Omnibus proteome-wide association study (PWAS-O) identified 43 risk genes for Alzheimer's disease dementia

Tingyang Hu[1,2], Randy L. Parrish[1,2], Qile Dai[1,2], Aron S. Buchman[3], Shinya Tasaki[3], David A. Bennett[3], Nicholas T. Seyfried[4], Michael P. Epstein[1], Jingjing Yang[1*]

1. Center for Computational and Quantitative Genetics, Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, 30322, USA
2. Department of Biostatistics and Bioinformatics, Emory University School of Public Health, Atlanta, GA, 30322, USA
3. Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, 60612, USA
4. Department of Biochemistry, Emory University School of Medicine, Atlanta, GA, 30322, USA

[*] Correspondence Author: jingjing.yang@emory.edu

Abstract

Proteome-wide association study (PWAS) integrating proteomics data with GWAS data is a powerful tool to identify risk genes for complex diseases, which can inform disease mechanisms with genetic effects mediated through protein abundance. We propose a novel omnibus method to improve PWAS power by modeling unknown genetic architectures with multiple statistical models. We applied TIGAR, PrediXcan, and FUSION to train protein abundance imputation models for 8,430 proteins from dorsolateral prefrontal cortex with whole genome sequencing data (n=355). Next, the trained models were integrated with GWAS summary data of Alzheimer's disease (AD) dementia (n=762,917) to conduct PWAS. Last, we employed the Aggregated Cauchy Association Test to obtain omnibus PWAS (PWAS-O) p-values from these three models. PWAS-O identified 43 risk genes of AD dementia including 5 novel risk genes that were interconnected through a protein-protein interaction network including *TOMM40*, *APOC1*, and *APOC2*. PWAS-O can be easily applied to study complex diseases.

Introduction

Recent large-scale genome-wide association studies (GWAS) of Alzheimer's disease (AD) dementia have identified 38 risk loci[1], but little is known about their underlying mechanisms. Integrating transcriptomic data with GWAS data through transcriptome-wide association study (TWAS) has been shown an effective approach for identifying potential causal risk genes of complex human diseases with biological interpretations[2-4]. TWAS can inform potential underlying disease mechanisms as the identified genetic effects may be potentially mediated via gene expression[5, 6]. Since many biological processes can affect protein abundance, gene expression levels might not be the optimal proxy for elucidating the role of proteins in disease mechanisms. The analogous proteome-wide association study (PWAS) approach integrating proteomic data with GWAS summary data[7-9] has the potential to identify novel genetic risk loci linked to disease via proteins that may be missed by GWAS and TWAS[2-4, 10].

Since gene expression and protein abundance are both molecular quantitative traits that might mediate genetic effects, existing TWAS tools can be used for PWAS. TWAS tools can train protein abundance imputation models using genetic variants as predictors from a reference panel with both proteomic and genetic data available. Analogously, genetic effect sizes in the imputation models could be viewed as effect sizes of "protein quantitative trait locus (pQTL)" in a broad sense (as these "pQTL" are not necessarily statistically significant), which would be taken as variant weights for gene-based association study (i.e., PWAS) with GWAS data.

Existing TWAS tools use various statistical regression models to train imputation models of quantitative molecular traits, and each tool has its own advantage for modeling a specific type of genetic architecture underlying the molecular trait of interest. For example, PrediXcan[11] uses a penalized linear regression model with Elastic-Net penalty[12] which has the advantage of modeling sparse genetic architectures. TIGAR[2] implements the nonparametric Bayesian latent Dirichlet process regression (DPR) which has been shown effective to model the infinitesimal genetic architecture[13]. FUSION[5] uses penalized linear regression model with Elastic-Net penalty and LASSO penalty[14], regular linear regression model with best unbiased linear predictor (BLUP), single variant model with Top eQTL or

54 pQTL (Top1), and Bayesian spare linear mixed model (BSLMM)[15], and then selects the best model for
55 gene-based association study based on cross-validation (CV) R^2 of these multiple models.
56 Even though the FUSION tool employs multiple statistical models, it only uses the model with
57 top performance based on reference data. Rather than selecting a single statistical method to estimate
58 molecular QTL weights, previous study[16] have shown that combining association test p-values based on
59 multiple statistical methods using the Aggregated Cauchy Association Test[17] (ACAT) can increase power
60 while controlling for type I error. Therefore, we apply ACAT to aggregate PWAS p-values obtained by
61 TIGAR, PrediXcan, and FUSION, which subsequently obtain an omnibus PWAS (PWAS-O) p-value.
62 PWAS-O is expected to improve PWAS power compared to PWAS based on a single imputation model
63 of protein abundances, as PWAS-O better accounts for the unknown genetic architecture of protein
64 abundances with multiple statistical models.
65 To evaluate the performance of PWAS-O, we conducted simulation studies using real genetic
66 data based on patterns observed in real PWAS. We showed that PWAS-O outperformed all three
67 individual methods with well calibrated type I error. We then applied PWAS-O to study AD dementia by
68 using the reference human proteome-wide data of dorsolateral prefrontal cortex (DLPFC) of 355 older
69 decedents with whole genome sequencing (WGS) genetic data, and summary data from the most recent
70 GWAS of AD dementia (n=762,917, excluding 23&me samples)[1]. Next, we examined the protein-protein
71 interaction (PPI) network among the identified PWAS risk genes of AD dementia. We identified 43 risk
72 genes of AD dementia by PWAS-O that were connected through a protein-protein interaction (PPI)
73 network involving the well-known risk genes of AD dementia. We validated the potential causality of 27
74 out of 43 (63%) PWAS risk genes by probabilistic Mendelian randomization (PMR-Egger)[18] and
75 summary data-based Mendelian randomization (SMR)[19, 20] tool.
76
77 **Results**
78
79 **<u>Omnibus PWAS (PWAS-O)</u>**
80

81 Similar to the standard two-stage PWAS/TWAS method (**Methods; Fig. 1**)[11, 21], protein
82 abundance imputation models are first trained in Stage I to obtain estimated effect sizes of genetic
83 variants/predictors (i.e., pQTL weights), by TIGAR[21] (nonparametric Bayesian DPR), PrediXcan[11]
84 (Elastic-Net), and FUSION[5] (Elastic-Net, LASSO, BLUP, Top1). Genetic variants within ±1MB around
85 the transcription starting and termination sites of the corresponding coding gene were considered as
86 predictors in the imputation models. In Stage II, these trained pQTL weights (coefficients of these genetic
87 variant predictors) are used as variant weights along with GWAS data to conduct gene-based association
88 tests. For each protein-coding gene, PWAS p-values will be respectively obtained by using the pQTL
89 weights estimated by TIGAR, PrediXcan, and FUSION, which will then be combined by ACAT to
90 generate the corresponding PWAS-O p-value (**Methods**). We present the PWAS-O framework in **Fig. 2**.
91 Both individual-level and summary-level GWAS data can be used in Stage II. A reference panel of the
92 same ancestry as the GWAS data can be used to provide LD information for gene-based association tests
93 in Stage II. The TIGAR[21] tool provides functions to calculate LD covariance matrices and conduct gene-
94 based association tests.
95
96 **Simulation Study**
97 Simulation studies were conducted to evaluate the performance of PWAS-O method. We use the
98 real genotype data from 1894 whole genome sequencing (WGS) samples from the Religious Orders
99 Study/Memory and Aging Project (ROS/MAP)[22] and Mount Sinai Brain Bank (MSBB)[23] study to
100 simulate protein abundance data. We randomly selected 500 protein-coding genes with various gene
101 length, and split samples into 568 training (30%) and 1326 testing (70%) samples to mimic a sample size
102 in real reference panels for training protein abundance imputation models. We considered 4 different
103 scenarios consisting of 2 proportions of causal cis-pQTL, $p_{causal} = (0.001, 0.01)$, and 2 proportions of
104 protein abundance variance attributed to causal pQTL, $h_p^2 = (0.01, 0.05)$.

We generated protein abundance (E_p) from the multivariable regression model $E_p = \mathbf{X}_g\mathbf{w} + \boldsymbol{\epsilon}_g$, using genotype matrix $\mathbf{X}_g$ of the randomly selected causal pQTL of protein p, $\boldsymbol{\epsilon}_g \sim N(0, (1 - h_p^2)\mathbf{I})$. For each protein-coding gene, we generated 10 replicates of pQTL effect sizes $\mathbf{w}$ from $N(0,1)$ per scenario, and rescaled these effect sizes to ensure the protein expression variance explained by causal variants is h_p^2. For each scenario, a total of 5000 replicates of protein traits were generated for selected 500 protein-coding genes. We then generated GWAS Z-scores from a multivariate normal distribution with $\mathbf{Z} \sim MVN\left(\boldsymbol{\Sigma}_g\mathbf{w}\sqrt{n_{gwas}\, h_{pheno}^2},\ \boldsymbol{\Sigma}_g\right)$ [24], where $\mathbf{w}$ denotes the true causal pQTL effect sizes, $\boldsymbol{\Sigma}_g$ is the correlation matrix of genotype $\mathbf{X}_{test}$ for testing samples, and n_{gwas} denotes GWAS sample size. We set $h_{pheno}^2 = 0.025$, denoting the amount of phenotypic variance explained by simulated genetically regulated protein abundance $\mathbf{X}_g\mathbf{w}$. We considered $n_{gwas} =$ (25K, 50K, 75K, 100K) for scenarios with $h_p^2 = 0.05$ and $n_{gwas} =$ (600K, 800K, 1000K, 1200K) with $h_p^2 = 0.01$. We generated 10 replicates of GWAS Z-scores per simulated protein trait per gene, resulting in a total of 50,000 PWAS simulation replicates.

In stage I of PWAS simulation analysis, we applied TIGAR, PrediXcan and FUSION to estimate pQTL weights from simulated protein traits of 568 training samples. In stage II, we integrated estimated pQTL weights with simulated GWAS Z-scores to conduct gene-based association tests. We further combined PWAS p-values obtained by these 3 tools using ACAT to generate PWAS-O p-values. We evaluated the training performance of these 3 individual methods by calculating the squared Pearson correlation coefficient between imputed genetically regulated protein abundances and simulated protein traits of 1326 testing samples, which was referred as test R^2. PWAS power was given by the proportion of 50,000 repeated simulations with PWAS p-value value $< 2.5 \times 10^{-6}$ (genome-wide significance threshold for gene-based association test). We compared PWAS power by PWAS-O to all 3 individual methods.

