Proteome-wide association studies using summary proteomic data identified 23 risk genes of Alzheimer's disease

Tingyang Hu[1,2], Qile Dai[1,3], Michael P. Epstein[1], Jingjing Yang[1,*]

Abstract

Characterizing the genetic mechanisms underlying Alzheimer's disease (AD) dementia is crucial for developing new therapeutics. Proteome-wide association study (PWAS) integrating proteomics data with genome-wide association study (GWAS) summary data was shown as a powerful tool for detecting risk genes. The identified PWAS risk genes can be interpretated as having genetic effects mediated through the genetically regulated protein abundances. Existing PWAS analyses of AD often rely on the availability of individual-level proteomics and genetics data of a reference cohort. Leveraging summary-level protein quantitative trait loci (pQTL) reference data of multiple relevant tissues is expected to improve PWAS findings for studying AD.

Here, we applied our recently developed OTTERS tool to conduct PWAS of AD dementia, by leveraging summary-level pQTL data of brain, cerebrospinal fluid (CSF), and plasma tissues, and multiple statistical methods. For each target protein, imputation models of the protein abundance with genetic predictors were trained from summary-level pQTL data, estimating a set of pQTL weights for considered genetic predictors. PWAS p-values were obtained by integrating GWAS summary data of AD dementia with estimated pQTL weights. PWAS p-values from multiple statistical methods were combined by the aggregated Cauchy association test to yield one omnibus PWAS p-value for the target protein. We identified significant PWAS risk genes through omnibus PWAS p-values and analyzed their protein-protein interactions using STRING. Their potential causal effects were assessed by the probabilistic Mendelian randomization (PMR-Egger).

As a result, we identified a total of 23 significant PWAS risk genes for AD dementia in brain, CSF, and plasma tissues, including 7 novel findings. We showed that 15 of these risk genes were interconnected within a protein-protein interaction network involving the well-known AD risk gene of *APOE* and 5 novel findings, and enriched in immune functions and lipids pathways

including positive regulation of immune system process, positive regulation of macrophage proliferation, humoral immune response, and high-density lipoprotein particle clearance. Existing biological evidence was found to relate our novel findings with AD. We validated the mediated causal effects of 14 risk genes (60.8%).

In conclusion, we identified both known and novel PWAS risk genes, providing novel insights into the genetic mechanisms in brain, CSF, and plasma tissues, and targeted therapeutics development of AD dementia. Our study also demonstrated the effectiveness of integrating public available summary-level pQTL data with GWAS summary data for mapping risk genes of complex human diseases.

Author affiliations:

1. Center for Computational and Quantitative Genetics, Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, 30322, USA
2. Division of Biostatistics and Bioinformatics, Department of Public Health Sciences, Pennsylvania State University College of Medicine, Hershey, PA, 17033, USA
3. Department of Biostatistics and Bioinformatics, Emory University School of Public Health, Atlanta, GA, 30322, USA

Correspondence to: Jingjing Yang
Center for Computational and Quantitative Genetics, Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, 30322, USA
jingjing.yang@emory.edu

Keywords: PWAS; pQTL summary data; GWAS; Alzheimer's disease dementia; OTTERS.

Introduction

Large-scale genome-wide association studies (GWAS) have successfully identified dozens of genetic risk loci related to Alzheimer's disease (AD) dementia[1–3]. However, the underlying molecular mechanisms of these GWAS risk genes of AD are still largely unknown. To gain biological insights into how associated risk genes might contribute to AD dementia, researchers have performed proteome-wide association studies (PWAS) that integrates reference proteomic data from a disease-related tissue with GWAS summary data of AD dementia to identify risk genes whose effects are mediated via genetically regulated protein abundance[4,5].

PWAS typically employs a two-stage framework: Stage I uses the genetics and proteomics data of the same reference cohort to train a protein abundance prediction model for each target protein, taking the protein abundance quantitative trait as the response variable and the cis-SNPs proximal to the protein coding gene as predictors. The estimated SNP coefficients from Stage I can be viewed as effect sizes of "protein quantitative trait loci (pQTL)" in a broad sense, as most cis-SNPs with non-zero effect sizes will not be statistically significant pQTL. Stage II proceeds by using the estimated pQTL effect sizes as variant weights to predict genetically regulated protein abundance in a GWAS cohort, and subsequently conducts a gene-based association test (of the corresponding protein coding gene) relating the predicted abundance of the target protein to phenotype.

Existing analytic tools derived for the analogous transcriptome-wide association studies (TWAS) have been used for PWAS. Most existing tools including TIGAR[6], PrediXcan[7], and FUSION[8], which utilize different statistical methods in Stage I to estimate the pQTL weights, require individual-level genetics and proteomics data of the reference cohort. For example, by PWAS analyses of AD dementia with the individual-level reference proteomics data of dorsolateral prefrontal cortex (DLPFC) tissue and whole genome sequencing (WGS) genotype data from samples in the Religious Orders Study and Memory and Aging Project (ROS/MAP)[9], Wingo et al.[4] detected 11 risk genes using the FUSION[8] tool and Hu et al.[10] identified 43 risk genes by leveraging all three tools of TIGAR[6], PrediXcan[7], and FUSION[8].

In this work, we utilized our recently developed OTTERS[11] tool to expand the PWAS analyses of AD dementia, by leveraging pQTL summary data of not only brain (parietal lobe cortex, n=380) but also other important tissues such as cerebrospinal fluid (CSF, n=835) and plasma (n=529) tissues[12]. Recent studies have shown that amyloid beta (Aβ)1-42/Aβ1-40 and

phosphorylated tau/Aβ1-42 ratios in CSF[13,14] and plasma[15,16] could be used as biomarkers for early diagnosis of AD. Thus, conducting PWAS with the recent GWAS summary data of AD dementia (n=~762K)[2] in all three tissues (brain, CSF, and plasma) is expected to identify more risk genes of AD whose genetic effects are potentially mediated through the genetically regulated protein abundances.

Materials and methods

OTTERS framework

In a two-stage PWAS framework with individual-level genetic and proteomic data from a reference dataset, Stage I involves fitting a multiple linear regression model (*Equation 1*) with protein abundance ($\mathbf{E}_p$) of a protein p as the outcome, genotype data ($\mathbf{X}$) of cis-SNPs of the corresponding protein coding gene (i.e., SNPs located within ±1Mb region around gene transcription starting/termination site) as predictors, and $\boldsymbol{w}$ denoting the pQTL weights to be estimated:

$$\boldsymbol{E}_p = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}; \qquad \boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 \mathbf{I}). \qquad (\textit{Equation 1})$$

Potential confounding covariates are assumed to be adjusted from the original protein abundance measures, resulting in the residuals of $\boldsymbol{E}_p$. Both $\boldsymbol{E}_p$ and columns of $\mathbf{X}$ are standardized with mean 0 and variance 1.

