Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Tingyang Hu Date

Integrate Proteomics Data with GWAS Summary data for Studying Alzheimer's Disease by Nonparametric Bayesian Method

By

Tingyang Hu
Master of Public Health

Department of Biostatistics and Bioinformatics

Jingjing Yang, Ph.D
(Thesis Advisor)

Michael P. Epstein, Ph.D
(Reader)

Integrate Proteomics Data with GWAS Summary data for Studying Alzheimer's Disease by Nonparametric Bayesian Method

By

Tingyang Hu
B.S.
Xiamen University, China
2020

Thesis Advisor: Jingjing Yang, Ph.D.

Reader: Michael P. Epstein, Ph.D.

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 Public Health
in Department of Biostatistics and Bioinformatics
2022

Abstract

Integrate Proteomics Data with GWAS Summary data for Studying Alzheimer's Disease by Nonparametric Bayesian Method
By Tingyang Hu

Background: Alzheimer's disease (AD) is a nerurogenerative disorder related to aging with polygenic inheritance. Genome-wide association studies (GWAS) of AD have identified many risk loci, but currently little is known about the underlying biological mechanism. Proteome-wide association study (PWAS) integrating proteomics data with GWAS summary data to identify risk genes associated with Alzheimer's disease, would provide novel insights to the impacts of genetic variation on AD potentially mediated through brain protein abundance.

Method: We conducted the weighted protein network analysis on the human proteomes from European ancestry of ROS/MAP (12691 proteins donated by 400 samples), to identify clusters of proteins with unsupervised hierarchical clustering and relate the protein modules to external clinical traits of AD. The PWAS was implemented with Transcriptome-Integrated Genetic Association Resource V2 (TIGAR-V2) tool in two stages. Firstly we applied either nonparametric Bayesian Direchlet Process Regression (DPR) or Elastic-Net penalized regression (as used by PrediXcan) to train protein abundance imputation models, taking proteomics abundance as the outcome and *cis*-SNPs as predictors. The protein quantitative trait locus (pQTL) effect sizes estimated from the protein abundance prediction models were integrated with AD GWAS summary level data to implement asscoation test using burden test statistics.

Results: Weighted protein network analysis of 8874 proteins after quality control identified 32 network modules, ranged in size from 33 to 2386 proteins. We observed 2 protein modules significantly associated with AD clinical traits. At training stage of PWAS, we obtained 6673 protein abundance prediction models trained by Bayesian DPR, which were all valid with 5-fold CV R^2 >0.005. Of 6389 protein abundance prediction models trained by Elastic net regression, only 1835 had 5-fold CV R^2 >0.005. Based on GWAS summary statistics of AD and Bayesian estimated pQTL weights, the PWAS has detected 13 genes were associated with at an FDR of P <0.05, with 3 genes previously known as GWAS risk gene of AD. Furthermore, We compared the PWAS results of AD using pQTL weights estimated by DPR with weights estimated by Elastic-Net method (PrediXcan function integrated in our TIGAR tool). PrediXican detected 7 significant genes at an FDR of P <0.05, of which 2 was also identified by TIGAR.

Conclusion: In this work, we detected PWAS risk genes for AD and demonstrated the usefulness of nonparametric Bayesian DPR method in PWAS for AD. We believe this approach can be applied widely to study other complex polygenic diseases and provide new insights into their pathogenesis.

Integrate Proteomics Data with GWAS Summary data for Studying Alzheimer's Disease by Nonparametric Bayesian Method

By

Tingyang Hu
B.S.
Xiamen University, China
2020

Thesis Advisor: Jingjing Yang, Ph.D.

Reader: Michael P. Epstein, Ph.D.

A thesis submitted to the Faculty of the
Rolllins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Department of Biostatistics and Bioinformatics
2022

Acknowledgments

I would like to thank all participants in ROS and MAP studies who contributed their data and agreed to autopsy at the time of their death for profiling the proteomics data. I am thankful to the staff and researchers at the Rush Alzheimer's Disease Center for making the proteomics data available to my thesis project.

I would like to thank my advisor Jingjing Yang, Ph.D. With her direction and guidance, I learned more than I ever expected about the application of statistical genetics in the real world of public health. It has always been my pleasure to work with her and the rest of the Yang lab members, and I was greatly supported by her proficiency and patience as well as the collaborative lab team during the thesis project. I would also like thank my thesis reader Michael P. Epstein, Ph.D, and the rest of people at the Center for Computational and Quantitative Genetics.

I would like to thank all the faculty and staff of Rollins School of Public Health Department of Biostatistics and Bioinformatics for their excellent lectures, continuous assistance and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Alzheimer's disease (AD) is a neurogenerative disorder and the sixth leading cause of death in the United States [1]. Genome-wide association studies (GWAS) of AD have identified 38 risk loci [2], but currently little is known about the underlying biological mechanism.

To implicate potential causal risk genes of AD with biological mechanistic interpretations, transcriptome-wide association study (TWAS) can impute the gene expression levels within AD GWAS using reference datasets to train gene expression models, and test how AD associate with gene expression levels predicted from genetic variants [3, 4]. Since gene expression is not considered as a perfect proxy for protein functionality, an enhancement to existing association study would be a protein-based method that considers the effects of genetics variants on the protein function alterations [5].

The proteome-wide association study (PWAS) integrates proteomics data with GWAS summary data to identify risk genes associated with Alzheimer's disease, can provide novel insights to the impacts of genetic variation on AD that are potentially mediated through brain protein abundance instead of gene expression at the transcript level [5]. The PWAS conducted by Wingo et al with FUSION method identified 13

risk genes whose *cis*-regulated brain protein levels were associated with AD [6].

Given that both gene expression and protein abundance are quantitative traits that might mediate genetic effects, the statistical tools for TWAS can be naturally applied to PWAS, by using trained protein quantitative trait locus (pQTL) effect sizes from a reference panel as SNP weights for gene-based association studies. There are various TWAS methods using different models to estimate the SNP weights from reference data. For example, PrediXcan can estimate pQTL effect sizes by a general linear regression model with Elastic-Net penalty [7]. FUSION can estimate pQTL weights by Elastic-Net, LASSO, sum of single effects and Bayesian sparse linear mixed model (BSLMM), and then select the best model for PWAS [8]. They both use parametric imputation models with limitations for modelling the complex genetic architecture and gene expression profiles. To make the model more flexible and general, a nonparametric Bayesian latent Dirichlet process regression (DPR) can be employed, in which the prior for effect sizes is nonparametric and can be estimated by assuming a Dirichlet process prior on variance of effect size [9].

The Transcriptome-Integrated Genetic Association Resource V2 (TIGAR-V2) [10] implemented this DPR method while including Elastic-Net (as used by PrediXcan [7]) and BSLMM (as used by FUSION [8]) as special cases, for gene expression imputation. Additionally, the TIGAR-V2 tool was shown to improve computation efficiency which directly reads Variant Call Format (VCF) files of genotyping and enables parallel computation, taking advantage of cloud computing clusters [10].

