Distribution Agreement

Signature:

________________________________ _______________

Xiaoran Meng Date

Development of Statistical Tool TIGAR for Transcriptome-Integrated Genetic Association Resource

By

Xiaoran Meng
Master of Science in Public Health

Department of Biostatistics and Bioinformatics

____

Hao Wu
Advisor

____

Jingjing Yang
Reader

Development of Statistical Tool TIGAR for Transcriptome-Integrated Genetic Association Resource

By

Xiaoran Meng
B.S, Shandong University, 2017

Thesis Committee Chair: Hao Wu, Associate Professor
Jingjing Yang, Associate Professor

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics
2019

Abstract

Development of Statistical Tool TIGAR for Transcriptome-Integrated Genetic Association Resource

By Xiaoran Meng

Transcriptome-wide association studies (TWAS) have been used to leverage reference data that have both transcriptomic and genetic profiles for the same samples in gene-based genome-wide association studies (GWAS). Basically, an imputation model for genetically regulated gene expression levels (GReX) per gene tissue type can be fitted by applying regression models on the reference data, where the effect-sizes of cis-expression quantitative trait loci (cis-eQTL) on expression levels will be estimated and used as variant weights in gene-based association studies. Many statistical tools have been developed for implementing TWAS, such as PrediXcan based on the Elastic-Net regression model and FUSION based on the Bayesian sparse linear mixed model (BSLMM). However, existing tools only implement parametric regression models to fit GReX imputation models, which have limitations to fully model the complex genetic architecture of transcriptome profiles. Recently proposed nonparametric Bayesian Dirichlet process regression (DPR) model has been shown improved the imputation accuracy of GReX over parametric regression models. Thus, my thesis is focused on developing a statistical tool to implement both parametric Elastic-Net model and nonparametric DPR model for fitting GReX imputation models and enable follow-up TWAS with both individual-level and summary-level GWAS data. To make the tool computationally efficient, I used advanced computational techniques such as multi-threading for parallel computation and TABIX for loading genotype data with memory efficiency. The tool is referred as Transcriptome-integrated Genetic Association Resource (TIGAR). In addition, to illustrate the advantages of TIGAR, I applied the tool on GTEx reference dataset to train GReX imputation models of brain frontal cortex tissue, and then conducted TWAS on ROS/MAP GWAS data for 4 different complex traits related to Alzheimer's Disease (AD) -- neurofibrillary tangle density, β-amyloid load, global AD pathology burden and final consensus cognitive diagnosis. Application results show that the DPR model obtained higher R^2 in both training and prediction data, and the Elastic-Net model lead to 3 potentially significant genes (with FDR 0.077) that might be associated with β-amyloid load. Overall, TIGAR is expected to provide a user-friendly, flexible, and computationally efficient tool for implementing TWAS.

Development of Statistical Tool TIGAR for Transcriptome-Integrated Genetic Association Resource

By

Xiaoran Meng
B.S, Shandong University, 2017

Thesis Committee Chair: Hao Wu, Associate Professor
Jingjing Yang, Associate Professor

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics
2019

Acknowledgments

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.

Contents

1 Introduction

The development of next-generation sequencing (NGS) [1, 2, 3] allows us to discover, sequence and genotype thousands of markers across any genome in a single step, which makes genome-wide association studies (GWASs) for organisms and wild populations possible. A GWAS [4, 5] is an observational study of genome-wide genetic variants across individuals to detect associations between these variants and phenotypic traits [6, 7]; e.g. associations between common single-nucleotide polymorphisms (SNPs) and a disease [8]. Results from previous GWASs indicated that a majority of genetic variants are found in non-coding regions, which have been shown to be enriched with expression of quantitative trait loci (eQTL) [9, 10].

Even though GWASs contribute a lot in detecting associations of genetic variants and complex traits, they still face two major challenges: (i) Relying on increasing sample size to improve statistical power for detecting expression-trait associations and (ii) SNPs identified by GWASs often reside in non-coding regions, which lead to difficulties in interpreting their functions and their associations with complex traits [8, 11, 12, 13]. An informative and easily measurable source of functional information is gene expression. Comparing the similarity of genes' expression profiles (co-expression) serves as a powerful means for interpreting GWAS candidate SNPs [14].

Gene expression involves two main steps – transcription and translation [15]. Transcription relates to the production of mRNA enzymes and RNA polymerase, and processing of mRNA molecules. Translation involves the use of mRNA to synthesize proteins and is followed by the post-translational processing of the protein molecules. Usually, variants within

1 Mb (megabase) on either side of the gene's transcribing start site (TSS) are called cis [16]. cis-eQTLs can potentially influence gene expression level by altering transcription factors (TFs) [17, 18, 19], which are proteins that regulate the proportion of transcription of genetic information from DNA to mRNA through binding to a specific DNA sequence.

In order to integrate transcriptomic data in GWAS and leverage reference data, recent studies have proposed transcriptome-wide association studies (TWASs) [20, 21, 22]. TWASs involve imputing genetic expression components in a large group of subjects from a relatively small set individuals with both gene-expression levels and genotype data known. TWAS is helpful in detecting an expression-trait association when individual-level GWAS is available. Many statistical tools have been developed for implementing TWAS, like PrediXcan [23, 24] and FUSION [20]. PrediXcan is based on the Elastic-Net regression model and FUSION is based on the Bayesian sparse linear mixed model (BSLMM). Elastic-Net assumes a mixed penalty of LASSO (L_1) and Ridge (L_2) in the linear regression model and BSLMM is a combination of Bayesian variable selection model (BVSR) and linear mixed model. In TWAS, PrediXcan and FUSION treat eQTL effect-size [25, 26] as SNP weight to make use of reference transcriptomic data in large GWAS. However, Elastic-Net and BSLMM assume parametric prior for cis-eQTL effect-sizes, which make it difficult to capture complex genetic architecture.

Previous studies have shown that the nonparametric Bayesian regression model is preferred for moding the complex genetic architecture of gene expression levels. Basically, the nonparametric Bayesian model assumes a Dirichlet process prior on the effect-size varaince of cis-eQTL [27]. DPR is a more generalized model that can include Elastic-Net and BSLMM as special cases. I developed a tool containing both Elastic-Net regression and DPR called

Transcriptome-Integrated Genetic Association Resource (TIGAR, `https://github.com/xmeng34/TIGAR`). TIGAR focuses on implementing both Elastic-Net and DPR models to impute transcriptomic data and run TWAS with individual and summary level GWAS data. To make the tool computationally efficient, I used advanced computational techniques such as multi-threading for parallel computation and TABIX for loading genotype data with memory efficiency. Some user-friendly options such as taking standard input files are also available for TIGAR. Generally, TIGAR can train genetically regulated gene expression (GReX) imputation models for Elastic-Net or DPR, along with TWAS for one gene in about 4 minutes. Comparing to similar existing tools that accept specific input files for data imputation and subsequent association studies, which require cumbersome data preparation, large memory space to loading genotype data, TIGAR not only takes care of tedious works to prepare input files, also provides options of imputation models and computation efficacy.

In this thesis, I will apply the tool on Genotype-Tissue Expression project (GTEx [28]) reference dataset to train GReX imputation models of brain frontal cortex tissue, and then conducted TWAS on Religious Orders Study (ROS [29]) and Rush Memory and Aging Project (MAP [29, 30]) GWAS data for 4 different complex traits including neurofibrillary tangle density, $\beta-$amyloid, global AD pathology burden and final consensus cognitive diagnosis related to Alzheimer's Disease (AD). The goal of the study is to test the performance of TIGAR on GTEx and ROS/MAP data and detect associations between specific AD indices and potential genes.

2 TIGAR

“TIGAR” stands for Transcriptome-Integrated Genetic Association Resource, which is developed using Python and BASH. TIGAR treats both Elastic-Net and Dirichlet Process Regression as training imputation models for transcriptomic data, following prediction of gene expression level and conduct genetic association tests using both individual-level and summary-level GWAS data for univariate and multivariate phenotypes. The main idea of developing TIGAR is to provide computational convenience with integrated functions for training imputation models and requirement of standard input files to run model training, prediction and conduct association study. To save calculation time, TIGAR also provides scalable muti-thread options. In general, with user-friendly inputs, TIGAR can provide training cis-eQTL effect-sizes, predicted gene expression level and TWAS for one gene in about 4 minutes.