128 We showed that optimal test R^2 and PWAS power based on one statistical model (**Fig. 3**) depended
129 on the underlying genetic architecture of protein abundance (p_{causal}) and protein heritability (h_p^2). When
130 protein heritability was small ($h_p^2 = 0.01$), PrediXcan had the lowest test R^2 and power. When protein
131 heritability increased ($h_p^2 = 0.05$) with sparse true cis-pQTLs ($p_{causal} = 0.001$), PrediXcan became the
132 optimal method with highest test R^2 and power. For a less sparse model with $p_{causal} = 0.01$, FUSION
133 generally performed best among the individual methods. We found that PWAS-O obtained the highest
134 power across all the scenarios by combining the PWAS p-values based on 3 methods.
135 We also evaluated type I errors of these individual methods and PWAS-O under one simulated
136 scenario with $h_p^2 = 0.05$ and $p_{causal} = 0.001$. We simulated 2×10^3 phenotype from $N(0,1)$ for each of
137 the 500 genes and permuted estimated pQTL weights to conduct a total of 10^6 null simulations. Both
138 individual methods and PWAS-O showed well calibrated type I errors, as shown by quantile-quantile (Q-
139 Q) plots of p-values (**Fig. S1**).
140
141 **ROS/MAP Proteomics and WGS data**
142
143 The human brain proteomic reference data used in the PWAS-O application study of AD
144 dementia in this study were profiled from the DLPFC of postmortem brain samples donated by 400
145 participants of European ancestry of the Religious Orders Study and Memory and Aging Project
146 (ROS/MAP)[7, 22]. These human proteome abundance data underwent quality control (**Methods; Fig. S2**)
147 that regressed out the effects of technical factors and confounding covariates including age at death, sex,
148 postmortem interval, and study (ROS or MAP). Proteome abundance data of 8,430 proteins of 355
149 individuals who also had WGS data profiled[25] were obtained after quality control and used as reference
150 data in the PWAS-O application study of AD dementia in this study. This ROS/MAP reference proteomic
151 data set is one of the largest available proteomic data sets for DLPFC tissue which is a relevant tissue for
152 studying AD dementia.
153

154 **PWAS-O results of studying AD dementia**

155 At training Stage I, five-fold cross-validation was conducted to evaluate the training performance
156 per statistical model. The protein abundance imputation models with 5-fold CV R^2 >0.5% were retained
157 and their estimated pQTL weights were used to conduct PWAS. As suggested by the TIGAR tool[21], the
158 threshold of 5-fold CV R^2 >0.5% was arbitrarily chosen to allow a greater number of protein-coding
159 genes to be tested by the follow-up gene-based association study (i.e., PWAS). The threshold of CV R^2
160 would not affect the statistical validity of the test statistic used in the follow-up PWAS.

161 As shown in **Table 1**, 6,673 protein abundance imputation models were retained by TIGAR with
162 median CV R^2 = 0.013; 1,835 protein abundance imputation models were retained by PrediXcan with
163 median CV R^2 = 0.0094; and 2,173 protein abundance imputation models were retained by FUSION (the
164 best model selected out of LASSO, Elastic-Net, BLUP, Top1 models with the highest CV R^2) with
165 median CV R^2 = 0.0093. Histograms of the CV R^2 obtained by these three tools were presented in **Fig.**
166 **S3**.

167 As described in the above about the PWAS-O framework, we first conducted PWAS by
168 employing TIGAR[21], PrediXcan[11], and FUSION[5] tools, using the ROS/MAP reference proteomics data of
169 DLPFC and the most recent GWAS[1] summary data of AD dementia (n=762,917, excluding 23&me
170 samples; **Methods**). That is, the estimated pQTL effect sizes in the retained protein abundance imputation
171 models by these three tools were taken as variant weights to conduct gene-based association tests along
172 with GWAS summary data of AD dementia (**Fig. S4**). Second, we implemented ACAT[17] to obtain
173 PWAS-O p-values from the PWAS p-values by these three tools. Q-Q plots of PWAS-O and PWAS
174 results of AD dementia by these three tools were presented in **Fig. S5**.

175 Our PWAS-O approach identified 43 significant genes with false discovery rate (FDR) (i.e., q-
176 value) <0.05, which could affect the risk of AD dementia through their genetic effects on the
177 corresponding protein abundances (**Fig. 4**). We curated 19 independent genes from these 43 significant
178 genes (**Table 2; Table S1**), by considering only the most significant genes among all significant genes

179 with shared test regions (±1MB around the test transcription starting and termination sites of the protein-
180 coding gene). In particular, we found all individual methods contributed to the PWAS-O results (**Table 2;**
181 **Table S1; Fig. S6**). Out of 43 PWAS risk genes, 33, 9, and 7 genes were detected by TIGAR, FUSION,
182 and PrediXcan, respectively.

183 We found 10 out of these 43 PWAS risk genes were detected by previous GWAS studies of AD
184 as shown in GWAS Catalogue[26], including *AGFG2*[27], *CLPTM1*[28], *PVR*[28], *PVRL2*[28], *APOC1*[29], *APOC2*[30],
185 *CKM*[28], *ERCC1*[31], *OPA3*[31], and *SYMPK*[31]. PWAS risk genes *CDC42*[32], *MAP3K7*[33], and *FNBP4*[32] were
186 found associated with cognitive ability in previous GWAS studies. PWAS risk genes *CCDC86*[34],
187 *KIF18B*[34], *PPM1N*[34], and *RTN2*[34] were previously found associated with neurofibrillary tangles
188 measurement (a key clinical indicator of AD) by GWAS studies. Additionally, 13 out of 43 significant
189 AD PWAS genes (*ZYX*[35], *C1QTNF4*[36], *FNBP4*[35], *PVR*[3], *PVRL2*[36], *TOMM40*[3], *APOC1*[4], *APOC2*[37],
190 *CLPTM1*[36], *ERCC1*[38], *OPA3*[38], *EML2*[4] and *DMWD*[35]) were significant risk genes of AD by previously
191 published TWAS.

192 Our PWAS-O findings not only replicated significant risk genes identified by previous GWAS
193 and TWAS, but also identify novel risk genes for AD dementia. PWAS-O identified 5 novel risk genes
194 that were neither detected by GWAS nor TWAS (*MGAT5, PACSIN3*, *PLA2G7, DIABLO, YWHAB*). The
195 novel gene *PLA2G7* was shown to have biological function involved in inflammatory and oxidative
196 activities[39], with a known association with AD and cognitive decline[40]. The novel gene *PACSIN3* has
197 been implicated in the disorders of the central nervous system such as schizophrenia[41].

198

199 **<u>pQTL weights by different regression models</u>**

200 To investigate the genetic architecture underlying protein abundances, we plotted the pQTL
201 weights estimated by TIGAR (DPR), PrediXcan (Elastic-Net), and FUSION (best model out of LASSO,
202 Elastic-Net, BLUP, Top1) for significant PWAS risk genes *PPM1N* and *DCAKD* in **Fig. 5**. Generally,
203 TIGAR and FUSION had non-zero pQTL weights estimated for almost all test SNPs within the test
204 region, for using a model assuming infinitesimal genetic architecture[13]. Whereas, PrediXcan had much

205 fewer non-zero pQTL weights estimated for assuming a sparse genetic architecture and enforcing variable
206 selection. For example, *PPM1N* was a significant risk gene identified by both TIGAR and FUSION
207 (**Table 2**), which has hundreds of test SNPs with non-zero pQTL weights (by both TIGAR and FUSION)
208 that were colocalized with significant GWAS signals (colored dots in **Fig. 5A**). Gene *DCAKD* was only
209 found significant by FUSION (p-value = 3.12×10^{-5}) but not by TIGAR (p-value = 3.62×10^{-3}), for
210 having fewer test SNPs of relatively large magnitude of pQTL weights that were colocalized with
211 significant GWAS signals (**Table2, Fig. 5B**).
212 Similarly, comparing pQTL weights obtained by TIGAR to PrediXcan for significant PWAS
213 genes *MBLAC1* and *MRPL16* (**Fig. S7**), we observed that less than 10 test SNPs had non-zero pQTL
214 weights by PrediXcan, versus thousands of test SNPs with non-zero pQTL weights by TIGAR. Because
215 PrediXcan estimated zero pQTL weights for most SNPs with significant GWAS p-values in the test
216 region of gene *MRPL16* (**Table 2**, **Fig. S7A**), *MRPL16* was only found significant by TIGAR (p-value =
217 4.33×10^{-15}), but not by PrediXcan (p-value = 1.31×10^{-2}). Whereas, when both TIGAR and
218 PrediXcan estimated non-zero pQTL weights for SNPs with significant GWAS p-values in the test region
219 of gene *MBLAC1* (**Table S1, Fig. S7B**), it was identified significant by both TIGAR (p-value =
220 7.31×10^{-5}) and PrediXcan (p-value = 3.49×10^{-4}).
221 These scatterplots of pQTL weights of PWAS risk genes showed that significant PWAS genes
222 were mainly driven by test SNPs that had non-zero pQTL weights colocalized with significant GWAS
223 signals. Because the underlying genetic architectures of protein abundances are complicated and unknown
224 in practice, no single statistical model would lead to superior PWAS power across genome-wide protein-
225 coding genes. Our proposed PWAS-O approach is expected to gain power in practice for considering
226 multiple statistical models per protein.
227
228 **PPI network among PWAS risk genes of AD dementia**
229 Next, we used STRING[42, 43] to investigate the PPI network among these 43 PWAS risk genes of
230 AD dementia, by using public data sources of known protein-protein interactions including evidences

231 from curated databases, experiments, text mining, co-expression, gene fusions, and protein homology[42]
232 (**Methods**). STRING clustered 23 out of these 43 PWAS risk genes into 4 communities (**Fig. 6**), where
233 different data sources of the protein-protein connection were depicted by differently colored connection
234 edges. It is interesting to see that the largest community of these PWAS risk genes is a network involving
235 genes *TOMM40*, *APOC1*, and *APOC2* that are located nearby the well-known risk gene *APOE* of AD
236 dementia, and are known to be associated with both AD dementia and low-density lipoprotein levels[44].
237 STRING also provided gene enrichment analysis results with respect to Gene Ontologies (GO)
238 and KEGG pathways. We observed that these 43 PWAS risk genes were enriched for the GO terms of
239 UV protection (*ERCC1, ERCC2, ERCC3*) with FDR = 0.010, transcription-coupled nucleotide-excision
240 repair (*ERCC1, ERCC2, ERCC3, PRPF19*) with FDR = 0.013, plasma lipoprotein particle remodeling
241 (*APOC1, APOC2, PLA2G7*) with FDR = 0.014, negative regulation of very-low-density lipoprotein
242 particle clearance (*APOC1, APOC2*) with FDR = 0.018, protein-containing complex subunit (*MRPL16,*
243 *ATPAF2, KIF18B, ERCC2, ERCC3, PRPF19, APOC1, APOC2, PLA2G7*) with FDR = 0.026, and the
244 KEGG pathway of adherens junction (*MAP3K7*, *CDC42*, *PVRL2*) with FDR = 0.012.
245 Interestingly, novel PWAS risk genes *ERCC3*, *ERCC2* and *PRPF19* are closely interconnected
246 with a known GWAS/TWAS risk gene *ERCC1* in the PPI network and enriched in nucleotide-excision
247 repair pathway. Novel PWAS risk gene *PLA2G7* connects with *APOC1* and *APOC2* in the PPI network
248 and involves in the plasma lipoprotein particle remodeling pathway.
249 These PPI network analysis results showed that the list of identified risk genes by PWAS-O were
250 connected with known risk genes of AD dementia and enriched in AD related pathways.
251
252 **Examine causal and pleiotropy effects of PWAS risk genes**
253 We applied both probabilistic Mendelian randomization (PMR-Egger)[18] and summary data-based
254 Mendelian randomization (SMR) methods to test if the causal genetic effects on AD dementia were
255 mediated through protein abundance in the possible presence of horizontal pleiotropy (**Methods**).
256 Horizontal pleiotropy means that genetic effects on AD dementia and on protein abundances originated