Here we applied the TIGAR-V2 tool for PWAS, taking protein abundance as the outcome of prediction models and *cis*-SNPs (SNPs within 1MB of the corresponding gene region) as predictors[10]. The standard PWAS comprised two stages with the TIGAR-V2 tool. Firstly, it can apply nonparametric Bayesian DPR or Elastic-Net penalized regression (as used by PrediXcan [7] to train protein abundance imputation models. The effect sizes of *cis*-SNPs estimated from the regression models were

treated as protein quantitative trait locus (pQTL) weights, which were then integrated with AD GWAS summary level data, i.e. Z-score statistics from GWAS tests of single-variants, to conduct gene-based association test using burden test statistics [11].

In this work, we used pQTL weights obtained from proteomics data of human brain from two prospective cohort studies — Religious Orders Study and Memory and Aging Project (ROS/MAP) [12] along with publicly available GWAS summary data of single variants to conduct PWAS for understanding AD, and identified 13 risk genes that confer AD risk through their effects on protein abundance.

To understand the protein abundance correlations and interactions which is biologically important, we also conducted weighted network analysis on the human brain proteomes, and found 32 correlated protein modules, of which 2 modules were significantly related to AD clinical traits.

In the following sections, we first describe the method of brain protein network analysis. We then outline the application of TIGAR-V2 to conduct the PWAS for AD. The observed protein networks and PWAS results will be explained and discussed.

Chapter 2

Method and Materials

2.1 Data Source

The human brain proteomes were generated from the dorsolateral prefrontal cortex of postmortem brain samples donated by 400 participants of European ancestry of the Religious Orders Study/Memory and Aging Project (ROS/MAP) [12]. Protein measurements were quantified with tandem mass spectrometry and generated as protein abundance. The protein abundance ratio was used for follow up analyses scaling each protein abundance with a sample-specific total protein abundance [13].

The proteomics profiles underwent quality control, which contained the protein abundance level for 12691 proteins corresponding to 400 samples. Proteins were included with missing values in more than 50% of the participants, whose abundance ratio related to baseline was calculated and log_2 transformed, with missing values imputed with the mean of protein level for 400 samples. The effects of clinical characteristics (i.e., sex, age at death, postmortem interval, and study type) and technical factors (i.e., sequencing batch and mass spectrometry reporter quantification mode) were regressed out. Poorly performing samples were then removed using iterative principal component analysis (PCA), i.e samples greater

than four standard deviations from the mean of either the first or second principal component.

The genotype data of ROS/MAP samples were profiled by whole genome sequencing (WGS) of frontal cortex on 1179 participants. Samples with missingness >5% were excluded. Variants were removed if they had significant deviation from Hardy-Weinberg equilibrium, missing values in genotype >5%, minor allele frequency <1% or not an SNP [6]. After quality control, 355 samples remained with both proteome and genome sequencing data for PWAS.

The AD summary level association statistics for single variants was obtained from the latest GWAS by Douglas et al [2], for 1,126,563 individuals from 13 European cohorts (mostly from 23andMe and UK Biobank). We will use the GWAS summary data of AD from meta-analysis of all cohorts except 23andme in our PWAS study.

2.2 Weighted Protein Network Analysis

Human brain protein network is biologically meaningful to investigate protein interactions and underlying mechanism. We can apply weighted gene co-expression network analysis (WGCNA) on the proteomics data due to its shared data characteristics with gene expression, which can be considered as ultimate products of gene regulation. Here we performed protein network analysis with R package "WGCNA" [14], which can be applied for constructing the gene co-expressed network using the correlation of gene expression levels as a measure of co-expression, identifying co-expressed modules based on the hierarchical clustering and relating co-expressed modules to external clinical traits.

The construction of weighted protein network comprised three steps. Firstly, the weighted correlation coefficients with a power function of protein abundance values were calculated, to obtain a scale free network and a weighted adjacency matrix. Then

the topological overlap (TO) based on the connection strengths of weighted adjacency matrix was generated, which indicated the similarity of protein abundance. Finally, the cluster dendrogram was obtained by using hierarchical clustering with 1-TO as the distance measure, and the correlated protein modules were identified using a dynamic tree-cutting algorithm [14].

To find clinically significant protein modules, the module-trait relationships between distinctive modules and clinical features of AD were characterized by correlating the module eigengens with random slope of cognition decline as well as AD cognition status [14].

Additionally, we investigated the modules with significant associations with AD clinical traits and find the key drivers that contributes most in these modules using PCA. The protein-protein interaction (PPI) enrichment in the protein modules of interest was analyzed with STRING tool, which aimed to collect public data sources of protein interaction information and characterize the connectivity network of proteins [15, 16]. The STRING was also used to construct the protein network for key drivers of significant modules, with Python application programming interface (API).

2.3 PWAS Framework

The TIGAR-V2 framework included training protein abundance imputation models from ROS/MAP proteomics data and WGS genotype data, generating reference LD covariance from reference panels and testing gene-based association with summary-level GWAS data (Figure 2.1).

The two-stage PWAS with TIGAR-V2 first trained the protein abundance imputation models with *cis*-genotype data as predictors ($\mathbf{X}$), as the following linear regression model for protein abundance level

$$\mathbf{E}_p = \mathbf{X}\mathbf{w} + \epsilon; \quad \epsilon \sim N(0, \sigma_\epsilon^2 \mathbf{I}) \tag{2.1}$$

where $E_{\boldsymbol{p}}$ denotes the protein abundance levels with respect to a target protein p after regressing out for individual characteristics and technical factors, $\mathbf{w}$ denotes the corresponding pQTL effect size vector and ϵ denotes the error term. With the TIGAR-V2 tool, the pQTL weights ($\mathbf{w}$) were estimated by either nonparametric Bayesian DPR or Elastic-Net penalty regression (as used by PrediXcan [7]).

The 5-fold cross-validation was also conducted for the prediction models by TIGAR with respect to each protein and generate an average training R^2 across 5 folds of validation data. Proteins with abundance prediction model 5-fold CV R^2 >0.005 would be considered as valid to be tested for follow-up PWAS.

At the stage of association test, we combined the summary-level genetic effect of AD (GWAS Z-score) with the protein weights to conduct the PWAS of AD. TIGAR would conduct association test using burden [8] statistics (S-PrediXcan).

$$Z_{p,S-PrediXcan} = \frac{\sum_{l=1}^{m} \hat{w}_l \hat{\sigma}_l Z_l}{\sqrt{\hat{\mathbf{w}}' \mathbf{V} \hat{\mathbf{w}}}} \tag{2.2}$$

Where $\hat{\mathbf{w}}$ denotes the pQTL effect size estimates from protein imputation models, Z_l denotes the Z-score statistic of single variant l by GWAS test and $\mathbf{V}$ denotes the linkage disequilibrium (LD) covariance matrix of test SNPs estimated from reference panels of the same ethnicity. The genotype variance $\hat{\sigma_l^2} = Var(x_l)$ can be estimated from a reference panel, given by $2f_l$(1-f_l) with minor allele frequency f_l.