2.1 cis-eQTL Effect-Sizes Calculation

Generally, SNPs within 1Mb of the gene boundary will be included in regression model and genetically regulated gene expression (GReX) can be imputed through $\widehat{GReX} = \mathrm{X}_{new}\hat{\mathrm{w}}$ with new genotype data X_{new}.

2.1.1 Elastic-Net Regression

Elastic-Net regression [31] method assumes linear regression model as follow:

$$\mathrm{E}_g = \mathrm{Xw} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, \boldsymbol{\sigma^2}) \tag{1}$$

$$\hat{\mathrm{w}} = \underset{\mathrm{w}}{\operatorname{argmin}}(\|\mathrm{E}_g - \mathrm{Xw}\|_2^2 + \lambda(\alpha\|\mathrm{w}\|_1 + \frac{1}{2}(1-\alpha)\|\mathrm{w}\|_2^2)), \alpha \in [0,1] \tag{2}$$

E_g represent gene expression level for specific gene g, usually corrected for confounding covariates like age, gender and genotype principle components. X is the genotype matrix, w denotes effect-size vector of corresponding SNPs and $\boldsymbol{\epsilon}$ is the error term. In this model, cis-eQTl effect-size w is estimated by adding a mixture of LASSO (L_1) and Ridge (L_2) penalties, where α denotes proportion of L_1 and L_2 penalty and λ is the penalty parameter. Specifically, PrediXcan assumes $\alpha = 0.5$ and picks λ by 5-folds cross validation.

2.1.2 Dirichlet Process Regression (DPR)

The linear regression model is quite similar as (1). According to latent Dirichlet process regression [32], the model assumes

$$\mathrm{E}_g = \mathrm{Xw} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, \boldsymbol{\sigma^2}), \sigma^2 \sim IG(a_\epsilon, b_\epsilon) \tag{3}$$

$$w_i \sim N(0, \sigma_w^2), \sigma_w^2 \sim D, D \sim DP(ID(a,b), \xi) \tag{4}$$

Where w_i denotes effect-size for each SNP within in gene g, which follows a normal distribution with mean 0 and variance σ^2 with Dirichlet process prior D that has base distribution inverse gamma IG(a,b) and concentration parameter ξ. After integrating out latent variable σ^2, an equivalent non-parametric prior distribution of w_i can be driven as follow:

$$w_i \sim \sum_{k=1}^{+\infty} \pi_k N(0, \sigma_k^2), \sigma_k^2 \sim IG(a_k, b_k), \pi_k = v_k \prod_{l=1}^{k-1}(1 - v_l), v_k \sim Beta(1, \xi) \tag{5}$$

Here, ξ means the same concentration parameter in (3) with a hyper prior $\xi{\sim}Gamma(a_\xi, b_\xi)$. DPR model is more robust in detect gene structure due to non-informative prior for σ_k^2,σ^2 and ξ, which usually assumes a_k, b_k,a_ϵ and b_ϵ as 0.1 and (a_ξ, b_ξ) as (1,0.1), then σ_k^2, σ^2 and ξ can be estimated through data and make w_i data-driven.

2.2 TWAS

2.2.1 Univariate Phenotype

With given weight (SNP effect-sizes) w, individual genotype X_{new}, single phenotype Y and covariance matrix C, the association test [33] of $\widehat{\mathrm{GReX}}$ and Y is conducted through linear regression model

$$f(E[Y|\mathrm{X},\mathrm{C}]) = \eta\mathrm{C} + \beta\widehat{\mathrm{GReX}} \tag{6}$$

$f(\cdot)$ is a pre-specified function and $H_0 : \beta = 0$ is the same with gene-based association test. TIGAR can also run association test through summary-level data when new genotype data is not provided. Let Z represent single-variance test for all cis-SNPs. Burden Z-score of association test is defined as

$$\widetilde{Z} = \frac{\mathrm{Z}\hat{\mathrm{w}}}{\sqrt{\hat{\mathrm{w}}^T\mathrm{V}\hat{\mathrm{w}}}} \tag{7}$$

Here, V denotes covariance matrix across training SNPs, which I can calculated through training genotype data.

2.2.2 Multivariate Phenotype

Association test for multivariate phenotype and imputed GReX is conducted through model as follow

$$\begin{aligned} Y_j &= \eta \mathrm{C} + \epsilon, j = 1, 2, ..., n \\ \widetilde{Y_j} &= Y_j - \hat{Y_j} \end{aligned} \tag{8}$$

$$\widehat{GReX}_g = \sum_{j=1}^{n} \beta_j \widetilde{Y_j} + \epsilon \tag{9}$$

Here $Y_j, j = 1, 2, ..., n$ represent n different phenotypes and C is a covariance matrix. In (8), TIGAR first adjust for covariates by calculating residual $\widetilde{Y_j}, j = 1, 2, ..., n$ for each phenotype. Association study is conducted base on R^2 from (9), which is the same as $H_0 : R^2 \neq 0$.

2.3 Computational Advantages

I implied the following functions in TIGAR to make it computationally efficient and users friendly.

(i) TIGAR accepts standard input files like vcf/dosages format for genotype data and PED for phenotype data. Original PrediXcan and DPR software require users to prepare training input files in specific formats, which are far from usually used standard files.

(ii) To calculate cis-eQTL effect-size, the original PrediXcan tool needs to run three python scripts for file preparation, model training and result generating. For DPR, training results are stored by genes, which is difficult for users to view results genome-wide. However, TIGAR can complete model training in only one command and collected output by chromosome.

(iii) Original PrediXcan tool fixed their default value in the scripts, like L_1 and L_2 penalties ratio for the Elastic-Net model. TIGAR make these default values users defined. Users can modify these values by adding up command when running TIGAR.

(iv) TIGAR reads in genotype data by TABIX for memory efficiency and provides a multi-thread option to execute multiple processes for computation efficiency.

(v) TIGAR provides minor allele frequency (MAF) and p-value for Hardy Weinberg Equilibrium exact test (HWE) calculation. Samples with $MAF > 0.01$ and $HWE > 0.001$ (thresholds can be users defined) will be used in model training, these values can help us exclude rare variances in a gene. In prediction, TIGAR will exclude samples with MAF different greater than 0.2 (threshold can be users defined) compare to training genotype data since large MAF different with the same SNP might indicate different races.

(vi) TIGAR runs 5-fold cross validation calculate average prediction R^2 before run model training with whole samples. If the average R^2 for the cross-validation is less than 0.01, TIGAR assumes that Elastic-Net and DPR model might not be valid for calculating cis-eQTL effect-sizes for this gene.

3 Data Description

3.1 ROS/MAP Data

ROS/MAP data were collected from participants of Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP), which are jointly designed prospective studies of aging and dementia with longitudinal cognitive phenotypes and structured neuropathologic examination after death [29, 30, 34]. MAP was designed as a complementary and extension of ROS, especially in organ donation. Study design for ROS and MAP are similar, both restricted to Catholic clergy and included participants who agree to annual clinical evaluation and organ donation, with a restricted range of life experiences and socioeconomic status and without dementia. However, comparing to ROS, MAP planned to enroll participants in a wider range [35]. Specifically, most of the subjects in ROS/MAP projects have agreed to annual clinical evaluation and brain donation at death, which overlap with brain tissue data record in GTEx. In this thesis, I'm mainly focusing on micro-array genotype data generated from 2,093 European participants [36] that are imputed to the 1,000 Genome Project Phase 3 in this analysis. The post-mortem brain samples (gray matter of the dorsolateral prefrontal cortex) were recorded for transcriptomic data by NGS from a subset of these participants [37]. I conducted gene-based association tests for four different phenotypes [29, 30]: (i) Neurofibrillary tangle density (tangles) is one of Alzheimer's Disease (AD) pathologic indice related to $\beta-$antibody immunostains that derived by Tau protein density from 8 brain regions; (ii) $\beta-$amyloid load (amyloid) is also an AD indice that quantifies average percent area of cortex contains $\beta-$amyloid protein within the same brain regions of tangles [29, 30]; (iii) Global AD pathology burden (gpath) is a quantitative summary of AD pathology calculated from three AD pathologies: neuritic plaques, diffuse plaques, and neurofibrillary tangles, as determined by microscopic examination of silver-stained slides from 5 regions: midfrontal cortex, midtemporal cortex,inferior parietal cortex, entorhinal cortex, and hippocampus; (iv) Final consensus cognitive diagnosis (cogdx) related to clinical consensus diagnosis of cognitive status at time of death.

