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: jingjing.yang@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, to improve on this, we propose to use a Bayesian method that assumes a data-driven nonparametric prior to impute gene expression. 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. In real applications, our nonparametric Bayesian method fitted transcriptomic imputation models for 57.6% more genes 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

89 the data by assuming a Dirichlet process prior on effect-size variance. This Bayesian model is
90 also known as latent Dirichlet process regression (DPR) model[26], which can flexibly model the
91 underlying complex genetic architecture of transcriptomes. Thus, DPR is a more generalized
92 model that includes Elastic-Net (implemented in PrediXcan[11]) and BSLMM (implemented in
93 FUSION[12]) as special cases. Consequently, DPR can robustly estimate cis-eQTLs and then
94 improve imputation R^2 (the squared Pearson correlation between the observed and imputed
95 values on test samples). Moreover, we will use a variational Bayesian algorithm[26-28] as an
96 alternative of Monte Carlo Markov Chain (MCMC)[29] to efficiently fit the Bayesian model.

97 Similar to PrediXcan[11] and FUSION[12] methods, we can use DPR to estimate cis-eQTLs
98 effect-sizes from a reference dataset, which can then be used for downstream TWAS using either
99 individual-level or summary-level GWAS data. In subsequent sections, we first describe the DPR
100 approach for estimating cis-eQTL effect-sizes from a reference dataset and how we can then use
101 these effect-sizes for downstream TWAS using either individual-level or summary-level GWAS
102 data. We then compare the performance of DPR with PrediXcan using both simulated data and
103 real SNP and transcriptome data from the Religious Orders Study and Rush Memory Aging
104 Project (ROS/MAP)[30-33] for studying Alzheimer's Disease (AD).

105 Our in-depth simulation studies demonstrated that the DPR method obtained higher
106 imputation R^2 on test samples, when the true expression heritability is <10% or when >1%
107 variants are true causal cis-eQTL. Consequently, better imputation R^2 resulted in improved power
108 for follow-up association studies. Meanwhile, application of DPR to the ROS/MAP study imputed
109 GReX for 57.6% more genes than PrediXcan. Using DPR, we also found a potentially associated
110 gene *TRAPPC6A* for AD pathology indices that was missed by PrediXcan. Further, by using the
111 transcriptomic imputation models fitted from ROS/MAP data and summary-level GWAS data
112 generated from the International Genomics of Alzheimer's Project (IGAP)[34], we identified 3 known
113 AD loci[34-38] that potentially affect the late-onsite AD risk through transcript abundance. We

114 conclude with a discussion of future topics and further describe our novel software tool TIGAR
115 (**T**ranscriptome-**I**ntegrated **G**enetic **A**ssociation **R**esource) implementing the nonparametric
116 Bayesian method for public use.
117

118 **Materials and Methods**

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

121 $$\boldsymbol{E_g} = \boldsymbol{Xw} + \boldsymbol{\epsilon}, \quad \epsilon \sim N(0, \sigma_\epsilon^2 I)\,; \qquad \textit{Equation 1}$$

122 where $\boldsymbol{E_g}$ denotes the gene expression levels (after corrections for confounding
123 covariates such as age, sex, and principal components (PCs)) for gene g, $\boldsymbol{X}$ denotes the
124 genotype matrix for all cis- genotypes (encoded as the number of minor alleles), $\boldsymbol{w}$ denotes the
125 corresponding cis-eQTL effect-size vector, and $\boldsymbol{\epsilon}$ denotes the error term. The intercept term is
126 dropped in *Equation 1* for assuming both $\boldsymbol{E_g}$ and $\boldsymbol{X}$ are centered at 0. Generally, SNPs within
127 1MB of the flanking 5' and 3' ends (cis-SNPs) are included in this regression model and non-zero
128 $\widehat{\boldsymbol{w}}$ will be used for follow-up analysis. In particular, the GReX will be imputed by

129 $$\widehat{\boldsymbol{GReX}} = \boldsymbol{X_{new}}\widehat{\boldsymbol{w}},$$

130 with cis-SNP data $\boldsymbol{X_{new}}$ for GWAS samples.
131

132 Nonparametric Bayesian Method

133 Following the latent Dirichlet process regression (DPR) model proposed in previous
134 studies for predicting complex phenotypes[26], we assume a normal prior $N(0, \sigma_w^2)$ for the cis-eQTL
135 effect-sizes $(w_i, i = 1, \ldots, p)$ and a Dirichlet process (DP) prior[39] for the effect-size variance σ_w^2 (as
136 in *Equation 1*):

137 $$w_i \sim N(0, \sigma_w^2),\ \sigma_w^2 \sim D,\ D \sim DP\ (IG(a, b), \xi). \quad \textit{Equation 2}$$

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.

