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], 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 summary data is a powerful tool to identify risk genes for complex diseases. These genes can inform disease mechanisms with genetic effects mediated through protein abundance. We propose a novel omnibus method to further 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=~70K) to conduct PWAS. Last, we employed the Aggregated Cauchy Association Test to obtain omnibus PWAS (PWAS-O) p-values from these three tools. PWAS-O identified 43 risk genes of AD dementia that were interconnected into a protein-protein interaction network including *TOMM40*, *APOC1*, and *APOC2*. It detected 5 novel genes missed by GWAS and TWAS. PWAS-O can be applied to characterize underlying genetic mechanisms for diverse diseases.

29 **Main**

30 Recent large-scale genome-wide association studies (GWAS) of Alzheimer's disease (AD)
31 dementia have identified 38 risk loci[1], but little is known about their underlying mechanisms that lead to
32 clinical Alzheimer's disease. Integrating transcriptomics with GWAS summary data has been shown an
33 effective approach for identifying potential causal risk genes of complex human diseases (e.g., AD
34 dementia). This approach can also inform on potential underlying disease mechanisms as the identified
35 genetic effects may be potentially mediated via gene expression[2-5]. Since many biological processes can
36 affect protein abundance, gene expression levels are not the optimal proxy for elucidating the role of
37 proteins in disease mechanisms. A protein-based approach that considers the effects of genetic variations
38 on alterations in protein abundance has the potential to identify novel genetic risk loci linked to disease
39 via proteins that may not be identified by GWAS and TWAS[6-9].

40 Proteome-wide association study (PWAS) that integrates proteomics data with GWAS summary
41 data has also shown promise for studying complex traits and diseases[10-12]. Since gene expression and
42 protein abundance are both molecular quantitative traits that might mediate genetic effects, existing
43 TWAS tools can be applied to train protein abundance imputation models using genetic variants as
44 predictors from a reference panel. Genetic effect sizes in the imputation models could be viewed as effect
45 sizes of "protein quantitative trait locus (pQTL)" in a broad sense (as these "pQTL" are not necessarily
46 statistically significant), which would be taken as variant weights for gene-based association study (i.e.,
47 PWAS) with GWAS data.

48 Existing TWAS tools use various statistical regression models to train imputation models of
49 quantitative molecular traits, and each tool has its own advantage for modeling a specific type of genetic
50 architecture underlying the molecular trait of interest. For example, PrediXcan[13] uses a penalized linear
51 regression model with Elastic-Net penalty[14] which has the advantage of modeling sparse genetic
52 architecture. TIGAR[7] implements the nonparametric Bayesian latent Dirichlet process regression (DPR)
53 which has been shown effective to model the infinitesimal genetic architecture[15]. FUSION[4] uses
54 penalized linear regression model with Elastic-Net penalty and LASSO penalty[16], regular linear

55 regression model with best unbiased linear predictor (BLUP), single variant model with Top eQTL or
56 pQTL (Top1), and Bayesian spare linear mixed model (BSLMM)[17], and then selects the best model for
57 PWAS based on cross-validation (CV) R^2 from model training.

58 Even though the FUSION tool employs multiple regression models, it only uses the model with
59 top performance based on reference data. Rather than considering a single statistical method to train
60 molecular QTL weights, there is strong evidence that combining association test p-values based on
61 multiple statistical methods using the Aggregated Cauchy Association Test[18] (ACAT) can increase power,
62 as demonstrated in recent TWAS[19] studies. Therefore, we applied ACAT to aggregate PWAS p-values
63 obtained by TIGAR, PrediXcan, and FUSION and subsequently derive an omnibus PWAS (PWAS-O) p-
64 value. PWAS-O is expected to improve PWAS power compared to PWAS based on a single imputation
65 model of protein abundances, since PWAS-O better accounts for the unknown genetic architecture of
66 protein abundances for a given gene.

67 To test our approach, we applied PWAS-O to study AD dementia by leveraging reference human
68 proteome-wide data of dorsolateral prefrontal cortex (DLPFC) of 355 cognitively well-characterized 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 into a protein-protein interaction (PPI) network
73 involving the well-known risk gene *APOE* of AD dementia. We validated the potential causality of 27 out
74 of 43 (63%) PWAS risk genes by probabilistic Mendelian randomization (PMR-Egger)[20] and summary
75 data-based Mendelian randomization (SMR)[21, 22] tool.

76

77 **Results**

78

79 **<u>Omnibus PWAS (PWAS-O)</u>**

80

81 Similar to the standard two-stage PWAS/TWAS method (**Fig. 1** and **Fig. 2**)[23], protein abundance
82 imputation models are first trained in Stage I to obtain effect size estimates of genetic variants/predictors

83 (i.e., pQTL weights), by TIGAR[23] (nonparametric Bayesian DPR), PrediXcan[13] (Elastic-Net), and
84 FUSION[4] (Elastic-Net, LASSO, BLUP, Top1). Genetic variants within ±1MB around the transcription
85 starting and termination sites of the corresponding coding gene were considered as predictors in the
86 imputation models. In Stage II, these trained pQTL weights (coefficients of these genetic variant
87 predictors) are used along with GWAS data to conduct gene-based association tests. For each protein-
88 coding gene, PWAS p-values will be respectively obtained by using the pQTL weights estimated by
89 TIGAR, PrediXcan, and FUSION, which will then be combined by ACAT to generate the corresponding
90 PWAS-O p-value. We present the PWAS-O framework in **Fig. 2**. Both individual-level and summary-
91 level GWAS data can be used in Stage II. A reference panel of the same ancestry as the GWAS data can
92 be used to provide LD information for gene-based association tests in Stage II. The TIGAR[23] tool
93 provides functions to calculate LD covariance matrices and conduct gene-based association tests.
94

95 **ROS/MAP Proteomics and WGS data**
96
97 The human brain proteomic profiles were generated from the DLPFC of postmortem brain
98 samples donated by 400 participants of European ancestry of the Religious Orders Study/Memory and
99 Aging Project (ROS/MAP)[10, 24]. The human proteome abundance data underwent quality control (**Fig. S1**)
100 that regressed out the effects of technical factors and confounding covariates including age at death, sex,
101 postmortem interval, and study (ROS or MAP) (see details in the Methods). As a result, we obtained the
102 proteome abundance data of 8,430 proteins. After quality control of these proteome data, 355 individuals
103 also had WGS data profiled[25]. The proteome abundance and WGS data of these 355 individuals were used
104 as a reference panel for our PWAS-O study of AD dementia.
105
106 **Protein abundance imputation models**
107 At training Stage I, five-fold cross-validation was conducted to evaluate the training
108 performance per statistical model. The protein abundance imputation models with 5-fold CV R^2 >0.5%
109 were retained and their estimated pQTL weights would be used to conduct PWAS. As suggested by the

110 TIGAR tool[23], the threshold of 5-fold CV R^2 >0.5% was arbitrarily chosen to allow a greater number of
111 genes to be tested by the follow-up gene-based association study (i.e., PWAS). The threshold of CV R^2
112 would not affect the statistical validity of the test statistic used in the follow-up PWAS.

113 As shown in **Table 1**, 6673 protein abundance imputation models were retained by TIGAR
114 with median CV R^2 = 0.013; 1835 protein abundance imputation models were retained by PrediXcan
115 with median CV R^2 = 0.0094; and 2173 protein abundance imputation models were retained by FUSION
116 (the best model selected out of LASSO, Elastic-Net, BLUP, Top1 models with the highest CV R^2) with
117 median CV R^2 = 0.0093. Histograms of the CV R^2 obtained by these three tools were presented in **Fig.**
118 **S2**.