3.2 GTEx Data

The Genotype-Tissue Expression (GTEx) [38] project was started in 2010 aiming at establishing a resource database and associated tissue bank to characterize human transcriptome within and across individuals. The GTEx project started with a 2-year pilot phase to establishing an autopsy program, which yields robust gene expression measurement. GTEx donors are identified through low PMI (post-mortem-interval) autopsy or organ and tissue transplantation settings [39]. The GTEx data resource contains whole-genome sequences and RNA-sequences from about 650 adult donors, with 54 tissue samples per donor (Figure 1).

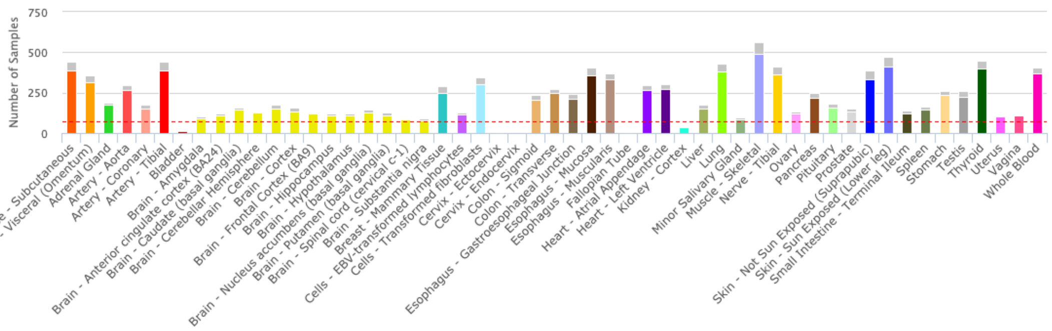


Figure 1: Overview of GTEx Project Samples [41] in 2019

To impute GTEx phenotype data, I first select genes with expression thresholds of > 0.1 RPKM (Reads Per Kilobase of transcript, per Million mapped reads is a normalized unit of transcript expression) in $>= 10$ samples. Then regressing out age, sex, first 4 PCs calculated from genotype data and peer factors calculated from top 10,000 genes that have higher expression level, using regress residuals as imputed gene expression level.

4 Application of TIGAR

4.1 Application Analysis Steps

I first compared the performance of the Elastic-Net and DPR model with respect to imputation R^2 in both training and testing data. I treated RNA-sequencing and genotype data from 129 GTEx participants recorded with Brain Frontal-Cortex(BA9) data as training data, and genotype data from 499 ROS/MAP participants as test data. The test samples were recorded with gene-expression levels for brain tissue, which helped a lot in comparing prediction R^2 for both models. The genotype and imputed genetics data for SNPs with MAF> 0.01 (European samples) and the p-value of Hardy-Weinberg test> 0.001 are included in the training model for each gene. In model training, I included a 5-fold cross validation before starting model training with whole samples. This procedure relates randomly split samples into five groups and treats one of the groups as test data and remaining as training data each time. Then running corresponding model (Elastic-Net/DPR) on training set following by prediction of gene expression level and calculate prediction R^2 based on the test set. For each gene, I calculated average prediction (or cross-validation) R^2 and set 0.01 as the threshold to decide whether Elastic-Net or DPR is valid in this case.

Next, I imputed GReX for all ROS/MAP samples using cis-eQTL effect-size from both Elastic-Net and DPR and calculate prediction R^2 based on 11846 genes recorded with true gene expression data.

Finally, I conducted gene-based association studies using all samples that have amyloid ($N = 1022$, N represents sample size), tangles ($N = 1024$), gpath ($N = 1053$) and cogdx ($N =$

1165) quantified. Confounding covariates include sex, age at death, smoking status, study (ROS or MAP) and top 3 genotype principle components were adjusted for association studies.

4.2 Application Result

As for training imputation R^2, a total of 7968 (31.5%) genes have significant imputation models with median average cross-validation R^2 3.6% and mean cross-validation R^2 8.0%. Meanwhile, a total number of 20208 (79.8%) genes have significant imputation models by DPR, with median average cross-validation R^2 4.2% and mean average cross-validation R^2 5.7% (Figure 2, Table 1). Specifically, when comparing 6771 (26.7%) genes that pass the threshold for both Elastic-Net and DPR, it turns out median average cross validation R^2 from Elastic-Net model is 5.4% versus DPR 3.6%, with mean average cross-validation R^2 7.5% versus 8.1% separately (Table 3). Although DPR fits significant imputation models for more number of genes, there is no significant difference in average cross-validation R^2 for genes that can be imputed by both models.

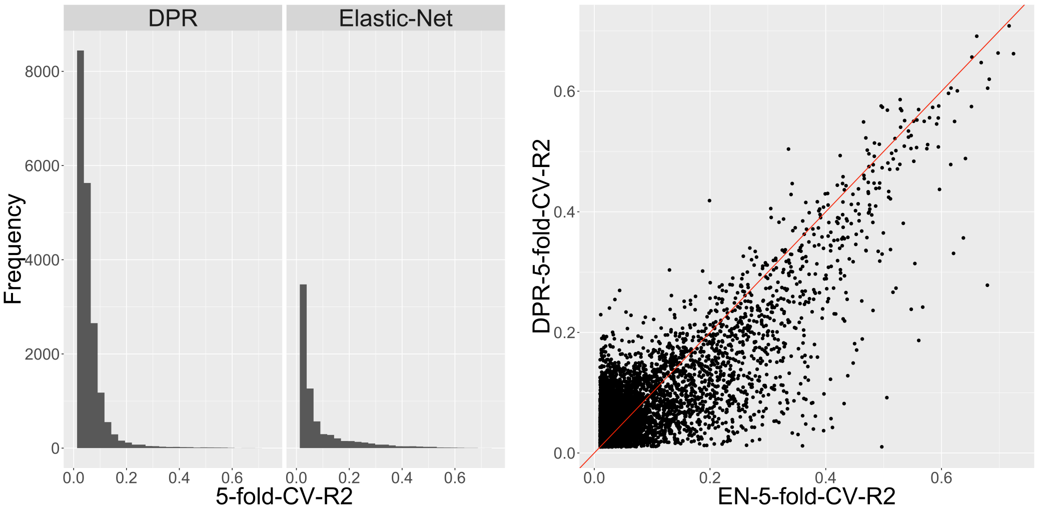


Figure 2: Average Cross-Validation R^2 Comparison

Model	Number of Significant Gene	Median R^2	Mean R^2
Elastic-Net	7968 (31.5%)	3.6%	8.0%
DPR	20208 (79.8%)	4.2%	5.7%

Table 1: Average Cross-Validation R^2 Comparison

In model prediction, Elastic-Net imputed 3545 (29.8%) genes with median prediction R^2 0.6% and mean prediction R^2 5.7%, versus 11197 (94.5%) genes with median prediction R^2 0.2% and average prediction R^2 1.7% by DPR (Figure 3, Table 2). When comparing 2885 (24.3%) genes that can be imputed by both Elastic-Net with median R^2 0.6% and mean R^2 4.2% and DPR with median R^2 0.7% and mean R^2 6.0% (Table 3). As a result, DPR gives higher prediction R^2 for those overlapped genes.