162 Elastic-Net and BSLMM Methods

163 The Elastic-Net model[17] (used by PrediXcan[11]) estimates the cis-eQTL effect-sizes $\hat{\boldsymbol{w}}$ in

164 *Equation 1* with a combination of L_1 (LASSO)[19] and L_2 (Ridge)[20] penalties by

165 $$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left(\left\| \boldsymbol{E_g} - \boldsymbol{X}\boldsymbol{w} \right\|_2^2 + \lambda \left(\alpha \|\boldsymbol{w}\|_1 + \frac{1}{2}(1-\alpha)\|\boldsymbol{w}\|_2^2 \right) \right),$$

166 where $\|\cdot\|_2$ denotes L_2 norm, $\|\cdot\|_1$ denotes L_1 norm, $\alpha \in [0,1]$ denotes the proportion of L_1

167 penalty, and λ denotes the penalty parameter. Particularly, PrediXcan[11] takes $\alpha = 0.5$ and tunes

168 the penalty parameter λ by a 5-fold cross validation.

169 As pointed out by previous studies[17; 21], the Elastic-Net model is equivalent to a Bayesian

170 model with a mixture Gaussian and Laplace (mixture normal) prior for $\boldsymbol{w}$, that is, $p(\boldsymbol{w}) \propto$

171 $\exp\left(-\lambda\left(\alpha\|\boldsymbol{w}\|_1 + \frac{1}{2}(1-\alpha)\|\boldsymbol{w}\|_2^2\right)\right)$. Whereas, the BSLMM[18] assumes a mixture of two normal as

172 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]

173 assumes all cis-SNPs have at least a small effect, which are normally distributed with variance

174 σ_2^2, and some proportion (π) of cis-SNPs have an additional effect, normally distributed with

175 variance σ_1^2. Particularly, with $\sigma_2^2 = 0$, BSLMM becomes BVSR[22], and with $\pi = 0$, the BSLMM

176 becomes the LMM[23]. Therefore, the DP normal mixture[26-28] as assumed by the DPR method

177 includes the parametric (mixture normal) priors for Bayesian Elastic-Net[21] and BSLMM[18] as

178 special cases, which is the main reason why DPR is a more generalized model including Elastic-

179 Net and BSLMM as special cases. Thus, we believe the DPR method can robustly model complex

180 genetic architecture and then improve the imputation R^2.

181

182 Association Study with Univariate Phenotype

183 Given individual-level GWAS data (genotype data $\boldsymbol{X}_{new}$, phenotype $\boldsymbol{Y}$, and covariant

184 matrix $\boldsymbol{C}$) and cis-eQTL effect-size estimates $\hat{\boldsymbol{w}}$, the follow-up TWAS (using a burden type gene-

185 based test[41]) is to test for association between $\widehat{\boldsymbol{GReX}} = \boldsymbol{X_{new}}\widehat{\boldsymbol{w}}$ and $\boldsymbol{Y}$ based on the following

186 generalized linear regression model

187 $$f(E[\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{C}]) = \eta\boldsymbol{C} + \beta\widehat{\boldsymbol{GReX}}\,. \qquad \textit{Equation 4}$$

188 Here, $f(\cdot)$ is a pre-specified link function, which can be set as identify function for quantitative

189 phenotype or set as logit function for dichotomous phenotype. The gene-based association test

190 is equivalent to test $H_0: \beta = 0$ in Equation 4.

191 If only summary-level data are available, we take the same approach as implemented by

192 the FUSION[12] method. Let $\boldsymbol{Z}$ denote the vector of Z-scores by single variant tests (Wald, likelihood

193 ratio, score tests, etc.) for all cis-SNPs. The burden Z-score for gene-based association test is

194 defined as

195 $$\tilde{Z} = \frac{\boldsymbol{Z}\widehat{\boldsymbol{w}}}{\sqrt{\boldsymbol{Z}\widehat{\boldsymbol{w}}}} = \frac{\boldsymbol{Z}\widehat{\boldsymbol{w}}}{\sqrt{\widehat{\boldsymbol{w}}'\boldsymbol{V}\widehat{\boldsymbol{w}}}}\ , \qquad \textit{Equation 5}$$

196 where $\boldsymbol{V}$ denotes the covariance matrix across study SNPs that can be estimated from training

197 data or reference panels such as 1000 Genome Project[42].

198

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

200 To test for association between multivariate phenotypes and imputed GReX of the focal

201 gene, we take a similar approach as the MultiPhen method[43] using individual-level data. For

202 example, consider two phenotypes $(\boldsymbol{Y_1}, \boldsymbol{Y_2})$ and a covariate matrix $\boldsymbol{C}$, we first adjust for the

203 covariates by taking the residuals $(\widetilde{\boldsymbol{Y_1}}, \widetilde{\boldsymbol{Y_2}})$ respectively from the linear regression models $\boldsymbol{Y_j} =$

204 $\boldsymbol{\eta C} + \boldsymbol{\epsilon},\ j = 1,2$. Then we test if the regression R^2 is significantly greater than zero ($H_0: R^2 = 0$) for

205 the following regression model

206 $$\widehat{\boldsymbol{GReX}}_g = \beta_1\widetilde{\boldsymbol{Y_1}} + \beta_2\widetilde{\boldsymbol{Y_2}} + \boldsymbol{\epsilon}\,. \qquad \textit{Equation 6}$$

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

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

213 $$\tilde{Z}_{joint} = \frac{\widetilde{\boldsymbol{Z}}\boldsymbol{J}}{\sqrt{\widetilde{\boldsymbol{Z}}\boldsymbol{J}}} = \frac{\widetilde{\boldsymbol{Z}}\boldsymbol{J}}{\sqrt{\boldsymbol{J}'\boldsymbol{V}_{\boldsymbol{Y}}\boldsymbol{J}}},\ \ \boldsymbol{J} = (1, \dots, 1)',$$