119
120 **PWAS-O results of AD dementia**
121
122 As described in the above paragraph about the PWAS-O framework, we first conducted PWAS
123 by employing TIGAR[23], PrediXcan[13], and FUSION[4] tools, using the ROS/MAP reference proteomics data
124 of DLPFC and the most recent GWAS[1] summary data of AD dementia (n=762,917). That is, the
125 estimated pQTL effect sizes in these trained protein abundance imputation models were taken as variant
126 weights to conduct gene-based association tests using the GWAS summary data (**Fig. S3**). Second, we
127 implemented ACAT[18] to obtain PWAS-O p-values from the PWAS p-values by these three tools. Our
128 PWAS-O approach identified 43 significant genes with false discovery rate (FDR) (i.e., q-value) <0.05,
129 which could affect the risk of AD dementia through their genetic effects on the corresponding protein
130 abundances (**Fig. 3, Fig. S4**). We curated 19 independent genes from these 43 significant genes (**Table 2,**
131 **Table S1**), by considering only the most significant genes among all significant genes with shared test
132 regions (±1MB around the test transcription starting and termination sites of the protein-coding gene).

133 We found 10 out of these 43 PWAS risk genes were detected by previous GWAS studies from
134 GWAS Catalogue[26], including *AGFG2*[27], *CLPTM1*[28], *PVR*[28], *PVRL2*[28], *APOC1*[29], *APOC2*[30], *CKM*[28],
135 *ERCC1*[31], *OPA3*[31], and *SYMPK*[31]. PWAS risk genes *CDC42*[32], *MAP3K7*[33], and *FNBP4*[32] were found

136 associated with cognitive ability in previous GWAS studies. PWAS risk genes *CCDC86*[34], *KIF18B*[34],
137 *PPM1N*[34], and *RTN2*[34] were found associated with neurofibrillary tangles (a key clinical indicator of AD)
138 measurement by GWAS studies. Additionally, 13 out of 43 significant AD PWAS genes (*ZYX*[35],
139 *C1QTNF4*[36], *FNBP4*[35], *PVR*[8], *PVRL2*[36], *TOMM40*[8], *APOC1*[9], *APOC2*[37], *CLPTM1*[36], *ERCC1*[38], *OPA3*[38],
140 *EML2*[9] and *DMWD*[35]) were significant risk genes of AD by previous TWAS.

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

147

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

149 To investigate the genetic architecture underlying protein abundances, we plotted the pQTL
150 weights estimated by TIGAR (DPR), PrediXcan (Elastic-Net), and FUSION (best model out of LASSO,
151 Elastic-Net, BLUP, Top1) for significant PWAS risk genes *PPM1N* and *DCAKD* in **Fig. 4**. Generally,
152 TIGAR and FUSION had non-zero pQTL weights estimated for almost all test SNPs within the test
153 region, for using a model assuming infinitesimal genetic architecture[15]. Whereas, PrediXcan had much
154 fewer non-zero pQTL weights estimated for assuming a sparse genetic architecture and enforcing variable
155 selection. For example, *PPM1N* was a significant risk gene identified by both TIGAR and FUSION
156 (**Table 2**), which has hundreds of test SNPs with non-zero pQTL weights (by both TIGAR and FUSION)
157 that were colocalized with significant GWAS signals (colored dots in **Fig. 4A**). Gene *DCAKD* was only
158 found significant by FUSION (p-value = 3.12×10^{-5}) but not by TIGAR (p-value = 3.62×10^{-3}), for
159 having fewer test SNPs of relatively large magnitude of pQTL weights that were colocalized with
160 significant GWAS signals (**Table2, Fig. 4B**).

161 Similarly, comparing pQTL weights obtained by TIGAR to PrediXcan for significant PWAS
162 genes *MBLAC1* and *MRPL16* (**Fig. S5**), we observed that less than 10 test SNPs had non-zero pQTL
163 weights by PrediXcan, versus thousands of test SNPs with non-zero pQTL weights by TIGAR. Because
164 PrediXcan estimated zero pQTL weights for most SNPs with significant GWAS p-values in the test
165 region of gene *MRPL16* (**Table 2**, **Fig. S5A**), *MRPL16* was only found significant by TIGAR (p-value =
166 4.33×10^{-15}). Whereas, when both TIGAR and PrediXcan estimated non-zero pQTL weights for SNPs
167 with significant GWAS p-values in the test region of gene *MBLAC1* (Table S1, Fig. S4B), it was
168 identified significant by both TIGAR (p-value = 7.31×10^{-5}) and PrediXcan (p-value = 3.49×10^{-4}).
169 These scatterplots of pQTL weights of PWAS risk genes showed that significant PWAS genes
170 were mainly driven by test SNPs that had colocalized pQTL weights and significant GWAS signals.
171 Because the underlying genetic architectures of protein abundances are unknown in practice, no single
172 statistical model would lead to superior PWAS power across genome-wide protein-coding genes. Our
173 proposed PWAS-O approach is expected to gain power in practice for considering multiple statistical
174 models per protein-coding gene.
175
176 **PPI network among PWAS risk genes of AD dementia**
177
178 Next, we used STRING[42, 43] to investigate the PPI network among these 43 PWAS risk genes of
179 AD dementia, by using public data sources of protein interactions. Based on evidences from curated
180 databases, experiments, text mining, co-expression, gene fusions, and protein homology[42], STRING
181 clustered 23 out of these 43 PWAS risk genes into 4 communities (**Fig. 5**). Different data sources of the
182 protein-protein connection were depicted by differently colored connection edges. It is interesting to see
183 that the largest community of these PWAS risk genes is a network involving genes *TOMM40*, *APOC1*,
184 and *APOC2* that are located nearby the well-known risk gene *APOE* of AD dementia, and are known to
185 be associated with both AD dementia and low-density lipoprotein levels[44].

186 STRING also provided gene enrichment analysis results with respect to Gene Ontologies (GO)
187 and KEGG. We observed that these 43 PWAS risk genes were enriched for the GO terms of UV
188 protection (*ERCC1, ERCC2, ERCC3*), transcription-coupled nucleotide-excision repair (*ERCC1, ERCC2,*
189 *ERCC3, PRPF19*), plasma lipoprotein particle remodeling (*APOC1, APOC2, PLA2G7*), negative
190 regulation of very-low-density lipoprotein particle clearance (*APOC1, APOC2*), protein-containing
191 complex subunit (*MRPL16, ATPAF2, KIF18B, ERCC2, ERCC3, PRPF19, APOC1, APOC2, PLA2G7*),
192 and the KEGG pathway of adherens junction (*MAP3K7*, *CDC42*, *PVRL2*).
193 Interestingly, novel PWAS risk genes *ERCC3*, *ERCC2* and *PRPF19* are closely interconnected
194 with a known GWAS/TWAS risk gene *ERCC1* in the PPI network (based on evidence from curated
195 databases, co-expression and textmining) and enriched in nucleotide-excision repair pathway. Novel
196 PWAS risk gene *PLA2G7* connects with *APOC1* and *APOC2* in the PPI network and involves in the
197 plasma lipoprotein particle remodeling pathway.
198 These PPI network analysis results showed that the list of identified risk genes by PWAS-O were
199 connected with known risk genes of AD dementia and enriched in AD related pathways.
200
201 **Examine causal and pleiotropy effects of PWAS risk genes**
202 We applied both probabilistic Mendelian randomization (PMR-Egger)[20] and summary data-based
203 Mendelian randomization (SMR)[45] methods to test if the causal genetic effects on AD dementia were
204 mediated through protein abundance in the possible presence of horizontal pleiotropy. Horizontal
205 pleiotropy means that genetic effects on AD dementia and on protein abundances were originated from
206 the same SNPs but through different pathways (i.e., with no mediation effects). PMR-Egger suggested
207 that 25 out of 43 PWAS risk genes have significant causal effects (FDR<0.05) on AD dementia after
208 controlling for potential horizontal pleiotropy effects, including 6 genes (*PVRL2*, *TOMM40, APOC1,*
209 *RTN2, OPA3, EML2*) nearby the *APOE* region that also have significant horizontal pleiotropy effects with
210 FDR<0.05. Similarly, the method of SMR with its accompanying heterogeneity in dependent instruments
211 (HEIDI) test found that 22 out of 43 PWAS risk genes have potential causal effects on AD dementia with