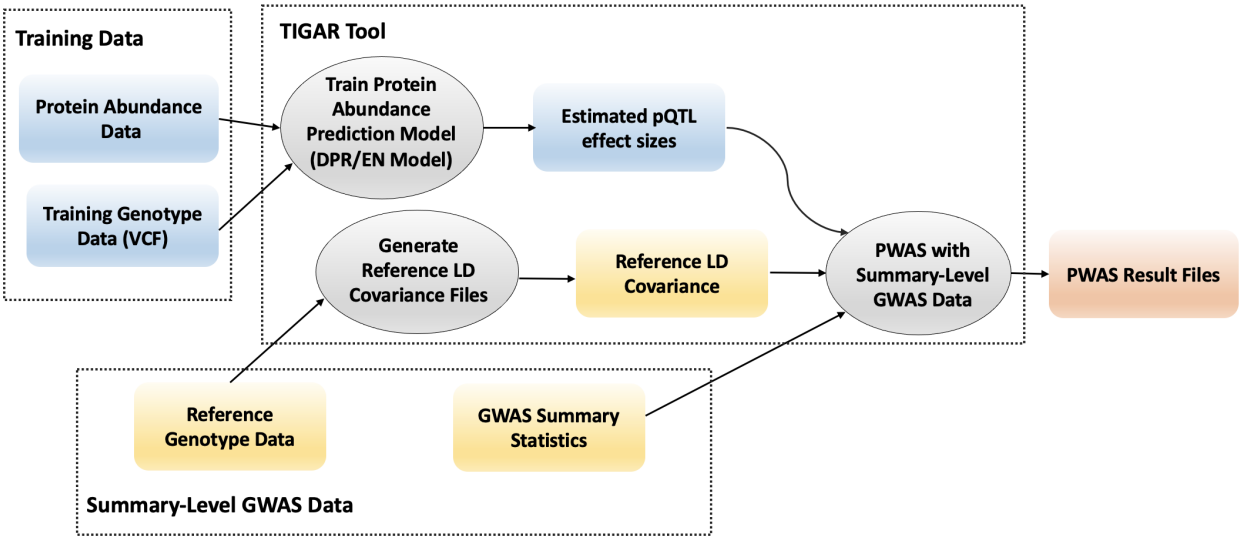

Figure 2.1: PWAS Framework

Chapter 3

Protein Network Analysis

3.1 Proteomics Data from ROS/MAP Studies

There were 8874 proteins remained after excluding proteins with missing values in more than 50% of the participants As a result of 2 rounds of PCA, 395 out of 400 samples remained. We plotted the first PC against the second PC for each round of PCA, with colored lines indicating the four standard deviations from the mean of PCs, thus dots outside the lines were identified as outliers (Figure 3.1). The principal components dispersed more randomly at the third round of PCA.

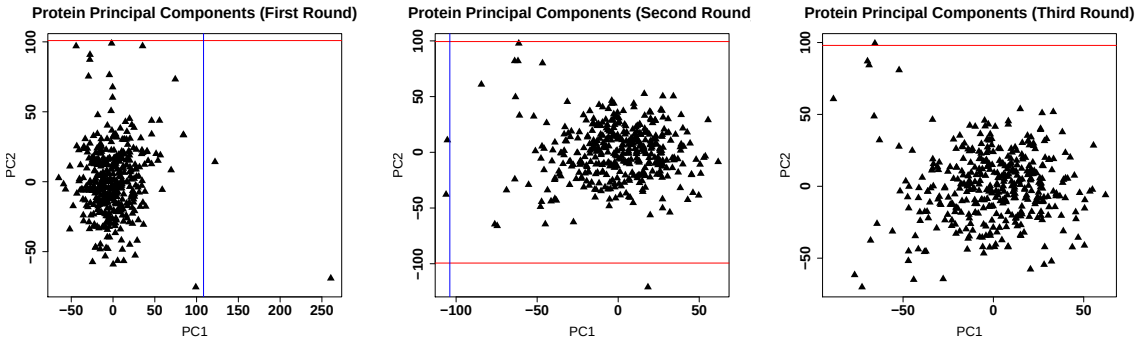

Figure 3.1: Principal Components of Protein Abundance Ratios, red lines coded with four standard deviations from the mean of PC1 and blue line coded with four standard deviations from the mean of PC2

3.2 Protein Modules/Networks

Weighted network analysis identified correlated protein network modules. There were 32 distinct network modules coded by different colors generated by unsupervised hierarchical clustering, ranged in size from 33 and 2386 proteins (Figure 3.2).

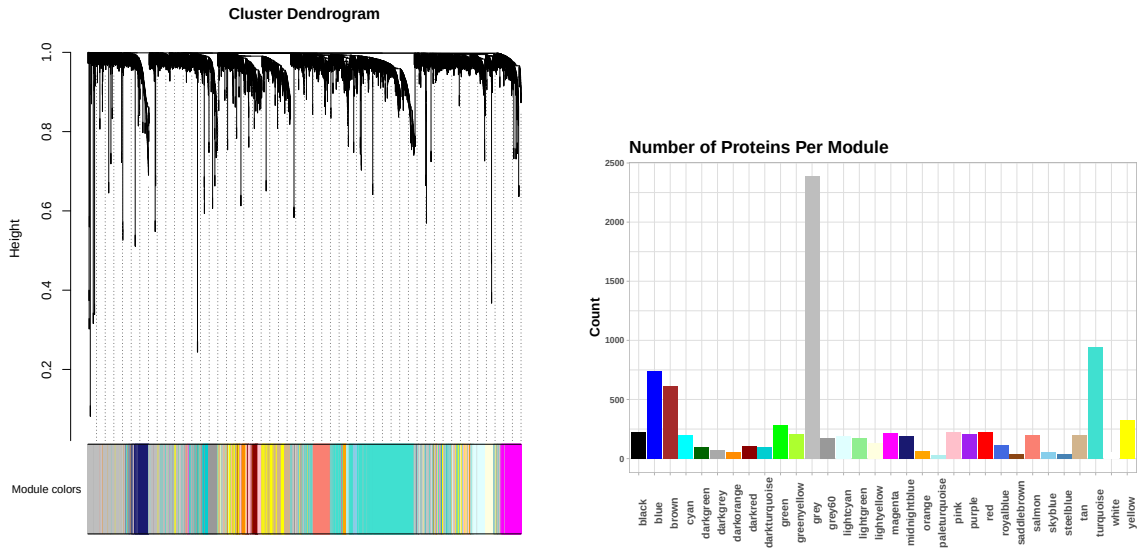

Figure 3.2: Protein Network Modules

3.2.1 Associations between Protein Modules and Clinical AD Traits

The association between protein modules and clinical traits were characterized by the biweight midcorrelation. Two protein network modules were significantly associated with clinical traits (the random of cognition decline and consensus diagnosis of AD status), excluding the effects of personal characteristics including sex, age, and education. (Figure 3.3). The protein module of black containing 223 proteins was significantly associated with cognition random slope with correlation coefficient of 0.12 and p-value = 0.02. Darkturquoise module with 94 proteins was

