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

Sini Nagpal[1], Xiaoran Meng[2,3], 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], Jingjing Yang[3*]

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

*Correspondence author: jyang51@emory.edu

Abstract

The transcriptome-wide association studies (TWAS) 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, TWAS facilitates 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, we propose an improved Bayesian method that assumes a data-driven nonparametric prior to impute gene expression. Our method is general and flexible and includes both 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. In real applications, our nonparametric Bayesian method fitted transcriptomic imputation models for 2X number of genes with 1.7X average regression R^2 over PrediXcan, thus improving the power of follow-up TWAS. Hence, the nonparametric Bayesian model is preferred for modeling the complex genetic architecture of transcriptomes and is expected to enhance transcriptome-integrated genetic association studies. We implement our Bayesian approach in a convenient software tool "**TIGAR**" (**T**ranscriptome-**I**ntegrated **G**enetic **A**ssociation **R**esource), which imputes transcriptomic data and performs subsequent TWAS using individual-level or summary-level GWAS data.

Introduction

Genome-wide association studies (GWAS) 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 the biological mechanisms of complex traits. Multiple studies have recently generated rich transcriptomic datasets for diverse tissues of the human body, 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 GWAS can help identify novel functional loci.

The majority of GWAS projects do not possess transcriptomic data and thus cannot directly conduct integrative analysis. 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], and then test for association between imputed GReX for

GWAS samples and the trait of interest — referred to as transcriptome-wide association studies (TWAS)[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 on reference dataset to estimate the cis-eQTL effect-sizes that are used to impute the GReX for GWAS samples, while the FUSION[12] tool implements a Bayesian sparse linear mixed model (BSLMM)[18] to estimate the cis-eQTL effect-sizes.

Basically, 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 prior[21]. Whereas, 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. However, a parametric prior is assumed for the cis-eQTL effect-sizes by both Elastic-Net and BSLMM, which 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]. As shown in the result section of this paper, similar underestimation of regression R^2 by the Elastic-Net model (used by PrediXcan[11]) was observed in both simulated and real datasets.

Therefore, to flexibly model cis-eQTL distributions, we propose to use a nonparametric Bayesian method, 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 the estimation of imputation R^2 (the squared correlation between the observed and imputed values on test samples). Moreover, we will use a variational Bayesian algorithm[26-28] 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 can use DPR to estimate cis-eQTLs effect-sizes from a reference dataset, which can then be used for downstream TWAS using either individual-level or summary-level GWAS data. In subsequent sections, we first describe the DPR approach for estimating cis-eQTL effect-sizes from a reference dataset and how we can then use these effect-sizes for downstream TWAS in a target GWAS using either individual-level or summary-level data. We then compare the performance of DPR with PrediXcan using both simulated data and real SNP and transcriptome data from the Religious Orders Study and Rush Memory Aging Project (ROS/MAP)[30-33] for studying Alzheimer's Disease (AD).

Our in-depth simulation studies demonstrated that the DPR method obtained higher imputation R^2 on test samples, when the true expression heritability is <10% or when >1% variants are true causal cis-eQTL. 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 twice the number of genes as PrediXcan with 70% higher regression R^2 on average. Using DPR, we also found a potentially associated gene *TRAPPC6A* for AD pathology indices that 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 7 out of 26 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 novel software tool TIGAR (**T**ranscriptome-

Integrated **G**enetic **A**ssociation **R**esource) implementing our nonparametric Bayesian method for public use.

Materials and Methods

We consider the following additive linear regression model for estimating the cis-eQTL effect-sizes from a reference study that has both genetic and transcriptomic data available,

$$\boldsymbol{E_g} = \boldsymbol{Xw} + \boldsymbol{\epsilon}, \quad \epsilon \sim N(0, \sigma_\epsilon^2 I)\,; \qquad \textit{Equation 1}$$

where $\boldsymbol{E_g}$ denotes the gene expression levels (after corrections for confounding covariates such as age, sex, and principal components (PCs)) for gene g, $\boldsymbol{X}$ denotes the genotype matrix for all cis- genotypes (encoded as the number of minor alleles), $\boldsymbol{w}$ denotes the corresponding cis-eQTL effect-size vector, and $\boldsymbol{\epsilon}$ 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 1MB 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. In particular, 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 latent Dirichlet process regression (DPR) model[26], we assume a normal prior $N(0, \sigma_w^2)$ for the cis-eQTL effect-sizes $(w_i, i = 1, \ldots, p)$ and a Dirichlet process (DP) prior[39] for the effect-size variance σ_w^2 (as in *Equation 1*):

$$w_i \sim N(0, \sigma_w^2),\ \ \sigma_w^2 \sim D,\ \ D \sim DP\ (IG(a, b), \xi). \qquad \textit{Equation 2}$$

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 on w_i, which is equivalent to a DP normal mixture model[26-28],

$$w_i \sim \sum_{k=0}^{+\infty} \pi_k N\left(0, \sigma_k^2\right),\ \sigma_k^2 \sim IG(a_k, b_k),\ \pi_k = \nu_k \prod_{l=0}^{k-1}(1-\nu_l),\ \nu_k \sim Beta(1,\ \xi). \qquad \textit{Equation 3}$$

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 $\left(\mathrm{N}\left(0, \sigma_{\mathrm{k}}^2\right), k = 0, \ldots, +\infty\right)$, corresponding weight π_k is determined by $(\nu_l, l = 0, \ldots, k)$ 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 $\sigma_\epsilon^2 \sim IG(a_\epsilon, b_\epsilon)$ are assumed.