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

215

216 <u>Simulation Study Design</u>

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

227 We performed comprehensive scenarios that varied the proportion of causal SNPs (out of
228 2,799 SNPs, influenced gene expression) among values in the vector $p_{causal} =$
229 $(0.01, 0.05, 0.1, 0.2)$. We varied the proportion of gene expression variance explained by causal
230 SNPs (i.e., expression heritability), along with the proportion of phenotypic variance explained by

231 simulated gene expression levels (i.e., phenotypic heritability), among values in the vector
232 $\left(h_e^2, h_p^2\right) = \left((0.05,0.8), (0.1, 0.5), (0.2,0.25), (0.5,0.1)\right)$. The phenotypic heritability was selected
233 arbitrarily with respect to expression heritability such that the follow-up association study power
234 fell within the range of $(25\%, 85\%)$. We also considered various training sample sizes (100, 300,
235 499) for simulation scenario with $p_{causal} = 0.2$ and $\left(h_e^2, h_p^2\right) = (0.2, 0.25)$.

236 With genotype matrix $\boldsymbol{X_g}$ of the randomly selected causal SNPs (according to p_{causal}), we
237 generated effect-sizes w_i from $N(0, 1)$ and then re-scaled the effect-sizes to ensure the targeted
238 h_e^2. Gene expression levels were generated by $\boldsymbol{E_g} = \boldsymbol{X_g w} + \epsilon$, with $\epsilon \sim N(0, (1 - h_e^2))$. Then the
239 phenotype values were generated by $\boldsymbol{Y} = \beta \boldsymbol{E_g} + \epsilon$, where β was selected with respect to h_p^2 and
240 $\epsilon \sim N(0, (1 - h_p^2))$.

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

248

249 <u>ROS/MAP Data</u>

250 Samples in the ROS/MAP data were collected from participants of the Religious Orders
251 Study (ROS) and the Rush Memory and Aging Project (MAP), which are prospective cohort
252 studies of studying aging and dementia[30; 31; 33]. The ROS/MAP study recruited senior adults
253 without known dementia at enrollment, who underwent annual clinical evaluation. Brain autopsy
254 was done at the time of death for each participant. All participants signed an informed consent

255 and Anatomic Gift Act, and the studies were approved by the Institutional Review Board of Rush
256 University Medical Center, Chicago, IL. Specifically, microarray genotype data generated for
257 2,093 European-decent participants[44], were further imputed to the 1000 Genome Project Phase
258 3[42] in our analysis. The post-mortem brain samples (gray matter of the dorsolateral prefrontal
259 cortex) from ~30% these participants were profiled for transcriptomic data by next-generation
260 RNA-seqencing[45]. In this paper, we conducted gene-based association studies for two important
261 indices of AD pathology that were quantified with β-antibody specific immunostains[30; 31; 33]:
262 neurofibrillary tangle density (tangles) with stereology and β-amyloid load (amyloid) with image
263 analysis. The neurofibrillary tangle density quantifies the average Tau tangle density within two
264 or more 20μm sections from eight brain regions — hippocampus, entorhinal cortex, midfrontal
265 cortex, inferior temporal, angular gyrus, calcarine cortex, anterior cingulate cortex, and superior
266 frontal cortex. The β-amyloid load quantifies the average percent area of cortex occupied by β-
267 amyloid protein in adjacent sections from the same eight brain regions.
268

269 **Results**

270 Simulation Studies

271 In the simulation studies, we observed that the DPR method performed robustly with
272 respect to different causal proportions and gene expression heritability. Especially, when
273 $p_{causal} > 0.01$ DPR outperformed PrediXcan across all expression heritability values, giving
274 higher imputation R^2 in test data (Figure 1A). For example, when $p_{causal} = 0.2$, the average
275 imputation R^2 of 1,000 simulations was estimated as 4.54% by using DPR versus 2.64% by using
276 PrediXcan with $h_e^2 = 0.1$, while the average imputation R^2 was estimated as 12.02% by using
277 DPR versus 9.13% by using PrediXcan with $h_e^2 = 0.2$ (Table 1). When $p_{causal} = 0.01$, DPR
278 performed similarly as PrediXcan with $h_e^2 = (0.2, 0.5)$ but slightly out-performed PrediXcan with
279 $h_e^2 = (0.05, 0.1)$ (Table 1, Figure 1).