significantly associated with both cognition random slope (correlation coefficient = -0.1, p-value = 0.04) and consensus diagnosis of AD status (correlation coefficient = 0.1, p-value = 0.04). Of the two modules associated with clinical traits, proteins that contributed most were detected with largest 1st principal component. As a result, we found gene HSPA1L and HIST1H2AG (both in chromosome 6) which were identified related to AD by previous protein network and pathway study for AD [17].

For the two WGCNA modules of protein significantly associated with AD clinical traits, their protein-protein associations were characterized using STRING tool [15, 16]. The black protein module had a potentially significant protein-protein interaction (PPI) enrichment p-value of 1.4×10^{-4}, while the PPI p-value for the darkturqiose protein module was 0.03. We observed 14 pairwise protein interactions which were experimentally confirmed (with medium or higher confidence experimental score) among 94 proteins in the darkturqiose module and 84 pairwise associations among 223 proteins in black module.

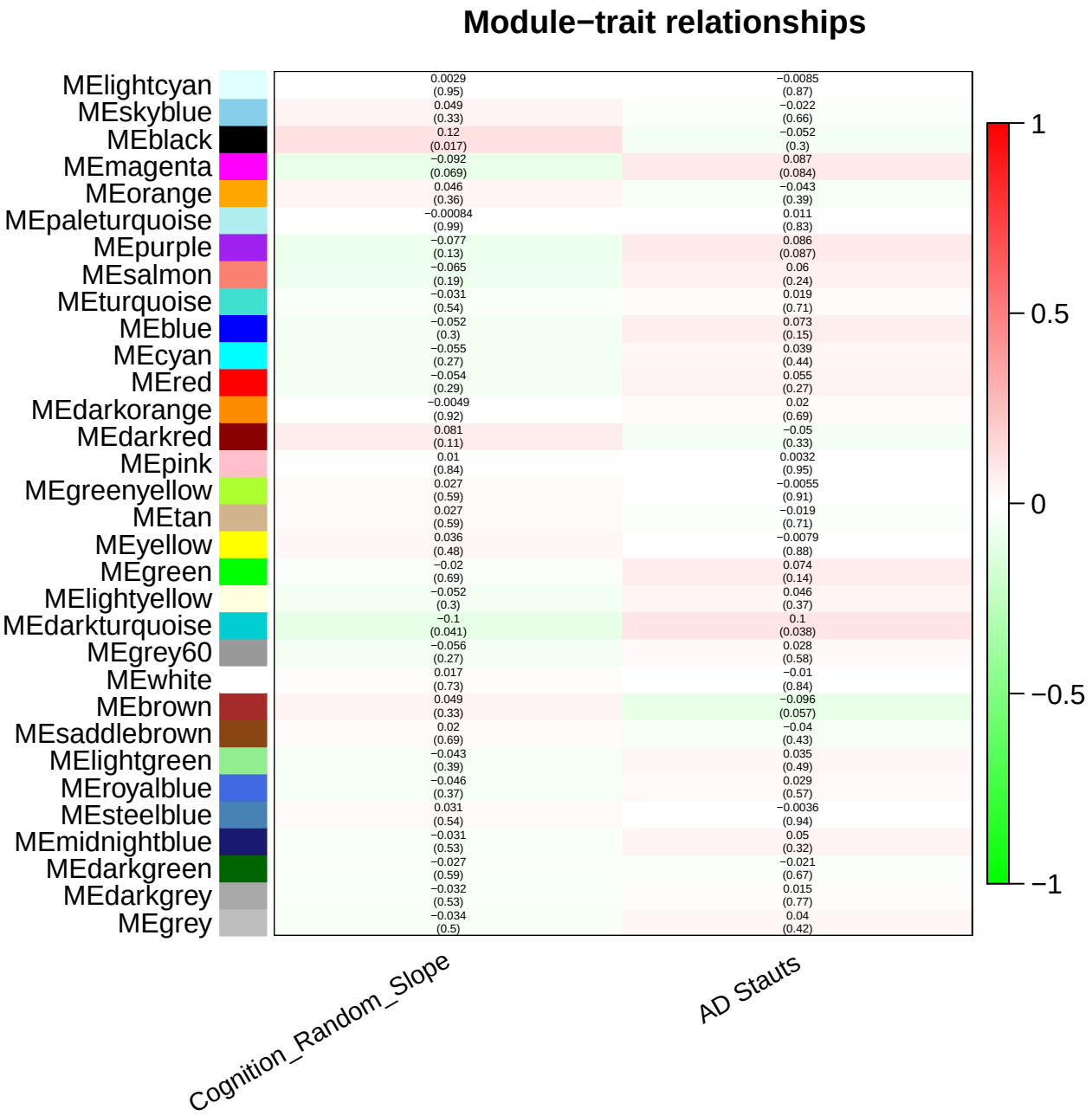

Figure 3.3: Relationships between Protein Modules and Traits

3.2.2 Protein Networks by STRING

The network was constructed for top significant proteins respectively with its 15 most confident interaction partners of both functional and physical protein associations, with line thickness indicating the strength of data support (Figure 3.4). The network of HSPA1L had the average local clustering coefficient of 0.84 and the PPI enrichment p-value $<1.0 \times 10^{-16}$, while the network HIST1H2AG had the average local clustering coefficient of 0.99 and the PPI enrichment p-value $<1.0 \times 10^{-16}$.