212 FDR<0.1, of which 20 genes were overlapped with 25 causal genes detected by PMR tool. These results
213 validated potential causal genetic effects that were mediated through protein abundances for 27 (63%)
214 PWAS risk genes identified by PWAS-O.
215
216 **Compare with TWAS results of AD dementia**
217 To determine whether the 43 PWAS risk genes had similar evidence at the transcript level, we
218 also conducted TWAS of AD dementia based on the ROS/MAP reference data using TIGAR and
219 PredXican. From TWAS p-values by these two tools, we obtained omnibus TWAS (TWAS-O) p-values
220 by ACAT. We found that 34 out of 43 PWAS risk genes (79%) were either detected by TWAS with
221 FDR<0.05, or located within 1MB of TWAS risk genes (**Table S2**). These results showed that majority
222 PWAS risk genes had genetic effects on phenotype that were mediated through both transcriptome and
223 proteome.
224
225 **Discussion**
226 In this work, we propose an omnibus PWAS approach (PWAS-O) that improves power for better
227 modeling of the unknown genetic architecture of proteome abundances. By aggregating PWAS p-values
228 obtained from several available tools — TIGAR, PrediXcan, and FUSION, our PWAS-O approach
229 identified 43 risk genes for AD dementia, by using the ROS/MAP reference proteomics data and the most
230 recent GWAS summary data of AD dementia. These identified PWAS risk genes may be linked to AD
231 dementia via protein abundances. From STRING analysis, we showed that out of these 43 PWAS
232 significant risk genes, 9 genes were enriched in the pathway of protein-containing complex subunit; 4
233 genes were enriched in the gene ontology pathway of transcription-coupled nucleotide-excision repair; 3
234 genes were enriched in the pathway of plasma lipoprotein particle remodeling. Compared to TWAS
235 results using the same GWAS summary data, 34 out of 43 significant PWAS risk genes were either
236 TWAS risk genes or within 1MB region of TWAS risk genes. These results showed that PWAS risk
237 genes identified by our PWAS-O approach are part of key biological pathways involving brain functions.

238 The genes and proteins identified may be high-value therapeutic targets for further drug discovery as well
239 as for mechanistic studies that can advance our understanding of the biology driving AD dementia in
240 aging brains.

241 Notable biological roles for these identified PWAS risk genes include lipoprotein metabolism,
242 inflammatory reactions, DNA repair, pre-mRNA splicing, and mitochondrial function. For example,
243 *APOC1* and *APOC2*, located in a gene cluster on chromosome 19 including the well-known AD risk gene
244 *APOE*[46-48], play important roles in the development of lipoprotein metabolism[44]. Especially, *APOE* is an
245 intensively studied risk gene of AD and plays an essential role in the central nervous system[49, 50].
246 Connected with *APOC1* and *APOC2,* as shown in the PPI of our identified PWAS risk genes, *PLA2G7* is
247 a novel AD risk gene identified by PWAS-O and associated with inflammatory oxidative activities, and
248 has been implicated in cognitive impairment and AD[39]. Additionally, *ERCC1, ERCC2,* and *ERCC3* are
249 involved in nucleotide excision repair pathway contributing to genome stability[51, 52], which plays a critical
250 role during the development of the central nervous system[53] and whose impairment is associated with
251 AD[54]. Gene *PRPF19* that is interconnected with *ERCC1, ERCC2*, and *ERCC3* encodes a pre-mRNA
252 processing factor involved in regulating pre-mRNA splicing[55]. Prior studies indicated that *PRPF19* had a
253 neuronal function and was implicated in neurological diseases[56, 57]; lower expression level of gene
254 *MRPL16* encoding a mitochondrial ribosome protein was observed in AD blood cells which may reflect a
255 general response to inflammatory environment in AD[58]; gene *ATPAF2* connected with *MRPL*16 encodes
256 a protein involved in mitochondrial function and lipid pathway, which has been implicated in dementia
257 pathogenesis[59].

258 PWAS-O still has its limitations. First, the CV R^2 is generally low for majority protein
259 abundance prediction models. This might be due to small sample sizes in the proteomics reference data,
260 or low heritability of protein abundances due to cis-SNPs. Only *cis*-SNPs within the $\pm 1MB$ region
261 around the test gene transcription starting and termination sites were considered as predictors for protein
262 abundances by PWAS-O. Trans-SNPs could also be considered as additional predictors in the protein
263 abundance prediction models, which might increase the CV R^2 of protein abundance imputation

264 models[8]. Second, the standard two-stage PWAS method cannot distinguish whether causal SNPs for the
265 phenotype of interest are only colocalized with pQTLs but have no mediation effect through the protein
266 abundances (i.e., horizontal pleiotropy) or indeed have genetic effects mediated by protein abundances
267 (i.e., vertical pleiotropy). Horizontal pleiotropy happens when genetic variants impact the outcome
268 through pathways different from or additional to the proteome abundances of the test gene[60]. Such an
269 issue can be alleviated by testing the vertical pleiotropy of PWAS risk genes by PMR-Egger tool[20], which
270 controls for possible horizontal pleiotropy.

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

277

278

279 **Methods**

280

281 **ROS/MAP Proteomics and WGS Data**

282 Human brain proteomic profiles were generated from the dorsolateral prefrontal cortex (DLPFC)
283 of postmortem brain samples donated by 400 participants of European ancestry of the Religious Orders
284 Study or Rush Memory and Aging Project (ROS/MAP)[10, 24] . Participants in ROS and MAP enroll without
285 known dementia and agree to annual clinical evaluation and organ donation at death. All participants sign
286 an informed consent, Anatomic Gift Act and Repository Consent. Both studies were reviewed by an
287 Institutional Review Board of Rush University Medical Center. Age was calculated from birth date and
288 date of death. Sex was self-report. Protein measurements from the dorsolateral prefrontal cortex (DLPFC)
289 were quantified by tandem mass spectrometry.

290 As described in previous publication[61], protein abundances were normalized by scaling each
291 protein abundance with respect to a sample-specific total protein abundance, and proteins with missing
292 values in more than 50% of the participants were excluded from the analyses. Protein abundance ratios
293 related to baseline (sample-specific total protein abundance) were $\log_2$ transformed, with missing values
294 imputed by the corresponding mean of 400 samples. Confounding effects of clinical characteristics
295 including sex, age at death, postmortem interval, and study type (ROS or MAP), as well as technical
296 factors including sequencing batch and MS2 versus MS3 mass spectrometry reporter quantification mode,
297 were regressed out from protein abundances. Outlier samples were removed using iterative principal
298 component analysis (PCA) as shown in Fig. S1. That is, samples with top 2 PCs greater than the
299 corresponding four times of the standard deviations from the mean of each PC[61].
300 The genotype data of ROS/MAP samples were profiled by whole genome sequencing (WGS)[25].
301 SNPs with Hardy-Weinberg equilibrium p-values $<10^{-5}$ and minor allele frequency >1% were used as
302 genetic predictors for training the protein abundance imputation models. After quality control of both
303 proteomic and genetic data, a total of 355 samples with both proteomic and WGS data were used as the
304 reference data for PWAS-O in this study.
305
306 **Summary-level GWAS data of AD dementia**
307 The summary-level GWAS data of AD dementia (i.e., single variant Z-score test statistics
308 obtained by meta-analysis) were generated by the latest GWAS by Douglas et al[1], including 762,917
309 individuals from 12 cohorts that did not include 23&Me. About 11.3% of participants had clinically
310 diagnosed AD dementia.
311
312 **Two-stage PWAS**
313
314 As shown in **Fig. 1**, the two-stage PWAS first trains protein abundance imputation models (stage
315 I) by taking *cis*-genotype data, ±1MB around the corresponding gene transcription starting and

316 termination sites, as predictors (X), protein abundance level ($\mathbf{E}_p$) as the outcome, as shown in the
317 following linear regression model:

318 $$\mathbf{E}_p = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon};\ \ \boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 \mathbf{I}).$$

319 Protein abundance level ($\mathbf{E}_p$) were residuals after regressing out individual characteristics and technical
320 factors from the profiled protein abundances (log2 transformed ratios as described in the above section).
321 The effect size vector $\mathbf{w}$ could be estimated by different statistical models as employed by TIGAR[23],
322 PrediXcan[13], and FUSION[4], which could be viewed as a broad sense of pQTL effect sizes (also referred
323 to as pQTL weights). Specifically, PrediXcan employs a penalized linear regression model with Elastic-
324 Net penalty[13]; FUSION[4] uses penalized linear regression model with Elastic-Net penalty and LASSO
325 penalty[16], regular linear regression model with best unbiased linear predictor (BLUP), single variant
326 model with Top pQTL (Top1), and Bayesian spare linear mixed model (BSLMM)[17], and then selects the
327 best model according to CV R^2; TIGAR[7] implements the nonparametric Bayesian latent DPR model.