280 Consequently, when $p_{causal} > 0.01$ and $h_e^2 < 0.5$, or $p_{causal} = 0.01$ and $h_e^2 < 0.2$, the
281 power of association studies was higher by using DPR versus using PrediXcan imputation models
282 (Figure 1B). When $h_e^2 = 0.5$, using both imputation models led to comparable power for
283 association studies (Figure 1B). Even though both methods had similar over-estimated training
284 R^2 (Figure S2), the DPR method still resulted in higher imputation R^2 for test data (Table 1;
285 Figures 1A) and higher power for association studies (Figure 1B). In addition, from the simulation
286 studies with various training sample sizes (100, 300, 499) and $p_{causal} = 0.2$ and $(h_e^2, h_p^2) =$
287 $(0.2, 0.25)$, the imputation R^2 and TWAS power increases as sample size increases while the DPR
288 method consistently outperforms PrediXcan (Figure 2). Overall, these results demonstrated the
289 advantages of the DPR method for modeling the complex genetic architecture of transcriptomes,
290 especially when the causal proportions > 0.01 or the expression heritability < 0.2.
291
292
293 Real Applications to ROS/MAP Data
294 To illustrate the performance of the DPR method in real studies, we applied both DPR and
295 PrediXcan on the ROS/MAP data (see Methods). We trained the gene expression imputation
296 models using 499 samples that have both transcriptomic data for frontal cortex tissues and
297 genotype data (imputed to 1000 Genome Phase 3, with MAF > 5%, Hardy-Weinberg p-value >
298 10^{-5}, and genotype imputation R^2 > 0.3). A total of 15,583 genes had gene expression levels
299 after standard RNA-sequencing quality control. The gene expression levels were first adjusted for
300 age at death, sex, postmortem interval, study (ROS or MAP), batch effects, RNA integrity number
301 scores, and cell type proportions (with respect to oligodendrocytes, astrocytes, microglia, neurons)
302 by linear regression models. For each gene, cis- SNPs within the 1MB of the flanking 5' and 3'
303 ends were used in the imputation models as predictors.

304 In this real study, we compared transcriptome-wide 5-fold cross validation (CV) regression
305 R^2 estimated by using both DPR and PrediXcan methods. Specifically, we randomly split 499
306 training samples into 5 folds, where the imputation R^2 of each fold was calculated using the model
307 trained with the other 4 folds samples. If the training model is null, we take the imputation R^2 as
308 0 and take the average imputation R^2 across all 5-fold test samples as 5-fold CV R^2. The
309 transcriptome-wide median of 5-fold CV R^2 is 0.013 by DPR versus 0.005 by PrediXcan. The 5-
310 fold CV R^2 was used as the criterion for selecting significant imputation models ($R^2 > 0.01$ as
311 used by previous studies[11; 46]). From Figure 3A, we can see that the DPR method obtained more
312 imputation models and higher imputation R^2 when 5-fold CV R^2 is in the range of (0.01, 0.05),
313 which is also consistent with our simulation studies. Overall, the DPR method obtained significant
314 imputation models for 8,742 genes versus 5,547 genes by PrediXcan (with 57.6% increases).
315 Thus, the DPR method featuring data-driven nonparametric prior for the cis-eQTL is preferred in
316 real studies for identifying more genes with imputable expression levels.

317 Next, we used all 499 training samples to fit imputation models for genes with respective
318 5-fold CV $R^2 > 0.01$ by both DPR and PrediXcan, and then used these models to impute the
319 GReX for all GWAS samples. We conducted univariate phenotype association studies (Methods)
320 using all GWAS samples (N = 1,164) that have the AD pathology indices (neurofibrillary tangle
321 density and β-amyloid load, with Pearson correlation 0.48) quantified. Possible confounding
322 covariates including age at death, sex, study (ROS or MAP), smoking, education, and first 3
323 genotype principle components were adjusted in the association studies. Interestingly, the
324 association studies for both AD pathology indices using the DPR imputation models identified the
325 same top significant gene *TRAPPC6A* (within the 2MB region from the major risk gene *APOE*,
326 encoding apolipoprotein E, but independent of *APOE*) with p-values 1.64×10^{-5} and 5.35×10^{-5}
327 (Figure S3A and Figure S4A). Moreover, the multivariate phenotype association studies (Methods)
328 for both AD pathology indices also identified *TRAPPC6A* as the most significant gene with p-value

329 5.81×10^{-6} and FDR 0.08 (Figure 3C). On the other hand, the PrediXcan failed to obtain a
330 transcriptomic imputation model for *TRAPPC6A* (Figure S3B, Figure S4B, and Figure S6).
331 Quantile-Quantile plots for these TWAS p-values were presented in Figure S5.

332 In addition, for 14 known common and rare loci of late-onset AD[34-38] with significant
333 imputation models, we conducted association studies using transcriptomic imputation models
334 (DPR and PrediXcan) fitted from ROS/MAP data and standard GWAS results from IGAP[34]. Using
335 the imputation models fit by DPR, we identified 3 significant loci with FDR < 0.05 (Figure 3B) —
336 *ADAM10, CD2AP,* and *TREM2* — that potentially affect late-onsite AD risk through transcript
337 abundance. Here, *TREM2* was also identified by using the PrediXcan imputation model (Figure
338 3B). Particularly, the PrediXcan method only imputed GReX for 5 out of these 14 loci. In summary,
339 these results show that the DPR method has superior power for follow-up TWAS.

340

341 **Discussion**

342 In this paper, we propose an improved nonparametric Bayesian method for transcriptomic
343 data imputation, accompanied with an integrated tool (freely available on GITHUB) that is referred
344 as **T**ranscriptome-**I**ntegrated **G**enetic **A**ssociation **R**esource (**TIGAR**). TIGAR integrates both
345 nonparametric Bayesian (DPR) and Elastic-Net models as two options for transcriptomic data
346 imputation, along with TWAS options using individual-level and summary-level GWAS data for
347 univariate and multi-variate phenotypes. We show the usefulness of the DPR method by both in-
348 depth simulations and real applications using individual-level ROS/MAP[30-33] and summary-level
349 IGAP[34] data. We demonstrate that DPR is preferred over PrediXcan because of the flexible
350 nonparametric modeling of cis-eQTL effect-sizes that results in improved imputation R^2 for gene
351 expression levels and higher power for TWAS.