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 $(\sigma_k^2, \sigma_\epsilon^2, \xi)$. That is, the parameters $(\sigma_k^2, \sigma_\epsilon^2, \xi)$ 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

$$P\left(\boldsymbol{w}, \boldsymbol{\pi}, \boldsymbol{\nu}, \xi, \sigma_\epsilon^2 | \boldsymbol{E_g}, \boldsymbol{X}\right) \propto$$

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

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 supplementary note for technical details of both MCMC sampling and variational inference algorithms for obtaining the Bayesian posterior estimates for the cis-eQTL effect-sizes.

<u>Elastic-Net and BSLMM Methods</u>

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

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

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, $p(\boldsymbol{w}) \propto \exp\left(-\lambda\left(\alpha\|\boldsymbol{w}\|_1 + \frac{1}{2}(1-\alpha)\|\boldsymbol{w}\|_2^2\right)\right)$. Whereas, 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 for 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. Thus, we believe the DPR method can robustly model complex genetic architecture and then improve the imputation R^2.

<u>Association Study with Univariate Phenotype</u>

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

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

Here, $f(\cdot)$ is a pre-specified link function, which can be set as identify 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 data are available, we take the same approach as implemented by the FUSION[12] method. Let $\boldsymbol{Z}$ denote the vector of Z-scores 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

$$\tilde{Z} = \frac{\boldsymbol{Z}\widehat{\boldsymbol{w}}}{\sqrt{\boldsymbol{Z}\widehat{\boldsymbol{w}}}} = \frac{\boldsymbol{Z}\widehat{\boldsymbol{w}}}{\sqrt{\widehat{\boldsymbol{w}}'\boldsymbol{V}\widehat{\boldsymbol{w}}}}\ , \qquad \textit{Equation 5}$$

where $\boldsymbol{V}$ denotes the covariance matrix across study SNPs that can be estimated from training data or reference panels such as 1000 Genome Project[42].

<u>Association Study with Multivariate Phenotype</u>

To test for association between multivariate phenotypes and imputed GReX of the focal gene, we take a similar approach as the MultiPhen method[43] using individual-level data. 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 $\boldsymbol{Y_j} = \boldsymbol{\eta C} + \boldsymbol{\epsilon},\ j = 1,2$. Then we test if the regression R^2 is significantly greater than zero (H_0: $R^2 = 0$) for the following regression model

$$\widetilde{\boldsymbol{GReX}}_{\boldsymbol{g}} = \beta_1\widetilde{\boldsymbol{Y_1}} + \beta_2\widetilde{\boldsymbol{Y_2}} + \boldsymbol{\epsilon}\,. \qquad \textit{Equation 6}$$

That is, we test if 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,

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

where $\boldsymbol{V}_{\boldsymbol{Y}}$ is the covariance matrix among study traits.

<u>Simulation Study Design</u>

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 association studies. Specifically, we used 499 ROS/MAP participants[44] with 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 gene *ABCA7* were used to simulate gene expression levels.

We performed comprehensive scenarios that varied the proportion of causal SNPs (out of 2,799 SNPs, influenced gene expression) among values in the vector $p_{causal} = (1\%, 5\%, 10\%, 20\%)$. 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 $(h_e^2, h_p^2) = ((5\%, 80\%), (10\%, 50\%), (20\%, 25\%), (50\%, 10\%))$. The phenotypic heritability was selected arbitrarily with respect to expression heritability such that the follow-up association study power fell within the range of $(25\%, 85\%)$. With genotype matrix $\boldsymbol{X}_{\boldsymbol{g}}$ 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_e^2. Gene expression levels were generated by $\boldsymbol{E_g} = \boldsymbol{X_g}\boldsymbol{w} + \epsilon$, with $\epsilon \sim N(0,(1-h_e^2))$. Then the phenotype values were generated by $\boldsymbol{Y} = \beta \boldsymbol{E_g} + \epsilon$, where β was selected with respect to h_p^2 and $\epsilon \sim N(0,(1-h_p^2))$.

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 hours 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 Genome 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 gene-based association studies 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 20μm 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. Especially, when $p_{causal} > 1\%$, DPR out-performed PrediXcan across all expression heritability values, giving higher imputation R^2 in test data (Figure 1). For example, when $p_{causal} = 20\%$, the average imputation R^2 of 1,000 simulations was estimated as 4.54% by using DPR versus 2.64% by using PrediXcan with $h_e^2 = 10\%$, while the average imputation R^2 was estimated as 12.02% by using DPR versus 9.13% by using PrediXcan with $h_e^2 = 20\%$ (Table 1). When $p_{causal} = 1\%$, DPR performed similarly as PrediXcan with $h_e^2 = 20\%, 50\%$ but slightly out-performed PrediXcan with $h_e^2 = 5\%, 10\%$ (Table 1, Figure 1).

Consequently, when $p_{causal} > 1\%$ and $h_e^2 < 50\%$, or $p_{causal} = 1\%$ and $h_e^2 < 20\%$, the power of association studies was higher by using DPR versus using PrediXcan imputation models (Figure 2). When $h_e^2 = 50\%$, using both imputation models led to comparable power for association studies (Figure 2). Interestingly, even when both methods had similar training regression R^2 (Figure S1), the DPR method still resulted in higher imputation R^2 for test data (Figures 1; Table 1) and higher power for association studies (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 $> 1\%$ or the expression heritability $< 10\%$.