Whereas, OTTERS[11] estimates $\mathbf{w}$ from the summary-level pQTL reference data that are assumed to be generated based on the following single variant linear regression models with standardized genotype vectors $\mathbf{x}_j$ for genetic variants $j = 1, \ldots, m$:

$$\boldsymbol{E}_p = \boldsymbol{x}_j w_j + \boldsymbol{\epsilon}_j, \qquad \boldsymbol{\epsilon}_j \sim N(0, \sigma_{\epsilon_j}^2 \boldsymbol{I}). \qquad (\textit{Equation 2})$$

Summary-level pQTL reference data include the marginal least squared effect estimates ($\widetilde{w}_j, j = 1, \ldots, m$), sample sizes, and linkage disequilibrium coefficients in the pQTL cohort (which can also be derived from an external reference panel with the same ancestry as the pQTL cohort). OTTERS employs five representative PRS methods, including the P-value Thresholding with linkage disequilibrium (LD) clumping (*P+T*)[17] with p-value thresholds of 0.05 and 0.001, frequentist LASSO regression (*lassosum*)[18,19], nonparametric Bayesian Dirichlet process regression (*SDPR*)[20,21], and Bayesian multiple linear regression with continuous shrinkage prior

(*PRS-CS*)[22]. These PRS methods can estimate five sets of pQTL weights ($\widehat{\boldsymbol{w}}$) for each protein coding gene.

In Stage II, OTTERS first uses these five sets of pQTL weights ($\widehat{\boldsymbol{w}}$) from Stage I as variant weights (*Equation 3*) to test gene-based association with respect to the phenotype in the summary-level GWAS test data. The test statistic can be written as

$$Z_p = \frac{\sum_{j=1}^{m}(\widehat{w}_j Z_j)}{\sqrt{\widehat{\boldsymbol{w}}'\boldsymbol{V}\widehat{\boldsymbol{w}}}}, \qquad (\textit{Equation 3})$$

where Z_j denotes the single variant Z-score test statistic in GWAS summary data for the jth genetic variant, and $\boldsymbol{V}$ denotes the genotype correlation matrix that could be obtained from an external reference panel of the same ancestry as the test GWAS data[8]. Such gene-based association test has been shown to be equivalent as testing the association between predicted genetically regulated protein abundances and the phenotype in the GWAS test data[7,8,23].

Since the performance of a PRS method depends on the unknown genetic architecture of protein abundances, OTTERS aggregates the PWAS p-values based on all five PRS methods by using the aggregated Cauchy association test[24]. An omnibus test p-value is derived for each protein coding gene, which is then used to identify significant PWAS risk genes (see details in the Supplementary Methods of the OTTERS paper[11]).

Apply OTTERS to conduct PWAS of AD dementia

We first applied OTTERS[11] to estimate pQTL weights from the recently released summary-level pQTL data of brain (n=380), CSF (n=835), and plasma (n=529)[12]. These summary-level pQTL data were generated by using proteomics data of individuals with AD and cognitively normal individuals of European ancestry profiled from an aptamer-based platform[25]. These summary-level pQTL data were generated for 1079 proteins in brain, 731 proteins in CSF, and 931 proteins in plasma, and ~14M genetic variants with minor allele frequency (MAF) ≥ 2%. Linear regression models with protein abundances as the response variable, genotype of a single genetic variant as the test covariate, and additional adjusting covariates of age, sex, first two genotype principal components factors, and genotype platform, were used to generate the summary-level pQTL data. Since the summary-level pQTL data of these three tissues were generated by using samples of European ancestry, the LD information obtained from the whole

genome sequencing data of European samples from the ROS/MAP study[9] was used to estimate pQTL weights, along with standardized marginal pQTL effect sizes and sample sizes.

For each available protein in these three tissue types, we obtained five sets of estimated pQTL weights by five PRS methods, which were used to conduct PWAS analyses with the recent GWAS summary data (n=~762K) of AD dementia[2]. The GWAS summary data of AD dementia[2] were generated by meta-analysis with 12 cohorts, excluding the 23&Me cohort. About 11.3% of the GWAS samples had clinically diagnosed AD dementia. Omnibus OTTERS p-values were obtained for each available protein in all three tissues. We corrected the omnibus OTTERS p-values per tissue by using the genomic control factor[26], to ensure that the median test p-value was adjusted to the expected value 0.5 under the null hypothesis. Then we used the adjusted nominal OTTERS p-values to calculate the false discovery rates (FDR, i.e., q-values) per tissue. We identified genes with q-values < 0.05 as significant PWAS risk genes for AD dementia in the corresponding tissue.

PPI network and enrichment analyses by STRING

The STRING[27–29] webtool integrates public data sources of protein infarction and analyzes the protein-protein interaction (PPI) network connectivity of proteins. Protein-protein edges represent the functional association, colored with six different connections — curated databases, experiments, textmining, co-expression, gene co-occurrence and protein homology. Gene co-occurrence association predictions are based on whole-genome comparisons. The STRING[27] webtool also provides gene enrichment analysis with respect to Gene Ontologies (GO)[30] annotations. Enrichment analysis aims to detect GO terms and pathways that are significantly enriched with genes in the network versus random genes. The enrichment strength is provided along with FDR, which indicates the ratio between the number of proteins in the network that are annotated with a term and the number of proteins that expected to be annotated with this term in a random network of the same size. In this study, we utilized the STRING webtool to conduct PPI network and enrichment analyses with the list of PWAS risk genes identified by OTTERS in all three tissues.

Examine causal effects of PWAS risk genes by PMR-Egger

As the two-stage PWAS framework does not distinguish genetic effects mediated through genetically regulated protein abundances (i.e., causal effects or vertical pleiotropy effects) versus effects through other pathways (i.e., horizontal pleiotropy effects), we further assessed the mediated causal effects of our identified significant PWAS risk genes by using the probabilistic Mendelian randomization (PMR-Egger) tool[31]. PMR-Egger can assess causal genetic effects while controlling for horizontal pleiotropy effects by using summary-level pQTL and GWAS data. The reference LD derived from the ROS/MAP WGS data was also used for implementing PMR-Egger.

Results

PWAS of AD dementia by OTTERS

By applying OTTERS[11] to the summary-level pQTL reference data of three tissues (brain, CSF and plasma)[5] and recent large-scale GWAS summary data of AD dementia, we obtained PWAS p-values with pQTL weights estimated by five complimentary PRS methods. Moderate inflation was observed in the Quantile-Quantile (Q-Q) plots of these proteome-wide p-values in all three tissues (**Fig. S1-S3**). Omnibus OTTERS p-values were obtained by combining the PWAS p-values across all 5 PRS methods[24], and then were adjusted by the genomic control factor[26]. FDR q-values were then obtained from the adjusted OTTERS p-values to account for multiple testing.

We identified 23 PWAS significant risk genes of AD dementia with FDR q-value < 0.05 by OTTERS, including 8, 9, and 10 genes respectively detected in brain, CSF, and plasma tissues (**Fig. 1; Table 1**), 4 detected in at least two tissues, and 2 independent genes (*APOM* and *BCAM*) detected in all three tissues. From these 23 significant PWAS risk genes, we curated 16 independent genes that do not have shared genetic variants in their test regions ($\pm$1MB around the test transcription starting/termination sites of the protein-coding gene).