352 With respect to user-friendly interface and computational efficiency, TIGAR can: (i) Take
353 standard input files such as genotype files in VCF and dosage formats, phenotype files in PED

354 format, and a combined file (tab delimited) for gene annotations and expression levels; (ii) Load
355 input data per gene by TABIX for memory efficiency; (iii) Filter SNPs by given MAF and Hardy-
356 Weinberg p-value thresholds; (iv) Provide options of training both Elastic-Net (using Python
357 scripts) and DPR (using the executable tool developed with C++[26]) imputation models with unified
358 output format; (v) Implement multi-threaded computation to take full advantage of multi-core
359 clusters. These features make TIGAR a preferred tool for saving tedious data preparation and
360 computation time. For example, TIGAR can complete training imputation models for ~20K genes
361 and ~1K samples within ~20 hours and TWAS within ~1hour with a 2.4GHz 16-core CPU. The
362 computation time is expected to increase linearly with respect to the sample size.

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

369 As observed in the real ROS/MAP studies, there remains a large gap between the 5-fold
370 CV R^2 using cis-eQTL predictors (~5%) and the average genome-wide heritability of gene
371 expression levels (21.8% estimated by GCTA[48] based on a LMM). This is likely due to the large
372 trans-acting contribution to transcript abundance documented for most genes. Thus, we
373 hypothesize that it is promising to further improve the imputation R^2 by fitting transcriptomic
374 imputation models with genome-wide variants as predictors. Scalable Bayesian inference
375 techniques such as the Expectation Maximization MCMC (EM-MCMC) algorithm[49] are required
376 for incorporating genome-wide variants.

377 Another limitation of existing TWAS methods is that the uncertainty of cis-eQTL effect-size
378 estimates has not been taken into accounted in the association studies. A Bayesian framework

379 can also be derived by taking the standard errors of these cis-eQTL effect-size estimates as prior
380 standard deviations, which is part of our continuing research.

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

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

391

392 **Supplemental Data**

393 Supplementary Text and Figures

394

395 **Acknowledgments**

396 J.Y. was supported by the startup funding from Emory University Department of Human Genetics.
397 A.P.W. and T.S.W. were supported by National Institutes of Health (NIH) R01AG056533.
398 M.P.E. was supported by NIH R01GM11796. L.C.T. was supported by the Dermatology
399 Foundation, the Arthritis National Research Foundation, the National Psoriasis
400 Foundation, and NIH K01AR072129. ROS/MAP study data were provided by the Rush
401 Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL. Data collection was
402 supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917,

403 R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public
404 Health, and the Translational Genomics Research Institute. In addition, we thank Thanneer
405 Perumal and Benjamin Logsdon for performing quality control of the ROS/MAP RNA-sequencing
406 data and for creating the brain cell type proportions.

407

408 **Web Resources**

409 TIGAR, https://github.com/yanglab-emory/TIGAR

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

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

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

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

414

415 **References**

416 1. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery.
417 American journal of human genetics 90, 7-24.
418 2. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., and
419 Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus,
420 uncertainty and challenges. Nature reviews Genetics 9, 356-369.
421 3. Huang, Q. (2015). Genetic study of complex diseases in the post-GWAS era. J Genet Genomics
422 42, 87-98.
423 4. Farh, K.K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton,
424 H., Ryan, R.J., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal
425 autoimmune disease variants. Nature 518, 337-343.
426 5. Tsoi, L.C., Stuart, P.E., Tian, C., Gudjonsson, J.E., Das, S., Zawistowski, M., Ellinghaus, E., Barker,
427 J.N., Chandran, V., Dand, N., et al. (2017). Large scale meta-analysis characterizes genetic
428 architecture for common psoriasis associated variants. Nat Commun 8, 15382.
429 6. GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. Nature
430 550, 204-213.
431 7. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-
432 associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.
433 PLoS genetics 6, e1000888.

8. Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (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 research 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 genetics 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., GTEx Consortium, Nicolae, D.L., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nature genetics 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. Nature genetics 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., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature genetics 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. American journal of human genetics 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. American journal of human genetics 102, 904-919.
16. Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S.M., Yu, Z., Li, B., Muchnik, S., Shi, Y., et al. (2018). A statistical framework for cross-tissue transcriptome-wide association analysis. bioRxiv.
17. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301-320.
18. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. PLoS genetics 9, e1003264.
19. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological) 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. Annals of Applied Statistics 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. (2005). A unified mixed-model method for