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 Methods). We trained the gene expression imputation models using 499 samples that have both transcriptomic data for frontal cortex tissues and genotype data (imputed to 1000 Genome Phase 3, with MAF > 5%, Hardy-Weinberg p-value $> 10^{-5}$, and 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 1MB of the flanking 5' and 3' ends were used in the imputation models as predictors.

In Figure 3, we compared transcriptome-wide regression R^2 estimated by using both DPR and PrediXcan methods. Specifically, DPR fitted imputation models for 14,214 genes with non-zero regression R^2 (average 10.3%), versus 6,549 genes with non-zero regression R^2 (average 6.05%) by PrediXcan. That is, DPR fitted imputation models for 2X genes with 1.7X average regression R^2 over PrediXcan. Thus, the DPR method featuring data-driven nonparametric prior for the cis-eQTL is indeed preferred in real studies.

Next, we imputed the GReX for all GWAS samples using the trained transcriptomic imputation models by both DPR and PrediXcan, and then conducted gene-based association studies (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 3 genotype principle components were adjusted 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 2MB 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 S2 and S3). Moreover, the multivariate phenotype association studies (Methods) for both AD pathology indices also identified *TRAPPC6A* as the most significant gene with p-value 5.81×10^{-6} and FDR 0.08 (Figure 4). On the other hand, the PrediXcan failed to obtain a transcriptomic imputation model for *TRAPPC6A* (Figures S4–S6).

In addition, for 26 known common and rare loci of late-onset AD[34-38], we conducted association studies using transcriptomic imputation models (DPR and PrediXcan) fitted from ROS/MAP data and standard GWAS results from IGAP[34]. We identified 7 significant loci (FDR < 0.05; Figure 5; Figure S7) — *ABCA1*, *ADAM10, CD2AP, HLA_DRB5, MAPT, MS4A6A*, and *TREM2* — that potentially affect late-onsite AD risk through transcript abundance. Specifically, 5 out of 7 significant loci were uniquely identified by using the DPR imputation models, 1 locus was uniquely identified by using the PrediXcan imputation model, while 1 locus was identified by using both imputation models (Figure 5). From the individual-level association studies with both AD pathology indices (ROS/MAP data) for these 26 loci (Figure S8), we see that the PrediXcan method only imputed GReX for 9 out of 26 loci with average regression $R^2 = 10.3\%$, while the DPR method imputed GReX for all 26 genes with average regression $R^2 = 16.7\%$. In summary, these results show that the DPR method has superior power for follow-up TWAS.

Discussion

In this paper, we propose an improved nonparametric Bayesian method for transcriptomic data imputation, accompanied with an integrated tool (freely available on GITHUB) that is referred as **T**ranscriptome-**I**ntegrated **G**enetic **A**ssociation **R**esource (**TIGAR**).

TIGAR integrates both nonparametric Bayesian (DPR) and Elastic-Net 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. We show the usefulness of the DPR method by both in-depth simulations and real applications using individual-level ROS/MAP[30-33] and summary-level IGAP[34] data. We demonstrate that DPR is preferred over PrediXcan because of the flexible nonparametric modeling of cis-eQTL effect-sizes that results in improved imputation R^2 for gene expression levels and higher power for TWAS.

With respect to user-friendly interface and computational efficiency, TIGAR can: (i) Take standard input files such as genotype files in VCF and dosage formats, phenotype files in PED format, and a combined file (tab delimited) for gene annotations and expression levels; (ii) Load input data per gene by TABIX for memory efficiency; (iii) Filter SNPs by given MAF and Hardy-Weinberg p-value thresholds; (iv) Provide options of training both Elastic-Net (using Python scripts) and DPR (using the executable tool developed with C++[26]) imputation models with unified output format; (v) 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 example, TIGAR can complete training imputation models for ~20K genes and ~1K samples within ~20 hours and TWAS within ~1hour with a 2.4GHz 16-core CPU. The computation time is expected to increase linearly with respect to the sample size.

However, as observed in the real ROS/MAP studies, there remains a large gap between the regression R^2 estimated by DPR using cis-eQTL predictors (10.3%) and the average genome-wide heritability of gene expression levels (21.8% estimated by GCTA[46] 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 should be possible to further improve the regression R^2 and imputation R^2 by fitting transcriptomic imputation models with genome-wide variants as predictors. To resolve the computational bottleneck of using genome-wide

variants, the current DPR model might be extended using a scalable Bayesian inference technique such as an Expectation Maximization MCMC (EM-MCMC) algorithm[47]. To provide more flexibility for the DPR model, the cis- and trans-acting genotype predictors could be modeled as two different categories with different nonparametric DP priors. As a result, we expect to further improve the prediction accuracy for GReX and then enhance the power of follow-up TWAS.

Besides the follow-up gene-based association studies (i.e., TWAS) 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.

Supplemental Data

Supplementary Text and Figures

Acknowledgments

J.Y. was supported by the startup funding from Emory University Department of Human Genetics. 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, 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.

Web Resources

TIGAR, https://github.com/xmeng34/TIGAR

PrediXcan, https://github.com/hakyim/PrediXcan

FUSION, http://gusevlab.org/projects/fusion/

ROS/MAP data, https://www.synapse.org/#!Synapse:syn3219045; http://www.radc.rush.edu/

IGAP data, http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php

References

1. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. American journal of human genetics 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. Nature reviews Genetics 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.