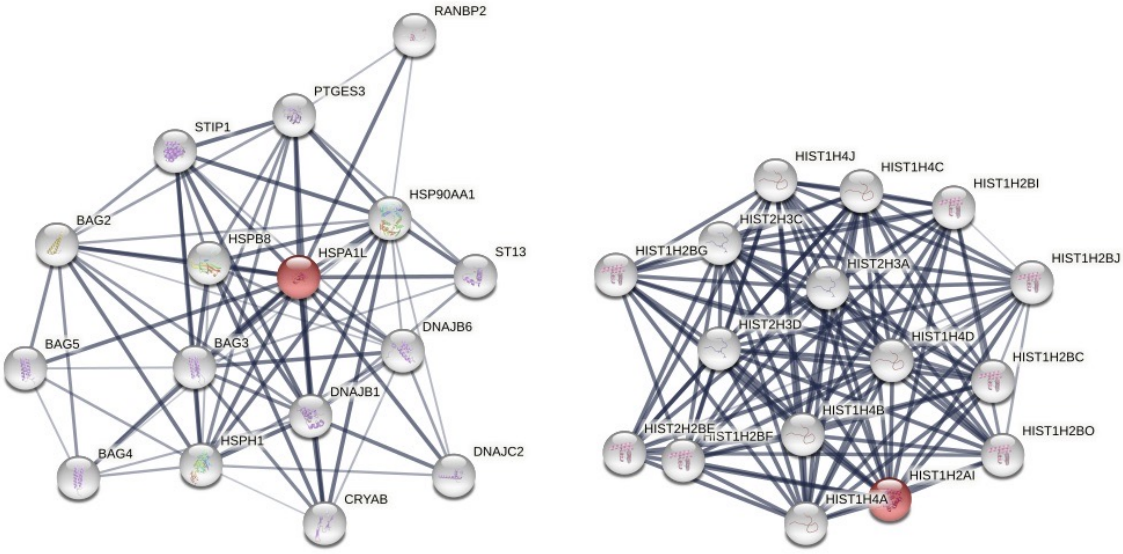

Figure 3.4: Interaction Network of Gene HSPA1L and HIST1H2AG

Chapter 4

PWAS of AD

4.1 Train Protein Abundance Prediction Models

At training stage, the protein abundance prediction model was trained with either nonparametric Bayesian Dirichlet process regression (DPR) or Elastic-net penalized regression. Five-fold cross-validation was conducted to evaluate the training performance via 5-fold CV R^2 per protein. The protein abundance prediction models with 5-fold CV R^2 >0.005 were retained and the estimated pQTL weights from these protein abundance prediction models would be used to conduct association studies. We obtained 6673 protein abundance prediction models trained by DPR, which were all valid with 5-fold CV R^2 >0.005. of 6389 protein abundance prediction models trained by Elastic net regression, 1835 had 5-fold CV R^2 >0.005. Only 1475 protein abundance imputation models were trained by FUSION and used by Wingo et al [6], of which 1158 protein abundance prediction models were also trained by TIGAR/DPR (1109 proteins were also trained by Elastic-Net model implemented in TIGAR), and thus 5515 more protein abundance prediction models were trained by TIGAR/DPR (5280 more proteins were trained by Elastic-Net implemented in TIGAR).

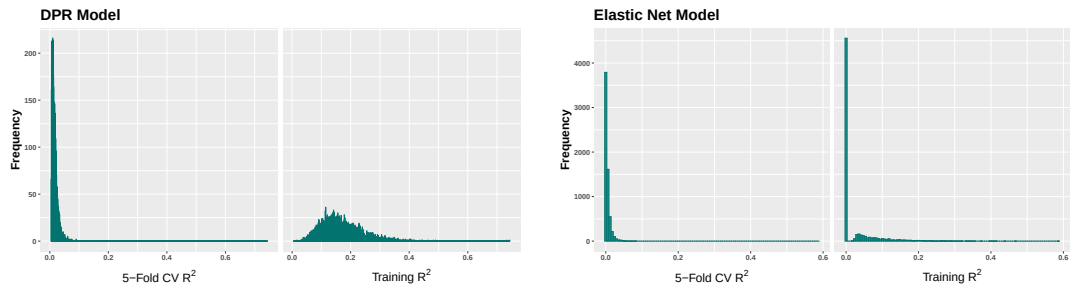

Figure 4.1: Histogram of CV R^2 and Training R^2 by TIGAR.

4.2 PWAS Results by TIGAR/DPR

Based on GWAS summary statistics of AD and Bayesian estimated pQTL weight, the PWAS has identified 13 significant genes at an FDR of P (q-value) <0.05 (Figure 4.2; Table 4.1).

The known GWAS risk genes are curated from GWAS Catalog containing at least one significant SNP within or ±1 Mb around the gene region [18]. Of the 13 risk genes identified by TIGAR, CCDC86 was known GWAS risk gene of AD which involved in the development of neurofibrillary tangles. AGFG2 was GWAS risk gene of family history of AD. FNBP4 was also reported for general cognitive ability. MBLAC1 (identified by both TIGAR and PrediXcan) and MAP3K7 has known biological functions involved in brain measurement such as cortical thickness. Interestingly, the most significant gene MRPL16 along with other 5 genes (C1QTNF4, DIABLO, ATPAF2, YWHAB, FNBP4) were previously found to have relationships with obesity and body shape.

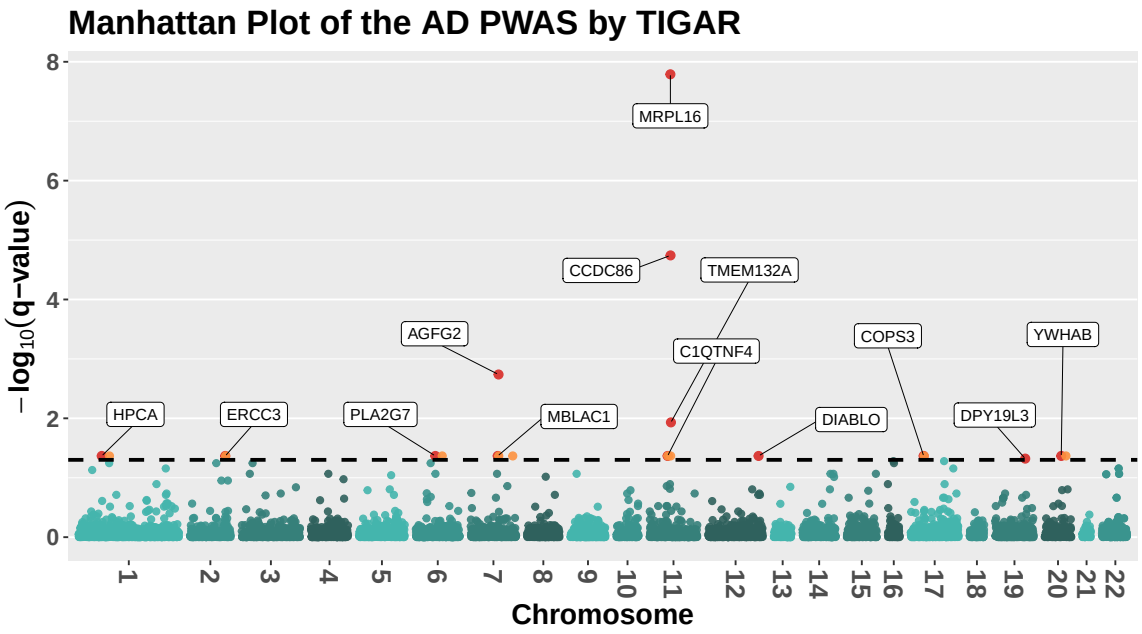

Figure 4.2: Manhattan Plot for the AD PWAS with FDR q-values by TIGAR.