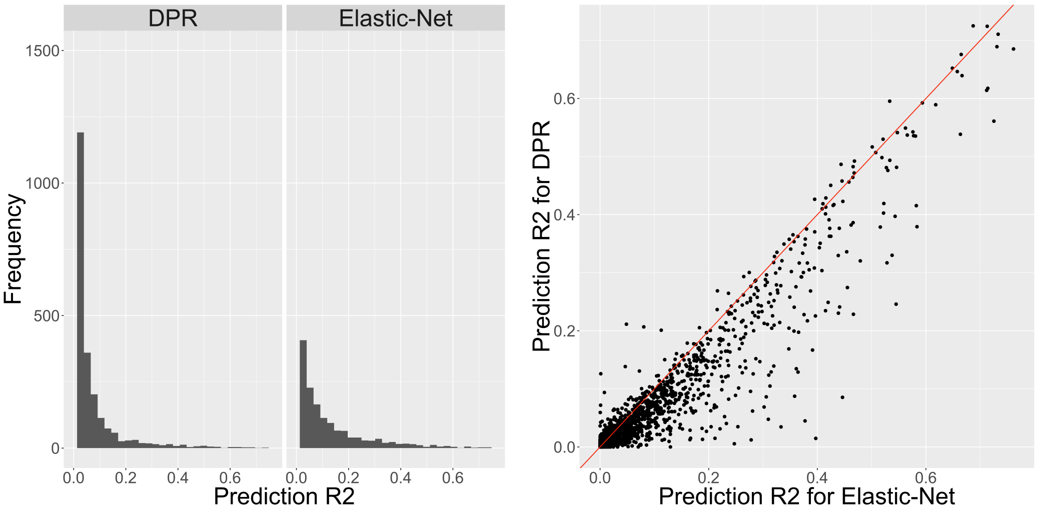


Figure 3: Prediction R^2 Comparison

Model	Number of Gene	Median R^2	Mean R^2
Elastic-Net	3534 (29.8%)	0.6%	5.7%
DPR	11197 (94.5%)	0.2%	1.7%

Table 2: Prediction R^2 Comparison

	Training		Prediction	
Number of Overlap Gene	6771 (26.7%)		2885 (24.3%)	
Median R^2	Elastic-Net	5.4%	Elastic-Net	0.6%
	DPR	3.6%	DPR	0.7%
Mean R^2	Elastic-Net	7.5%	Elastic-Net	4.2%
	DPR	8.1%	DPR	6.0%

Table 3: R^2 Comparison for Overlapping Genes

Finally, Manhattan plots with genome-wide significant threshold 2.5×10^{-6} and Q-Q plots for p-values of TWAS by Elastic-Net and DPR model (Figure 4-15) show that no genes in both models pass the significant threshold. Meanwhile, Q-Q plots show little deflation (p-values are systematically less significant than the expected distribution) and $\lambda_{GC} < 1.1$ (genomic control factor) as usual [40]. Then I calculate FDR adjusted p-value with significant threshold 0.1 to identified significant genes. The Elastic-Net model identified 3 significant loci (Table 4) – RP11-769N22.1, SBDS, and AC004951.5 with all FDR 0.077 that potentially affect amyloid traits through transcriptomes. No significant gene is identified by DPR. This might cause by the fact that no gene pass genome-wide significant threshold 2.5×10^{-6} in Manhattan plots for both model and genes that identified by Elastic-Net have FDR p-value less than 0.05, which suggest these genes are no significant enough to capture by DPR.

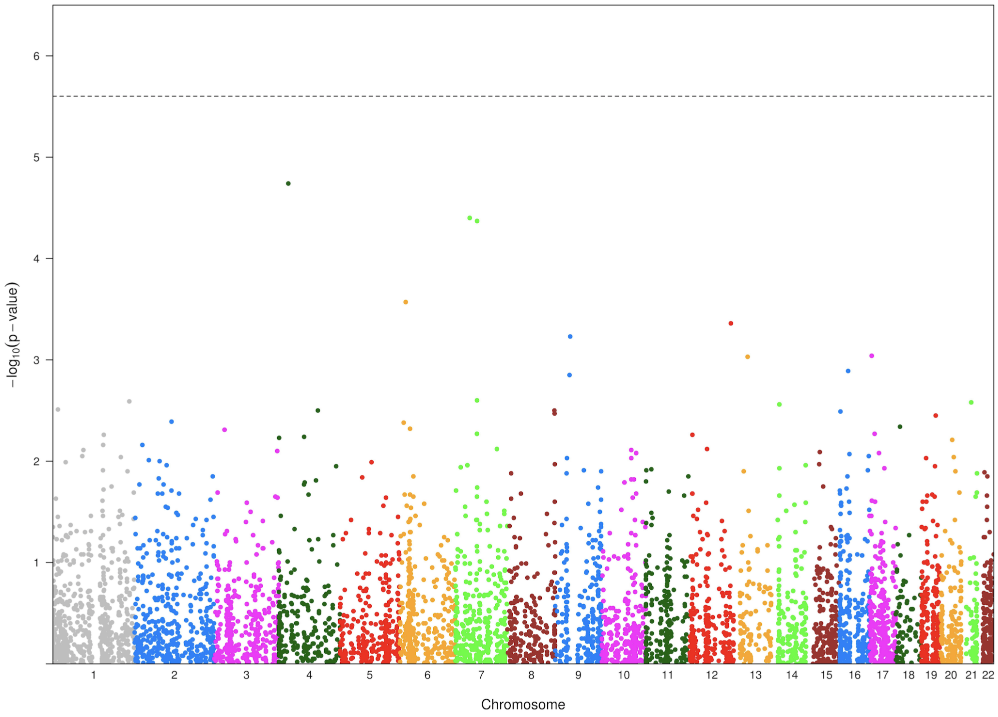


Figure 4: Manhattan Plot for TWAS p-values of amyloid traits by Elastic-Net

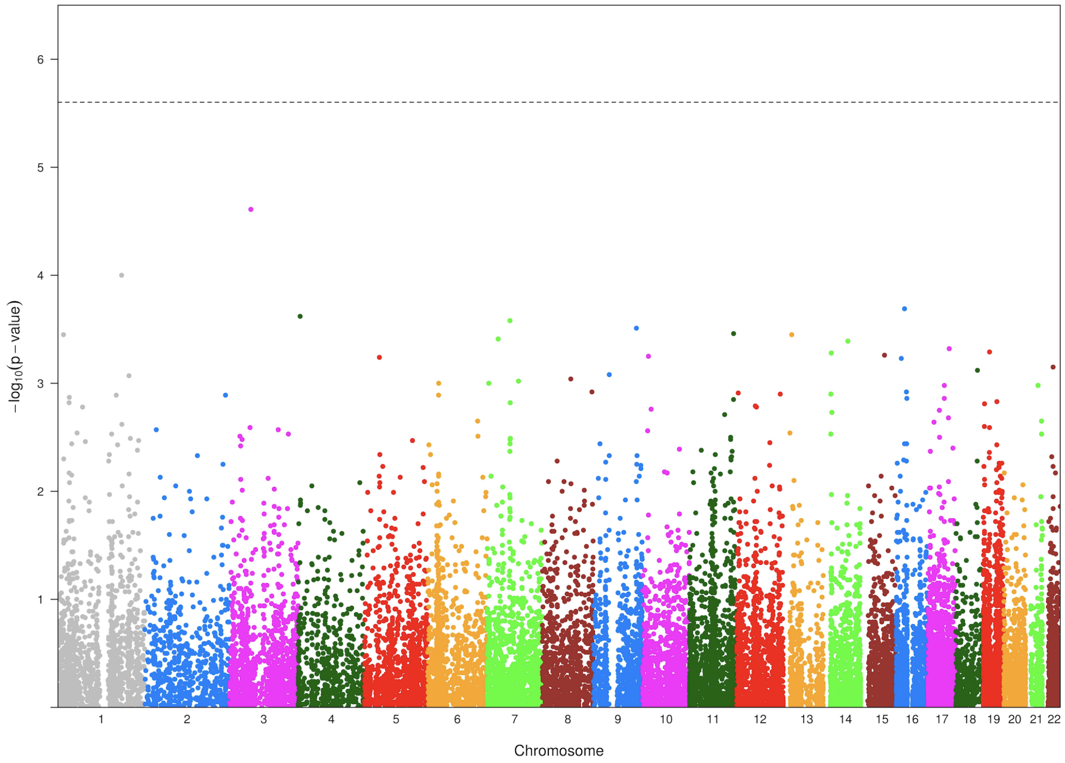
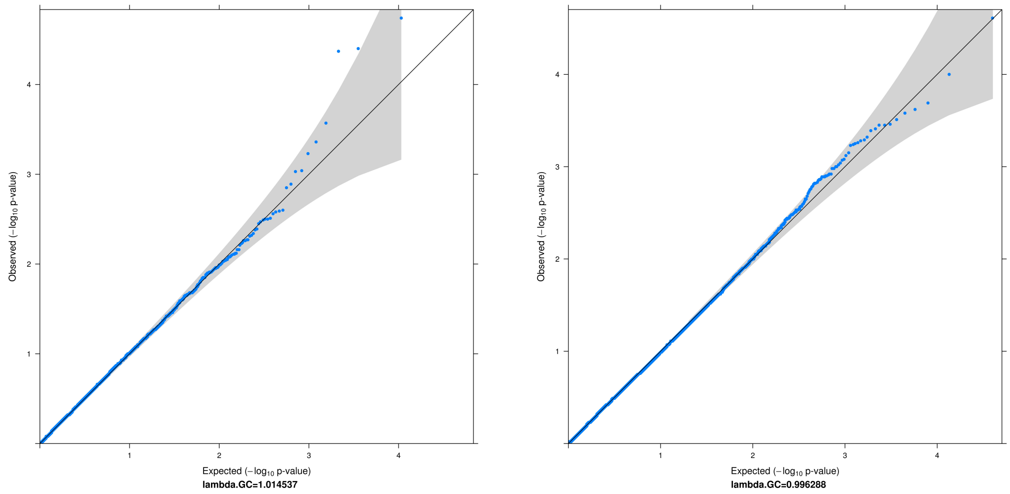


