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

Sini Nagpal, Georgia Institute of Technology
Xiaoran Meng, Emory University
Michael Epstein, Emory University
Lam C. Tsai, University of Michigan
Matthew Patrick, University of Michigan
Greg Gibson, Georgia Institute of Technology
Philip L. De Jager, Columbia University
David A. Bennett, Rush University
Aliza Wingo, Emory University
Thomas Wingo, Emory University

Only first 10 authors above; see publication for full author list.

Journal Title: The American Journal of Human Genetics
Volume: Volume 105, Number 2
Publisher: Cell Press | 2019-08-01, Pages 258-266
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1016/j.ajhg.2019.05.018
Permanent URL: https://pid.emory.edu/ark:/25593/vm80x

Final published version: http://dx.doi.org/10.1016/j.ajhg.2019.05.018

Copyright information:

© 2019 American Society of Human Genetics.

This is an Open Access work distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Accessed November 11, 2022 12:29 AM EST
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

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 $\leq 0.2$. 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. Recent studies have shown that these associated regions were enriched for regulatory elements such as enhancers (H3K27ac marks) and expression of quantitative trait loci (eQTL), 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, Genetic European Variation in Health and Disease (GEUVADIS) for lymphoblastoid cell lines, Depression Genes and Networks (DGN) for whole-blood samples, and the North American Brain Expression Consortium (NABEC) for cortex tissues. Previous studies 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 have shown that one can impute the genetically regulated gene expression (GReX) within such GWAS projects by using reference datasets like GTEx and GEUVADIS 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). 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 method uses an Elastic-Net variable selection model and the FUSION tool implements a Bayesian sparse linear mixed model (BSLMM) 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 model used by PrediXcan assumes a combination of LASSO ($L_1$) and Ridge ($L_2$) penalties on the cis-eQTL effect sizes, which is equivalent to a Bayesian model with a mixture Gaussian and Laplace
prior. In contrast, the BSLMM used by FUSION is a combination of Bayesian variable selection model (BVSR) and linear mixed model (LMM) for 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 have also shown that both PrediXcan and FUSION estimated the average regression $R^2$ (i.e., the percentage of gene expression variation that can be explained by cis-ge- notypes) as $\sim 5\%$ 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. Therefore, to flexibly model cis-eQTL distributions, we use a nonparametric Bayesian method that was originally proposed for genetic prediction of complex traits, 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, 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) and BSLMM (implemented in FUSION) 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 can be employed as an alternative of Monte Carlo Markov Chain (MCMC) to efficiently fit the model.

Similar to PrediXcan and FUSION methods, we employ DPR to estimate cis-eQTL 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 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) for studying Alzheimer disease (AD).

Our in-depth simulation studies demonstrated that the DPR method obtained higher imputation $R^2$ on test samples, when $\geq 1\%$ cis-SNPs are true causal and the true expression heritability is $\leq 0.2$. 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 TRAPP64A 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), we identified three known AD loci that potentially affect the late-onset 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,

$$E_g = Xw + \varepsilon$$

(Equation 1)

where $E_g$ denotes the gene expression levels (after corrections for confounding covariates such as age, sex, and principal components) for gene $g$, $X$ denotes the genotype matrix for all cis-genotypes (encoded as the number of minor alleles or genotype dosages), $w$ denotes the corresponding cis-eQTL effect-size vector, and $\varepsilon$ denotes the error term. The intercept term is dropped in Equation 1 for assuming both $E_g$ and $X$ 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 $w$ will be used for follow-up analysis. The GReX will be imputed by

$$\text{GReX} = Xw$$

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

**Nonparametric Bayesian Method**

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

$$w_l \sim N(0, \sigma_n^2), \sigma_n^2 \sim \text{D. D} \sim \text{DP}(\text{IG}(\mathbf{a}, \mathbf{b}), \xi).$$

(Equation 2)

The prior distribution $D$ deviates from the DP with base distribution as an inverse gamma (IG) distribution and concentration parameter $\xi$. Note that $\sigma_n^2$ can be viewed as a latent variable and integrating out $\sigma_n^2$ will induce a nonparametric prior distribution for $w_l$, which is equivalent to a DP normal mixture model.