257 from the same SNPs but through different pathways (i.e., with no mediation effects). PMR-Egger
258 suggested that 25 out of 43 PWAS risk genes have significant causal effects (FDR<0.05) on AD dementia
259 after controlling for potential horizontal pleiotropy effects, including 6 genes (*PVRL2*, *TOMM40, APOC1,*
260 *RTN2, OPA3, EML2*) nearby the *APOE* region that also have significant horizontal pleiotropy effects with
261 FDR<0.05. Similarly, the method of SMR with its accompanying heterogeneity in dependent instruments
262 (HEIDI) test found that 22 out of 43 PWAS risk genes have potential causal effects on AD dementia with
263 FDR<0.1, of which 20 genes were overlapped with 25 causal genes detected by PMR tool. These results
264 validated potential causal genetic effects that were mediated through protein abundances for 27 (63%)
265 PWAS risk genes identified by PWAS-O (**Table 2; Table S1**).
266
267 **Compare with TWAS results of AD dementia**
268 To determine whether the 43 PWAS risk genes had similar evidence at the transcript level, we
269 also conducted TWAS of AD dementia using ROS/MAP transcriptomic reference data of DLPFC tissue
270 and the same GWAS summary data of AD dementia by TIGAR and PredXican. From TWAS p-values by
271 these two tools, we obtained omnibus TWAS (TWAS-O) p-values by ACAT (**Methods**). We found that
272 34 out of 43 PWAS risk genes (79%) were either detected by TWAS-O or located within 1MB of the
273 transcription starting sites and termination sites of the significant risk genes by TWAS-O with FDR<0.05
274 (**Table S2, Fig. S8**). These results showed that the majority of PWAS risk genes had genetic effects on
275 phenotype that were mediated through both transcriptome and proteome.
276
277 **Discussion**
278 In this work, we propose an omnibus PWAS approach (PWAS-O) that improves power for better
279 modeling of the unknown genetic architecture of proteome abundances by multiple statistical models. By
280 aggregating PWAS p-values obtained from several available tools (i.e., TIGAR, PrediXcan, and
281 FUSION) that used different statistical models, our PWAS-O approach improves power over using one
282 individual tool. Our simulation studies showed that the performance of each of these three tools depended

283 on the underlying genetic architecture for protein abundance and protein heritability. We found PrediXcan
284 performed better for sparse architecture and higher heritability, whereas TIGAR and FUSION had better
285 performance for denser architecture. PWAS-O was demonstrated with optimal performance across all
286 considered simulation scenarios with well calibrated type I error. Since the genetic architecture of protein
287 abundance is unknown, the practical use of PWAS-O is well justified.

288 By using the ROS/MAP reference proteomic data and the most recent GWAS summary data of
289 AD dementia, PWAS-O identified 43 risk genes of AD dementia. All three individual methods
290 contributed to the PWAS-O results. Notable biological roles were shown by previous studies for these
291 identified PWAS risk genes include lipoprotein metabolism, inflammatory reactions, DNA repair, pre-
292 mRNA splicing, and mitochondrial function. For example, *APOC1* and *APOC2*, located in a gene cluster
293 on chromosome 19 including the well-known AD risk gene *APOE*[45, 46], play important roles in the
294 development of lipoprotein metabolism[44]. Especially, *APOE* is an intensively studied risk gene of AD and
295 plays an essential role in the central nervous system[47, 48]. Connected with *APOC1* and *APOC2,* as shown
296 in the PPI of our identified PWAS risk genes, *PLA2G7* is a novel AD risk gene identified by PWAS-O
297 and associated with inflammatory oxidative activities, and has been implicated in cognitive impairment
298 and AD[39].

299 Additionally, *ERCC1, ERCC2,* and *ERCC3* are involved in nucleotide excision repair pathway
300 contributing to genome stability[49, 50], which plays a critical role during the development of the central
301 nervous system[51] and whose impairment is associated with AD[52]. Gene *PRPF19* that is interconnected
302 with *ERCC1, ERCC2*, and *ERCC3* encodes a pre-mRNA processing factor involved in regulating pre-
303 mRNA splicing[53]. Prior studies indicated that *PRPF19* had a neuronal function and was implicated in
304 neurological diseases[54, 55]; lower expression level of gene *MRPL16* encoding a mitochondrial ribosome
305 protein was observed in AD blood cells which may reflect a general response to inflammatory
306 environment in AD[56]; gene *ATPAF2* connected with *MRPL*16 encodes a protein involved in
307 mitochondrial function and lipid pathway, which has been implicated in dementia pathogenesis[57].

308 By STRING analysis, we showed that out of these 43 PWAS significant risk genes, 9 genes were
309 enriched in the pathway of protein-containing complex subunit; 4 genes were enriched in the gene
310 ontology pathway of transcription-coupled nucleotide-excision repair; 3 genes were enriched in the
311 pathway of plasma lipoprotein particle remodeling. Compared to TWAS-O results using the same GWAS
312 summary data, 34 out of 43 significant PWAS risk genes were either TWAS-O risk genes or within 1MB
313 test region of TWAS risk genes. These results showed that PWAS risk genes identified by our PWAS-O
314 approach are part of key biological pathways involving brain functions. The genes and proteins identified
315 may be high-value therapeutic targets for further drug discovery as well as for mechanistic studies that
316 can advance our understanding of the biology driving AD dementia in aging brains.
317 PWAS-O still has its limitations. First, the CV R^2 is generally low for majority protein
318 abundance prediction models. This might be due to small sample sizes in the reference proteomic data, or
319 low heritability of protein abundances due to cis-SNPs. Only *cis*-SNPs within the $\pm 1MB$ region around
320 the transcription starting and termination sites of the corresponding protein-coding gene were considered
321 as predictors for protein abundances by PWAS-O. Trans-SNPs could also be considered as additional
322 predictors in the protein abundance prediction models, which might increase the CV R^2 of protein
323 abundance imputation models but will be subject to high computation burden[3]. Second, the standard two-
324 stage PWAS method cannot distinguish whether causal SNPs for the phenotype of interest are only
325 colocalized with pQTLs but have no mediation effect through the protein abundances (i.e., horizontal
326 pleiotropy) or indeed have genetic effects mediated by protein abundances (i.e., vertical pleiotropy).
327 Horizontal pleiotropy happens when genetic variants impact the outcome through pathways different from
328 or additional to the proteome abundances of the test gene[58]. Here, we applied PMR-Egger[18] and SMR[59]
329 tool to test the causal effects of 43 PWAS risk genes while controlling for horizontal pleiotropy. We
330 validated the potential causality 27 genes out of 43 genes identified by our PWAS-O. Third, a cluster of
331 significant PWAS risk genes are likely to be identified as shown in the locus around *APOE* for AD
332 dementia for sharing test SNPs in the overlapped test gene region. Even though the PMR-Egger and SMR

tools provide some insights into the potential causality/mediation role of the genetically regulated protein abundances for the phenotype of interest, true causal genes will still need to be replicated by further biological experiments.

In summary, PWAS-O represents the first omnibus PWAS method that increases power by leveraging multiple statistical models. PWAS-O can be easily applied to integrating reference proteomic data with publicly available large-scale GWAS summary data, for studying complex polygenic diseases. We believe that PWAS-O would be useful for mapping risk genes with genetic effects potentially mediated through proteome abundances for complex diseases. Identified PWAS risk genes would shed light upon future studies of disease pathogenesis and therapeutics.

Methods

Two-stage PWAS

As shown in **Fig. 1**, the two-stage PWAS first trains protein abundance imputation models (Stage I) by taking *cis*-genotype data, ±1MB around the transcription starting and termination sites of the corresponding protein-coding gene, as predictors (X), protein abundance level ($\mathbf{E}_p$) as the outcome, as shown in the following linear regression model:

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

The effect size vector $\mathbf{w}$ could be estimated by different statistical models as employed by TIGAR[21], PrediXcan[11], and FUSION[5], which could be viewed as a broad sense of pQTL effect sizes (also referred to as pQTL weights). Specifically, PrediXcan employs a penalized linear regression model with Elastic-Net penalty[11]; FUSION[5] uses penalized linear regression model with Elastic-Net penalty and LASSO penalty[14], regular linear regression model with best unbiased linear predictor (BLUP), single variant model with Top pQTL (Top1), and Bayesian spare linear mixed model (BSLMM)[15], and then selects the best model according to CV R^2; TIGAR[2] implements the nonparametric Bayesian latent DPR model.

360 In Stage II, with individual-level GWAS data, the protein abundances will first be imputed by
361 using the trained imputation models and individual-level genotype data of test samples. Then PWAS will
362 test the association between the imputed protein abundances and the phenotype of interest. With summary
363 GWAS data, PWAS would take the pQTL weights estimated in Stage I as variant weights to conduct
364 gene-based association test with GWAS Z-scores and reference LD, which has been shown to be
365 equivalent as testing the association between imputed genetically regulated protein abundances and the
366 phenotype of interest[60]. That is, the PWAS test is equivalent to a burden test with the following test
367 statistic:

368 $$\tilde{Z}_{p,S\text{-}PrediXcan} = \frac{\sum_{l=1}^{m}(\widehat{w_l}\widehat{\sigma_l}Z_l)}{\sqrt{\widehat{\mathbf{w}}'\mathbf{V}\widehat{\mathbf{w}}}}, \quad \widehat{\sigma_l^2} = \mathrm{Var}(\mathbf{X_{0,l}}), \quad \mathbf{V} = \mathrm{Cov}(\mathbf{G}_0).$$

369

370 Here, $\widehat{w_l}$ denotes the pQTL effect size estimates from protein imputation models; Z_l denotes the Z-score
371 statistics of single variant l by GWAS; and $\mathbf{V}$ denotes the LD covariance matrix from a reference panel of
372 the same ancestry as the test GWAS cohort. The genotype variance $\widehat{\sigma_l}$ of test SNPs can also be estimated
373 from a reference panel of the same ancestry as the test cohort.