Comparing our PWAS results to previous GWAS findings in GWAS Catalog[32] (**Table 1**), we found 6 out of 23 PWAS significant genes (*SEZ6L2*[33], *APOE*[34], *HAVCR2*[2], *BCAM*[1], *POMC*[35], *IL34*[3]) were detected by previous GWAS of AD, or GWAS of AD pathological hallmarks. Comparing to previous TWAS findings, we found 10 of our identified PWAS risk genes (*MST1*[36], *C4A*[36], *CXCL16*[23], *APOE*[23], *MAPKAPK2*[37], *APOM*[36], *BCAM*[23], *POMC*[23], *AIF1*[36],

F2[37]) were either known TWAS risk genes or located within the shared (1MB) test region of previously detected TWAS risk genes. We also found 7 out of these 23 risk genes (*C4A*, *APOE*, *APOM*, *BCAM*, *AIF1*, *F2*) were identified by previous PWAS of AD dementia using proteomics data of the DLPFC tissue[10]. Moreover, we identified 7 novel risk genes that were not reported in previous GWAS, TWAS, or PWAS of AD dementia: *LTA* (p-value = 1.70e-05), *HSPA1A* (p-value = 1.85e-04), *CLIC1* (p-value = 1.85e-04), *FOLH1* (p-value = 9.51e-21), *NCR3* (p-value = 1.50e-12), *CD177* (p-value = 4.36e-04), and *PLAUR* (p-value = 2.64e-11).

Comparison of different PRS methods

As described in the Methods section, OTTERS leverages all 5 complimentary PRS methods (*P+T (0.001)*, *P+T (0.05)*, *lassosum*, *SDPR*, and *PRS-CS*) to account for complex genetic architecture underlying the protein abundance quantitative traits, thus improving the PWAS power for studying complex diseases. Here, we compared the omnibus OTTERS p-values to the PWAS p-values obtained by each individual PRS methods (**Table 1**). Similar as reported in the OTTERS paper, we found that all individual PRS methods contributed to the omnibus OTTERS results. For example, the well-known risk gene *APOE* in brain was only detected by *lassosum* and *P+T (0.001)*, and gene *BCAM* in CSF that is proximate to *APOE* was detected by *lassosum*, *PRS-CS*, and *SDPR*. To investigate how individual PRS methods contribute to the final OTTERS results, we plotted the pQTL weights estimated by all five PRS methods for three example PWAS risk genes (*APOE*, *BCAM*, and *CD177*) that were detected in all three tissues by OTTERS and identified by two or three individual PRS methods (**Fig. 2**). We plotted these pQTL weights with color coded corresponding to $-log10$ (GWAS p-values).

In **Fig. 2A**, we showed pQTL weights for gene *APOE* (with pQTL summary data of brain), which was found significant by *lassosum* (p-value = 3.38e-102) and *P+T (0.001)* (p-value <1e-300). We found that the significance of gene *APOE* was driven by this one GWAS significant SNP (rs157580, chr19:44892009, located in the intron region of *TOMM40* and downstream of *NECTIN2*) with non-zero pQTL weights estimated by *P+T (0.001)* and *lassosum*. When the pQTL weights of this driving SNP were estimated as 0 or near 0 by *P+T (0.001)*, *SDPR*, and *PRS-CS*, PWAS p-values by these PRS methods were less significant (1.84e-4 by *P+T (0.05)*, 9.13e-2 by *PRS-CS*, 1.72e-4 by *SDPR*). Interestingly, genes *TOMM40* and *NECTIN2* are proximal to *APOE*. *TOMM40* is a known risk gene of AD dementia[1] and found to

be associated with family history of AD[38]. *NECTIN2* was previously reported as a GWAS risk gene for beta-amyloid 1-42[39] , and the interaction of low-density lipoprotein cholesterol levels and short total sleep time[40].

In **Fig. 2B**, we showed pQTL weights of gene *BCAM* (with pQTL summary data of CSF), which was detected by *lassosum* (p-value = 2.16e-65), *PRS-CS* (p-value = 8.35e-07), and *SDPR (*p-value = 2.81e-07). Compared to the *P+T* methods, these PRS methods all estimated non-zero pQTL weights for more GWAS significant SNPs in the test region. In **Fig 2C**, we showed pQTL weights of gene *CD177* (with pQTL summary data of plasma), which was only detected by *lassosum* (p-value = 1.17e-05). We can see that both *SDPR* and *PRS-CS* estimated near-zero pQTL weights for "significant" GWAS SNPs (GWAS p-values < 1e-5, colored in the plots) in the test region, and most "significant" GWAS SNPs were filtered out by the *P+T* method due to their pQTL p-values >0.05. Only *lassosum* estimated non-zero pQTL weights for these driving GWAS "significant" SNPs.

Additionally, we plotted the pQTL weights of another 3 significant PWAS risk genes *C4A* (**Fig. S4)**, *APOM* **(Fig. S5**), and *AIF1* **(Fig. S6)**. Genes *C4A* (in brain) and *APOM* (in plasma) were found significant by all PRS methods, which all estimated non-zero pQTL weights for "significant" GWAS SNPs. Gene *AIF1* (in CSF) was found significant by *P+T (0.005)* (p-value = 7.21e-08) and *PRS-CS* (p-value = 4.06e-06), which estimated pQTL weights in relatively higher magnitude for "significant" GWAS SNPs.

Overall, these plots of pQTL weights demonstrated that significant PWAS risk genes were mainly driven by test genetic variants that had non-zero pQTL weights colocalized with "significant" GWAS p-values, and that OTTERS leverages the strength of all complement PRS methods to achieve higher power.

PPI network and enrichment analyses

By using the STRING[27] webtool (Methods), we found 15 proteins out of these 23 significant protein-coding genes identified by OTTERS were interconnected in a network, including *APOE*, a well-known risk genes of AD dementia (**Fig. 3**). The edges of the network were colored according to the protein-protein interactions based on different data sources. The STRING webtool also provided gene enrichment analyses results (**Fig. 3**). We found that the genes in this network were enriched in GO pathways of positive regulation of immune system

process (*CD177, HSPA1A, IL34, LTA, NCR3, AIF1, HAVCR2, C4A, MAPK3, POMC*) with enrichment FDR = 2.78e-07 and strength = 1.1, response to stress (*CD177, HSPA1A, IL34, LTA, NCR3, AIF1, HAVCR2, C4A, MAPK3, POMC*) with enrichment FDR = 3.7e-05 and strength = 0.64, positive regulation of macrophage proliferation (*MAPK3, IL34*) with enrichment FDR = 2.9e-03 and strength = 2.64, high-density lipoprotein particle clearance (*APOE, APOM*) with enrichment FDR = 7.5e-03 and strength = 2.64, humeral immune response (*POMC, F2, C4A LTA*) with enrichment FDR = 1.2e-02 and strength = 1.2, as well as fibrinolysis (*PLAUR, F2*) with enrichment FDR = 1.8e-02 and strength = 2.04.

Particularly, the detected PPI network showed that the novel PWAS risk genes identified by OTTERS were closely interconnected with known risk genes of AD dementia. For example, novel PWAS risk genes *LTA* and *NCR3* were found connected with known GWAS and TWAS risk genes *AIF1* and *HAVCR2*, and were enriched in the GO pathways of positive regulation of immune system process and response to stress. In addition, the known PWAS risk gene of AD dementia *MAPK3* was found connected with 3 novel risk genes *HSPA1A, CD177* and *PLAUR*, and they were enriched in the pathway of positive regulation of immune system process.