477 association mapping that accounts for multiple levels of relatedness. Nature genetics 38,
478 203.
479 24. Huan, T., Liu, C., Joehanes, R., Zhang, X., Chen, B.H., Johnson, A.D., Yao, C., Courchesne, P.,
480 O'Donnell, C.J., Munson, P.J., et al. (2015). A systematic heritability analysis of the human
481 whole blood transcriptome. Human genetics 134, 343-358.
482 25. Lloyd-Jones, L.R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A.,
483 Metspalu, A., Dermitzakis, M., et al. (2017). The Genetic Architecture of Gene Expression
484 in Peripheral Blood. American journal of human genetics 100, 371.
485 26. Zeng, P., and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent
486 Dirichlet process regression models. Nat Commun 8, 456.
487 27. Blei, D.M., and Jordan, M.I. (2006). Variational inference for Dirichlet process mixtures.
488 Bayesian Anal 1, 121-143.
489 28. Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational Inference: A Review for
490 Statisticians. Journal of the American Statistical Association 112, 859-877.
491 29. Casella, G. (2001). Empirical Bayes Gibbs sampling. Biostatistics 2, 485-500.
492 30. Bennett, D.A., Schneider, J.A., Arvanitakis, Z., and Wilson, R.S. (2012). Overview and findings
493 from the religious orders study. Curr Alzheimer Res 9, 628-645.
494 31. Bennett, D.A., Schneider, J.A., Buchman, A.S., Barnes, L.L., Boyle, P.A., and Wilson, R.S. (2012).
495 Overview and findings from the rush Memory and Aging Project. Curr Alzheimer Res 9,
496 646-663.
497 32. Ng, B., White, C.C., Klein, H.U., Sieberts, S.K., McCabe, C., Patrick, E., Xu, J., Yu, L., Gaiteri, C.,
498 Bennett, D.A., et al. (2017). An xQTL map integrates the genetic architecture of the human
499 brain's transcriptome and epigenome. Nat Neurosci 20, 1418-1426.
500 33. Bennett, D.A., Buchman, A.S., Boyle, P.A., Barnes, L.L., Wilson, R.S., and Schneider, J.A. (2018).
501 Religious Orders Study and Rush Memory and Aging Project. J Alzheimers Dis 64, S161-
502 S189.
503 34. Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano,
504 A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al. (2013). Meta-analysis of 74,046
505 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature genetics
506 45, 1452-1458.
507 35. Reitz, C. (2014). Genetic loci associated with Alzheimer's disease. Future Neurol 9, 119-122.
508 36. Reitz, C. (2015). Novel susceptibility loci for Alzheimer's disease. Future Neurol 10, 547-558.
509 37. Sims, R., van der Lee, S.J., Naj, A.C., Bellenguez, C., Badarinarayan, N., Jakobsdottir, J., Kunkle,
510 B.W., Boland, A., Raybould, R., Bis, J.C., et al. (2017). Rare coding variants in PLCG2, ABI3,
511 and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. Nature
512 genetics 49, 1373-1384.
513 38. Yuan, X.Z., Sun, S., Tan, C.C., Yu, J.T., and Tan, L. (2017). The Role of ADAM10 in Alzheimer's
514 Disease. J Alzheimers Dis 58, 303-322.
515 39. Muller, P., and Mitra, R. (2013). Bayesian Nonparametric Inference - Why and How. Bayesian
516 Anal 8.
517 40. Carbonetto, P., and Stephens, M. (2012). Scalable Variational Inference for Bayesian Variable
518 Selection in Regression, and Its Accuracy in Genetic Association Studies. Bayesian Analysis
519 7, 73-107.

520 41. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common
521 diseases: application to analysis of sequence data. American journal of human genetics
522 83, 311-321.
523 42. 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092
524 human genomes. Nature 491, 56-65.
525 43. O'Reilly, P.F., Hoggart, C.J., Pomyen, Y., Calboli, F.C., Elliott, P., Jarvelin, M.R., and Coin, L.J.
526 (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS.
527 PloS one 7, e34861.
528 44. De Jager, P.L., Shulman, J.M., Chibnik, L.B., Keenan, B.T., Raj, T., Wilson, R.S., Yu, L., Leurgans,
529 S.E., Tran, D., Aubin, C., et al. (2012). A genome-wide scan for common variants affecting
530 the rate of age-related cognitive decline. Neurobiol Aging 33, 1017 e1011-1015.
531 45. De Jager, P.L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L.C., Yu, L., Eaton, M.L.,
532 Keenan, B.T., Ernst, J., McCabe, C., et al. (2014). Alzheimer's disease: early alterations in
533 brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat Neurosci 17, 1156-
534 1163.
535 46. Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., Bolla, M.K., Shu, X.O., Lu, Y., Cai,
536 Q., et al. (2018). A transcriptome-wide association study of 229,000 women identifies new
537 candidate susceptibility genes for breast cancer. Nature genetics 50, 968-978.
538 47. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J.,
539 Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic
540 Risk Prediction across Diverse Populations. American journal of human genetics 100, 635-
541 649.
542 48. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide
543 complex trait analysis. American journal of human genetics 88, 76-82.
544 49. Yang, J., Fritsche, L.G., Zhou, X., Abecasis, G., and International Age-Related Macular
545 Degeneration Genomics, C. (2017). A Scalable Bayesian Method for Integrating Functional
546 Information in Genome-wide Association Studies. American journal of human genetics
547 101, 404-416.
548