411 6. Consortium, G.T., Laboratory, D.A., Coordinating Center -Analysis Working, G., Statistical
412 Methods groups-Analysis Working, G., Enhancing, G.g., Fund, N.I.H.C., Nih/Nci,
413 Nih/Nhgri, Nih/Nimh, Nih/Nida, et al. (2017). Genetic effects on gene expression across
414 human tissues. Nature 550, 204-213.
415 7. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-
416 associated SNPs are more likely to be eQTLs: annotation to enhance discovery from
417 GWAS. PLoS genetics 6, e1000888.
418 8. Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A.,
419 Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013).
420 Transcriptome and genome sequencing uncovers functional variation in humans. Nature
421 501, 506-511.
422 9. Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild,
423 C.D., Beckman, K.B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of
424 transcriptome diversity through RNA-sequencing of 922 individuals. Genome research
425 24, 14-24.
426 10. Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.L., Arepalli,
427 S., Dillman, A., Rafferty, I.P., Troncoso, J., et al. (2010). Abundant quantitative trait loci
428 exist for DNA methylation and gene expression in human brain. PLoS genetics 6,
429 e1000952.
430 11. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J.,
431 Eyler, A.E., Denny, J.C., Consortium, G.T., Nicolae, D.L., et al. (2015). A gene-based
432 association method for mapping traits using reference transcriptome data. Nature
433 genetics 47, 1091-1098.
434 12. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J.,
435 Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale
436 transcriptome-wide association studies. Nature genetics 48, 245-252.
437 13. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W.,
438 Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary data
439 from GWAS and eQTL studies predicts complex trait gene targets. Nature genetics 48,
440 481-487.
441 14. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017).
442 Integrating Gene Expression with Summary Association Statistics to Identify Genes
443 Associated with 30 Complex Traits. American journal of human genetics 100, 473-487.
444 15. Su, Y.R., Di, C., Bien, S., Huang, L., Dong, X., Abecasis, G., Berndt, S., Bezieau, S., Brenner, H.,
445 Caan, B., et al. (2018). A Mixed-Effects Model for Powerful Association Tests in
446 Integrative Functional Genomics. American journal of human genetics 102, 904-919.
447 16. Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S.M., Yu, Z., Li, B., Muchnik, S., Shi, Y., et al.
448 (2018). A statistical framework for cross-tissue transcriptome-wide association analysis.
449 bioRxiv.
450 17. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net.
451 Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301-320.
452 18. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse
453 linear mixed models. PLoS genetics 9, e1003264.

454 19. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal
455 Statistical Society Series B (Methodological) 58, 267-288.
456 20. Hoerl, A.E., and Kennard, R.W. (2000). Ridge Regression: Biased Estimation for
457 Nonorthogonal Problems. Technometrics 42, 80-86.
458 21. Li, Q., and Lin, N. (2010). The Bayesian elastic net. Bayesian Anal 5, 151-170.
459 22. Guan, Y.T., and Stephens, M. (2011). Bayesian Variable Selection Regression for Genome-
460 Wide Association Studies and Other Large-Scale Problems. Annals of Applied Statistics 5,
461 1780-1815.
462 23. Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D.,
463 Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2005). A unified mixed-model method for
464 association mapping that accounts for multiple levels of relatedness. Nature genetics 38,
465 203.
466 24. Huan, T., Liu, C., Joehanes, R., Zhang, X., Chen, B.H., Johnson, A.D., Yao, C., Courchesne, P.,
467 O'Donnell, C.J., Munson, P.J., et al. (2015). A systematic heritability analysis of the
468 human whole blood transcriptome. Human genetics 134, 343-358.
469 25. Lloyd-Jones, L.R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A.,
470 Metspalu, A., Dermitzakis, M., et al. (2017). The Genetic Architecture of Gene Expression
471 in Peripheral Blood. American journal of human genetics 100, 371.
472 26. Zeng, P., and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with
473 latent Dirichlet process regression models. Nat Commun 8, 456.
474 27. Blei, D.M., and Jordan, M.I. (2006). Variational inference for Dirichlet process mixtures.
475 Bayesian Anal 1, 121-143.
476 28. Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational Inference: A Review for
477 Statisticians. Journal of the American Statistical Association 112, 859-877.
478 29. Casella, G. (2001). Empirical Bayes Gibbs sampling. Biostatistics 2, 485-500.
479 30. Bennett, D.A., Schneider, J.A., Arvanitakis, Z., and Wilson, R.S. (2012). Overview and findings
480 from the religious orders study. Curr Alzheimer Res 9, 628-645.
481 31. Bennett, D.A., Schneider, J.A., Buchman, A.S., Barnes, L.L., Boyle, P.A., and Wilson, R.S.
482 (2012). Overview and findings from the rush Memory and Aging Project. Curr Alzheimer
483 Res 9, 646-663.
484 32. Ng, B., White, C.C., Klein, H.U., Sieberts, S.K., McCabe, C., Patrick, E., Xu, J., Yu, L., Gaiteri, C.,
485 Bennett, D.A., et al. (2017). An xQTL map integrates the genetic architecture of the
486 human brain's transcriptome and epigenome. Nat Neurosci 20, 1418-1426.
487 33. Bennett, D.A., Buchman, A.S., Boyle, P.A., Barnes, L.L., Wilson, R.S., and Schneider, J.A.
488 (2018). Religious Orders Study and Rush Memory and Aging Project. J Alzheimers Dis 64,
489 S161-S189.
490 34. Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C.,
491 DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al. (2013). Meta-analysis
492 of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature
493 genetics 45, 1452-1458.
494 35. Reitz, C. (2014). Genetic loci associated with Alzheimer's disease. Future Neurol 9, 119-122.
495 36. Reitz, C. (2015). Novel susceptibility loci for Alzheimer's disease. Future Neurol 10, 547-558.
496 37. Sims, R., van der Lee, S.J., Naj, A.C., Bellenguez, C., Badarinarayan, N., Jakobsdottir, J.,
497 Kunkle, B.W., Boland, A., Raybould, R., Bis, J.C., et al. (2017). Rare coding variants in