374

375 **PWAS-O Method**

376 As shown in **Fig. 2**, with PWAS p-values obtained by two-stage PWAS tools including TIGAR[21],
377 PrediXcan[11], and FUSION[5], PWAS-O uses ACAT[17] to aggregate these three PWAS p-values to generate
378 an omnibus test p-value per protein-coding gene. The ACAT test statistic is given by

379 $$T = \frac{1}{K}\sum_{k=1}^{K}\tan\{(0.5 - p_k)\}\pi\} , \qquad \pi \approx 3.14;$$

380 where p_k is the PWAS p-value obtained by standard two-stage PWAS tools. Under null hypothesis, the p-
381 value p_k is uniformly distributed in $[0, 1]$ and the transformation $\tan\{(0.5 - p_k)\}$ is Cauchy distributed.
382 Thus, the null distribution of the test statistic T can be approximated by a Cauchy distribution[18] and the
383 corresponding PWAS-O p-value can be approximated by

384 $$pvalue_T \approx \frac{1}{2} - \frac{\{\arctan(T)\}}{\pi}, \qquad \pi \approx 3.14.$$

385
386
387 **ROS/MAP Proteomics and WGS Data**

388 Human brain proteomic profiles were generated from the dorsolateral prefrontal cortex (DLPFC)
389 of postmortem brain samples donated by 400 participants of European ancestry of the Religious Orders
390 Study or Rush Memory and Aging Project (ROS/MAP)[7, 22] . Participants in ROS and MAP enroll without
391 known dementia and agree to annual clinical evaluation and organ donation at death. All participants sign
392 an informed consent, Anatomic Gift Act and Repository Consent. Both studies were reviewed by an
393 Institutional Review Board of Rush University Medical Center. Age was calculated from birth date and
394 date of death. Sex was self-reported. Protein measurements from DLPFC were quantified by tandem mass
395 spectrometry.

396 As described in previous publication[61], protein abundances were normalized by scaling each
397 protein abundance with respect to a sample-specific total protein abundance, and proteins with missing
398 values in more than 50% of the participants were excluded from the analyses. Protein abundance ratios
399 related to baseline (sample-specific total protein abundance) were $\log_2$ transformed, with missing values
400 imputed by the corresponding mean in the cohort. Confounding covariates of clinical characteristics
401 including sex, age at death, postmortem interval, and study type (ROS or MAP), as well as technical
402 factors including sequencing batch and MS2 versus MS3 mass spectrometry reporter quantification mode,
403 were regressed out from protein abundances. Outlier samples were removed using iterative principal
404 component analysis (PCA) as shown in **Fig. S2**. That is, samples with one of the top 2 principal
405 components (PCs) greater than the corresponding four times of the standard deviations from the mean of
406 the corresponding PC[61].

407 The genotype data of ROS/MAP samples were profiled by whole genome sequencing (WGS)[25].
408 SNPs with Hardy-Weinberg equilibrium p-values $<10^{-5}$ and minor allele frequency >1% were used as
409 genetic predictors for training the protein abundance imputation models. After quality control of both

410 proteomic and genetic data, a total of 355 samples with both proteomic and WGS data were used as the
411 reference data for the application study of PWAS-O in this study.
412
413 **Summary-level GWAS data of AD dementia**
414 The summary-level GWAS data of AD dementia (i.e., single variant Z-score test statistics
415 obtained by meta-analysis) were generated by the latest GWAS by Wightman et al[1], including 762,917
416 individuals from 12 cohorts except 23&Me samples. About 11.3% of participants had clinically
417 diagnosed AD dementia.
418
419 **PWAS of AD dementia**
420 In this study, we used the ROS/MAP proteomics and WGS data of 355 samples as reference data
421 to train protein abundance imputation models. For each protein, we first used TIGAR tool to train
422 nonparametric Bayesian DPR model and Elastic-Net regression model (as used by PrediXcan), and then
423 used FUSION[5] tool to train LASSO, Elastic-Net, BLUP, and Top1 models and then selected the best
424 model according to the CV R^2. We did not train a BSLMM model due to its high computation cost for
425 WGS data, and we expect that the nonparametric Bayesian DPR model would account for the mixed
426 sparse and infinitesimal genetic architecture. The 5-fold CV was conducted for all protein abundance
427 imputation models. The average CV R^2 across 5 folds of validation data per protein was used to select the
428 best performing FUSION model as well as filter protein abundance imputation models for the follow-up
429 PWAS.
430 Second, protein abundance imputation models with 5-fold CV R^2 >0.005 were considered as
431 valid, and the corresponding pQTL weights were used along with summary-level GWAS data of AD
432 dementia to test for PWAS risk genes[21]. Last, PWAS-O p-values were generated by aggregating PWAS p-
433 values obtained by using pQTL weights estimated by TIGAR (DPR), PrediXcan (Elastic-Net), and
434 FUSION (best model) by ACAT.
435

436 **PPI network and pathway analysis**

437

438 To investigate the PPI network among the identified PWAS risk genes of AD dementia, we used
439 STRING[42, 43], a website for PPI network analysis which integrates public data sources of protein
440 interaction information and investigates the connectivity network of proteins. Proteins are considered as
441 nodes in the PPI network. Protein-protein edges represent the predicted functional associations, colored
442 differently to indicate seven categories — co-expression, text-mining, experiments (biochemical/genetic
443 data), databases (previously curated pathway and protein complex information), gene co-occurrence, gene
444 fusion and neighborhood. Gene co-occurrence, fusion, and neighborhood represent association
445 predictions are based on whole-genome comparisons.

446 STRING also enables gene-set enrichment analysis with respect to Gene Ontology (GO)
447 annotations[62], KEGG pathways[63], Reactome pathways[64], and UniProt keywords[65]. Enrichment analysis is
448 to detect GO terms, pathways, or UniProt keywords that are significantly enriched with genes in the
449 network versus random genes, adjusting for FDR. STRING also shows significant association between
450 input proteins and human phenotypes based on Monarch Initiative[66], a system to organize genotype-
451 phenotype connections from multiple data sources.

452

453 **Test causal and pleiotropy effects**

454 We used the PMR-Egger[18] tool to test the vertical pleiotropy effect (i.e., causal mediation effect)
455 between the protein abundance level of PWAS risk genes and AD dementia, while controlling for
456 horizontal effect. The Multi-SNP-based SMR[59, 67] software and its accompanying heterogeneity in
457 dependent instruments (HEIDI) were also implemented to test the causal effect, using multiple *cis*-pQTL
458 SNPs. Reference LD derived from ROS/MAP WGS data were used to implement both PMR-Egger and
459 SMR-HEIDI methods. Test cis-SNPs were pruned according to their LD correlation matrix. The threshold
460 of $R^2 = 0.95$ was used for implementing PMR-Egger and $R^2 = 0.99$ was used for implementing SMR-

461 HEIDI. For each protein-coding gene, only SNPs that have nominal eQTL p-value < 0.05 were tested by
462 SMR-HEIDI.
463

464 **TWAS-O of AD dementia**

465 TWAS p-values were first obtained by using reference transcriptomic data of DLPFC tissue from
466 ROS/MAP samples and GWAS summary data of AD dementia by TIGAR and PrediXcan, and then used
467 to derive TWAS-O p-values by ACAT. The same GWAS summary data of AD dementia were used in
468 both PWAS-O and TWAS-O analysis in this study. The reference transcriptomic data include 465
469 ROS/MAP samples with both gene expression data for DLPFC tissue (by RNA sequencing) and WGS
470 genotype data profiled. Gene expression data in the format of Transcripts Per Million (TPM) were log2
471 transformed and adjusted for age at death, sex, postmortem interval, study (ROS or MAP), batch effects,
472 RNA integrity number scores, estimated cell type proportions (with respect to oligodendrocytes,
473 astrocytes, microglia, neurons), top five genotype PCs, and top probabilistic estimation of expression
474 residuals (PEER) factors by linear regression models. The cell type proportions were estimated by using
475 CIBERSORT pipeline[68] with single-cell RNA-seq transcriptome profiles from human brain tissues as the
476 reference[69] to de-convolute bulk RNA-seq data[70]. Only genes with >0.1 TPM in ≥10 samples were
477 considered. Cis-SNPs within a 1MB window of each gene that met the MAF threshold >1% and Hardy-
478 Weinberg Equilibrium (HWE) threshold of p-value>10^{-5} were used in the imputation model as predictors.
479 Only genes with an average 5-fold cross validation $R^2 \geq 0.005$ in the gene expression imputation models
480 were analyzed in TWAS.
481

482 **Acknowledgements**

483 We are thankful to all participants and the staff at the Rush Alzheimer's Disease Center for making the
484 proteomics and WGS data of ROS/MAP samples available to the research community. This work was
485 supported by the National Institute of Health R35GM138313 (T.H., R.P., Q.D., J.Y.). The ROS/MAP

studies are supported by P30AG10161, K01AG054700, R01AG15819, R01AG17917, R01AG56352; the Illinois Department of Public Health; and the Robert C. Borwell Endowment Fund. The funding organizations had no role in the design or conduct of the study; collection, management, analysis, or interpretation of the data; or preparation, review, or approval of the manuscript.

Author contributions

T.H. conducted data analysis and drafted the manuscript. J.Y. conceptualized and led the project, and edited the manuscript. R.P. and Q.D. contributed to data analysis. A.S.B., S.T., D.A.B., N.T.S. contributed to ROS/MAP data and edited the manuscript. M.P.E consulted data analysis results and edited the manuscript.

Data availability

Proteomic, transcriptomic, and WGS data of ROS/MAP samples used in this manuscript are available from Synapse (https://doi.org/10.7303/syn3219045). These data are available via the AD Knowledge Portal (https://adknowledgeportal.org), subject to requirements for data access and data attribution as provided in https://adknowledgeportal.org/DataAccess/Instructions. GWAS summary data of AD dementia is available from https://ctg.cncr.nl/software/summary_statistics. Results of the pQTL weights and PWAS summary data generated in this study will be deposited to Synapse after publication of this work. Instructions for implementing PWAS-O are available through PWAS-O Github page (https://github.com/tingyhu45/PWAS-O).

Ethics declarations

Both MAP and ROS studies were approved by the Institutional Review Board of Rush University Medical Center. All individual-level data analyzed in this work were de-identified. Written informed consent was obtained from all study participants as was an Anatomical Gift Act for organ donation. ROSMAP resources can be requested at www.radc.rush.edu.