Figure 5: Manhattan Plot for TWAS p-values of amyloid traits by DPR

CHROM	GeneName	GeneID	P-value	FDR Adjusted P-value
4	RP11-769N22.1	ENSG00000249228.1	1.85e-05	0.077
7	SBDS	ENSG00000126524.5	4.28e-05	0.077
7	AC004951.5	ENSG00000239556.2	3.80e-05	0.077

Table 4: Significant Genes for TWAS of amyloid traits by Elastic-Net



(a) Q-Q Plot by Elastic-Net

(b) Q-Q Plot by DPR

Figure 6: Q-Q Plot for TWAS p-values of amyloid traits by Elastic-Net and DPR

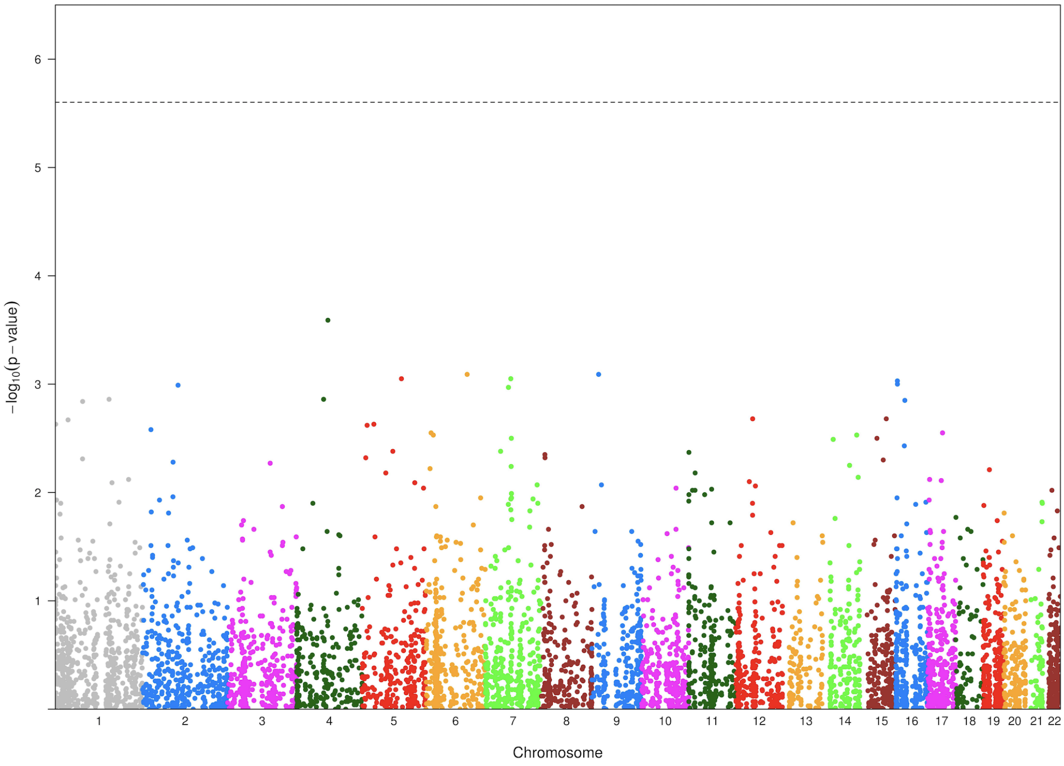


Figure 7: Manhattan Plot for TWAS p-values of tangles traits by Elastic-Net

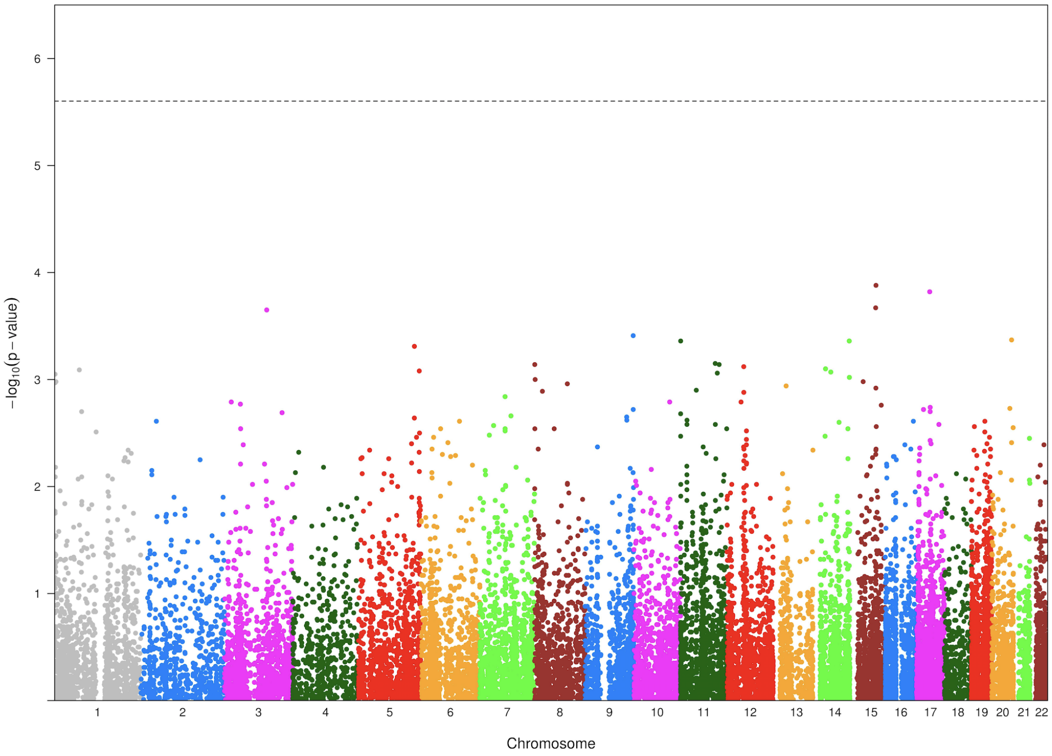
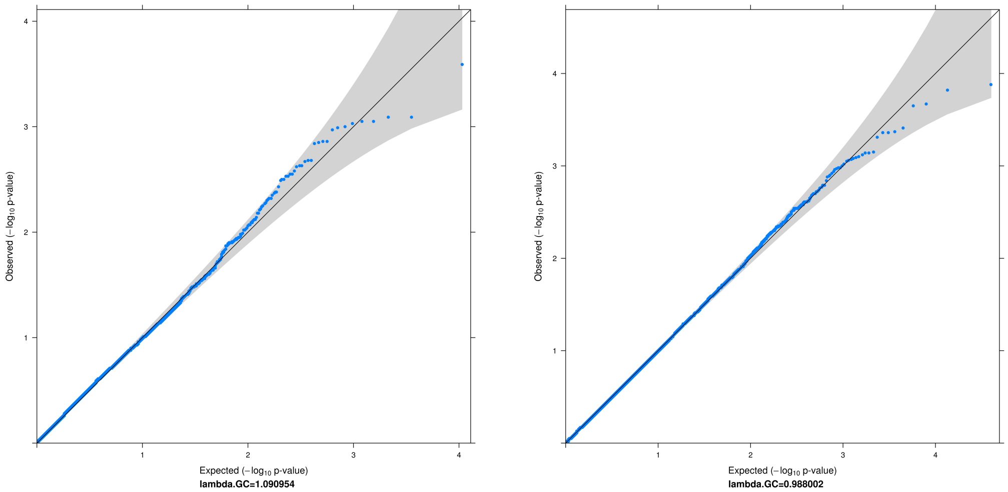