328 In stage II, with individual-level GWAS data, the protein abundances will first be imputed by
329 using the trained imputation models and individual-level genotype data of test samples. Then PWAS will
330 test the association between the imputed protein abundances and the phenotype of interest. With summary
331 GWAS data, PWAS would take the pQTL weights estimated in stage I as variant weights to conduct
332 gene-based association test with GWAS Z-scores and reference LD. The PWAS test is equivalent to a
333 burden test with the following test statistic:

334 $$\tilde{Z}_{p,S\text{-}PrediXcan} = \frac{\sum_{l=1}^{m}(\widehat{w_l}\widehat{\sigma_l}Z_l)}{\sqrt{\widehat{\mathbf{w}}'\mathbf{V}\widehat{\mathbf{w}}}}, \quad \widehat{\sigma_l^2} = \mathrm{Var}(\mathbf{X}_{\mathbf{0},\mathbf{l}}), \quad \mathbf{V} = \mathrm{Cov}(\mathbf{G}_0).$$

335

336 Here, $\widehat{w_l}$ denotes the pQTL effect size estimates from protein imputation models; Z_l denotes the Z-score
337 statistics of single variant l by GWAS test; and $\mathbf{V}$ denotes the LD covariance matrix from a reference
338 panel. The genotype variance $\widehat{\sigma_l}$ of test SNPs can also be estimated from a reference panel. Here the
339 reference panel is required to be of the same ancestry as the GWAS summary data, and could be the same
340 reference genotype data as used for training pQTL weights.

341 **PWAS-O Method**

342 As shown in **Fig. 2**, with PWAS p-values obtained by two-stage PWAS tools including TIGAR[23],
343 PrediXcan[13], and FUSION[4], PWAS-O uses the ACAT-O test[18] to aggregate these PWAS p-values to
344 generate an omnibus test p-value per protein/gene. The ACAT-O test statistic is given by

345 $$T = \frac{1}{K}\sum_{k=1}^{K} \tan\{(0.5 - p_k)\}\pi\} \ , \qquad \pi \approx 3.14;$$

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

350 $$pvalue_T \approx \frac{1}{2} - \frac{\{\arctan(T)\}}{\pi}.$$

351

352 **PWAS of AD dementia**

353 In this study, we used the ROS/MAP proteomics and WGS data of 355 samples as reference data
354 to train protein abundance imputation models. For each protein, we first used TIGAR tool to train
355 nonparametric Bayesian DPR model and Elastic-Net regression model (as used by PrediXcan), and then
356 used FUSION[4] tool to train LASSO, Elastic-Net, BLUP, and Top1 models and then selected the best
357 model according to the CV R^2. We did not train a BSLMM model due to its high computation cost for
358 WGS data, and we expect that the nonparametric Bayesian DPR model would account for the mixed
359 sparse and infinitesimal genetic architecture. The 5-fold CV was conducted for all protein abundance
360 imputation models. The average CV R^2 across 5 folds of validation data was used to select the best
361 performing FUSION model as well as filter protein abundance imputation models for the follow-up
362 PWAS.

363 Second, protein abundance imputation models with 5-fold CV R^2 >0.005 were considered as
364 valid, and the corresponding pQTL weights were used along with summary-level GWAS data of AD

365 dementia to test for PWAS risk genes[23]. Last, PWAS-O p-values were generated by aggregating PWAS p-
366 values obtained by using pQTL weights estimated by TIGAR (DPR), PrediXcan (Elastic-Net), and
367 FUSION (best model).
368
369 **PPI network and pathway analysis**
370
371 To investigate the PPI network among the identified PWAS risk genes of AD dementia, we used
372 STRING[42, 43], a website for PPI network analysis which integrates public data sources of protein
373 interaction information and investigates the connectivity network of proteins. Proteins are considered as
374 nodes in the PPI network. Protein-protein edges represent the predicted functional associations, colored
375 differently to indicate seven categories — co-expression, text-mining, experiments (biochemical/genetic
376 data), databases (previously curated pathway and protein complex information), gene co-occurrence, gene
377 fusion and neighborhood. Gene co-occurrence, fusion, and neighborhood represent association
378 predictions are based on whole-genome comparisons.
379 STRING also enables gene-set enrichment analysis with respect to Gene Ontology (GO)
380 annotations[62], KEGG pathways[63], Reactome pathways[64], and UniProt keywords[65]. Enrichment analysis is
381 to detect GO terms, pathways, or UniProt keywords that are significantly enriched with genes in the
382 network versus random genes, adjusting for FDR. STRING also shows significant association between
383 input proteins and human phenotypes based on Monarch Initiative[66], a system to organize genotype-
384 phenotype connections from multiple data sources.
385
386 **Test Pleiotropy effects**
387 We used the PMR-Egger[20] tool to test the vertical pleiotropy effect between the protein
388 abundance level of PWAS risk genes and AD dementia, while controlling for horizontal effect. The
389 Multi-SNP-based SMR[45, 67] software and its accompanying heterogeneity in dependent instruments
390 (HEIDI) were also implemented to test the causal effect, using multiple *cis*-pQTL SNPs. Reference LD

391 derived from ROS/MAP WGS data were used to implement both PMR-Egger and SMR-HEIDI methods.
392 Test cis-SNPs were pruned according to their LD correlation matrix. The threshold of $R^2 = 0.95$ was used
393 for implementing PMR-Egger and $R^2 = 0.99$ was used for implementing SMR-HEIDI. For each protein-
394 coding gene, only SNPs that have nominal eQTL p-value < 0.05 were tested by SMR-HEIDI.
395

396 **TWAS-O of AD Dementia**

397 TWAS p-values were first obtained by using reference transcriptomic data of ROS/MAP samples
398 and GWAS summary data by TIGAR and PrediXcan, and then used to derive TWAS-O p-values by
399 ACAT. The reference transcriptomic data include 465 ROS/MAP samples with both gene expression
400 data for DLPFC tissue (by RNA sequencing) and WGS genotype data profiled. Gene expression data in
401 the format of Transcripts Per Million (TPM) were log2 transformed and adjusted for age at death, sex,
402 postmortem interval, study (ROS or MAP), batch effects, RNA integrity number scores, cell type
403 proportions (with respect to oligodendrocytes, astrocytes, microglia, neurons), top five genotype principal
404 components, and top probabilistic estimation of expression residuals (PEER) factors by linear regression
405 models. The cell type proportions were derived by using CIBERSORT pipeline[68] with single-cell RNA-
406 seq transcriptome profiles from human brain tissues as the reference[69] to de-convolute bulk RNA-seq
407 data[70]. Only genes with >0.1 TPM in ≥10 samples were considered. Cis-SNPs within a 1MB window of
408 each gene that met the MAF threshold >1% and Hardy-Weinberg Equilibrium (HWE) threshold of p-
409 value>10^{-5} were used in the imputation model as predictors. Only genes with an average 5-fold cross
410 validation $R^2 \geq 0.005$ in the gene expression imputation models were analyzed in TWAS.
411

412 **Acknowledgments**

413 We are thankful to all participants and the staff at the Rush Alzheimer's Disease Center for making the
414 proteomics and WGS data of ROS/MAP samples available to the research community. This work was
415 supported by the National Institute of Health R35GM138313 (T.H., 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.

Data availability

Proteomics, transcriptomics, 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.

Competing interests

The authors declare no competing interests.

Supplementary information

The supplementary file contains 2 supplementary tables and 5 supplementary figures.