512 **Competing interests**

513 The authors declare no competing interests.

514

515 **Supplementary information**

516 The supplementary file contains 2 supplementary tables and 8 supplementary figures.

517

References

1. Wightman, D.P., *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nature genetics* **53**, 1276-1282 (2021).
2. Nagpal, S., *et al.* TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. *Am J Hum Genet* **105**, 258-266 (2019).
3. Luningham, J.M., *et al.* Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. *American journal of human genetics* **107**, 714-726 (2020).
4. Tang, S., *et al.* Novel Variance-Component TWAS method for studying complex human diseases with applications to Alzheimer's dementia. *PLoS genetics* **17**, e1009482 (2021).
5. Gusev, A., *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics* **48**, 245-252 (2016).
6. Gockley, J., *et al.* Brain transcriptome wide association study (TWAS) implicates 8 genes across 6 loci in Alzheimer's disease: Genetics: Genetics and omics of AD I. *Alzheimer's & Dementia* **16**, e044839 (2020).
7. Wingo, A.P., *et al.* Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. *Nature genetics* **53**, 143-146 (2021).
8. Wingo, T.S., *et al.* Brain proteome-wide association study implicates novel proteins in depression pathogenesis. *Nat Neurosci* **24**, 810-817 (2021).
9. Zhang, J., *et al.* Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat Genet* **54**, 593-602 (2022).
10. Brandes, N., Linial, N. & Linial, M. PWAS: proteome-wide association study-linking genes and phenotypes by functional variation in proteins. *Genome biology* **21**, 173 (2020).
11. Gamazon, E.R., *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* **47**, 1091-1098 (2015).
12. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301-320 (2005).
13. Boyle, E.A., Li, Y.I. & Pritchard, J.K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186 (2017).
14. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267-288 (1996).
15. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics* **9**, e1003264 (2013).
16. Dai, Q., *et al.* OTTERS: a powerful TWAS framework leveraging summary-level reference data. *Nature Communications* **14**, 1271 (2023).
17. Liu, Y., *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *American journal of human genetics* **104**, 410-421 (2019).
18. Yuan, Z., *et al.* Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nat Commun* **11**, 3861 (2020).
19. Zhu, Z., *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-487 (2016).
20. Wu, Y., *et al.* Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Commun* **9**, 918 (2018).

21. Parrish, R.L., Gibson, G.C., Epstein, M.P. & Yang, J. TIGAR-V2: Efficient TWAS tool with nonparametric Bayesian eQTL weights of 49 tissue types from GTEx V8. *HGG Adv* **3**, 100068 (2022).
22. Bennett, D.A.*, et al.* Religious Orders Study and Rush Memory and Aging Project. *J Alzheimers Dis* **64**, S161-S189 (2018).
23. Wang, M.*, et al.* The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Sci Data* **5**, 180185 (2018).
24. Feng, H.*, et al.* Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association studies. *PLoS Genet* **17**, e1008973 (2021).
25. De Jager, P.L.*, et al.* A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Scientific data* **5**, 1-13 (2018).
26. Buniello, A.*, et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-d1012 (2019).
27. Marioni, R.E.*, et al.* GWAS on family history of Alzheimer's disease. *Transl Psychiatry* **8**, 99 (2018).
28. Jansen, I.E.*, et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature genetics* **51**, 404-413 (2019).
29. Kim, S.*, et al.* Genome-wide association study of CSF biomarkers Abeta1-42, t-tau, and p-tau181p in the ADNI cohort. *Neurology* **76**, 69-79 (2011).
30. Schwartzentruber, J.*, et al.* Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nature genetics* **53**, 392-402 (2021).
31. Gouveia, C.*, et al.* Genome-wide association of polygenic risk extremes for Alzheimer's disease in the UK Biobank. *Sci Rep* **12**, 8404 (2022).
32. Davies, G.*, et al.* Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat Commun* **9**, 2098 (2018).
33. Lee, J.J.*, et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics* **50**, 1112-1121 (2018).
34. Wang, H.*, et al.* Genome-wide interaction analysis of pathological hallmarks in Alzheimer's disease. *Neurobiol Aging* **93**, 61-68 (2020).
35. Hao, S., Wang, R., Zhang, Y. & Zhan, H. Prediction of Alzheimer's Disease-Associated Genes by Integration of GWAS Summary Data and Expression Data. *Front Genet* **9**, 653 (2018).
36. Gockley, J.*, et al.* Multi-tissue neocortical transcriptome-wide association study implicates 8 genes across 6 genomic loci in Alzheimer's disease. *Genome Med* **13**, 76 (2021).
37. Gerring, Z., Lupton, M., Edey, D., Gamazon, E. & Derks, E. An analysis of genetically regulated gene expression across multiple tissues implicates novel gene candidates in Alzheimer's disease. *Alzheimer's Research & Therapy* **12**, 43 (2020).
38. Sun, Y.*, et al.* A transcriptome-wide association study of Alzheimer's disease using prediction models of relevant tissues identifies novel candidate susceptibility genes. *Genome Medicine* **13**, 141 (2021).

608 39. Davidson, J.E., *et al.* Plasma lipoprotein-associated phospholipase A2 activity in Alzheimer's
609 disease, amnestic mild cognitive impairment, and cognitively healthy elderly subjects: a
610 cross-sectional study. *Alzheimers Res Ther* **4**, 51 (2012).
611 40. Drenos, F., *et al.* Integrated associations of genotypes with multiple blood biomarkers linked
612 to coronary heart disease risk. *Hum Mol Genet* **18**, 2305-2316 (2009).
613 41. Dumont, V. & Lehtonen, S. PACSIN proteins in vivo: Roles in development and physiology.
614 *Acta Physiol (Oxf)* **234**, e13783 (2022).
615 42. Szklarczyk, D., *et al.* STRING v11: protein–protein association networks with increased
616 coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic*
617 *Acids Research* **47**, D607-D613 (2018).
618 43. Szklarczyk, D., *et al.* The STRING database in 2021: customizable protein-protein networks,
619 and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*
620 **49**, D605-d612 (2021).
621 44. Jong, M.C., Hofker, M.H. & Havekes, L.M. Role of ApoCs in Lipoprotein Metabolism.
622 *Arteriosclerosis, Thrombosis, and Vascular Biology* **19**, 472-484 (1999).
623 45. Myklebost, O. & Rogne, S. The gene for human apolipoprotein CI is located 4.3 kilobases
624 away from the apolipoprotein E gene on chromosome 19. *Hum Genet* **73**, 286-289 (1986).
625 46. Smit, M., van der Kooij-Meijs, E., Frants, R.R., Havekes, L. & Klasen, E.C. Apolipoprotein
626 gene cluster on chromosome 19. Definite localization of the APOC2 gene and the
627 polymorphic Hpa I site associated with type III hyperlipoproteinemia. *Human genetics* **78**,
628 90-93 (1988).
629 47. Saunders, A.M., *et al.* Association of apolipoprotein E allele ϵ4 with late-onset familial and
630 sporadic Alzheimer's disease. *Neurology* **43**, 1467-1467 (1993).
631 48. Panza, F., *et al.* Lipid metabolism in cognitive decline and dementia. *Brain Res Rev* **51**, 275-
632 292 (2006).
633 49. Wood, R.D., Mitchell, M., Sgouros, J. & Lindahl, T. Human DNA repair genes. *Science* **291**,
634 1284-1289 (2001).
635 50. Weeda, G., *et al.* A presumed DNA helicase encoded by ERCC-3 is involved in the human
636 repair disorders xeroderma pigmentosum and Cockayne's syndrome. *Cell* **62**, 777-791
637 (1990).
638 51. Fishel, M.L., Vasko, M.R. & Kelley, M.R. DNA repair in neurons: so if they don't divide
639 what's to repair? *Mutat Res* **614**, 24-36 (2007).
640 52. Weissman, L., *et al.* Defective DNA base excision repair in brain from individuals with
641 Alzheimer's disease and amnestic mild cognitive impairment. *Nucleic Acids Research* **35**,
642 5545-5555 (2007).
643 53. Ajuh, P., *et al.* Functional analysis of the human CDC5L complex and identification of its
644 components by mass spectrometry. *Embo j* **19**, 6569-6581 (2000).
645 54. Urano, Y., *et al.* Involvement of the mouse Prp19 gene in neuronal/astroglial cell fate
646 decisions. *J Biol Chem* **281**, 7498-7514 (2006).
647 55. Yamada, T., *et al.* The U-box-type ubiquitin ligase PRP19β regulates astrocyte
648 differentiation via ubiquitination of PTP1B. *Brain Res* **1524**, 12-25 (2013).
649 56. Lunnon, K., *et al.* Mitochondrial genes are altered in blood early in Alzheimer's disease.
650 *Neurobiol Aging* **53**, 36-47 (2017).
651 57. Reynolds, C.A., *et al.* Analysis of lipid pathway genes indicates association of sequence
652 variation near SREBF1/TOM1L2/ATPAF2 with dementia risk. *Human molecular genetics*
653 **19**, 2068-2078 (2010).

654 58. Verbanck, M., Chen, C.Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy
655 in causal relationships inferred from Mendelian randomization between complex traits and
656 diseases. *Nature genetics* **50**, 693-698 (2018).
657 59. Zhu, Z.*, et al.* Integration of summary data from GWAS and eQTL studies predicts complex
658 trait gene targets. *Nature Genetics* **48**, 481-487 (2016).
659 60. Barbeira, A.N.*, et al.* Exploring the phenotypic consequences of tissue specific gene
660 expression variation inferred from GWAS summary statistics. *Nature Communications* **9**,
661 1825 (2018).
662 61. Wingo, A.P.*, et al.* Shared proteomic effects of cerebral atherosclerosis and Alzheimer's
663 disease on the human brain. *Nature Neuroscience* **23**, 696-700 (2020).
664 62. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330-
665 d338 (2019).
666 63. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference
667 resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462 (2016).
668 64. Jassal, B.*, et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* **48**, D498-d503
669 (2020).
670 65. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506-d515 (2019).
671 66. Mungall, C.J.*, et al.* The Monarch Initiative: an integrative data and analytic platform
672 connecting phenotypes to genotypes across species. *Nucleic Acids Res* **45**, D712-d722
673 (2017).
674 67. Wu, Y.*, et al.* Integrative analysis of omics summary data reveals putative mechanisms
675 underlying complex traits. *Nature Communications* **9**, 918 (2018).
676 68. Newman, A.M.*, et al.* Robust enumeration of cell subsets from tissue expression profiles.
677 *Nature methods* **12**, 453-457 (2015).
678 69. Darmanis, S.*, et al.* A survey of human brain transcriptome diversity at the single cell level.
679 *Proc Natl Acad Sci U S A* **112**, 7285-7290 (2015).
680 70. Wingo, T.S.*, et al.* Brain microRNAs associated with late-life depressive symptoms are also
681 associated with cognitive trajectory and dementia. *NPJ Genom Med* **5**, 6 (2020).
682
683
684

Tables

Table 1. The total number of tested proteins with CV R^2 > 0.005 for the corresponding protein abundance imputation models trained using ROS/MAP reference data by TIGAR, PrediXcan, and FUSION tools. Median CV R^2 of tested proteins are presented in the third column.