Figure 8: Manhattan Plot for TWAS p-values of tangles traits by DPR



(a) Q-Q Plot by Elastic-Net

(b) Q-Q Plot by DPR

Figure 9: Q-Q Plot for TWAS p-values of tangles traits by Elastic-Net and DPR

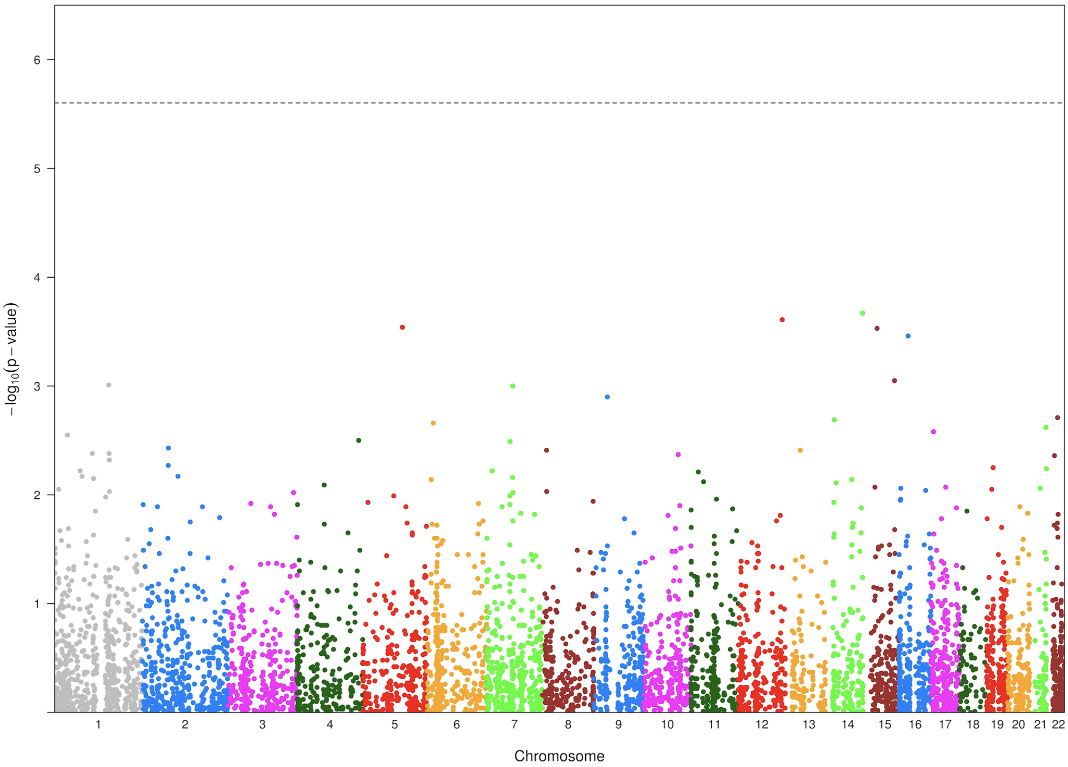


Figure 10: Manhattan Plot for TWAS p-values of gpath traits by Elastic-Net

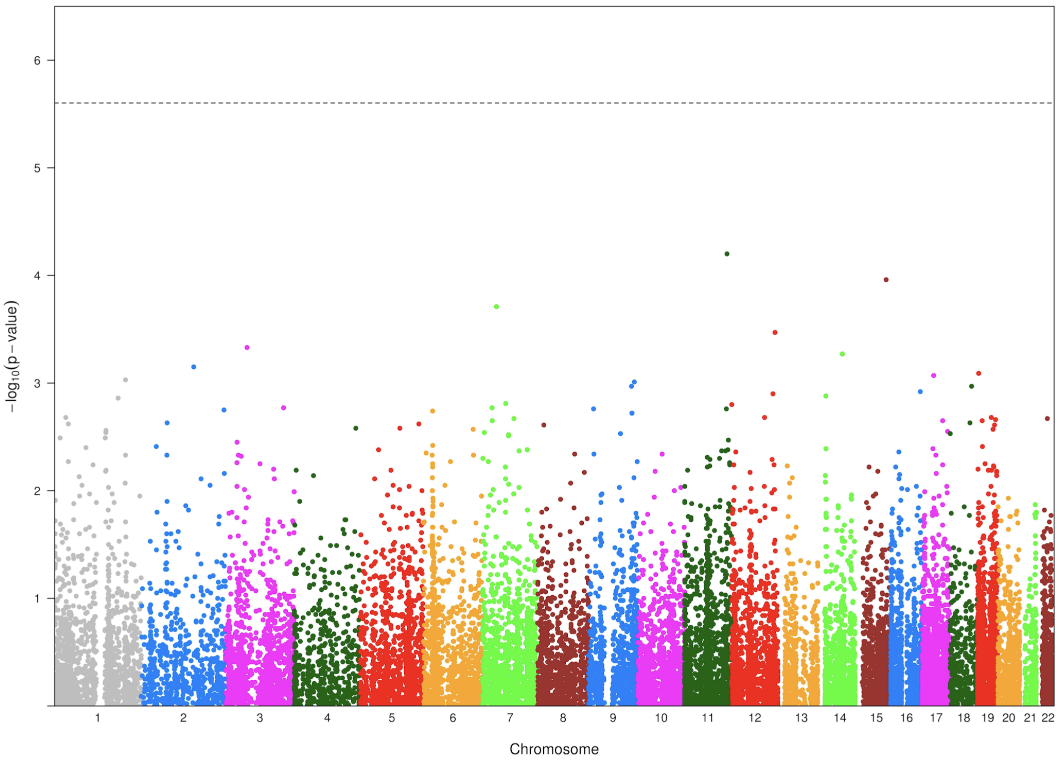
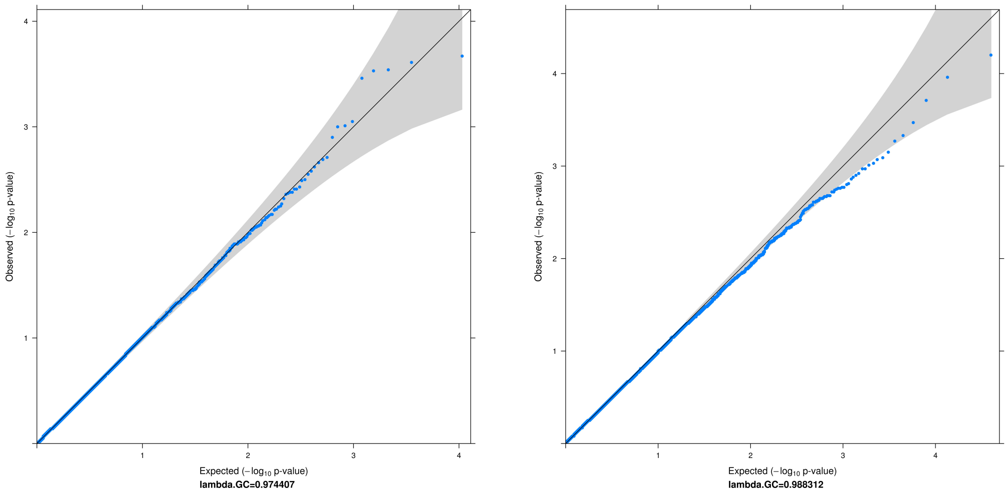