Assess mediated causal effects of PWAS risk genes

We employed the PMR-Egger[31] tool to assess if the genetic effects of these 23 PWAS risk genes were mediated through genetically regulated protein abundances and causal for AD dementia, while accounting for possible horizontal pleiotropy effects (Methods). As shown in **Table 1**, we found that 5 out of 8 (62.5%) PWAS risk genes in brain, 7 out of 9 (77.8%) PWAS risk genes in CSF, 6 out of 10 (60%) PWAS risk genes in plasma tissues had significant mediated causal genetic effects with p-values <0.002 (with Bonferroni adjustment for testing 23 genes), while no significant horizontal pleiotropy effects. Additionally, we found 3 PWAS risk genes in plasma with significant mediated causal genetic effects that also had significant horizontal pleiotropy effects with Bonferroni corrected p-values <0.002. These PMR-Egger analyses validated causal genetic effects that were mediated through genetically regulated protein abundances for 14 (60.8%) out of 23 PWAS risk genes in brain, CSF, and plasma tissues.

Discussion

We utilized our recently developed OTTERS[11] tool to conduct PWAS of AD dementia, by leveraging the recently released data resources of summary-level pQTL data of brain, CSF, and plasma tissues[12] as well as GWAS summary data of AD dementia[2]. We identified 23 PWAS risk genes whose genetic effects were potentially mediated through genetically regulated protein abundances, including 8 in brain, 9 in CSF and 10 in plasma tissues. We found OTTERS gained power by leveraging multiple complementary PRS methods to estimate pQTL weights, and by considering reference pQTL data of multiple tissues. Specifically, we showed that each PRS method made distinct and considerable contributions to the final omnibus PWAS results by OTTERS. Through PPI network and enrichment analyses, we found 15 out of these 23 PWAS risk genes were interconnected into one community, including both known AD risk genes and 5 novel PWAS genes. These genes were enriched in important biological pathways associated with AD, including pathways of response to stress, positive regulation of immune system process, positive regulation of macrophage proliferation, high-density lipoprotein particle clearance, and fibrinolysis which was revealed to be abnormal in AD mice[41].

Comparing our PWAS results to previous ones by Wingo et al.[4], which were obtained by using FUSION tool[8] with individual-level proteomics reference data of DLPFC of ROS/MAP cohort[9] and GWAS summary data of AD dementia released in 2019[1], we did not find any overlapped PWAS risk genes. Only ~1,000 proteins were tested by Wingo et al.[4] Next, we compared our findings to the PWAS results by Hu et. al.[10] which were aggregated from the PWAS results by using TIGAR[6], PrediXcan[7], and FUSION[8] tools, with the same individual-level proteomics reference data of DLPFC of ROS/MAP cohort[9] and the same GWAS summary data of AD dementia analyzed in this study. Besides statistical methods used by the FUSION[8] tool, additional statistical methods including the penalized regression with Elastic-Net by PrediXcan[7] and DPR by TIGAR[6] were also considered by Hu et. al.[10] Different from FUSION tool that only selects one best performing statistical method, Hu et. al.[10] aggregated the PWAS results by all three tools for a total of 6,673 proteins. We found 7 overlapped risk genes, *C4A* and *APOE* in brain, *AIF1* and *F2* in CSF, *APOM* in plasma, and *BCAM* in both CSF and plasma, which were also identified in DLPFC by Hu et. al.[10]

The distinction of our PWAS analyses from previous ones lies in the use of summary-level pQTL reference data from multiple tissues (brain, CSF, and plasma) related to

neurodegenerative disorders. CSF and plasma are two bodily fluids believed to contain the richest source of biomarkers of AD and play important roles in research of AD pathology[42]. CSF surrounds the central nervous system (CNS), and is a highly representative and obtainable fluid for detecting brain pathologies. Blood plasma contains proteins that affect brain functions from the periphery, as well as proteins exported from the brain[42]. Especially, recent studies have shown that amyloid beta and phosphorylated tau presenting in CSF and plasma could be used as biomarkers for detecting AD dementia in early stages[13–16]. Therefore, our PWAS results leveraging proteomics data of plasma and CSF tissues in addition to brain are expected to reveal important risk genes of AD mediated through protein abundances in biofluids, providing valuable insights into future biomarker discovery of AD dementia.

In this work, we provide a list of 23 potential risk genes of AD dementia whose genetic effects are mediated through their protein abundances in at least one of three tissues (8 in brain, 9 in CSF, and 10 in plasma). Previous studies have highlighted notable biological roles for these PWAS risk genes, including functions in the immune system processes and lipoprotein metabolism. For example, *APOE*, known as a major risk for AD dementia[43] was identified by our PWAS, which has a crucial function in the central nervous system[44]. The risk gene *APOM* is an apolipoprotein associated with AD dementia and has been implicated in the lipid processing pathway[45]. The risk gene *C4A*, an immune gene, exhibited increased expression in the cortex of the mouse model as Aβ amyloidosis progressed, which suggests an association between *C4A* and AD progression[46]. The risk gene *AIF1* has been associated with the activation of microglia (a type of immune cell localized throughout the central nervous system)[47], which is a key player in the response to central nervous disorders such as AD[48]. The risk gene *MAPK3* involved in the immune system process was found activated in AD brains and involved in the pathogenesis of AD including tau phosphorylation and amyloid deposition[49].

Importantly, our PWAS identified 7 novel genes that are not previously undetected by GWAS, TWAS, or PWAS — *LTA* and *HSPA1A* in brain; *CLIC1* and *FOLH1* in CSF; *NCR3*, *CD177*, and *PLAUR* in plasma. The inflammatory protein TNFB encoded by the novel PWAS risk gene *LTA* has been associated with cognitive function and risks for AD dementia[50]. A study of brain proteomics data discovered that the novel PWAS risk gene *HSPA1A* might serve as a potential biomarker in monitoring the progression of mild cognitive impairment to AD[51]. A study highlighted the role of *CLIC1* in the neurodegenerative process through the regulation of

microglial activation and oxidative stress, which are key factors in the pathogenesis of AD[52]. The PWAS risk gene *FOLH1* has been associated with AD by the Summary Mendelian Randomization (SMR) test[53]. The increased expression of gene *CD177* was found in mild AD patients, which played a role in neutrophil activation[54]. An inverse correlation between soluble PLAUR levels and AD, along with brain atrophy has been observed[55]. Four of these novel PWAS risk genes were validated by PMR-Egger (*HSPA1A, CLIC1, CD177,* and *PLAUR),* with casual genetic effects of AD dementia mediated through protein abundances. Five of these novel PWAS risk genes were identified within in the PPI network (*LTA, HSPA1A, NCR3, CD177,* and *PLAUR*), and were enriched in the pathway of response to stress.

These previous findings, Mendelian Randomization analysis results by PMR-Egger, and PPI network analysis results demonstrated the significance of our identified PWAS risk genes in three tissues. Further experimental studies about the functions of our findings are essential but out of the scope of this work.

The PWAS analysis by the OTTERS tool still has its limitations. First, we only consider *cis*-pQTL within the $\pm 1Mb$ region around the transcription starting/termination sites of the corresponding protein-coding gene. Second, the two-stage PWAS cannot account for possible horizontal pleiotropy genetic effects (those directly affecting the phenotype of interest), when testing if the genetic effects are mediated through genetically regulated protein abundances. Although the PMR-Egger tool can account for horizontal pleiotropy genetic effects, the computation burden of the PMR-Egger tool impedes its applications to test the protein-coding genes of proteome-wide proteins (average 60 vs. 2 CPU minutes per protein coding gene by OTTERS). Thus, we only applied the PMR-Egger[31] tool to our identified PWAS risk genes, and found that 5 out of 8 (62.5%) risk genes in brain, 7 out of 9 (77.8%) risk genes in CSF, 6 out of 10 (60%) risk genes in plasma had significant causal genetic effects mediated through genetically regulated protein abundances.