***Table 1**. Average imputation R^2 in the test data under various simulation scenarios, when 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 imputation R^2 in the test samples are bold.*

h_e^2	Causal Proportion 0.01		Causal Proportion 0.2	
	DPR	PrediXcan	DPR	PrediXcan
0.05	**1.60%**	1.12%	**1.54%**	0.76%
0.1	**4.44%**	4.13%	**4.54%**	2.64%
0.2	12.26%	**12.29%**	**12.02%**	9.13%
0.5	39.30%	**42.05%**	**39.32%**	36.04%

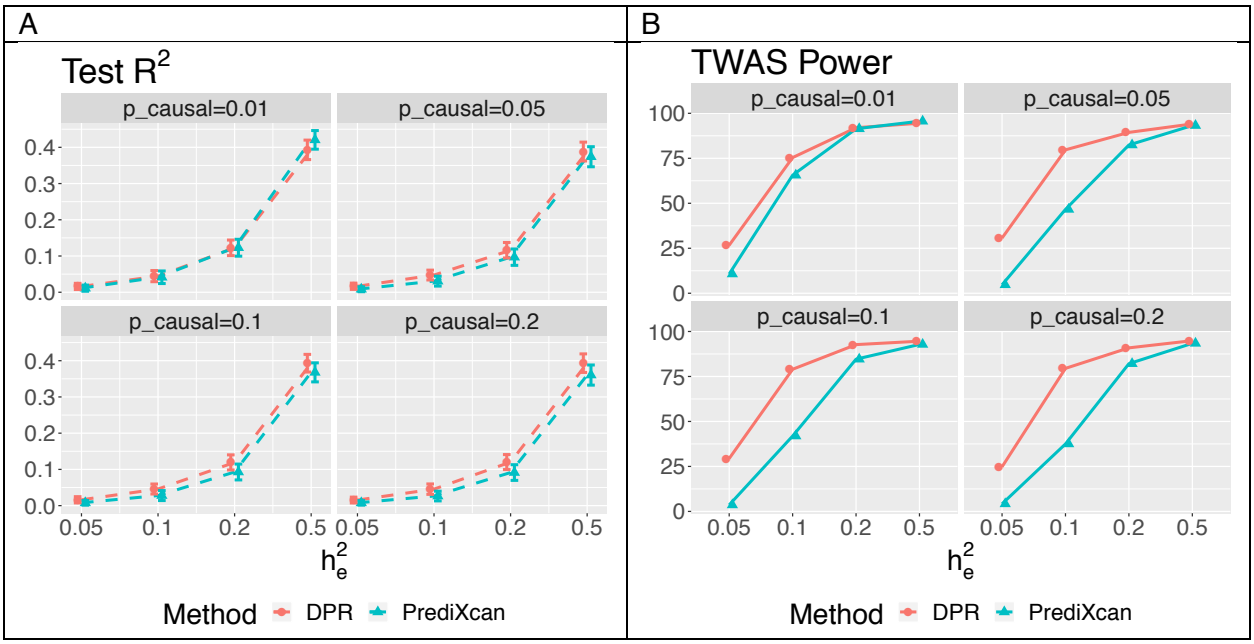

***Figure 1**. Plots of average imputation R^2 (**A**) and TWAS power (**B**) in test samples by DPR and PrediXcan, with various proportions of true causal SNPs $p_{causal} = (0.01, 0.05, 0.1, 0.2)$ and true expression heritability $h_e^2 = (0.05, 0.1, 0.2, 0.5)$. TWAS Power was evaluated with paired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