Figure 11: Manhattan Plot for TWAS p-values of gpath traits by DPR



(a) Q-Q Plot by Elastic-Net

(b) Q-Q Plot by DPR

Figure 12: Q-Q Plot for TWAS p-values of gpath traits by Elastic-Net and DPR

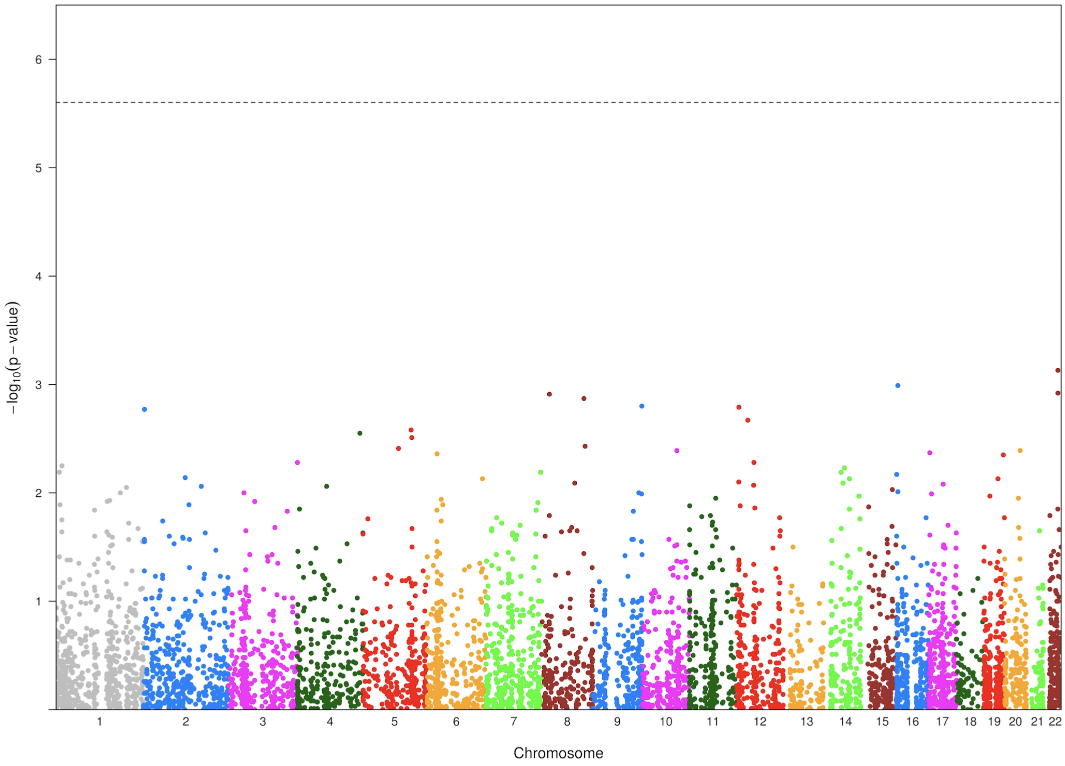


Figure 13: Manhattan Plot for TWAS p-values of cogdx traits by Elastic-Net

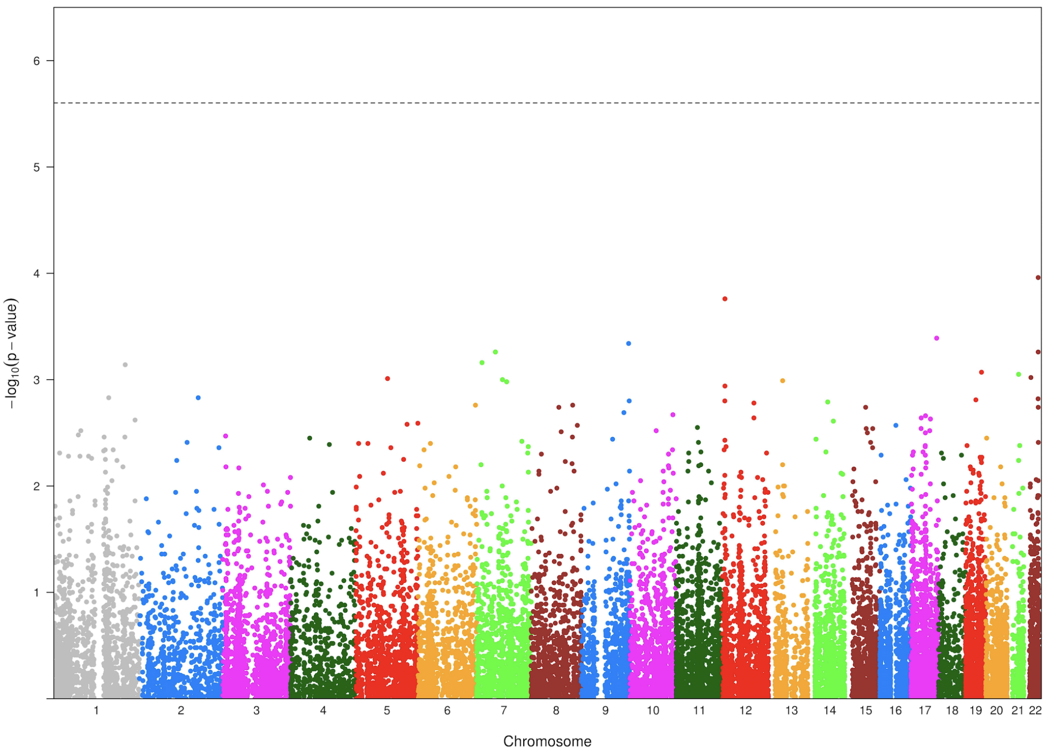
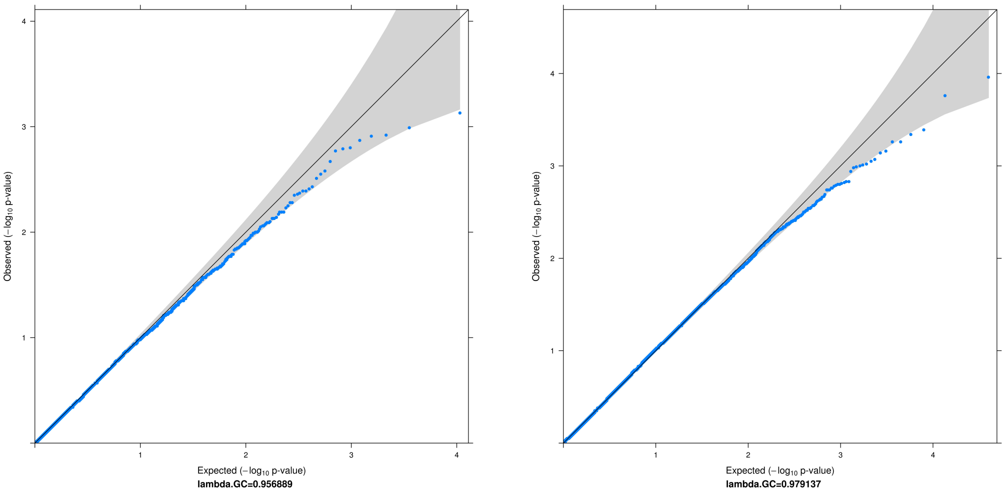


Figure 14: Manhattan Plot for TWAS p-values of cogdx traits by DPR



(a) Q-Q Plot by Elastic-Net

(b) Q-Q Plot by DPR

Figure 15: Q-Q Plot for TWAS p-values of cogdx traits by Elastic-Net and DPR

5 Discussion

In this thesis, I apply TIGAR on GTEx Brain Frontal-Cortex data, following by gene-based association study with ROS/MAP data for amyloid, tangles, gpath, and cogdx traits. TIGAR contains the Elastic-Net and DPR model for transcriptomic data imputation, with options of conduct gene-based association studies using individual-level and summary-level GWAS data for univariate and multivariate phenotype with corresponding imputation model. Advantages for using TIGAR includes taking standard input files like vcf/dosage format for genotype data, calculating MAF and HWE by default, users defined parameter within each model, computation and memory efficacy. Generally, TIGAR can finish the above procedure for one gene in about 4 minutes. Specifically, for the same input files, DPR runs faster comparing to Elastic-Net. Comparing to similar existing tools (PrediXcan/DPR) that accept specific input files for data imputation and subsequent association studies with various output files, which require cumbersome data preparation, large memory space to loading genotype data and works to organize output files, TIGAR not only takes care of tedious works to prepare input files and organize output files for users, also provides options of imputation models and computation efficacy.