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. *Nat Genet* **53**, 1276-1282 (2021).
2. 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**, 180142 (2018).
3. Liu, B., Gloudemans, M.J., Rao, A.S., Ingelsson, E. & Montgomery, S.B. Abundant associations with gene expression complicate GWAS follow-up. *Nat Genet* **51**, 768-769 (2019).
4. Gusev, A., *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245-252 (2016).
5. 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).
6. Brandes, N., Linial, N. & Linial, M. PWAS: proteome-wide association study-linking genes and phenotypes by functional variation in proteins. *Genome Biol* **21**, 173 (2020).
7. 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).
8. Luningham, J.M., *et al.* Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. *Am J Hum Genet* **107**, 714-726 (2020).
9. Tang, S., *et al.* Novel Variance-Component TWAS method for studying complex human diseases with applications to Alzheimer's dementia. *PLoS Genet* **17**, e1009482 (2021).
10. Wingo, A.P., *et al.* Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. *Nat Genet* **53**, 143-146 (2021).
11. Wingo, T.S., *et al.* Brain proteome-wide association study implicates novel proteins in depression pathogenesis. *Nat Neurosci* **24**, 810-817 (2021).
12. 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).
13. Gamazon, E.R., *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091-1098 (2015).
14. 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).
15. Boyle, E.A., Li, Y.I. & Pritchard, J.K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186 (2017).
16. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267-288 (1996).
17. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics* **9**, e1003264 (2013).
18. Liu, Y., *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet* **104**, 410-421 (2019).
19. Dai, Q., *et al.* OTTERS: A powerful TWAS framework leveraging summary-level reference data. *bioRxiv*, 2022.2003.2030.486451 (2022).
20. Yuan, Z., *et al.* Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nat Commun* **11**, 3861 (2020).

487 21. Zhu, Z., *et al.* Integration of summary data from GWAS and eQTL studies predicts
488 complex trait gene targets. *Nat Genet* **48**, 481-487 (2016).
489 22. Wu, Y., *et al.* Integrative analysis of omics summary data reveals putative mechanisms
490 underlying complex traits. *Nat Commun* **9**, 918 (2018).
491 23. Parrish, R.L., Gibson, G.C., Epstein, M.P. & Yang, J. TIGAR-V2: Efficient TWAS tool
492 with nonparametric Bayesian eQTL weights of 49 tissue types from GTEx V8. *HGG Adv* **3**,
493 100068 (2022).
494 24. Bennett, D.A., *et al.* Religious Orders Study and Rush Memory and Aging Project. *J*
495 *Alzheimers Dis* **64**, S161-s189 (2018).
496 25. De Jager, P.L., *et al.* A multi-omic atlas of the human frontal cortex for aging and
497 Alzheimer's disease research. *Scientific data* **5**, 1-13 (2018).
498 26. Buniello, A., *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide
499 association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-
500 d1012 (2019).
501 27. Marioni, R.E., *et al.* GWAS on family history of Alzheimer's disease. *Transl Psychiatry*
502 **8**, 99 (2018).
503 28. Jansen, I.E., *et al.* Genome-wide meta-analysis identifies new loci and functional
504 pathways influencing Alzheimer's disease risk. *Nat Genet* **51**, 404-413 (2019).
505 29. Kim, S., *et al.* Genome-wide association study of CSF biomarkers Abeta1-42, t-tau, and
506 p-tau181p in the ADNI cohort. *Neurology* **76**, 69-79 (2011).
507 30. Schwartzentruber, J., *et al.* Genome-wide meta-analysis, fine-mapping and integrative
508 prioritization implicate new Alzheimer's disease risk genes. *Nat Genet* **53**, 392-402 (2021).
509 31. Gouveia, C., *et al.* Genome-wide association of polygenic risk extremes for Alzheimer's
510 disease in the UK Biobank. *Sci Rep* **12**, 8404 (2022).
511 32. Davies, G., *et al.* Study of 300,486 individuals identifies 148 independent genetic loci
512 influencing general cognitive function. *Nat Commun* **9**, 2098 (2018).
513 33. Lee, J.J., *et al.* Gene discovery and polygenic prediction from a genome-wide association
514 study of educational attainment in 1.1 million individuals. *Nat Genet* **50**, 1112-1121 (2018).
515 34. Wang, H., *et al.* Genome-wide interaction analysis of pathological hallmarks in
516 Alzheimer's disease. *Neurobiol Aging* **93**, 61-68 (2020).
517 35. Hao, S., Wang, R., Zhang, Y. & Zhan, H. Prediction of Alzheimer's Disease-Associated
518 Genes by Integration of GWAS Summary Data and Expression Data. *Front Genet* **9**, 653 (2018).
519 36. Gockley, J., *et al.* Multi-tissue neocortical transcriptome-wide association study
520 implicates 8 genes across 6 genomic loci in Alzheimer's disease. *Genome Med* **13**, 76 (2021).
521 37. Gerring, Z., Lupton, M., Edey, D., Gamazon, E. & Derks, E. An analysis of genetically
522 regulated gene expression across multiple tissues implicates novel gene candidates in
523 Alzheimer's disease. *Alzheimer's Research & Therapy* **12**, 43 (2020).
524 38. Sun, Y., *et al.* A transcriptome-wide association study of Alzheimer's disease using
525 prediction models of relevant tissues identifies novel candidate susceptibility genes. *Genome*
526 *Medicine* **13**, 141 (2021).
527 39. Davidson, J.E., *et al.* Plasma lipoprotein-associated phospholipase A2 activity in
528 Alzheimer's disease, amnestic mild cognitive impairment, and cognitively healthy elderly
529 subjects: a cross-sectional study. *Alzheimers Res Ther* **4**, 51 (2012).
530 40. Drenos, F., *et al.* Integrated associations of genotypes with multiple blood biomarkers
531 linked to coronary heart disease risk. *Hum Mol Genet* **18**, 2305-2316 (2009).

532 41. Dumont, V. & Lehtonen, S. PACSIN proteins in vivo: Roles in development and
533 physiology. *Acta Physiol (Oxf)* **234**, e13783 (2022).
534 42. Szklarczyk, D.*, et al.* STRING v11: protein–protein association networks with increased
535 coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids*
536 *Research* **47**, D607-D613 (2018).
537 43. Szklarczyk, D.*, et al.* The STRING database in 2021: customizable protein-protein
538 networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic*
539 *Acids Res* **49**, D605-d612 (2021).
540 44. Jong, M.C., Hofker, M.H. & Havekes, L.M. Role of ApoCs in Lipoprotein Metabolism.
541 *Arteriosclerosis, Thrombosis, and Vascular Biology* **19**, 472-484 (1999).
542 45. Zhu, Z.*, et al.* Integration of summary data from GWAS and eQTL studies predicts
543 complex trait gene targets. *Nature Genetics* **48**, 481-487 (2016).
544 46. Myklebost, O. & Rogne, S. The gene for human apolipoprotein CI is located 4.3
545 kilobases away from the apolipoprotein E gene on chromosome 19. *Hum Genet* **73**, 286-289
546 (1986).
547 47. Smit, M., van der Kooij-Meijs, E., Frants, R.R., Havekes, L. & Klasen, E.C.
548 Apolipoprotein gene cluster on chromosome 19. Definite localization of the APOC2 gene and
549 the polymorphic Hpa I site associated with type III hyperlipoproteinemia. *Hum Genet* **78**, 90-93
550 (1988).
551 48. Lauer, S.J.*, et al.* Two copies of the human apolipoprotein C-I gene are linked closely to
552 the apolipoprotein E gene. *J Biol Chem* **263**, 7277-7286 (1988).
553 49. Saunders, A.M.*, et al.* Association of apolipoprotein E allele ϵ4 with late-onset familial
554 and sporadic Alzheimer's disease. *Neurology* **43**, 1467-1467 (1993).
555 50. Panza, F.*, et al.* Lipid metabolism in cognitive decline and dementia. *Brain Res Rev* **51**,
556 275-292 (2006).
557 51. Wood, R.D., Mitchell, M., Sgouros, J. & Lindahl, T. Human DNA repair genes. *Science*
558 **291**, 1284-1289 (2001).
559 52. Weeda, G.*, et al.* A presumed DNA helicase encoded by ERCC-3 is involved in the
560 human repair disorders xeroderma pigmentosum and Cockayne's syndrome. *Cell* **62**, 777-791
561 (1990).
562 53. Fishel, M.L., Vasko, M.R. & Kelley, M.R. DNA repair in neurons: so if they don't divide
563 what's to repair? *Mutat Res* **614**, 24-36 (2007).
564 54. Weissman, L.*, et al.* Defective DNA base excision repair in brain from individuals with
565 Alzheimer's disease and amnestic mild cognitive impairment. *Nucleic Acids Research* **35**, 5545-
566 5555 (2007).
567 55. Ajuh, P.*, et al.* Functional analysis of the human CDC5L complex and identification of its
568 components by mass spectrometry. *Embo j* **19**, 6569-6581 (2000).
569 56. Urano, Y.*, et al.* Involvement of the mouse Prp19 gene in neuronal/astroglial cell fate
570 decisions. *J Biol Chem* **281**, 7498-7514 (2006).
571 57. Yamada, T.*, et al.* The U-box-type ubiquitin ligase PRP19β regulates astrocyte
572 differentiation via ubiquitination of PTP1B. *Brain Res* **1524**, 12-25 (2013).
573 58. Lunnon, K.*, et al.* Mitochondrial genes are altered in blood early in Alzheimer's disease.
574 *Neurobiol Aging* **53**, 36-47 (2017).
575 59. Reynolds, C.A.*, et al.* Analysis of lipid pathway genes indicates association of sequence
576 variation near SREBF1/TOM1L2/ATPAF2 with dementia risk. *Human Molecular Genetics* **19**,
577 2068-2078 (2010).