498 PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's
499 disease. Nature genetics 49, 1373-1384.
500 38. Yuan, X.Z., Sun, S., Tan, C.C., Yu, J.T., and Tan, L. (2017). The Role of ADAM10 in Alzheimer's
501 Disease. J Alzheimers Dis 58, 303-322.
502 39. Muller, P., and Mitra, R. (2013). Bayesian Nonparametric Inference - Why and How.
503 Bayesian Anal 8.
504 40. Carbonetto, P., and Stephens, M. (2012). Scalable Variational Inference for Bayesian
505 Variable Selection in Regression, and Its Accuracy in Genetic Association Studies.
506 Bayesian Analysis 7, 73-107.
507 41. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for
508 common diseases: application to analysis of sequence data. American journal of human
509 genetics 83, 311-321.
510 42. Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M.,
511 Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map
512 of genetic variation from 1,092 human genomes. Nature 491, 56-65.
513 43. O'Reilly, P.F., Hoggart, C.J., Pomyen, Y., Calboli, F.C., Elliott, P., Jarvelin, M.R., and Coin, L.J.
514 (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS.
515 PloS one 7, e34861.
516 44. De Jager, P.L., Shulman, J.M., Chibnik, L.B., Keenan, B.T., Raj, T., Wilson, R.S., Yu, L., Leurgans,
517 S.E., Tran, D., Aubin, C., et al. (2012). A genome-wide scan for common variants
518 affecting the rate of age-related cognitive decline. Neurobiol Aging 33, 1017 e1011-1015.
519 45. De Jager, P.L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L.C., Yu, L., Eaton, M.L.,
520 Keenan, B.T., Ernst, J., McCabe, C., et al. (2014). Alzheimer's disease: early alterations in
521 brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat Neurosci 17, 1156-
522 1163.
523 46. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide
524 complex trait analysis. American journal of human genetics 88, 76-82.
525 47. Yang, J., Fritsche, L.G., Zhou, X., Abecasis, G., and International Age-Related Macular
526 Degeneration Genomics, C. (2017). A Scalable Bayesian Method for Integrating
527 Functional Information in Genome-wide Association Studies. American journal of human
528 genetics 101, 404-416.
529

Table 1. Average regression R^2 in the training data and imputation R^2 in the test data under various simulation scenarios, when the proportion of true causal SNPs is 1% and 20% with expression heritability $h_e^2 = (5\%, 10\%, 20\%, 50\%)$. The best imputation R^2 in the test samples are bold.

h_e^2		Causal Proportion 1%		Causal Proportion 20%	
		DPR	PrediXcan	DPR	PrediXcan
5%	Training R^2	17.60%	12.38%	17.60%	12.54%
	Test R^2	**1.60%**	1.12%	**1.54%**	0.76%
10%	Training R^2	21.57%	17.48%	21.33%	18.24%
	Test R^2	**4.44%**	4.13%	**4.54%**	2.64%
20%	Training R^2	31.80%	30.27%	31.48%	31.30%
	Test R^2	12.26%	**12.29%**	**12.02%**	9.13%
50%	Training R^2	60.50%	58.40%	60.71%	61.91%
	Test R^2	39.30%	**42.05%**	**39.32%**	36.04%