TIGAR has still had some limitations: (i) TIGAR is only suitable for cis-eQTL effect-size calculation (within 1Mb on ether gene's TSS). It might reach memory limit by including a wider range of SNPs in one gene; (ii) cross-validation steps for identified significant model and parameter selection in Elastic-Net model will increase the computation burden; (iii) TIGAR called the original DPR tool for training imputation instead of re-writing it in python. Some errors from the original DPR tool might disturb model training; (iv) Only Elastic-Net and

DPR are available for TIGAR in training imputation; (v) Although TIGAR uses 5-fold cross-validation to identified significant imputation model for each gene, overall training (with whole samples) R^2 for a gene can still be 0, i.e. imputation model is still not significant.

To draw a conclusion, TIGAR is expected to provide a computational convenience and powerful tool in transcriptomic data imputation and conduct TWAS. Our application shows that DPR has advantages when the underlying gene expression heritability is relatively lower, e.g., < 0.2, whereas the Elastic-Net is preferred when the gene expression heritability is ¿0.2. This shows that an "optimal" model might be chosen with respect to each gene by comparing the cross validation R^2 from both models.

References

[1] Asude Alpman Durmaz, Emin Karaca, Urszula Demkow, and et al. Evolution of genetic techniques: Past, present, and beyond. *BioMed Research International*, (90):1–7, 2015.

[2] Schuster SC. Next-generation sequencing transforms today's biology. *Nat Methods*, (5):8–16, 2008.

[3] Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, and Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet*, (12):499–510, 2011.

[4] Visscher, P.M., Brown, M.A., McCarthy, M.I., Yang, and J. Five years of gwas discovery. *American journal of human genetics*, (90):7–24, 2012.

[5] McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Hirschhorn, and J.N. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews Genetics*, (9):356–369, 2008.

[6] Yang J and et al. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nat Genet*, (44):369–S3, 2012.

[7] Wood AR and et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*, (46):1173–86, 2014.

[8] Croteau-Chonka DC1, Rogers AJ2, and et al. Expression quantitative trait loci information improves predictive modeling of disease relevance of non-coding genetic variation. *PLoS One*, (10):1–20, 2015.

[9] Farhad Hormozdiari, Martijn van de Bunt, and et al. Colocalization of gwas and eqtl signals detects target genes. *Am J Hum Genet*, (99):1245–1260, 2016.

[10] Alexandra C. Nica, Emmanouil T. Dermitzakis, and et al. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci*, (19):1–6, 2013.

[11] Li L, Kabesch M, Bouzigon E, and et al. Using eqtl weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Front Genet*, (4):1–111, 2013.

[12] Dan L. Nicolae, Eric Gamazon, Wei Zhang, and et al. Trait-associated snps are more likely to be eqtls: Annotation to enhance discovery from gwas. *PLoS Genet*, (6):1–10, 2010.

[13] Xin He, Chris K. Fuller, Yi Song, and et al. Sherlock: Detecting gene-disease associations by matching patterns of expression qtl and gwas. *Am J Hum Genet*, (92):667–680, 2013.

[14] Robert JS. and et al. Integrating coexpression networks with gwas to prioritize causal genes in maize. *The Plant Cell*, (30):2922–2942, 2018.

[15] Rudolf Jaenisch and Adrian Bird. Epigenetics regulation of gene expression: how the genome integrates intrsinsic and environmental signals. *Nature Genetics Supplement*, (45):246–254, 2003.

[16] Harm-Jan Westra, Marjolein J. Peters, and et al. Systematic identification of trans-eqtls as putative drivers of known disease associations. *HHS Public Access*, (45):1238–1243, 2013.

[17] Sunil Kumar, Giovanna Ambrosini, and Philipp Bucher. Snp2tfbs – a database of regulatory snps affecting predicted transcription factor binding site affinity. *Nucleic Acids Research*, (45):139–144, 2017.

[18] Chao Cheng, Roger Alexander, and Renqiang Min. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Cenome Research*, (22):1658–1667, 2012.

[19] Chao Cheng, Roger Alexander, and Renqiang Min. Variation in transcription factor binding among humans. *HHS Public Access*, (328):232–235, 2010.

[20] Alexander Gusev, Arthur Ko, Huwenbo Shi, and et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, (48):245–252, 2016.

[21] Torres JM and et al. Cross-tissue and tissue-specific eqtls: partitioning the heritability of a complex trait. *Am J Hum Genet*, (95):521–34, 2014.

[22] Lappalainen T and et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, (501):506–11, 2013.

[23] Gamazon E.R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, (9):1091–1098, 2015.

[24] Binglan Li, Shefali S. Verma, Yogasudha C. Veturi, and et al. Evaluation of predixcan for prioritizing gwas associations and predicting gene expression. *Pac Symp Biocomput*, (23):448–459, 2018.

[25] Gamazon E.R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, (9):1091–1098, 2015.

[26] Wu C and Pan W. Integrating eqtl data with gwas summary statistics in pathway-based analysis with application to schizophrenia. *Genet Epidemiol*, (42):303–316, 2018.

[27] Jingjing Yang and et al. Tigar: An improved bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. *bioRxiv*, 2018.

[28] GTEx Consortium. The genotype-tissue expression (gtex) project. *Nat. Genet*, (45):580–585, 2013.

[29] Bennett, D.A, Schneider, J.A., Arvanitakis, Z., Wilson, and R.S. Overview and findings from the religious orders study. *Curr Alzheimer Res*, (9):682–645, 2012.

[30] Bennett, D.A., Schneider, J.A., Buchman, A.S., Barnes, L.L., Boyle, P.A., Wilson, and R.S. Overview and findings from the rush memory and aging project. *Curr Alzheimer*, (9):646–663, 2012.

[31] Zou, H., Hastie, and T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, (67):267–288, 2005.

[32] Peng Zeng. et al. Non-parametric genetic prediction of complex traits with latent dirichlet process regression models. *Nature Communications*, (8):1–11, 2017.

[33] Li and B. et al. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American journal of human genetics*, (83):311–321, 2008.

[34] Bennett DA, Wilson RS, and et al. Selected findings from the religious orders study and rush memory and aging project. *J Alzheimers Dis*, (33):397–403, 2013.

[35] David A. Bennett, Aron S. Buchman, Patricia A. Boyle, and et al. Religious orders study and rush memory and aging project. *J Alzheimers Dis*, (64):S161–S189, 2018.

[36] De Jager PL, Shulman JM, Chibnik LB, and et al. A genome-wide scan for common variants affecting the rate of age-related cognitive decline. *Neurobiol Aging*, (33):1–15, 2012.

[37] De Jager PL, Srivastava G, Lunnon K, and et al. Alzheimer's disease:early alterations in brain dna methylation at ank1, bin1, rhbdf2 and other loci. *Nat Neurosci*, (9):1156–63, 2014.

[38] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, (550):1–39, 2017.

[39] GTEx Consortium. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Human Genomics*, (348):648–660, 2015.

[40] Michael N Weedon Shaun Purcell Jian Yang and et al. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, (9):807–812, 2011.

Website

[41] `https://gtexportal.org/home/tissueSummaryPage`