578 60. Verbanck, M., Chen, C.Y., Neale, B. & Do, R. Detection of widespread horizontal
579 pleiotropy in causal relationships inferred from Mendelian randomization between complex traits
580 and diseases. *Nat Genet* **50**, 693-698 (2018).
581 61. Wingo, A.P.*, et al.* Shared proteomic effects of cerebral atherosclerosis and Alzheimer's
582 disease on the human brain. *Nature Neuroscience* **23**, 696-700 (2020).
583 62. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**,
584 D330-d338 (2019).
585 63. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a
586 reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462 (2016).
587 64. Jassal, B.*, et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* **48**, D498-
588 d503 (2020).
589 65. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506-d515
590 (2019).
591 66. Mungall, C.J.*, et al.* The Monarch Initiative: an integrative data and analytic platform
592 connecting phenotypes to genotypes across species. *Nucleic Acids Res* **45**, D712-d722 (2017).
593 67. Wu, Y.*, et al.* Integrative analysis of omics summary data reveals putative mechanisms
594 underlying complex traits. *Nature Communications* **9**, 918 (2018).
595 68. Newman, A.M.*, et al.* Robust enumeration of cell subsets from tissue expression profiles.
596 *Nat Methods* **12**, 453-457 (2015).
597 69. Darmanis, S.*, et al.* A survey of human brain transcriptome diversity at the single cell
598 level. *Proc Natl Acad Sci U S A* **112**, 7285-7290 (2015).
599 70. Wingo, T.S.*, et al.* Brain microRNAs associated with late-life depressive symptoms are
600 also associated with cognitive trajectory and dementia. *NPJ Genom Med* **5**, 6 (2020).
601
602
603

Tables

Table 1. Trained protein abundance imputation models with CV R^2 > 0.005 from ROS/MAP reference data, by TIGAR, PrediXcan, and FUSION tools.

Model	N	CV R^2 median
TIGAR	6673	0.013
PrediXcan	1835	0.0094
FUSION	2173	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 (because of 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.

Figures Titles and Legends

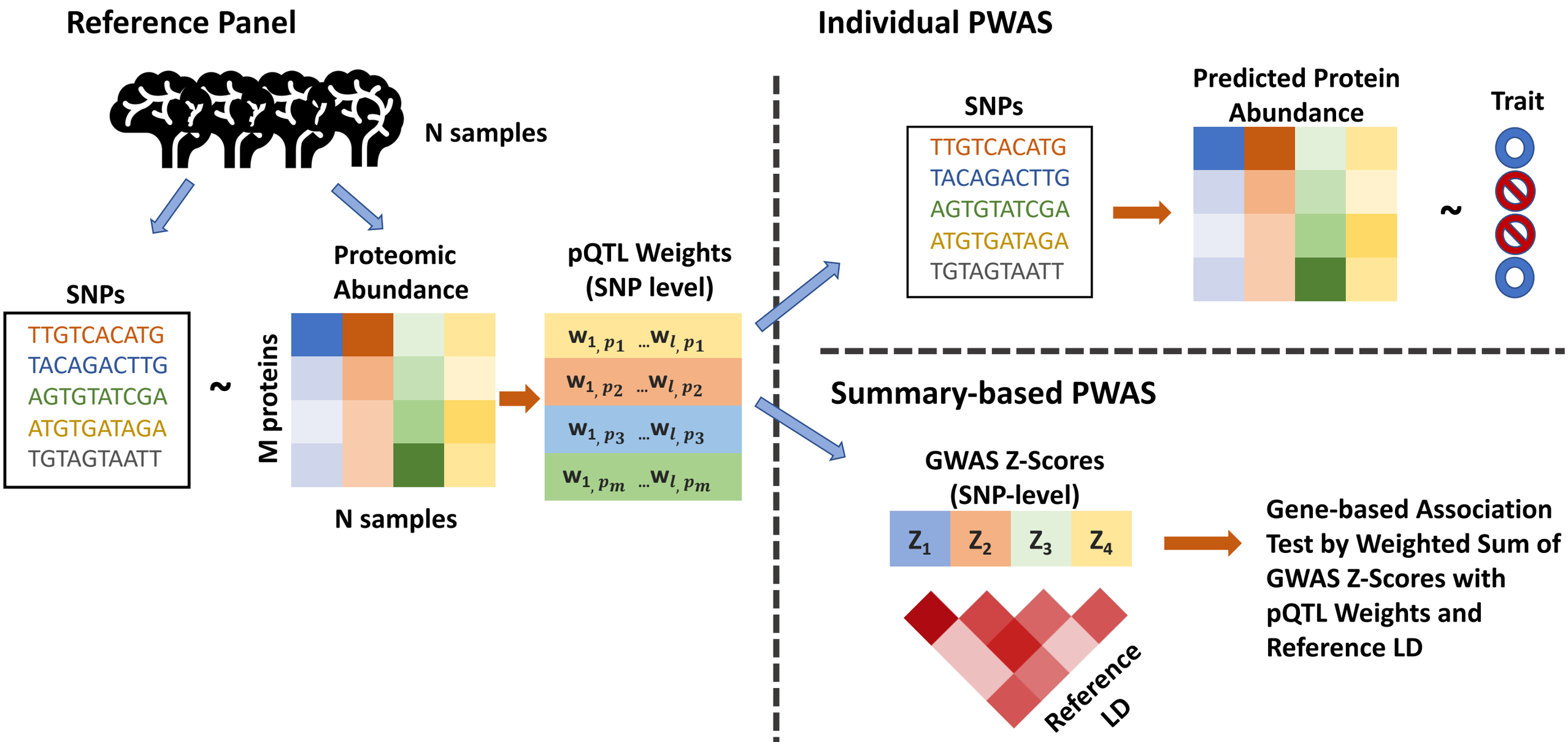
Fig.1: Two-stage PWAS framework.
Stage I: Protein abundance imputation models are trained using *cis*-SNPs as predictors and proteomic abundance as the outcome, which estimates pQTL effect sizes (weights) per protein-coding gene. Stage II: Individual PWAS predicts the protein abundances using the trained 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 gene-based association test with GWAS Z-scores and reference LD.