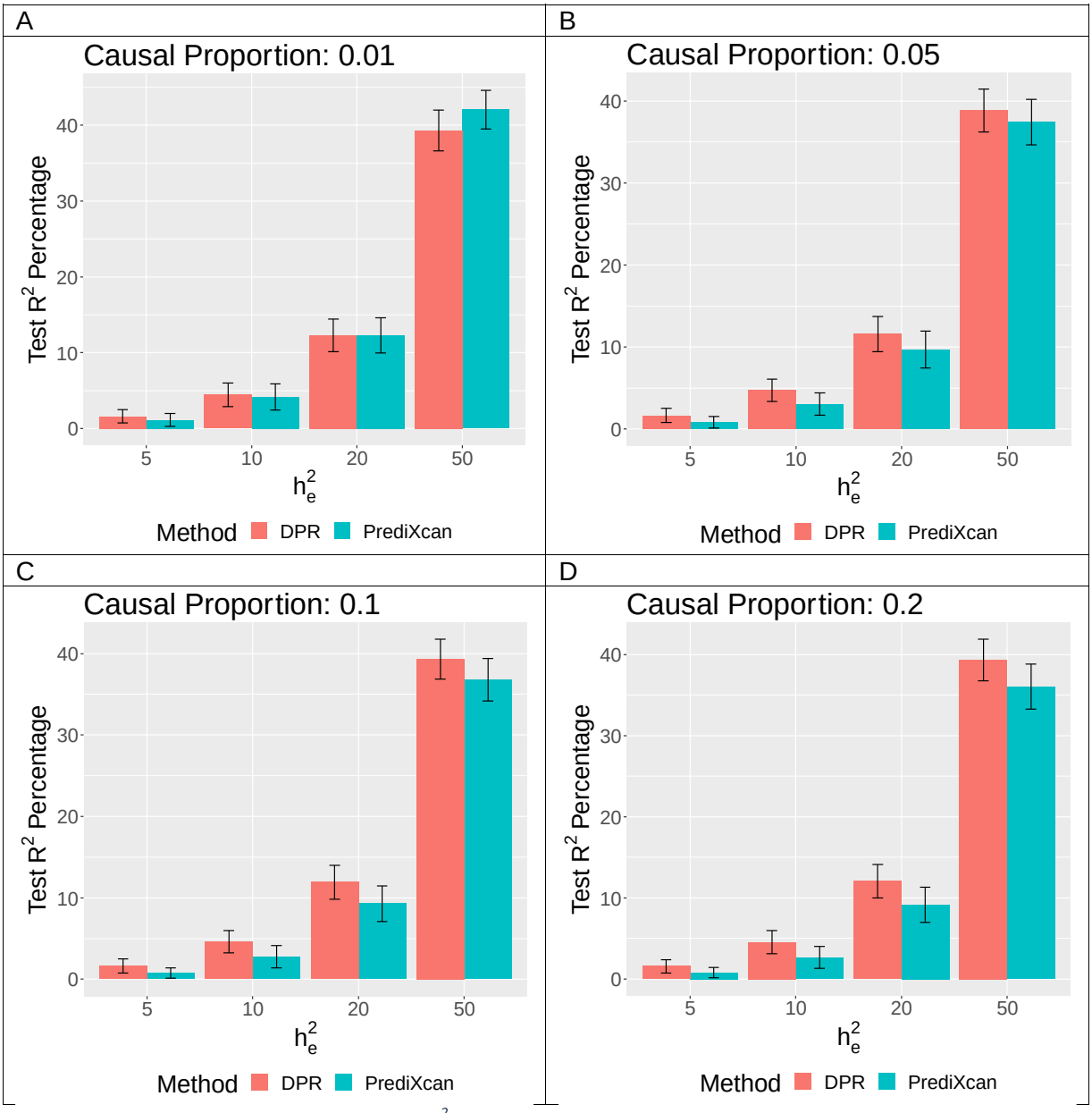

Figure 1. Boxplots of average imputation R^2 in test samples by DPR and PrediXcan under various simulation scenarios with various proportions of true causal SNPs $(1\%, 5\%, 10\%, 20\%)$ *and true expression heritability* $h_e^2 = (5\%, 10\%, 20\%, 50\%)$.