In conclusion, we presented the first PWAS analysis of AD dementia utilizing the summary-level pQTL reference data of multiple tissues related to neurodegenerative disorders. Our identified PWAS risk genes provide candidates in biofluids such as CSF and plasma and brain tissues for follow-up functional experiments and targeted therapeutic developments of AD dementia. Additionally, this study showed the practical usefulness of the OTTERS tool for

leveraging publicly available summary-level pQTL data and GWAS data resources to conduct PWAS of complex diseases.

Data availability

GWAS summary data of AD dementia is available from[2]. Summary-level pQTL data of brain, CSF, and plasma tissues can be accessed by emailing niagads@pennmedicine.upenn.edu to set up an FTP transfer of the data. OTTERS tool is available from https://github.com/daiqile96/OTTERS. PMR tool is available from https://github.com/yuanzhongshang/PMR. The code used in this study for conducting PWAS of AD dementia are available from GitHub https://github.com/tingyhu45/PWAS_OTTERS. Trained pQTL weights by five PRS methods and our PWAS summary data will be deposited to SYNAPSE once this work is accepted.

Acknowledgements

The authors would like to thank the ROS/MAP studies for providing whole genome sequencing data (available with approved access from https://doi.org/10.7303/syn10901595) that are used to generate reference LD matrices for implementing the OTTERS tool. All data used in this study are de-identified and summary-level data, which are not considered as human data per NIH guidelines.

Funding

This work was supported by the National Institutes of Health (NIH), National Institute of General Medical Sciences (NIGMS, R35GM138313, for T.H., Q.D., and J.Y), and National Institute on Aging (NIA, AG071170, for Q.D. and M.P.E.).

Competing interests

The authors declare no competing financial interests relative to the present study.

References

1. Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet*. 2019;51(3):404-413.

2. Wightman DP, Jansen IE, Savage JE, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat Genet*. 2021;53(9):1276-1282.

3. Bellenguez C, Küçükali F, Jansen IE, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet*. 2022;54(4):412-436.

4. Wingo AP, Liu Y, Gerasimov ES, et al. Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. *Nat Genet*. 2021;53(2):143-146.

5. Brandes N, Linial N, Linial M. PWAS: proteome-wide association study—linking genes and phenotypes by functional variation in proteins. *Genome Biol*. 2020;21(1):173.

6. Nagpal S, Meng X, Epstein MP, et al. TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. *Am J Hum Genet*. 2019;105(2):258-266.

7. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47(9):1091-1098.

8. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. 2016;48(3):245-252.

9. Bennett DA, Buchman AS, Boyle PA, et al. Religious Orders Study and Rush Memory and Aging Project. Perry G, Avila J, Moreira PI, Sorensen AA, Tabaton M, eds. *J Alzheimers Dis*. 2018;64(s1):S161-S189.

10. Hu T, Parrish RL, Dai Q, et al. *Omnibus Proteome-Wide Association Study (PWAS-O) Identified 43 Risk Genes for Alzheimer's Disease Dementia*. medrxiv; 2022.

11. Dai Q, Zhou G, Zhao H, et al. OTTERS: a powerful TWAS framework leveraging summary-level reference data. *Nat Commun*. 2023;14(1):1271.

12. Yang C, Farias FHG, Ibanez L, et al. Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders. *Nat Neurosci*. 2021;24(9):1302-1312.

13. Paraskevas GP, Kapaki E. Cerebrospinal Fluid Biomarkers for Alzheimer's Disease in the Era of Disease-Modifying Treatments. *Brain Sci*. 2021;11(10):1258.

14. Bouwman FH, Frisoni GB, Johnson SC, et al. Clinical application of CSF biomarkers for Alzheimer's disease: From rationale to ratios. *Alzheimers Dement Diagn Assess Dis Monit*. 2022;14(1):e12314.

15. Pais MV, Forlenza OV, Diniz BS. Plasma Biomarkers of Alzheimer's Disease: A Review of Available Assays, Recent Developments, and Implications for Clinical Practice. *J Alzheimers Dis Rep*. 2023;7(1):355-380.

16. Altomare D, Stampacchia S, Ribaldi F, et al. Plasma biomarkers for Alzheimer's disease: a field-test in a memory clinic. *J Neurol Neurosurg Psychiatry*. 2023;94(6):420-427.

17. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009;460(7256):748-752.

18. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. *Genet Epidemiol*. 2017;41(6):469-480.

19. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Ser B Methodol*. 1996;58(1):267-288.

20. Zhou G, Zhao H. A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics. Speed D, ed. *PLOS Genet*. 2021;17(7):e1009697.

21. Zeng P, Zhou X. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nat Commun*. 2017;8(1):456.

22. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun*. 2019;10(1):1776.

21. Tang S, Buchman AS, De Jager PL, et al. Novel Variance-Component TWAS method for studying complex human diseases with applications to Alzheimer's dementia. *PLOS Genet*. 2021;17(4):e1009482.

24. Liu Y, Chen S, Li Z, et al. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet*. 2019;104(3):410-421.

25. Gold L, Ayers D, Bertino J, et al. Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. Gelain F, ed. *PLoS ONE*. 2010;5(12):e15004.

26. Devlin B, Roeder K, Wasserman L. Genomic Control, a New Approach to Genetic-Based Association Studies. *Theor Popul Biol*. 2001;60(3):155-166.

27. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607-D613.

28. Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. 2021;49(D1):D605-D612.

29. Szklarczyk D, Kirsch R, Koutrouli M, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res*. 2023;51(D1):D638-D646.

30. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*. 2019;47(D1):D330-D338.

31. Yuan Z, Zhu H, Zeng P, et al. Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nat Commun*. 2020;11(1):3861.

32. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(D1):D1005-D1012.

33. Chung J, Das A, Sun X, et al. Genome-wide association and multi-omics studies identify MGMT as a novel risk gene for Alzheimer's disease among women. *Alzheimers Dement J Alzheimers Assoc*. 2022;19(3):896-908.

34. Jansen IE, Van Der Lee SJ, Gomez-Fonseca D, et al. Genome-wide meta-analysis for Alzheimer's disease cerebrospinal fluid biomarkers. *Acta Neuropathol (Berl)*. 2022;144(5):821-842.

35. Wang H, Yang J, Schneider JA, et al. Genome-wide interaction analysis of pathological hallmarks in Alzheimer's disease. *Neurobiol Aging*. 2020;93:61-68.

36. Luningham JM, Chen J, Tang S, et al. Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. *Am J Hum Genet*. 2020;107(4):714-726.

37. 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*. 2019;9:653.

38. Marioni RE, Harris SE, Zhang Q, et al. GWAS on family history of Alzheimer's disease. *Transl Psychiatry*. 2018;8(1):99.

39. Chung J, Wang X, Maruyama T, et al. Genome-wide association study of Alzheimer's disease endophenotypes at prediagnosis stages. *Alzheimers Dement*. 2018;14(5):623-633.