$$w_l \sim \sum_{k=0}^{\infty} \pi_k N(0, \sigma_k^2), \sigma_k^2 \sim \text{IG}(\mathbf{a}_k, \mathbf{b}_k), \pi_k = \nu_k \prod_{l=0}^{k-1} (1 - \nu_l), \nu_k \sim \text{Beta}(1, \xi).$$

(Equation 3)

Here, the nonparametric prior distribution on $w_l$ 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 $\nu_k$ is determined by $\nu_k = \frac{1}{k}$ with a Beta prior, and $\xi$ 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 $\xi \sim \text{Gamma}(a_1, b_1)$ and $\sigma_k^2 \sim \text{IG}(a_k, b_k)$ are assumed.
Generally, the hyper parameters \(a_i, b_i, a_s, b_s\) in the inverse gamma distributions can be set as 0.1 and \((a_i, b_i)\) in the gamma distribution can be set as \((1, 0.1)\) to induce non-informative priors for \((\sigma_i^2, \sigma_s^2, \xi)\). That is, the parameters \((\sigma_i^2, \sigma_s^2, \xi)\) will be adaptively estimated from the data and the nonparametric prior on \(w_i\) will be data driven. The posterior estimates for \(w\) can be obtained by the MCMC\(^{29}\) or variational Bayesian algorithm\(^{28,40}\) from the following joint conditional posterior distribution

\[
P(w, \pi, \xi, \sigma_i^2, \sigma_s^2, \xi_E | \mathbf{E}_X) = \frac{P(\mathbf{E}_X | w, \sigma_i^2)P(w | \pi, \sigma_i^2, \ldots, \sigma_s^2, \ldots)\prod_{k=0}^{p}(P(\sigma_i^2 | a_i, b_i) \times P(\pi | \nu)P(\xi | a_s, b_s)P(\sigma_s^2 | a_s, b_s))}{\int P(\mathbf{E}_X | w, \sigma_i^2)P(w | \pi, \sigma_i^2, \ldots, \sigma_s^2, \ldots)\prod_{k=0}^{p}(P(\sigma_i^2 | a_i, b_i) \times P(\pi | \nu)P(\xi | a_s, b_s)P(\sigma_s^2 | a_s, b_s)) dw}\]

where \(\mathbf{E}_X\) is the follow-up TWAS (using a burden type gene-based test\(^{41}\)).

Particularly, the variational Bayesian algorithm\(^{28,40}\) is an approximation 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 \(\mathbf{w}\) in Equation 1 with a combination of \(L_1\) (LASSO)\(^{19}\) and \(L_2\) (Ridge)\(^{20}\) penalties by

\[
\mathbf{w} = \arg \min_{\mathbf{w}} \left( \| \mathbf{E}_X - \mathbf{Xw} \|_2^2 + \lambda \left( a \| \mathbf{w} \|_1 + \frac{1}{2} (1 - a) \| \mathbf{w} \|_2^2 \right) \right),
\]

where \(\| \cdot \|_2\) denotes \(L_2\) norm, \(\| \cdot \|_1\) denotes \(L_1\) norm, \(a \gtrless 0.1\) denotes the proportion of \(L_1\) penalty, and \(\lambda\) denotes the penalty parameter. Particularly, PrediXcan\(^{11}\) takes \(a = 0.5\) and tunes the penalty parameter \(\lambda\) 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 \(w\), that is, \(p(w) \propto \exp\left( -\lambda \left( a \| \mathbf{w} \|_1 + \frac{1}{2} (1 - a) \| \mathbf{w} \|_2^2 \right) \right)\). In contrast, the BSLMM\(^{18}\) assumes a mixture of two priors for the cis-eQTL effect sizes, \(w_i \sim \mathcal{N}(0, \sigma_i^2) + (1 - \pi)(\mathcal{N}(0, \sigma_s^2))\). That is, the BSLMM\(^{18}\) assumes all cis-SNPs have at least a small effect, which are normally distributed with variance \(\sigma_s^2\) and some proportion \((\pi)\) of cis-SNPs have an additional effect, normally distributed with variance \(\sigma_i^2\). Particularly, with \(\sigma_s^2 = 0\), BSLMM becomes BVSN\(^{22}\) 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 \((\text{genotype data } X_{\text{new}}\text{ phenotype } \mathbf{Y}, \text{covariant matrix } \mathbf{C})\) and cis-eQTL effect size estimates \(\mathbf{w}\), the follow-up TWAS (using a burden type gene-based test\(^{24}\)) is to test the association between \(\mathbf{GReX} = X_{\text{new}} \mathbf{w}\) and \(\mathbf{Y}\) based on the following generalized linear regression model