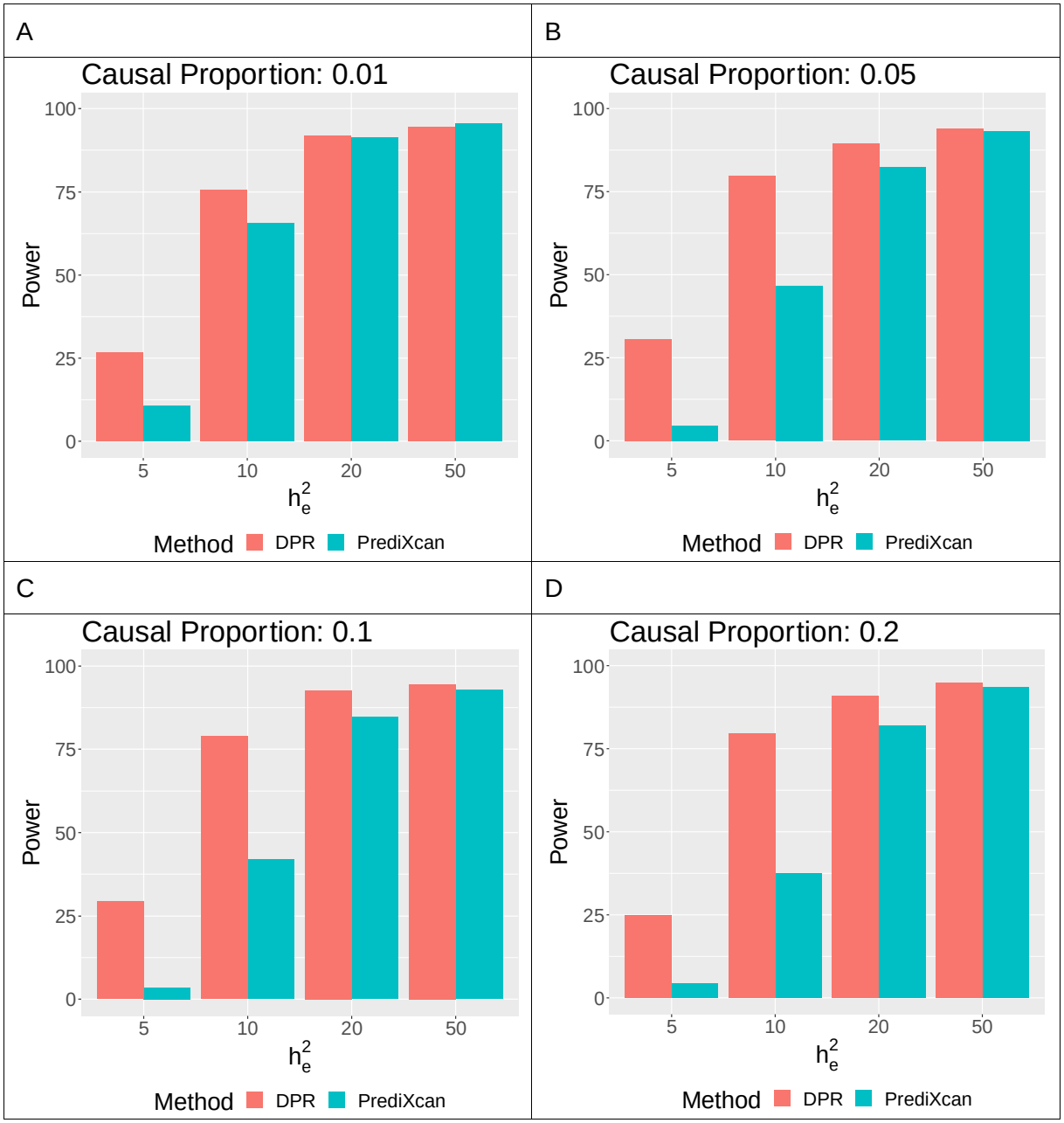

Figure 2. Power of genetic association studies by using DPR and PrediXcan imputation models, under various simulation scenarios with proportions of true causal SNPs $(1\%, 5\%, 10\%, 20\%)$*, expression and phenotype heritability* $(h_e^2, h_p^2) = ((5\%, 80\%), (10\%, 50\%), (20\%, 25\%), (50\%, 10\%))$.

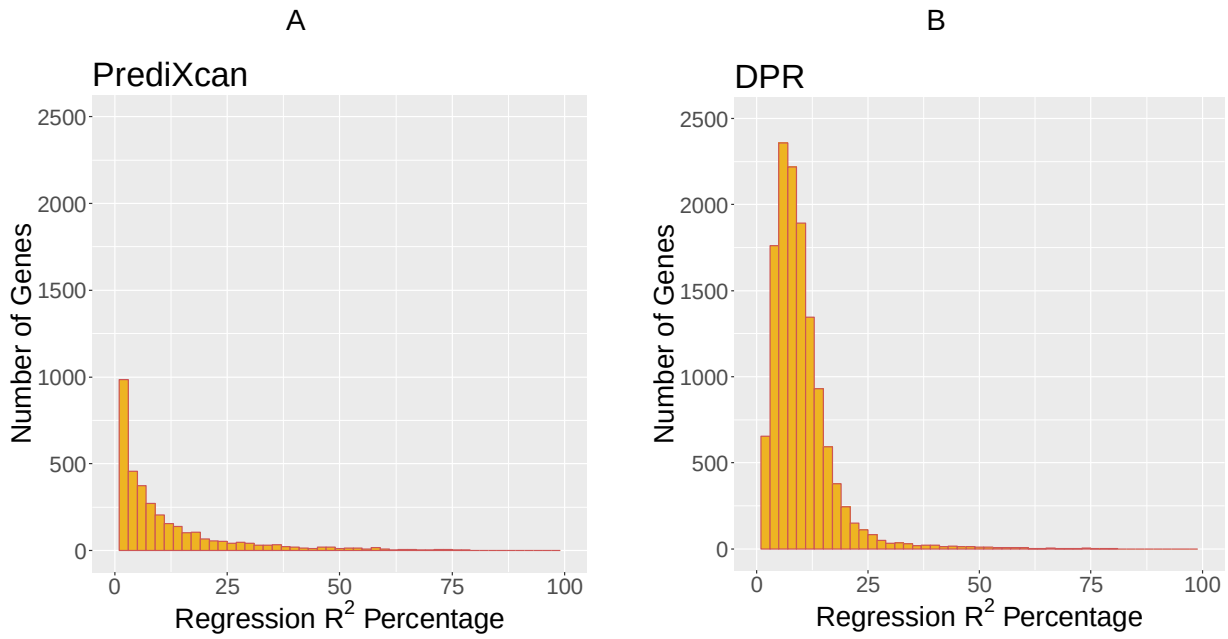

Figure 3. Histogram plots of transcriptome-wide regression R^2 by DPR and PrediXcan.