40. Noordam R, Bos MM, Wang H, et al. Multi-ancestry sleep-by-SNP interaction analysis in 126,926 individuals reveals lipid loci stratified by sleep duration. *Nat Commun*. 2019;10(1):5121.

41. Cortes-Canteli M, Paul J, Norris EH, et al. Fibrinogen and β-Amyloid Association Alters Thrombosis and Fibrinolysis: A Possible Contributing Factor to Alzheimer's Disease. *Neuron*. 2010;66(5):695-709.

42. Aluise CD, Sowell RA, Butterfield DA. Peptides and proteins in plasma and cerebrospinal fluid as biomarkers for the prediction, diagnosis, and monitoring of therapeutic efficacy of Alzheimer's disease. *Biochim Biophys Acta BBA - Mol Basis Dis*. 2008;1782(10):549-558.

43. Kim J, Basak JM, Holtzman DM. The Role of Apolipoprotein E in Alzheimer's Disease. *Neuron*. 2009;63(3):287-303.

44. Saunders AM, Strittmatter WJ, Schmechel D, et al. Association of apolipoprotein E allele 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*. 1993;43(8):1467-1467.

45. Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet*. 2019;51(3):414-430.

46. Castillo E, Leon J, Mazzei G, et al. Comparative profiling of cortical gene expression in Alzheimer's disease patients and mouse models demonstrates a link between amyloidosis and neuroinflammation. *Sci Rep*. 2017;7(1):17762.

47. De Leon-Oliva D, Garcia-Montero C, Fraile-Martinez O, et al. AIF1: Function and Connection with Inflammatory Diseases. *Biology*. 2023;12(5):694.

48. Woodburn SC, Bollinger JL, Wohleb ES. The semantics of microglia activation: neuroinflammation, homeostasis, and stress. *J Neuroinflammation*. 2021;18(1):258.

49. Zhu X, Lee H gon, Raina AK, Perry G, Smith MA. The Role of Mitogen-Activated Protein Kinase Pathways in Alzheimer's Disease. *Neurosignals*. 2002;11(5):270-281.

50. Chen J, Doyle MF, Fang Y, et al. Peripheral inflammatory biomarkers are associated with cognitive function and dementia: Framingham Heart Study Offspring cohort. *Aging Cell*. 2023;22(10):e13955.

51. Muraoka S, Jedrychowski MP, Yanamandra K, et al. Proteomic Profiling of Extracellular Vesicles Derived from Cerebrospinal Fluid of Alzheimer's Disease Patients: A Pilot Study. *Cells*. 2020;9(9):1959.

52. Averaimo S, Milton RH, Duchen MR, Mazzanti M. Chloride intracellular channel 1 (CLIC1): Sensor and effector during oxidative stress. *FEBS Lett*. 2010;584(10):2076-2084.

53. Nazarian A, Arbeev KG, Yashkin AP, Kulminski AM. Genetic heterogeneity of Alzheimer's disease in subjects with and without hypertension. *GeroScience*. 2019;41(2):137-154.

54. Le Page A, Lamoureux J, Bourgade K, et al. Polymorphonuclear Neutrophil Functions are Differentially Altered in Amnestic Mild Cognitive Impairment and Mild Alzheimer's Disease Patients. Fiala M, ed. *J Alzheimers Dis*. 2017;60(1):23-42.

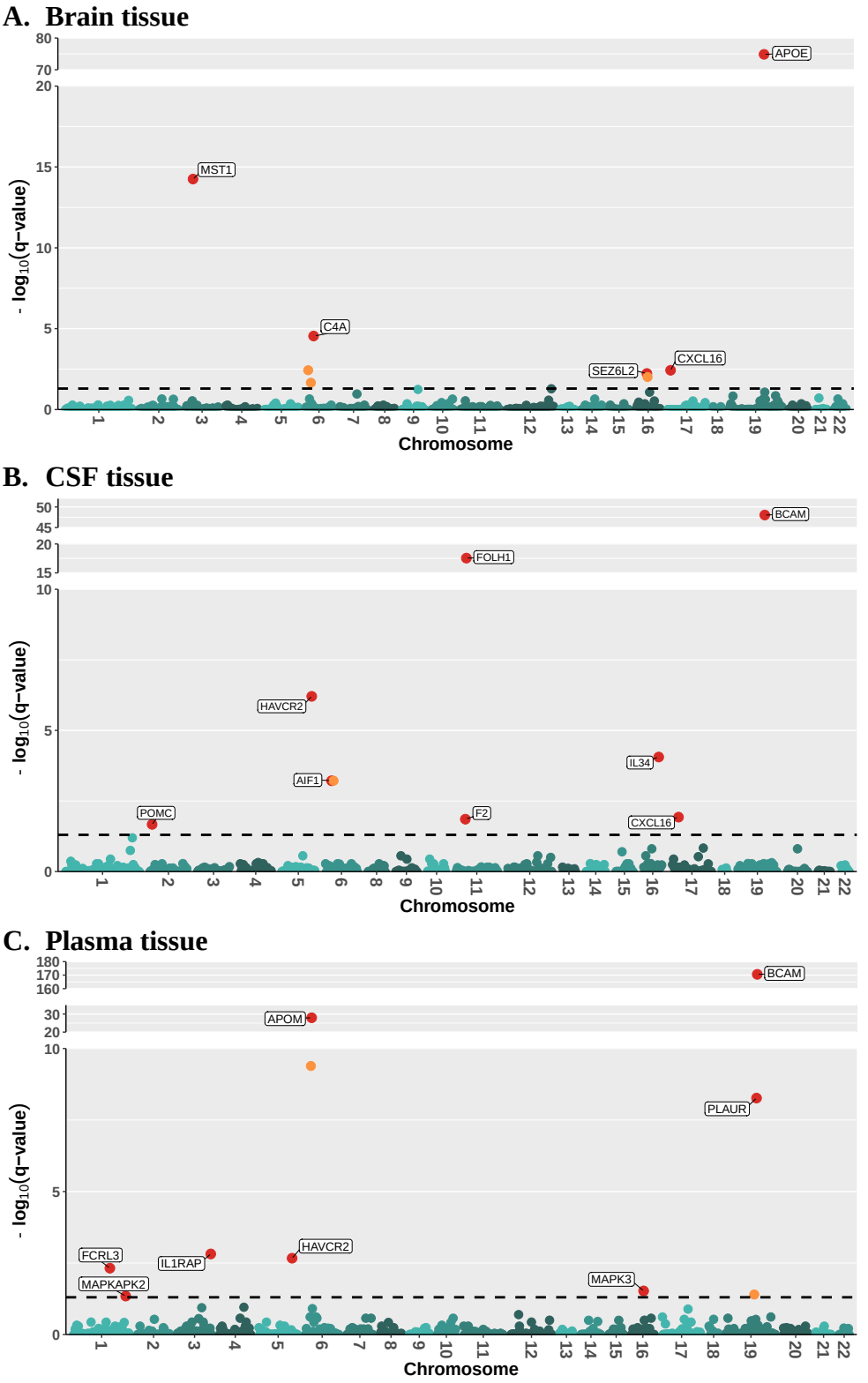
55. Greco I, Day N, Riddoch-Contreras J, et al. Alzheimer's disease biomarker discovery using in silico literature mining and clinical validation. *J Transl Med*. 2012;10(1):217.

