5] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet*, 9(2):e1003264, 02 2013.

[6] G. Casella. Empirical bayes gibbs sampling. *Biostatistics*, 2(4):485–500, 2001.

[7] Richard A Levine, Zhaoxia Yu, William G Hanley, and John J Nitao. Implementing random scan gibbs samplers. *Computational Statistics*, 20(1):177–196, 2005.

[8] Richard A Levine and George Casella. Optimizing random scan gibbs samplers. *Journal of Multivariate Analysis*, 97(10):2071–2100, 2006.

[9] John T Ormerod and Matt P Wand. Explaining variational approximations. *The American Statistician*, 64(2):140–153, 2010.

[10] Justin Grimmer. An introduction to bayesian inference via variational approximations. *Political Analysis*, 19(1):32–47, 2011.

[11] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition, 2007.