\[
f(f(Y | X, C)) = \eta C + \beta \mathbf{GReX}.
\]

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\(^{27}\) method. Let \(\mathbf{Z}\) 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

\[
\mathbf{Z} = \frac{\mathbf{Z}^T \mathbf{w}}{\sqrt{\mathbf{w}^T \mathbf{V}_W \mathbf{w}}}
\]

where \(\mathbf{V}_W\) denotes the covariance matrix of analyzed SNPs that can be estimated from training data or reference panels such as 1000 Genomes Project\(^{25}\) (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 \((Y_1, Y_2)\) and a covariate matrix \(\mathbf{C}\), we first adjust for the covariates by taking the residuals \(( \mathbf{Y}_1, \mathbf{Y}_2)\) respectively from the linear regression models \(Y_j = \eta C + \epsilon, j = 1, 2\). Then we test whether the regression \(R^2\) is significantly greater than zero \((H_0: R^2 = 0)\) for the following regression model

\[
\mathbf{GReX}_g = \beta_1 \mathbf{Y}_1 + \beta_2 \mathbf{Y}_2 + \epsilon.
\]

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., \(\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)\) with two phenotypes. Then, a similar burden approach can be used to obtain a joint Z-score for multi-phenotype test,

\[
\mathbf{Z}_{\text{joint}} = \frac{\mathbf{Z}_j}{\sqrt{\mathbf{Z}_j^T \mathbf{V}_Y \mathbf{Z}_j}} = \frac{\mathbf{Z}_j}{\sqrt{\mathbf{V}_Y J} j = (1, \ldots, 1)},
\]

where \(\mathbf{V}_Y\) 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\(^{34}\) which contains both RNA-sequencing and genotype data as training data, and genotype data from an additional 1,200 ROS/MAP participants\(^{34}\) 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.
We performed comprehensive scenarios that varied the proportion of causal SNPs (out of 2,799 SNPs, influenced gene expression) among values in the vector \( p_{\text{causal}} = (0.001, 0.01, 0.1, 0.2) \). 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 \( (h^2_G, h^2_Y) = ((0.05, 0.8), (0.1, 0.5), (0.2, 0.25), (0.5, 0.1)) \). 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 \( p_{\text{causal}} = 0.2 \) and \( (h^2_G, h^2_Y) = (0.2, 0.25) \).

With genotype matrix \( X_g \) of the randomly selected causal SNPs (according to \( p_{\text{causal}} \)), we generated effect sizes \( w_i \) from \( N(0, 1) \) and then re-scaled the effect sizes to ensure the targeted \( h^2_Y \). Gene expression levels were generated by \( X_
 \beta \times \varepsilon \), where \( \beta \) was scaled with respect to \( h^2_Y \) and \( \varepsilon \sim N(0, 1 - h^2_Y) \). Then the phenotype values were generated by \( Y = \beta \times \varepsilon \), where \( \beta \) was selected as 0.5 DPR outperformed PrediXcan across all expression heritability values, giving higher imputation \( R^2 \) in test data (Figure 1A). For example, when \( p_{\text{causal}} = 0.2 \), the average imputation \( R^2 \) of 1,000 simulations was estimated as 4.55% by using DPR versus 2.64% by using PrediXcan with \( h^2_Y = 0.1 \), while the average imputation \( R^2 \) was estimated as 12.02% by using DPR versus 9.13% by