Figure legends

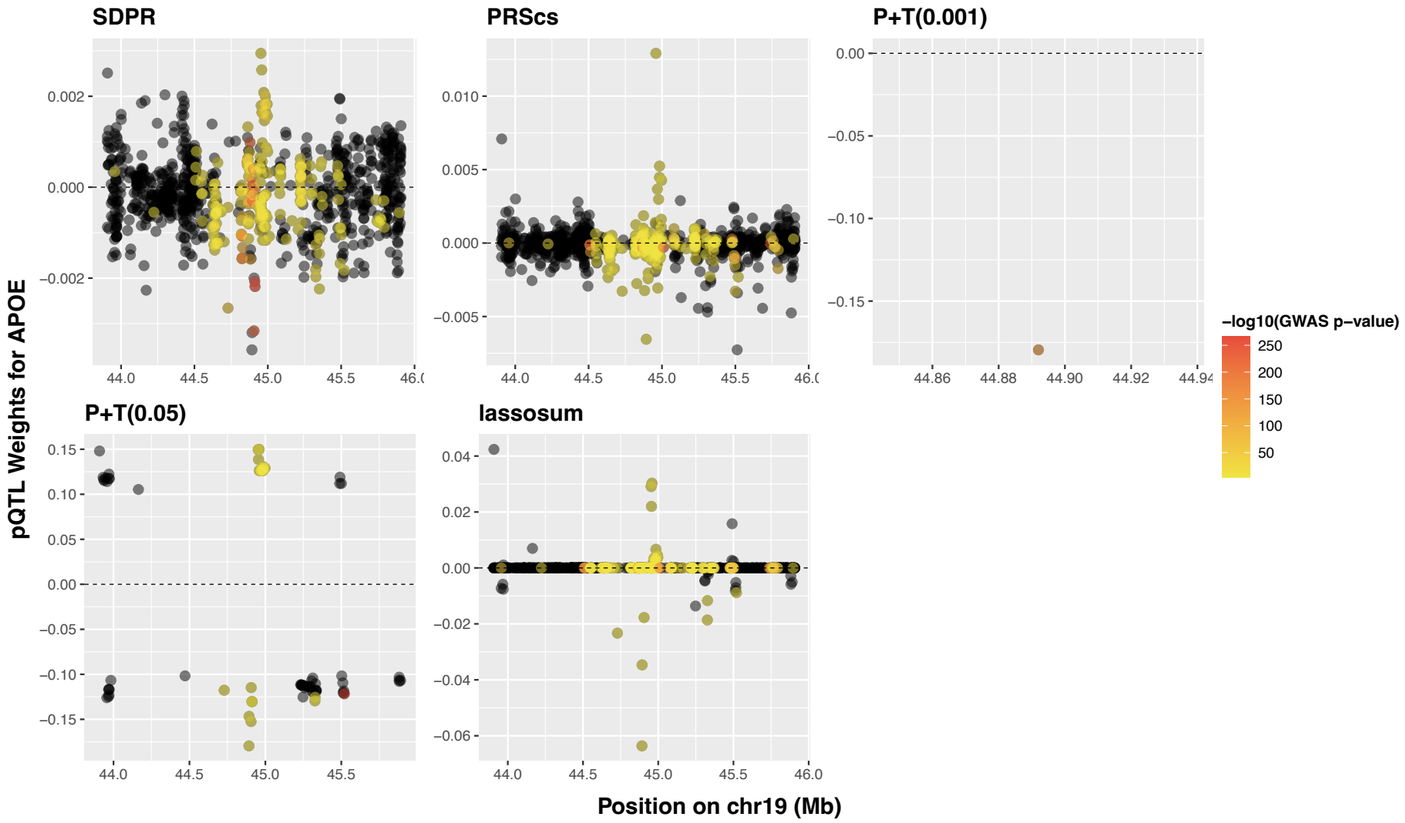
Figure 1 Manhattan plots of PWAS results (FDR q-values) of AD dementia by OTTERS in Brain (A), Plasma (B), and CSF (C) tissues. The -log10(q-values) were plotted on the y-axis, and -log10(0.05) was plotted as the dashed horizontal line. Independent significant genes are labeled.

Figure 2 Scatter plots of pQTL weights of example PWAS risk genes of *APOE* in brain, *BCAM* in CSF, and *CD177* in plasma that were estimated by individual PRS methods. The pQTL weights were plotted in the y-axis for all test genetic variants in the test gene region, with color-coded with respect to -log10 (GWAS p-value). Test SNPs with GWAS p-value $<10^{-5}$ were colored.

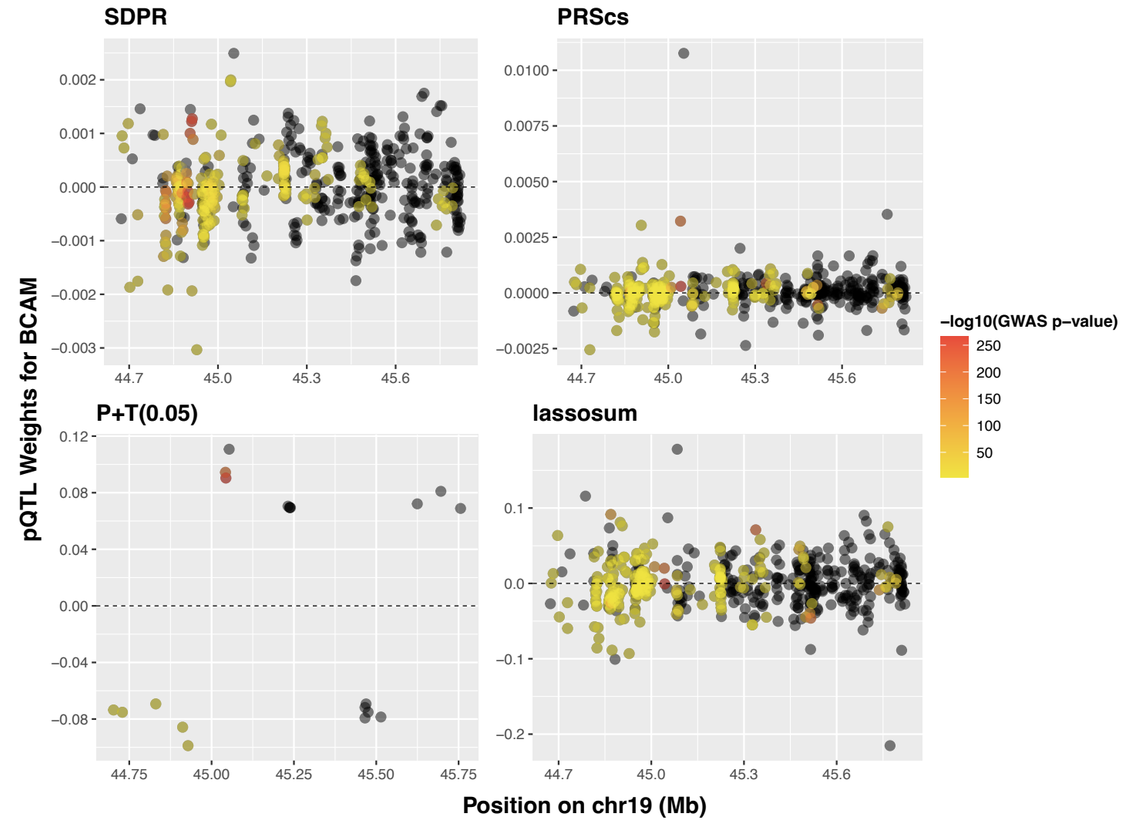
Figure 3 PPI network and enrich analyses results with 23 PWAS risk genes of AD dementia by STRING. Edges represent physical PPI, with different colors representing different sources of connection evidences. Node colors represent different enriched GO terms with FDR < 0.05.