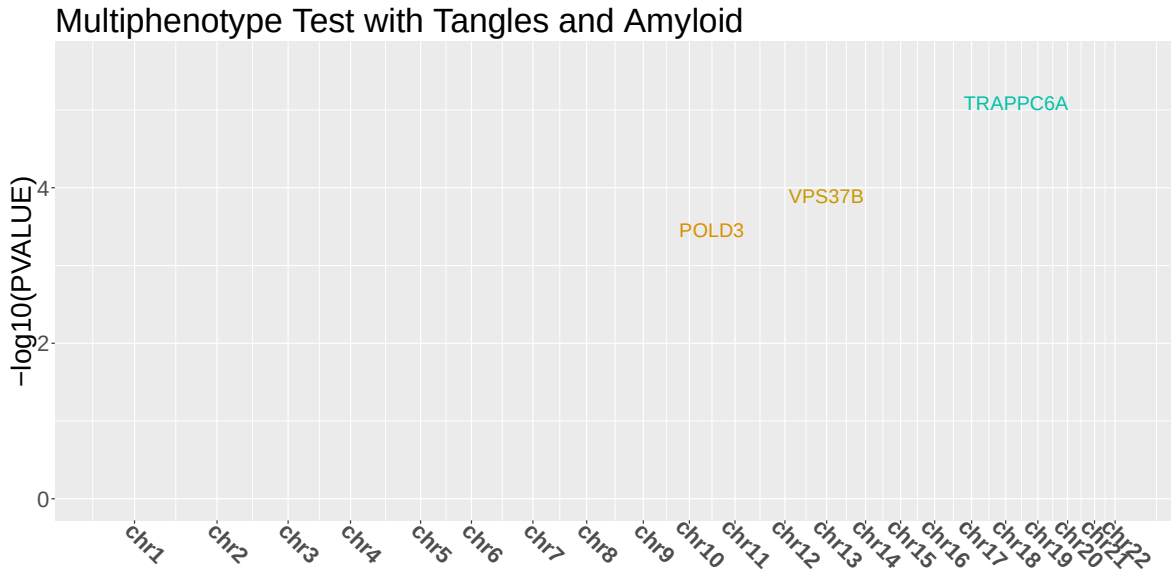

Figure 4. Manhattan plot for the multivariate phenotype associations studies using imputed GReX by DPR and two AD pathologic indices, neurofibrillary tangle density (Tangles) and β-amyloid load (Amyloid), from individual-level ROS/MAP data.

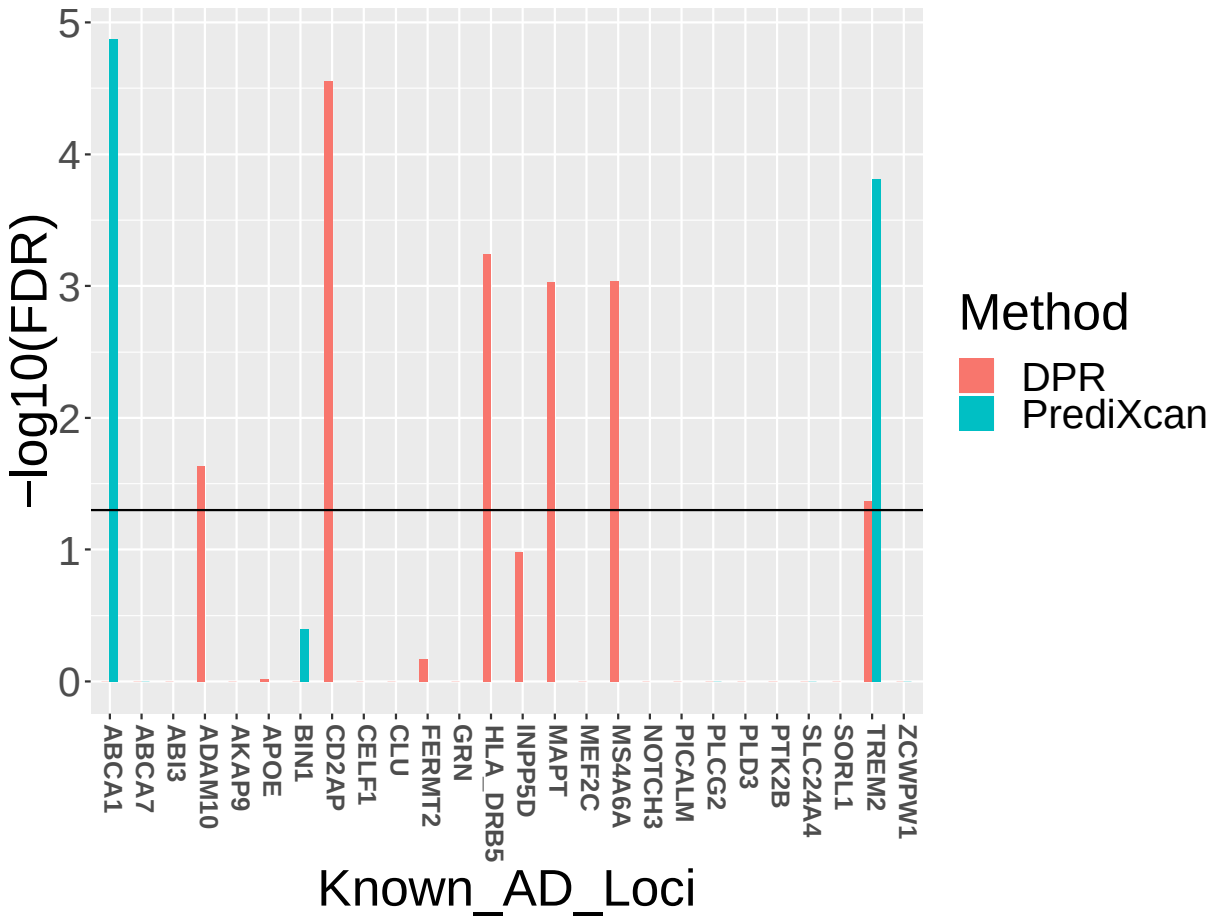

Figure 5. Bar plot of -log10(FDR) of gene-based association studies for known AD loci using summary-level GWAS statistics from IGAP and the transcriptomic imputation models by DPR and PrediXcan from the ROS/MAP data. The horizontal line shows the significance level with FDR=0.05. The missing values are due to either zero regression R^2 by PrediXcan or large FDR values (~1) using the DPR imputation models.