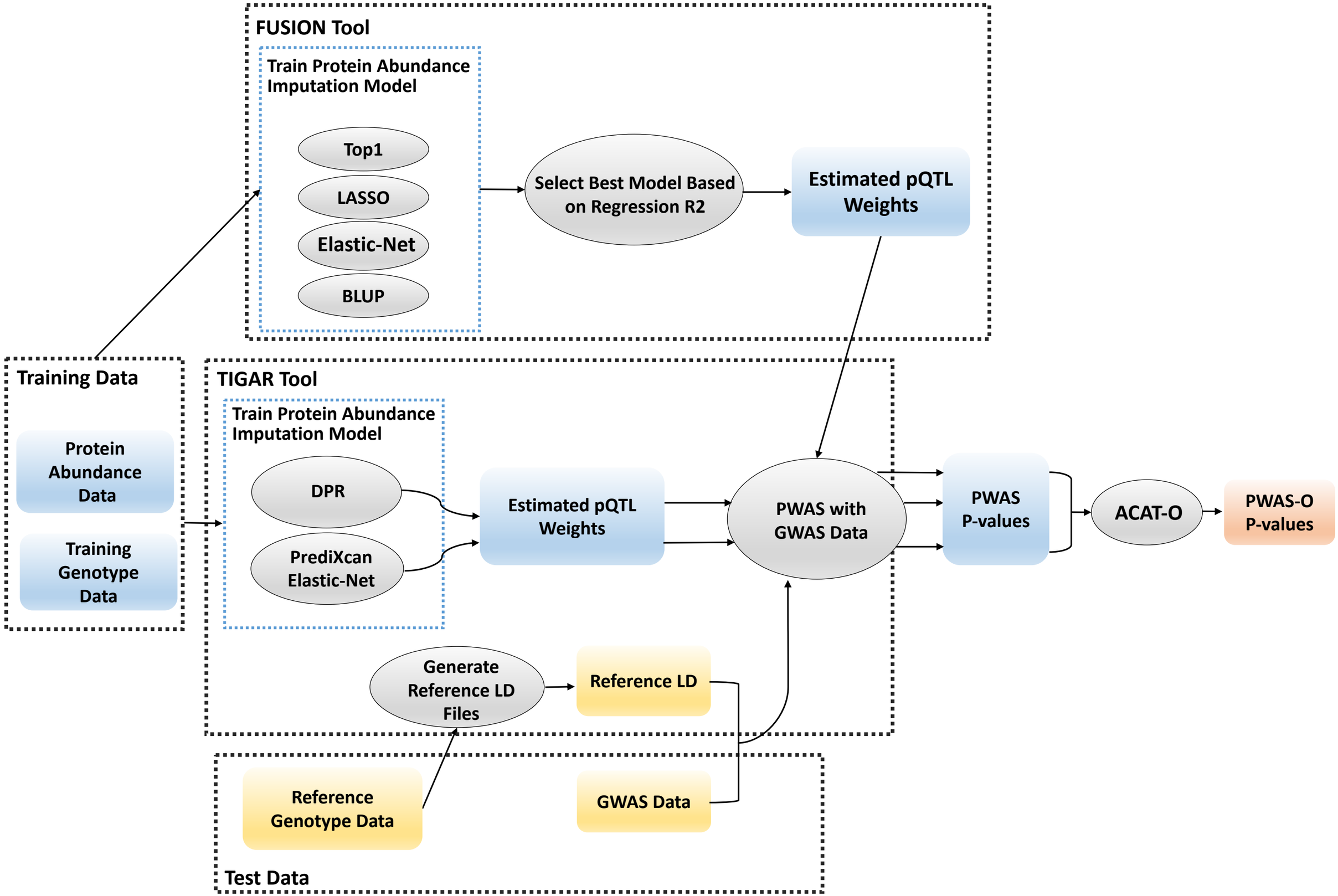
Fig. 2: PWAS-O framework.
First, train protein abundance imputation models using reference data and conduct two-stage PWAS (**Fig. 1**) using GWAS data by TIGAR, PrediXcan, and FUSION tools. Second, implement ACAT-O with all PWAS p-values to obtain a PWAS-O p-value per protein-coding gene.

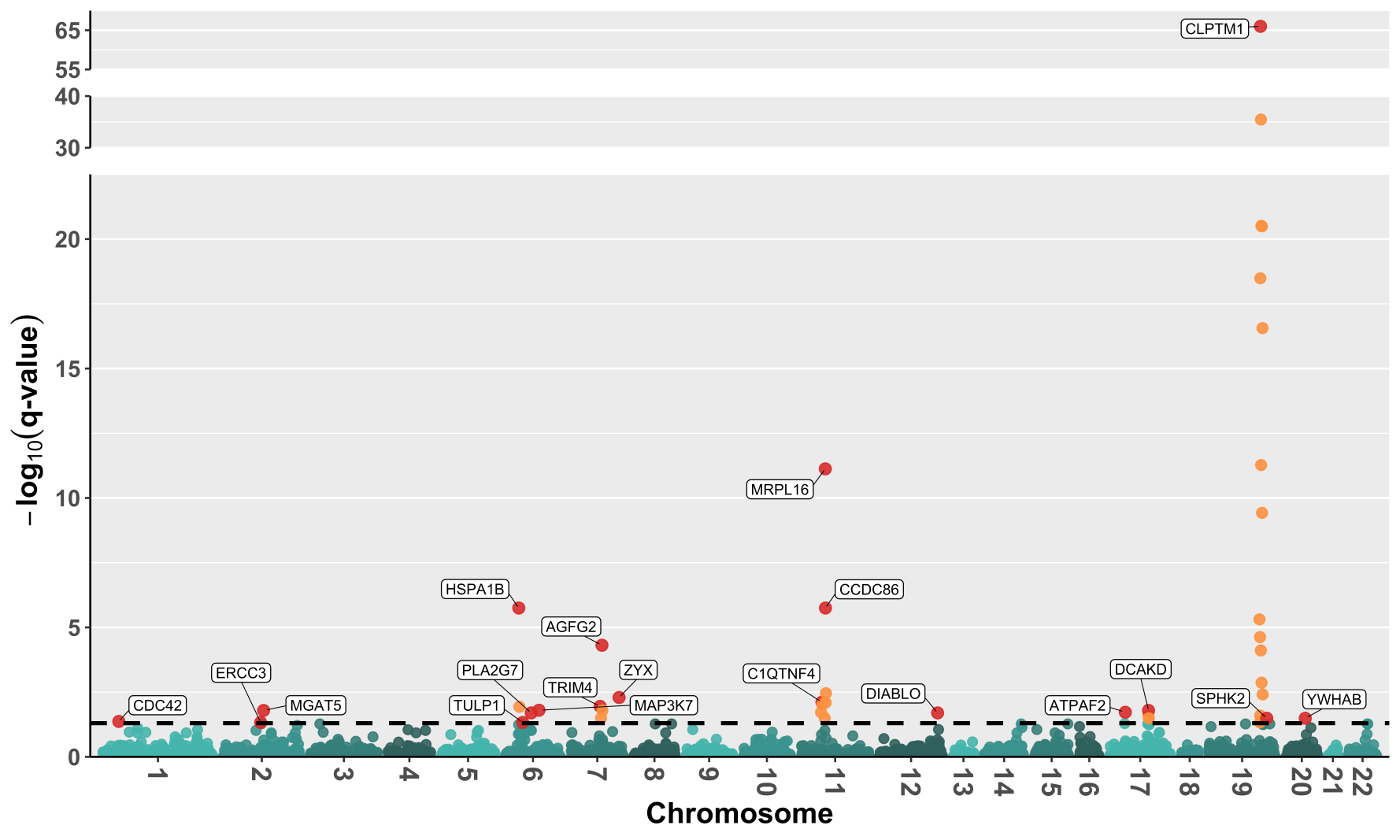
Fig. 3: Manhattan plot of PWAS-O results for AD dementia.
A total of 43 significant risk genes were identified with FDR q-value < 0.05. A total of 19 independently significant genes were labeled. The -log10(q-values) were plotted in the y-axis, and -log10(0.05) was plotted as the dashed horizontal line.