Model	Number of Proteins	Median CV R^2
TIGAR	6,673	0.013
PrediXcan	1,835	0.0094
FUSION	2,173	0.0093

Table 2. Independent PWAS-O risk genes of AD dementia. Missing P-values denoted by dashes are due to either CV R^2< 0.5% for the corresponding protein abundance imputation models or no trained protein abundance imputation model by PrediXcan due to variable selection.

Gene	CHR	PWAS-O P-value	TIGAR P-value	PrediXcan P-value	FUSION P-value
CDC42[c]	1	2.47×10^{-4}	2.47×10^{-4}	-	-
ERCC3[b]	2	2.91×10^{-4}	1.80×10^{-4}	-	7.54×10^{-4}
MGAT5	2	6.38×10^{-5}	7.62×10^{-4}	2.19×10^{-5}	2.97×10^{-2}
HSPA1B[b]	6	2.64×10^{-9}	-	-	2.64×10^{-9}
TULP1[b]	6	2.88×10^{-4}	4.14×10^{-1}	-	1.44×10^{-4}
PLA2G7[c]	6	9.03×10^{-5}	9.03×10^{-5}	-	-
MAP3K7	6	5.71×10^{-5}	5.71×10^{-5}	-	-
TRIM4[b]	7	3.72×10^{-5}	3.72×10^{-5}	-	-
AGFG2[a,b,c]	7	9.86×10^{-8}	1.24×10^{-7}	-	8.20×10^{-8}
ZYX[b,c]	7	1.39×10^{-5}	7.91×10^{-6}	5.86×10^{-5}	-
C1QTNF4[b,c]	11	2.24×10^{-5}	2.24×10^{-5}	-	-
MRPL16[b,c]	11	8.66×10^{-15}	4.33×10^{-15}	1.31×10^{-2}	-
CCDC86[b]	11	2.83×10^{-9}	1.42×10^{-9}	-	1.59×10^{-1}
DIABLO	12	9.26×10^{-5}	9.26×10^{-5}	-	-
ATPAF2[b]	17	8.20×10^{-5}	8.20×10^{-5}	-	-
DCAKD[b]	17	6.19×10^{-5}	3.62×10^{-3}	-	3.12×10^{-5}
CLPTM1[a,b,c]	19	1.44×10^{-70}	1.49×10^{-21}	7.21×10^{-71}	-
SPHK2[c]	19	1.76×10^{-4}	4.87×10^{-2}	8.82×10^{-5}	-
YWHAB	20	1.71×10^{-4}	1.71×10^{-4}	-	-

a Known GWAS risk genes of AD dementia.
b Known TWAS risk genes of AD dementia, or test region overlapped with TWAS risk genes.
c Significant mediated causal genetic effects by PMR-Egger or SMR.

Figure titles and legends

Fig.1: Two-stage PWAS framework.
Stage I: Protein abundance imputation models are trained using *cis*-SNPs as predictors and proteomic abundance as the outcome, which estimates pQTL effect sizes (weights) per protein-coding gene. Stage II: Individual PWAS predicts the protein abundances using the trained imputation models and individual-level genotype data of test samples, and then tests the association between the predicted protein abundance and the trait of interest. Summary-based PWAS takes the pQTL effect sizes as variant weights to conduct an equivalent gene-based association test with GWAS Z-scores and reference LD.

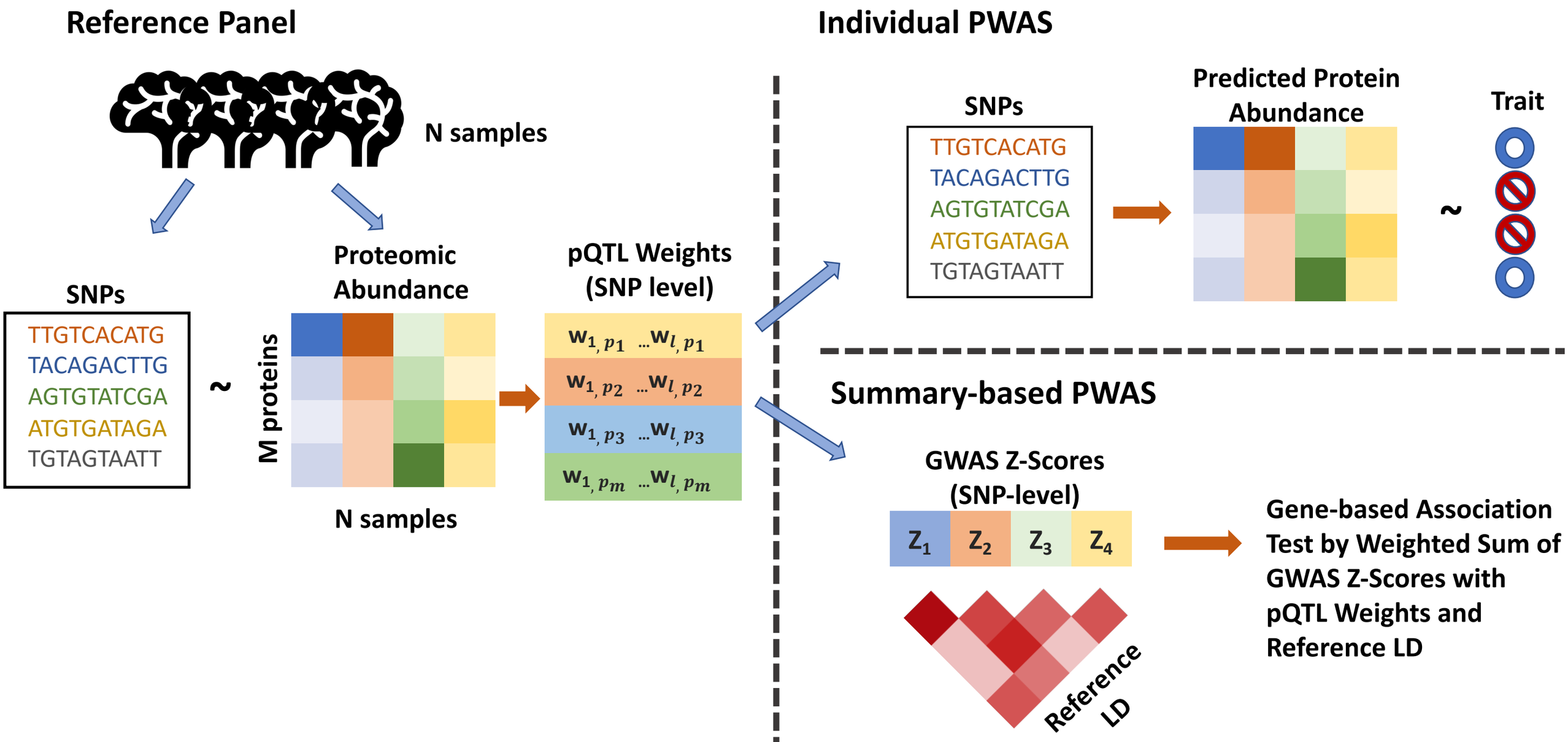
Fig. 2: PWAS-O framework.
First, train protein abundance imputation models using reference data and conduct two-stage PWAS (**Fig. 1**) using GWAS data by TIGAR, PrediXcan, and FUSION tools. Second, obtain a PWAS-O p-value per protein-coding gene by aggregating all PWAS p-values by ACAT.

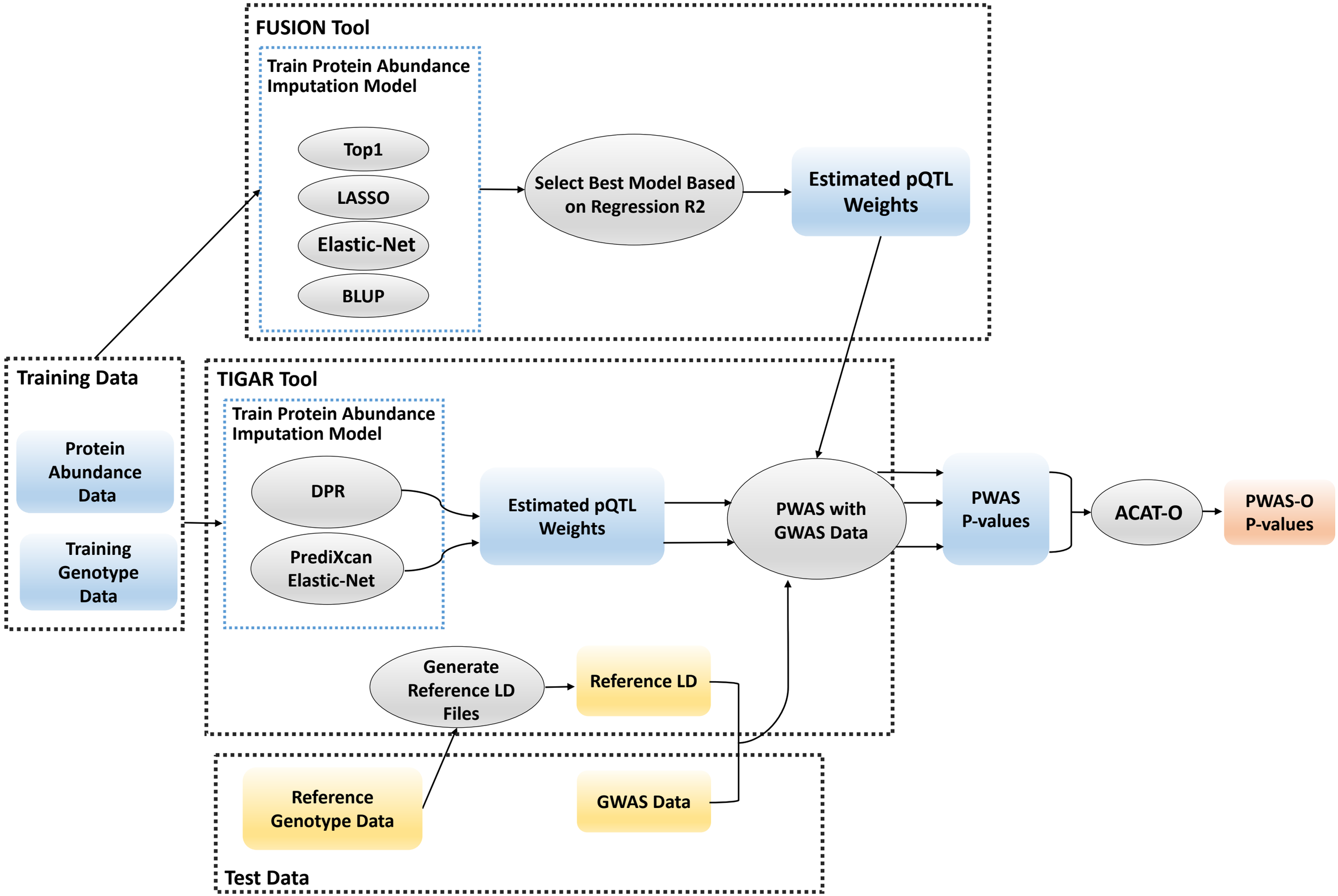
Fig. 3: Test R^2 (A) and PWAS power (B) comparison in simulation studies
Different scenarios with proportions of true causal cis-pQTL P_{causal} = (0.001,0.01) and protein expression heritability h_p^2 = (0.01, 0.05) were considered in the simulation studies. Boxplots of test R^2 in 50,000 simulations per scenario for 3 PWAS tools were presented (A). Any simulation that was unsuccessful in training protein imputation models would contribute a zero value to the test R^2 score. The median was depicted as a black bar, while the lower and upper hinges represented the 25th and 75th percentiles, respectively. Additionally, the whiskers extended from the hinges to a maximum of 1.5 times the interquartile range. Any data points beyond the whiskers were plotted individually. The power of 50,000 simulations were plotted with respect to various GWAS sample sizes, comparing PWAS-O to 3 PWAS tools (B).