Table 4.1: PWAS risk genes of AD with nonparamtric DPR method

Gene	Chr	Start	End	PWAS z-score	PWAS P	PWAS q-value
HPCA	1	32886497	32893862	3.81	1.41×10^{-4}	4.30×10^{-2}
ERCC3[a]	2	127257290	127294081	-3.81	1.40×10^{-4}	4.30×10^{-2}
PLA2G7	6	46704316	46735693	3.84	1.24×10^{-4}	4.30×10^{-2}
MBLAC1[a]	7	100126697	100128196	-3.85	1.20×10^{-4}	4.30×10^{-2}
AGFG2	7	100539211	100564991	-4.91	9.05×10^{-7}	1.83×10^{-3}
C1QTNF4	11	47589664	47594409	-4.10	4.16×10^{-5}	4.30×10^{-2}
MRPL16	11	59806135	59810658	6.99	2.68×10^{-12}	1.62×10^{-8}
CCDC86	11	60841956	60850325	5.82	5.98×10^{-9}	1.81×10^{-5}
TMEM132A	11	60924441	60936907	4.47	7.72×10^{-6}	1.17×10^{-2}
DIABLO	12	122207662	122227534	-3.84	1.21×10^{-4}	4.30×10^{-2}
COPS3	17	17246624	17281186	-3.79	1.49×10^{-4}	4.30×10^{-2}
DPY19L3	19	32405749	32482240	-3.76	1.73×10^{-4}	4.30×10^{-2}
YWHAB	20	44885599	44906438	-3.80	1.47×10^{-4}	4.30×10^{-2}

[a] PWAS risk gene identified by PrediXcan

The PWAS p values of genes within a ±1 MB region of top significant gene MRPL16 were plotted against their position on chromosome 11, and genes were color-

coded with respect to their correlation R^2 of predicted genetically regulated protein level with MRPL16. CCDC86 and TMEM132A located within the 1MB around MRPL16 were also identified as PWAS risk genes by TIGAR/DPR, considered as independent of predicted protein abundance of MRPL16 (R^2 <0.2). The heatmap indicates the pairwise protein abundance R^2, with bright red denoting R^2 close to 1 and white denoting R^2 close to 0. Genes with nearby test regions had correlated predicted protein abundance values.

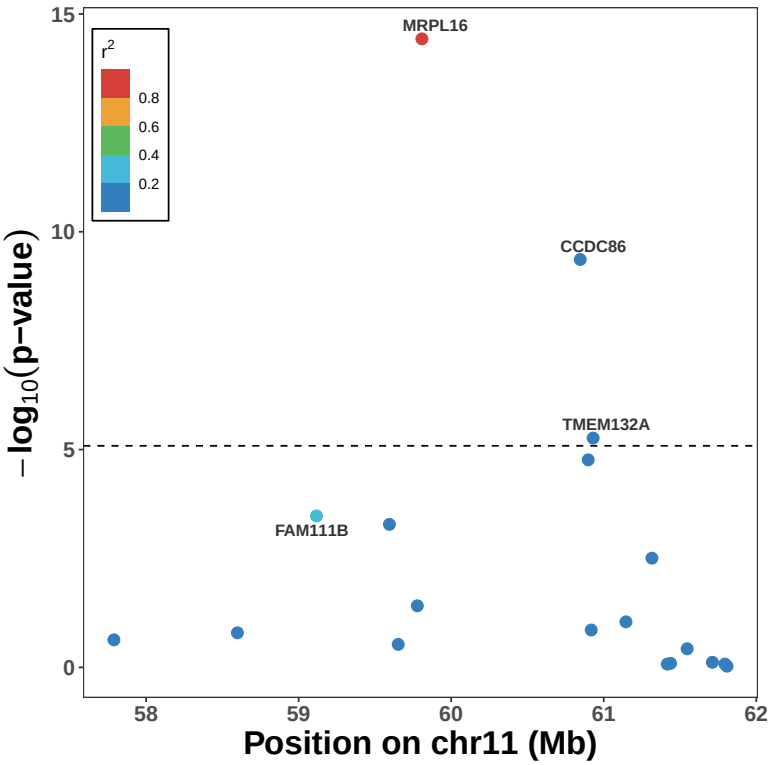

Figure 4.3: LocusZoom Plot for Genes within 1MB around the Most Significant Gene

4.3 Compare with PWAS results by PrediXcan

We also trained Elastic-Net penalized regression models as implemented by PrediXcan [7], which was integrated in TIGAR tool. We compared the PWAS results of AD using pQTL weights estimated by DPR with weights estimated by Elastic-Net method. PrediXican detected 7 significant genes at an FDR of P (q-value) <0.05 (Figure 4.4; Table 4.2). Gene MBLAC1 located on chromosome 7 was also identified by TIGAR. GRP17 located within 1MB around ERCC3 detected by TIGAR could be also seen as a shared risk gene.

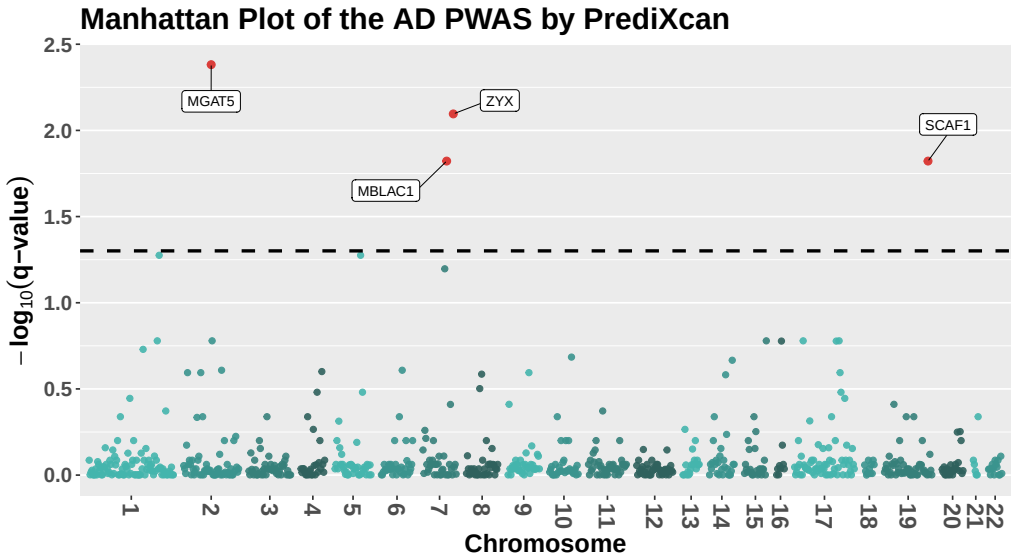

Figure 4.4: Manhattan Plot of the AD PWAS with FDR q-values by PrediXcan.