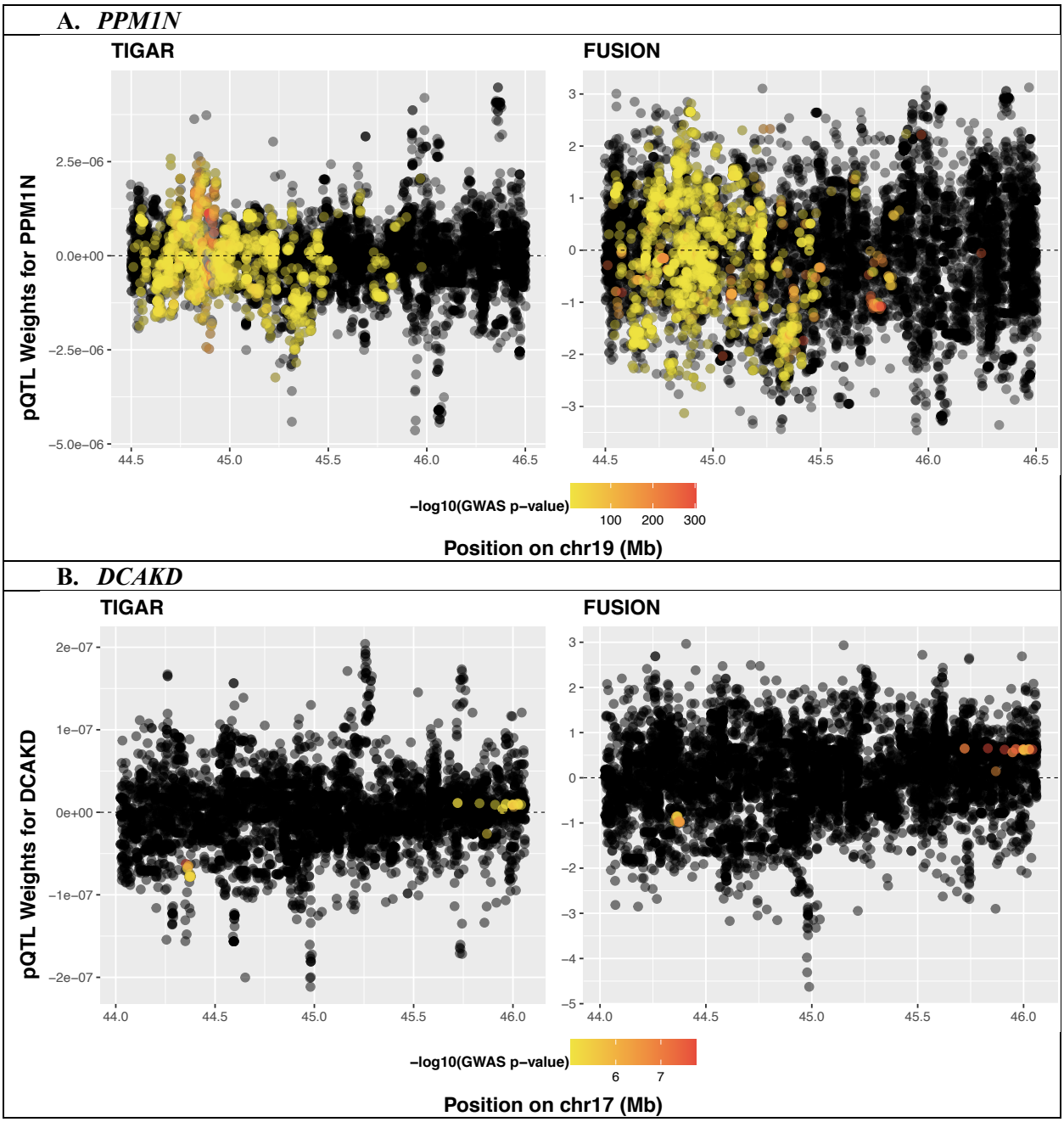
Fig. 4: pQTL weights estimated by TIGAR (DPR) and FUSION (best model of LASSO, Elastic-Net, BLUP, Top1) for example significant genes *PPM1N* (A) and *DCAKD* (B) by PWAS-O.
The pQTL weights were plotted in the y-axis for all test SNPs in the test gene region, with color coded with respect to their -log10 (GWAS p-value). SNPs with -log10(GWAS p-value) < 10^{-5} were plotted in yellow color. Gene *PPM1N* 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* 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.

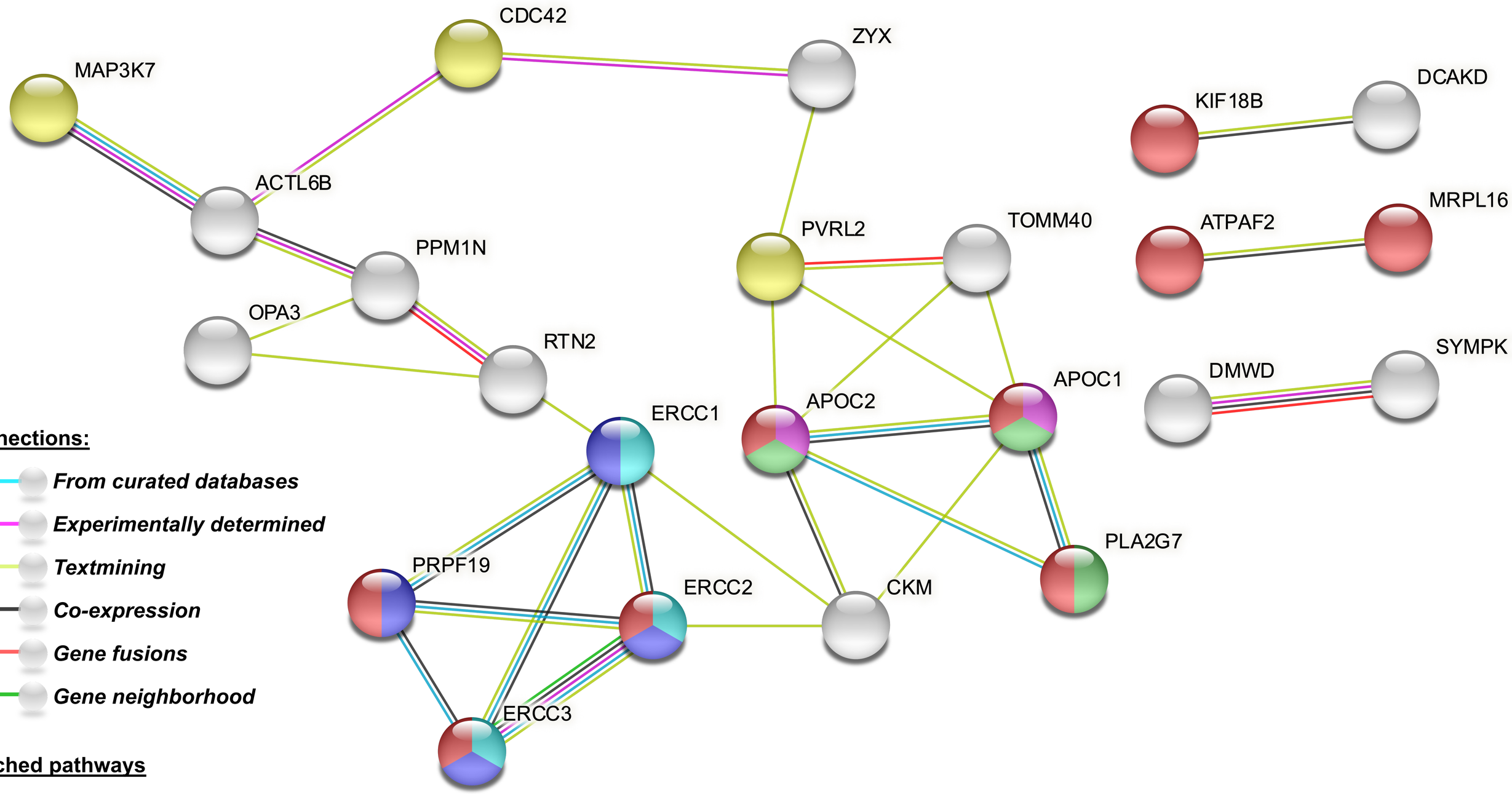
Fig. 5: PPI networks and enriched pathways among 43 PWAS risk genes of AD dementia by STRING.
Edges represent physical PPI, with different colors representing different sources of connection evidence. Node colors represent different enriched GO terms or KEGG pathways with FDR < 0.05.


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


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)*