A. Brain tissue
APOE
MST1
C4A
SEZ6L2
CXCL16
- log10(q−value)
Chromosome
B. CSF tissue
BCAM
FOLH1
HAVCR2
IL34
AIF1
POMC
F2
CXCL16
- log10(q−value)
Chromosome
C. Plasma tissue
BCAM
APOM
PLAUR
FCRL3
IL1RAP
HAVCR2
MAPKAPK2
MAPK3
- log10(q−value)
Chromosome

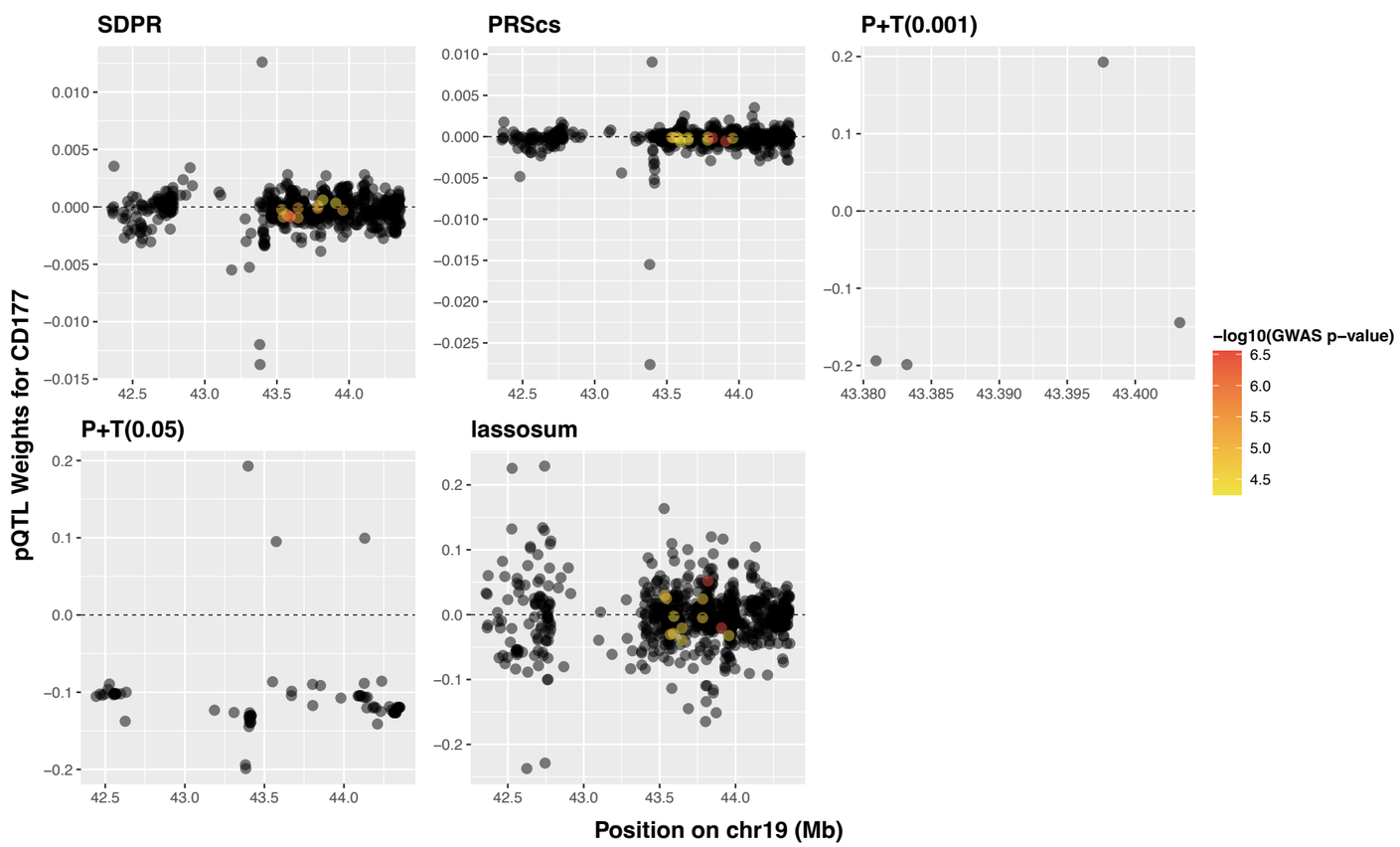
A. *APOE* (brain)



B. *BCAM* (CSF)

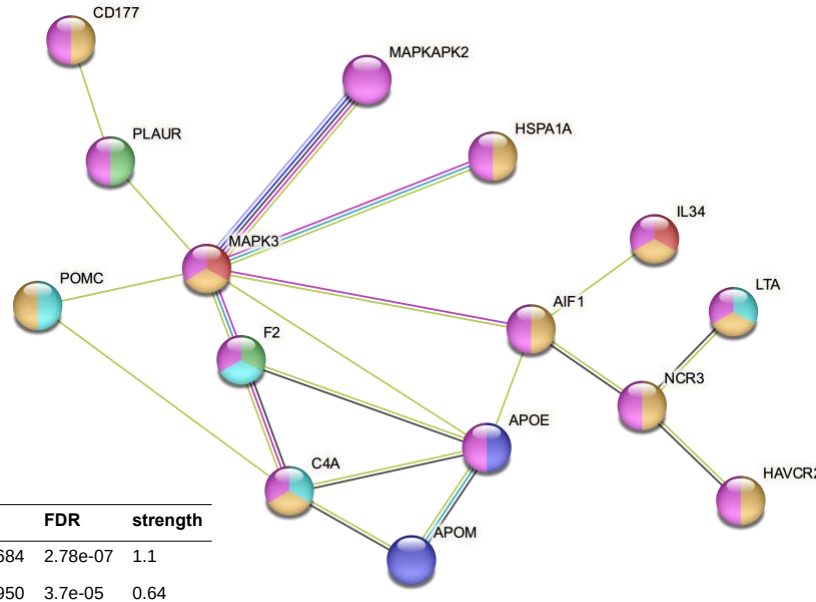


C. *CD177* (plasma)



Connections:

- *From curated databases*
- *Experimentally determined*
- *Textmining*
- *Co-expression*
- *Gene co-occurrence*
- *Protein homology*



Gene Ontology pathways

Biological process	GO term	FDR	strength
positive regulation of immune system process	GO: 0002684	2.78e-07	1.1
response to stress	GO: 0006950	3.7e-05	0.64
positive regulation of macrophage proliferation	GO: 0120041	2.9e-03	2.64
high-density lipoprotein particle clearance	GO: 0034384	7.5e-03	2.34
humoral immune response	GO: 0006959	1.2e-02	1.2
fibrinolysis	GO: 0042730	1.8e-02	2.04