Table 4.2: PWAS risk genes of AD with Elastic-Net method

Gene	Chr	Start	End	PWAS z-score	PWAS P	PWAS q-value
GPR17[a]	2	127645864	127651755	-10.84	0.00	0.00
MGAT5	2	134254259	134448847	4.24	2.19×10^{-5}	4.15×10^{-3}
MBLAC1[a]	7	100126697	100128196	-4.04	1.26×10^{-4}	1.51×10^{-2}
ZYX	7	143381267	143390682	-4.04	5.28×10^{-5}	8.02×10^{-3}
CLPTM1	19	44954585	44992897	-17.79	0.00	0.00
ERCC2	19	45351391	45370540	-12.58	0.00	0.00
SCAF1	19	49642125	49658399	3.81	1.39×10^{-4}	1.51×10^{-2}

[a] PWAS risk gene shard with TIGAR (Bayesian weights)

Additionally, we plotted the pQTL weights respectively estimated by the Bayesian DPR method (TIGAR) and Elastic-Net regression (PrediXcan) versus position for gene MBLAC1 detected by both TIGAR and PrediXcan as well as the top significant gene MRPL16 only identified by TIGAR, color-coded with respect to log_{10} (p value) by GWAS of test SNPs. Bayesian DPR estimates generally had non-zero values for all SNPs within the test region, while Elastic-Net estimates had substantially less non-zero values within the test region that had pQTL weights of relatively larger magnitudes. We observed the test SNPs with non-zero weights estimated by DPR method (TIGAR) had more significant GWAS p values for top gene MRPL16 only detected by TIGAR compared to gene MBLAC1, since their PWAS associations are mainly driven by GWAS significant SNPs.

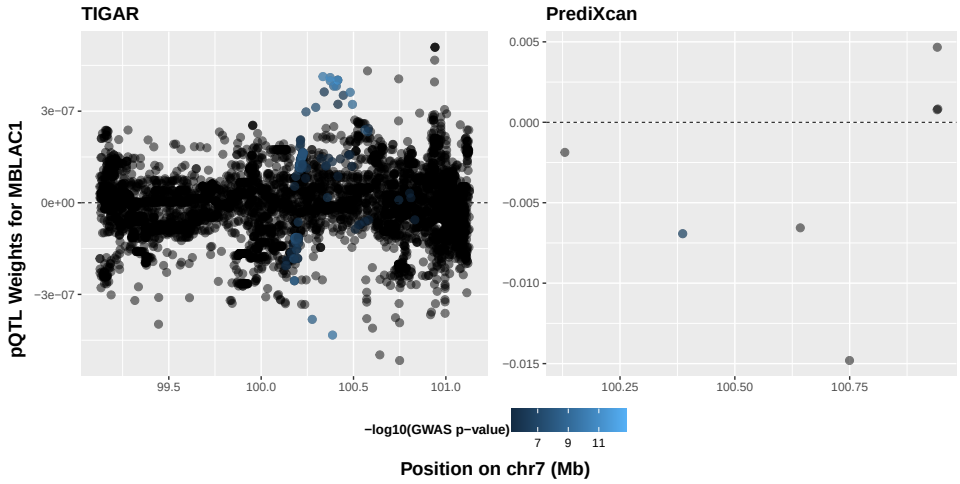

Figure 4.5: pQTL Weights Estimated by TIGAR/DPR and PrediXcan/Elastic-Net for MBLAC1, Color Coded with Respect to log_{10} (p value) by GWAS

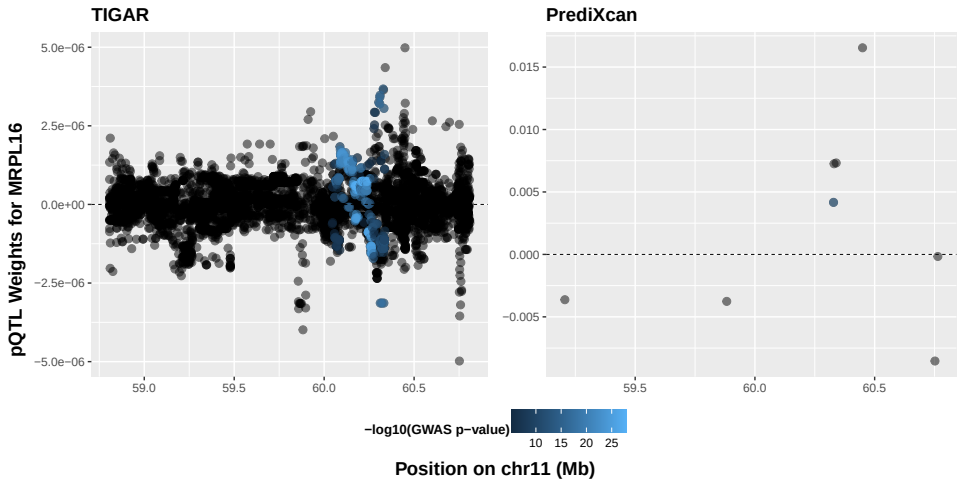

Figure 4.6: pQTL Weights Estimated by TIGAR/DPR and PrediXcan/Elastic-Net for MRPL16, Color Coded with Respect to log_{10} (p value) by GWAS

4.4 Compare with PWAS Results by FUSION as in Wingo's Paper

The significant genes from previous PWAS published in Wingo et al [6] using the same set of human proteomes were examined in our PWAS associations (Table 4.3). Gene STX6 was also identified by TIGAR at a p value adjusting for multiple testing. Gene ICA1L and LACTB were potential risk genes by TIGAR tool at a p value <0.05, with protein abundance prediction CV R^2 >0.005 trained by DPR. ACE was potential risk gene detected by PrediXican with valid training CV R^2. For genes with protein abundance models CV R^2 <0.005 trained by Elastic-Net method, PWAS p value was not reported by PredXican which was enabled by TIGAR tool.

We also applied the TIGAR tool to conduct the association study integrating the pQTL weights from FUSION as used by Wingo et al [6] with the updated AD GWAS summary statistics as our study. As a result, 4 genes (STX6, ICAL1, STX4, PVR) were identified as PWAS risk genes at an FDR of P (q-value) <0.05.

The PWAS risk genes identified by study of Wingo et al [6] using the same set of proteomics data were not originally detected by our TIGAR tool at an FDR of P <0.05. We could not replicate wingo's findings potentially due to different training models, updated GWAS summary data and different association test statistics (S-PrediXcan vs FUSION). Another reason might be because we did not adjust for AD status from our training protein abundance traits as Wingo's did.