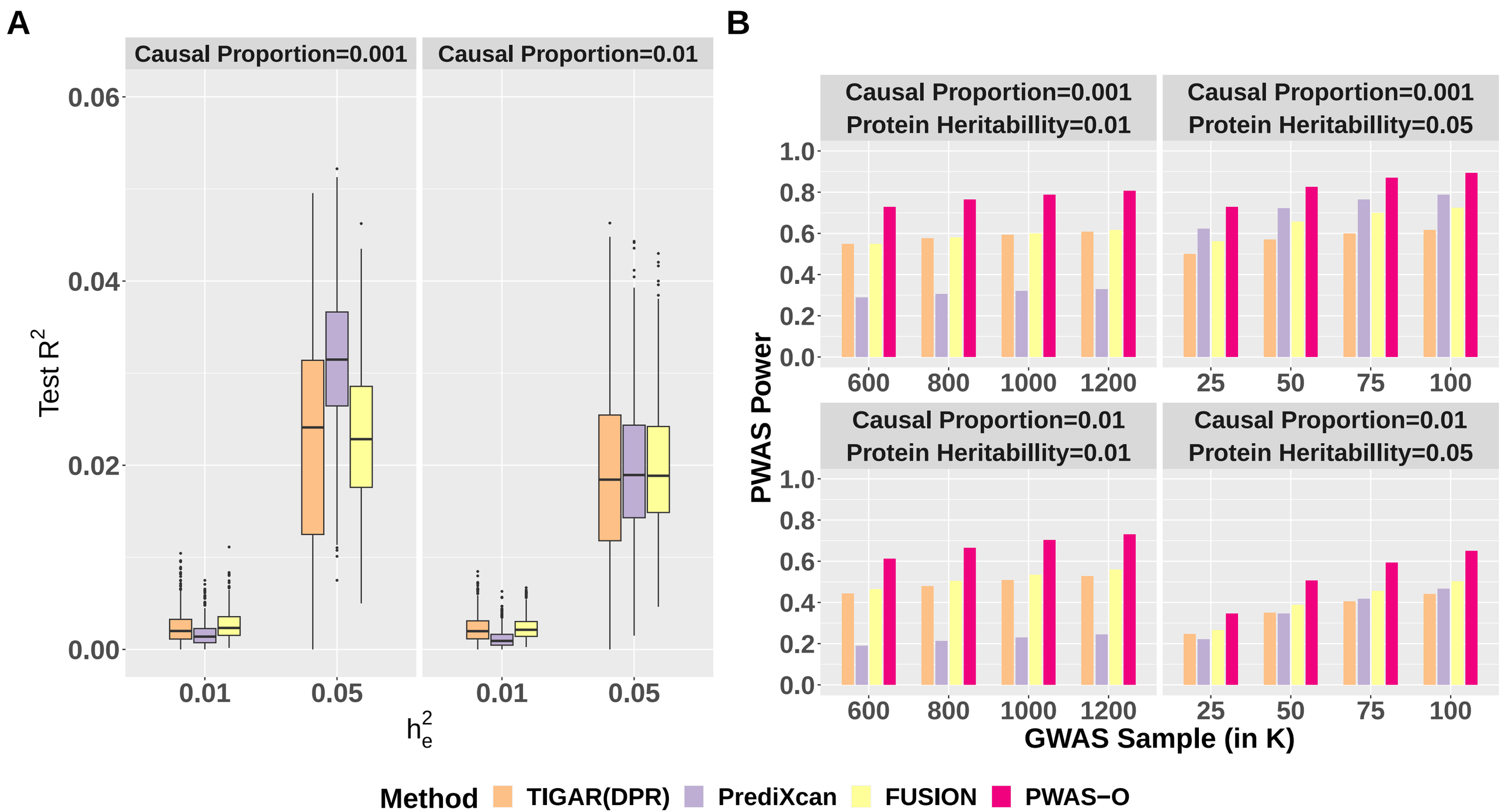
Fig. 4: Manhattan plot of PWAS-O results for AD dementia.
A total of 43 significant risk genes were identified with FDR q-value < 0.05. A total of 19 independently significant genes were labeled. The -log10(q-values) were plotted in y-axis, and -log10(0.05) was plotted as the dashed horizontal line.

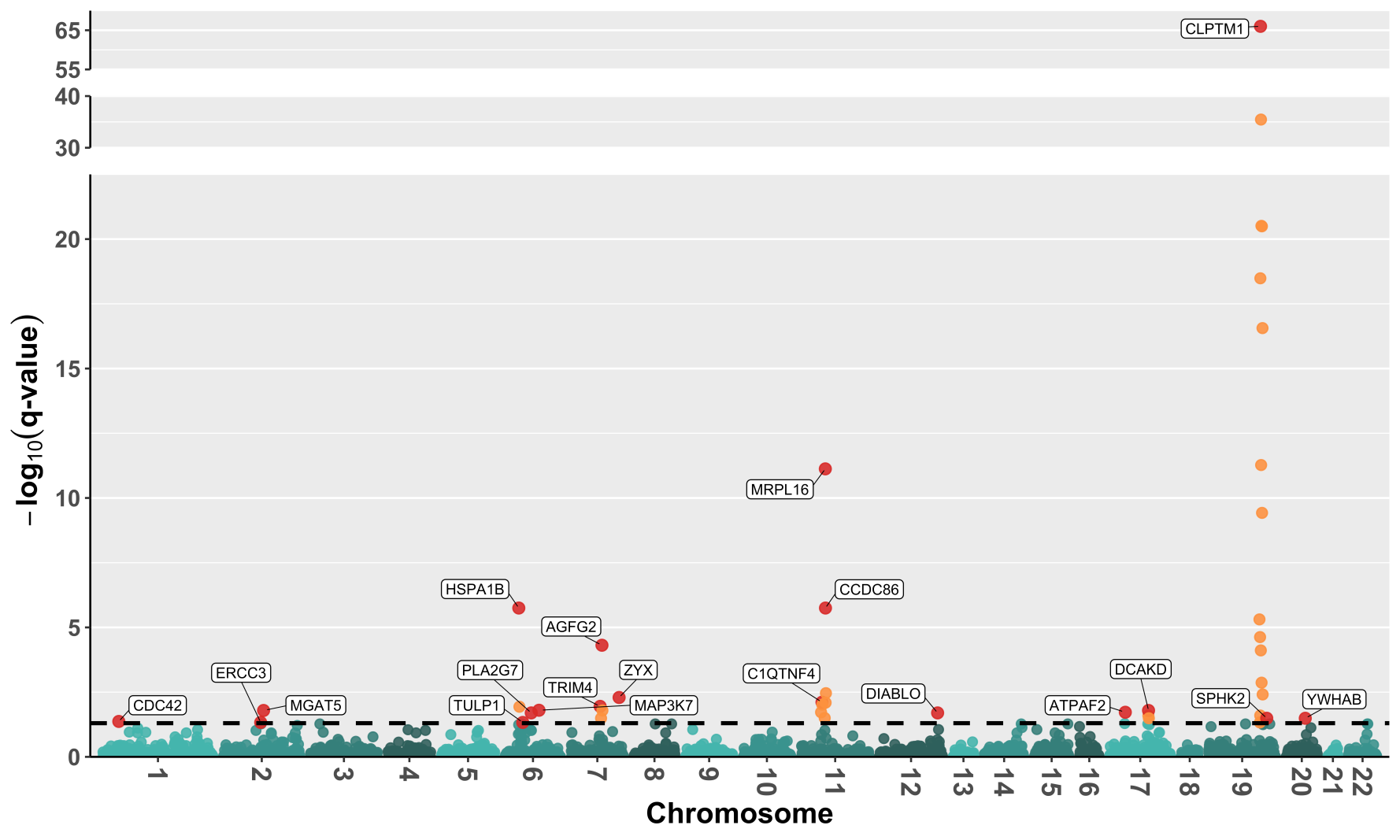
Fig. 5: pQTL weights estimated by TIGAR (DPR) and FUSION (best model of LASSO, Elastic-Net, BLUP, Top1) for example significant genes *PPM1N* (A) and *DCAKD* (B) by PWAS-O.
The pQTL weights were plotted in the y-axis for all test SNPs in the test gene region, with color coded with respect to their -log10 (GWAS p-value). SNPs with GWAS p-value < 10^{-5} were plotted in yellow color. Gene *PPM1N* (A) was found significant by both TIGAR and FUSION, as multiple test SNPs with non-zero pQTL weights (by TIGAR and FUSION) were colocalized with significant GWAS SNPs (yellow dots). Whereas, the significance of *DCAKD* (B) was mainly driven by FUSION, for having fewer SNPs with significant GWAS p-values that are also colocalized with a relatively large magnitude of pQTL weights.