**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 \( p_{\text{causal}} = (0.001, 0.01, 0.1, 0.2) \) and true expression heritability \( h^2_Y = (0.05, 0.1, 0.2, 0.5) \). TWAS power was evaluated with paired expression and phenotype heritability \( (h^2_G, h^2_Y) = ((0.05, 0.8), (0.1, 0.5), (0.2, 0.25), (0.5, 0.1)) \).
using PrediXcan with $h^2_e = 0.2$ (Table 1). When $p_{\text{causal}} = 0.01$, DPR performed slightly out-performed PrediXcan with $h^2_e = (0.05, 0.1, 0.2)$ and PrediXcan outperformed DPR with $h^2_e = 0.5$ (Table 1, Figure 1). On the other hand, under a sparse cis-eQTL causality model with $p_{\text{causal}} = 0.001$ (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 $p_{\text{causal}} \geq 0.01$ and $h^2_e \leq 0.2$, the power of association studies was higher by using DPR than using PrediXcan imputation models (Figure 1B). When $h^2_e = 0.5$, 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 $p_{\text{causal}} \geq 0.01$ and $h^2_e \leq 0.2$ (Figure 1B). In addition, from the simulation studies with various training sample sizes (100, 300, 499), $p_{\text{causal}} = 0.2$, and $(h^2_e, h^2_e) = (0.2, 0.25)$, 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-folds. 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 studies11,46). 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 prediction $R^2$ 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 ß-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.

### Table 1. Simulation Prediction $R^2$ Comparison

<table>
<thead>
<tr>
<th>Causal Proportion</th>
<th>$0.01$</th>
<th>$0.2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h^2_e$</td>
<td>DPR</td>
<td>PrediXcan</td>
</tr>
<tr>
<td>$0.05$</td>
<td>1.60%*</td>
<td>1.12%</td>
</tr>
<tr>
<td>$0.1$</td>
<td>4.54%*</td>
<td>4.13%</td>
</tr>
<tr>
<td>$0.2$</td>
<td>12.54%*</td>
<td>12.29%</td>
</tr>
<tr>
<td>$0.5$</td>
<td>39.31%</td>
<td>42.05%*</td>
</tr>
</tbody>
</table>

Various simulation scenarios were considered, with the proportion of true causal SNPs $p_{\text{causal}} = (0.01, 0.2)$ and expression heritability $h^2_e = (0.05, 0.1, 0.2, 0.5)$. The best prediction $R^2$ per scenario is indicated with asterisk (*)
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 $1.64 \times 10^{-5}$ and $5.35 \times 10^{-5}$ (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 $5.81 \times 10^{-6}$ 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 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. 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 and summary-level IGAP GWAS data, we demonstrated that the nonparametric Bayesian DPR method is preferred for imputing gene expression when the proportion of causal cis-eQTL $\geq 0.01$ and the true gene expression heritability $\leq 0.2$. 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++) 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). 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.

As observed in the real ROS/MAP studies, there remains a large gap between the 5-fold CV $R^2$ using cis-eQTL predictors ($\sim 5\%$) and the average genome-wide heritability of gene expression levels (21.8% estimated by GCTA 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
transcriptomic imputation models with genome-wide variants as predictors. Scalable Bayesian inference techniques such as the Expectation Maximization MCMC (EM-MCMC) algorithm 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 account 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.

**Supplemental Data**

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

---

**Table 2. Real Study Prediction $R^2$ Comparison**

<table>
<thead>
<tr>
<th>Number of Genes</th>
<th>DPR</th>
<th>PrediXcan</th>
</tr>
</thead>
<tbody>
<tr>
<td>8,752$^a$</td>
<td>0.011</td>
<td>0.003</td>
</tr>
<tr>
<td>5,547$^b$</td>
<td>0.026</td>
<td>0.026</td>
</tr>
<tr>
<td>4,819$^c$</td>
<td>0.033</td>
<td>0.036</td>
</tr>
</tbody>
</table>

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.

---

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

References

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—NDR1; Biospecimen Collection Source Site—RPC1; 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.