Table 4.3: AD risk genes identified by previous PWAS

Gene	Chr	PWAS P (Wingo et al)	FUSION (Published weights)	TIGAR	TIGAR CV R^2	PrediXcan	PrediXcan CV R^2
$STX6^{b}$	1	1.3×10^{-4}	8.4×10^{-1}	1.2×10^{-3}	5.2×10^{-3}	-	7.9×10^{-9}
$ICA1L^{a,b}$	2	1.1×10^{-4}	8.1×10^{-3}	2.9×10^{-2}	-	-	
EPHX2	8	4.7×10^{-8}	7.0×10^{-2}	5.5×10^{-2}	8.5×10^{-3}	-	4.7×10^{-3}
PLEKHA1	10	1.1×10^{-5}	-	-	-	-	-
SNX32	11	2.8×10^{-6}	2.9×10^{-2}	7.7×10^{-1}	2.0×10^{-2}	-	8.4×10^{-4}
$LACTB^{b}$	15	1.7×10^{-4}	5.2×10^{-2}	1.2×10^{-2}	1.5×10^{-2}	-	5.2×10^{-3}
CTSH	15	2.9×10^{-6}	1.2×10^{-2}	3.4×10^{-2}	5.2×10^{-3}	-	3.1×10^{-3}
$DOC2A^{a}$	16	6.4×10^{-6}	1.6×10^{-6}	-	-	-	-
CARHSP1	16	2.6×10^{-4}	2.5×10^{-2}	-	-	-	-
$STX4^{a}$	16	6.2×10^{-5}	4.6×10^{-5}	4.8×10^{-1}	7.2×10^{-3}	-	2.9×10^{-3}
ACE^{c}	17	8.5×10^{-8}	-	-	-	8.8×10^{-3}	5.5×10^{-3}
PVR^{a}	19	7.1×10^{-28}	0	-	-	-	-
RTFDC1	20	2.1×10^{-5}	-	-	-	-	-

[a] PWAS risk genes by TIGAR with same pQTL weights as the previous study
[b] PWAS risk genes by TIGAR that were identified by previous PWAS
[c] PWAS risk genes by PrediXcan that were identified by previous PWAS

Chapter 5

Conclusion

In this study, we first conducted the weighted network analysis to identify the proteins network modules and found protein modules that were significantly associated with AD clinical traits. The key drivers of these significant modules had known associations with AD by previous correlation analysis.

Then we applied the TIGAR-V2 tool to conduct PWAS test for AD, using either nonparametric Bayesian DPR or Elastic-Net methods (used by PrediXcan) to train protein abundance prediction models taking genotype data as predictors for protein expression levels, as well as test the gene-based association using summary-level GWAS Z-score for single variants. The associated tests were implemented with burden statistics (S-PrediXcan).

We observed the Bayesian DPR provided more valid protein abundance prediction models (i.e 5-fold CV R^2 >0.005) than Elastic -Net method, which matched with the assumption that the nonparametric Bayesian method is more flexible and general. With larger number of valid protein abundance imputation models, the PWAS using Bayesian pQTL weights would detect more risk genes than Elastic-Net weights.

We identified PWAS risk genes for AD consistent with previous GWAS studies and potential risk genes that had known biological functions involved in the brain

measurement and cognition ability.

Our work still has limitations. For example, the CV R^2 is generally low for most proteins because *cis*-SNPs were considered as predictors and training sample size is small (n=355). Since standard TWAS methods such as TIGAR fail to account for horizontal pleiotropy effect between proteins and phenotypes [19], the PWAS significant proteins by TIGAR might only share the same causal SNPs with the phenotype of interest but not having any mediation effect. We will work on apply the PMR-Egger tool to the ROS/MAP data in our ongoing research.

In conclusion, we identified 13 brain genes that confer AD risk through affecting the protein abundance level for future studies of AD pathogenesis and therapeutics. We believe PWAS with TIGAR-V2 tool can be useful and crucial for mapping risk genes of complex disease.

Bibliography

[1] Alzheimer's disease facts and figures. Alzheimer's Dementia, 2020.

[2] D. P. Wightman, I. E. Jansen, J. E. Savage, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for alzheimer's disease. Nature genetics, 2021.

[3] A. Gusev, A. Ko, H. Shi, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics, 2016.

[4] S. M. Kelse, William L. P., L. Yue, et al. Brain transcriptome wide association study (twas) implicates 8 genes across 6 loci in alzheimer's disease. Alzheimer's Dementia, 2020.

[5] N. Brandes, N. Linial, and M. Linial. Pwas: proteome-wide association study-linking genes and phenotypes by functional variation in proteins. Genome biology, 2020.

[6] A. P. Wingo, Y. Liu, E. S. Gerasimov, et al. Integrating human brain proteomes with genome-wide association data implicates new proteins in alzheimer's disease pathogenesis. Nature genetics, 2021.

[7] E. Gamazon, H. Wheeler, K. Shah, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature genetics, 2015.

[8] A. Gusev, A. Ko, and H. Shi. Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics, 2016.

[9] S. Nagpal, X. Meng, M. P. Epstein, et al. Tigar: An improved bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. American journal of human genetics, 2019.

[10] R. L. Parrish, G. C. Gibson, M. P. Epstein, and J. Yang. Tigar-v2: Efficient twas tool with nonparametric bayesian eqtl weights of 49 tissue types from gtex v8. Human Genetics and Genomics Advances, 2022.

[11] B. Li and S. M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. American journal of human genetics, 2008.

[12] D. A. Bennett, A. S. Buchman, P. A. Boyle, et al. Religious orders study and rush memory and aging project. Journal of Alzheimer's disease, 2018.

[13] A. P. Wingo, W. Fan, D. M. Duong, et al. Shared proteomic effects of cerebral atherosclerosis and alzheimer's disease on the human brain. Nature neuroscience, 2020.

[14] P. Langfelder and S Horvath. Wgcna: an r package for weighted correlation network analysis. BMC Bioinformatics, 2008.

[15] D. Szklarczyk, A. L. Gable, D. Lyon, et al. String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research, 2019.

[16] D. Szklarczyk, A. L. Gable, K. C. Nastou, et al. The string database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research, 2021.

[17] Y. S. Hu, J. Xin, Y. Hu, et al. Analyzing the genes related to alzheimer's disease via a network and pathway-based approach. Alzheimer's research therapy, 2017.

[18] A. Buniello, J. MacArthur, M. Cerezo, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research, 2019.

[19] Z. Yuan, H. Zhu, P. Zeng, et al. Testing and controlling for horizontal pleiotropy with probabilistic mendelian randomization in transcriptome-wide association studies. Nature Communications volume, 2020.