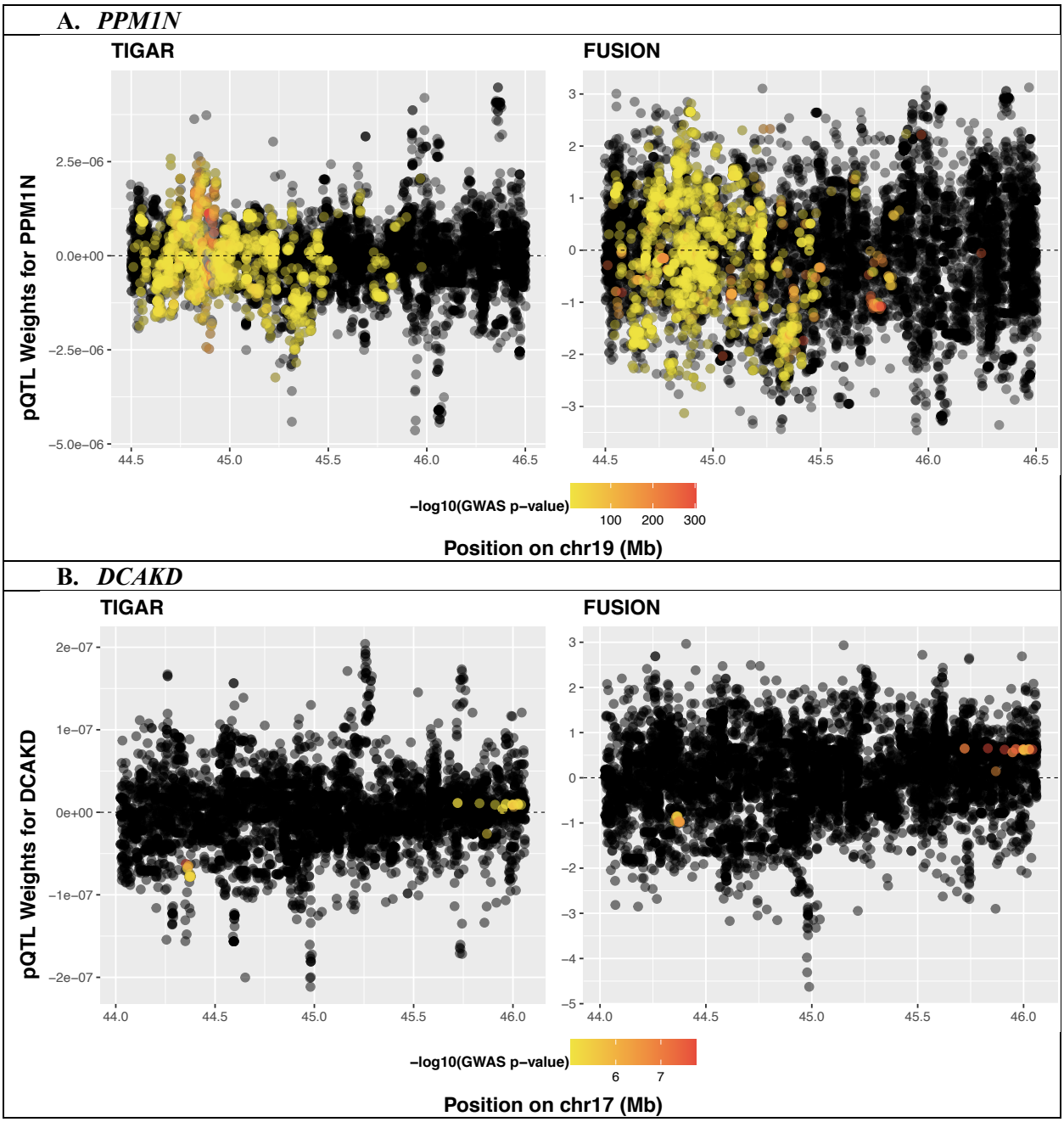
Fig. 6: PPI networks and enriched pathways among 43 PWAS risk genes of AD dementia by STRING.
Edges represent physical PPI, with different colors representing different sources of connection evidence. Node colors represent different enriched GO terms or KEGG pathways with FDR < 0.05.

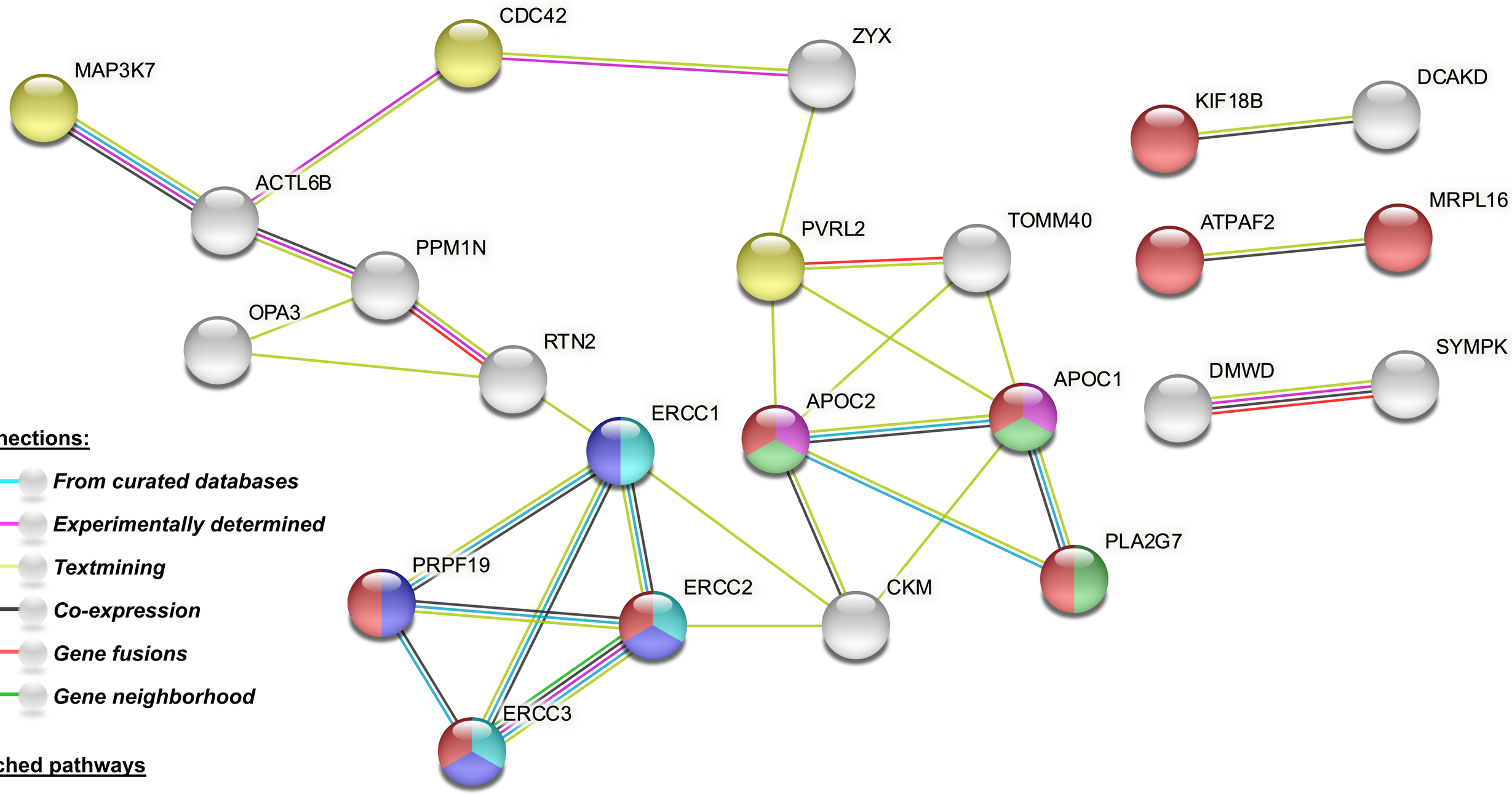

Reference Panel
N samples
SNPs
TTGTCACATG
TACAGACTTG
AGTGTATCGA
ATGTGATAGA
TGTAGTAATT
~
Proteomic Abundance
M proteins
N samples
pQTL Weights (SNP level)
w_{1,p_1} ... w_{l,p_1}
w_{1,p_2} ... w_{l,p_2}
w_{1,p_3} ... w_{l,p_3}
w_{1,p_m} ... w_{l,p_m}
Individual PWAS
SNPs
TTGTCACATG
TACAGACTTG
AGTGTATCGA
ATGTGATAGA
TGTAGTAATT
Predicted Protein Abundance
~
Trait
Summary-based PWAS
GWAS Z-Scores (SNP-level)
Z_1 Z_2 Z_3 Z_4
Reference LD
Gene-based Association Test by Weighted Sum of GWAS Z-Scores with pQTL Weights and Reference LD


FUSION Tool
Train Protein Abundance Imputation Model
Top1
LASSO
Elastic-Net
BLUP
Select Best Model Based on Regression R2
Estimated pQTL Weights
Training Data
Protein Abundance Data
Training Genotype Data
TIGAR Tool
Train Protein Abundance Imputation Model
DPR
PrediXcan Elastic-Net
Estimated pQTL Weights
PWAS with GWAS Data
PWAS P-values
ACAT-O
PWAS-O P-values
Generate Reference LD Files
Reference LD
Reference Genotype Data
GWAS Data
Test Data


A
Causal Proportion=0.001
Causal Proportion=0.01
Test R^2
0.06
0.04
0.02
0.00
0.01
0.05
0.01
0.05
h_e^2
B
Causal Proportion=0.001
Protein Heritabillity=0.01
Causal Proportion=0.001
Protein Heritabillity=0.05
Causal Proportion=0.01
Protein Heritabillity=0.01
Causal Proportion=0.01
Protein Heritabillity=0.05
PWAS Power
1.0
0.8
0.6
0.4
0.2
0.0
600
800
1000
1200
25
50
75
100
GWAS Sample (in K)
Method
TIGAR(DPR)
PrediXcan
FUSION
PWAS−O


65
55
40
30
20
15
10
5
0
−log10(q-value)
CLPTM1
MRPL16
CCDC86
HSPA1B
AGFG2
ZYX
ERCC3
PLA2G7
TRIM4
C1QTNF4
DCAKD
CDC42
MGAT5
TULP1
MAP3K7
DIABLO
ATPAF2
SPHK2
YWHAB
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome


A. PPM1N
TIGAR
FUSION
pQTL Weights for PPM1N
2.5e−06
0.0e+00
−2.5e−06
−5.0e−06
44.5
45.0
45.5
46.0
46.5
3
2
1
0
−1
−2
−3
−log10(GWAS p−value)
100
200
300
Position on chr19 (Mb)
B. DCAKD
TIGAR
FUSION
pQTL Weights for DCAKD
2e−07
1e−07
0e+00
−1e−07
−2e−07
44.0
44.5
45.0
45.5
46.0
3
2
1
0
−1
−2
−3
−4
−5
−log10(GWAS p−value)
6
7
Position on chr17 (Mb)



Connections:

- *From curated databases*
- *Experimentally determined*
- *Textmining*
- *Co-expression*
- *Gene fusions*
- *Gene neighborhood*

Enriched pathways

- *UV protection (GO:0009650, FDR = 0.0101)*
- *Adherens junction (hsa04520, FDR = 0.0124)*
- *Transcription-coupled nucleotide-excision repair (GO:0006283, FDR = 0.0131)*
- *Plasma lipoprotein particle remodeling (GO:0034369, FDR = 0.0138)*
- *Negative regulation of very-low-density lipoprotein particle clearance (GO:0043933, FDR = 0.0181)*
- *Protein-containing complex subunit organization (GO:0043933, FDR = 0.0255)*