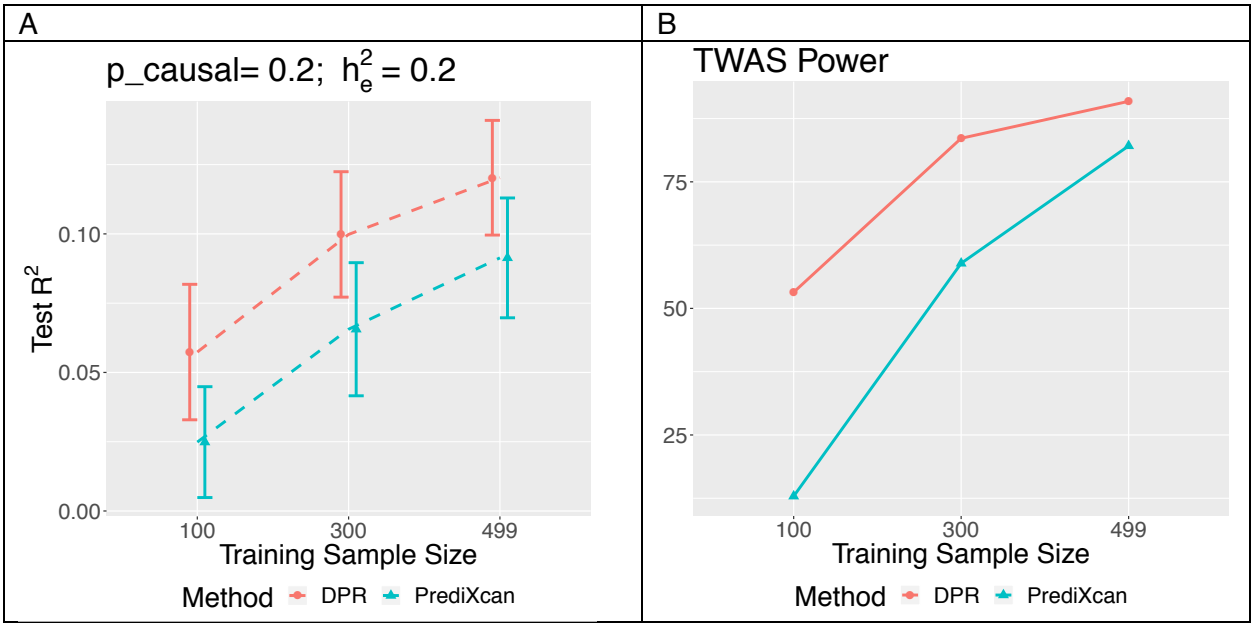

Figure 2. *Test* R^2 (***A***) *and TWAS power* (***B***) *from simulation studies with causal proportion* $p_{causal} = 0.2$, *expression heritability and phenotype heritability* $(h_e^2, h_p^2) = (0.2, 0.25)$, *and various training sample sizes* $(100, 300, 499)$.

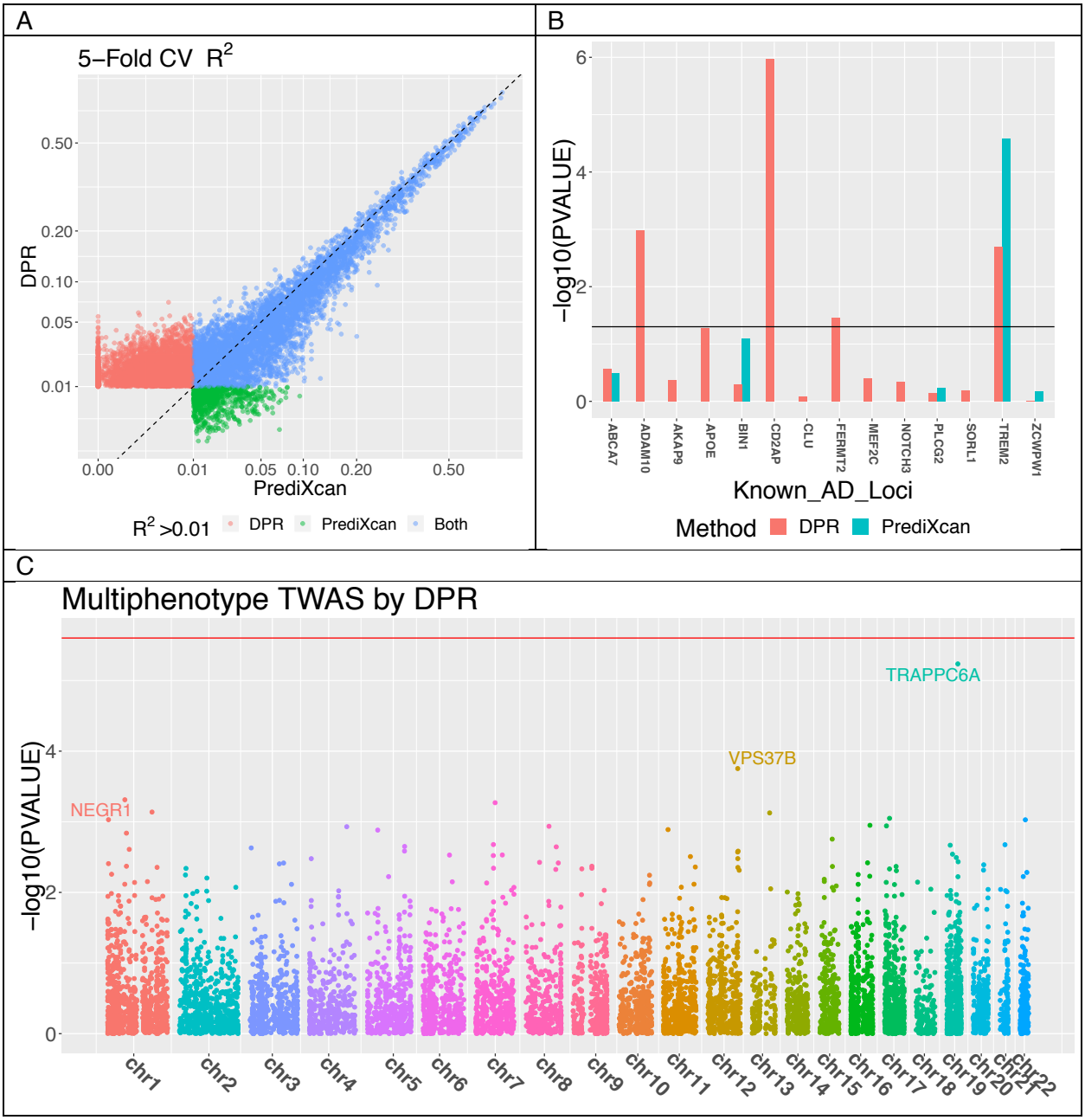

Figure 3. *Transcriptome-wide 5-fold cross validation R^2 (**A**) by PrediXcan and DPR with 499 ROS/MAP training samples, with different colors denoting whether the imputation $R^2 > 0.01$ by DPR, PrediXcan, or both methods (genes with $R^2 < 0.01$ by both DPR and PrediXcan were excluded from the plot). 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.*