CSF proteogenomics implicates novel proteins and humoral immunity in Alzheimer's disease risk

Authors

Aliza P. Wingo[1,2,#], Yue Liu[3], Zhen Mei[3], Michael Wang[2], Selina M. Vattathil[3], Ekaterina S. Gerasimov[2], Anantharaman Shantaraman[4,5], Fang Wu[4,5], Duc M. Duong[4,5], Edward J. Fox[4,5], Alzheimer's Disease Neuroimaging Initiative[*], David A. Bennett[6], Allan I. Levey[5,7,8], Nicholas T. Seyfried[4,5,7,8], Thomas S. Wingo[3,9,#]

Affiliations

[1]VA Northern California Healthcare System, Sacramento, CA, USA

[2]Department of Psychiatry, University of California, Davis, Sacramento, CA, USA

[3]Department of Neurology, University of California, Davis, Sacramento, CA, USA

[4]Department of Biochemistry, Emory University School of Medicine, Atlanta, GA, USA

[5]Center for Neurodegenerative Disease, Emory University School of Medicine, Atlanta, GA, USA

[6]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois, USA

[7]Goizueta Alzheimer's Disease Research Center, Emory University School of Medicine, Atlanta, GA, USA

[8]Department of Neurology, Emory University School of Medicine, Atlanta, GA, USA

[9]Alzheimer's Disease Research Center, University of California, Davis, Sacramento, USA

[#]Corresponding author

[*]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Corresponding Authors

[#]Corresponding Authors, Aliza P. Wingo, apwingo@health.ucdavis.edu and Thomas S. Wingo, twingo@health.ucdavis.edu

Abstract

We profiled 2,961 cerebrospinal fluid (CSF) proteins in 1,005 participants of the Alzheimer's Disease Neuroimaging Initiative (ADNI), including 1,066 proteins not measured in prior studies, using mass spectrometry (MS). We mapped protein quantitative trait loci (pQTLs) in CSF, compared them with brain and plasma pQTLs, and integrated them with Alzheimer's disease (AD) genome-wide association study (GWAS) data. We identified 1,417 index *cis* pQTLs for 654 unique genes and 130 index *trans* pQTLs for 94 unique genes. Cross-tissue and cross-proteomic-platform comparisons show broad consistency between MS-based CSF pQTLs and MS-based brain pQTLs as well as affinity-based CSF and plasma pQTLs. Lastly, through integrating CSF pQTLs with the largest AD GWAS, we identified 24 candidate AD causal proteins in CSF, including 10 novel and 14 previously identified in either brain, CSF, or plasma using similar approaches. These CSF AD candidate causal proteins are involved in immune response – notably humoral immunity (3 of 24) – that expands the role of the immune system in AD beyond innate immunity, as well as lysosomal function and neurovascular growth and remodeling. Together, our findings provide novel insights into AD biology and new targets for biomarker and therapeutic development.

Introduction

Genome-wide association studies (GWAS) have identified many regions associated with diseases, yet for most loci, the underlying biological mechanisms remain unknown. A prevalent hypothesis supported by considerable evidence[1] is that GWAS variants influence disease by altering gene expression. Functional genomics tests this hypothesis by mapping quantitative trait loci (QTLs) – genetic variants that regulate molecular phenotypes like gene expression (eQTLs) or protein abundance (pQTLs). Integration of QTL data with GWAS results can reveal whether disease associations at a locus are explained by effects on gene expression. Approaches like Summary-based Mendelian Randomization (SMR)[2] and colocalization[3,4] can identify genes whose genetically regulated expression is consistent with a causal role in disease from those with coincidental overlap due to linkage disequilibrium (LD). These integration methods point us to future testable hypotheses about molecular function and therapeutic targets for disease.

Tissue-specific functional genomic studies provide important insight into GWAS results because genetic regulation of gene expression varies across tissues, cell types, and developmental stages. While plasma proteomic studies have profiled tens of thousands of individual samples and provide exceptional power to resolve pQTLs[5,6], plasma likely offers limited insights into brain-specific processes. Conversely, functional proteomics of brain tissue have illuminated novel molecular mechanisms in Alzheimer's disease and other neurological conditions[7–10], but the sample sizes are constrained by availability[9,11]. Cerebrospinal fluid (CSF) represents an intermediate for investigating brain-related traits and diseases because it directly interfaces with brain tissue, reflects neurological and neurodegenerative disease processes, and can be obtained from living individuals.

Large-scale functional genomic studies of CSF proteomics to date have relied predominantly on affinity-based platforms – either aptamer-based (SomaLogic)[9,12,13] or antibody-based (Olink)[14] – that target predefined proteins. By contrast, bottom-up mass spectrometry-based (MS) proteomics enables untargeted protein identification and quantification, revealing proteins not captured by existing affinity panels.

Here, we present a large-scale functional genomics study of CSF using tandem mass tag mass spectrometry (MS) based proteomics in 1,005 individuals (**Figure 1**). Using an untargeted approach, we identified 2,961 proteins, including 1066 not measured in prior affinity-based CSF studies. We performed *cis* and *trans* pQTL mapping, sex-stratified pQTL testing, and fine mapping of *cis* pQTLs that allow us to compare MS-based CSF pQTLs with MS-based brain pQTLs and affinity-based CSF and plasma pQTLs. Given the proximity of CSF to brain and ability to collect CSF in living participants, we compared differential expression of MS-based proteins between control and Alzheimer's disease (AD) dementia in CSF and brain. Finally, integrating MS-based CSF pQTLs with AD GWAS using proteome-wide association study (PWAS)[4] and SMR Mendelian randomization[2] revealed 24 putatively AD causal proteins, of which 14 were also identified in analogous studies in brain, CSF, or plasma, and 10 were novel, providing new insights into AD pathogenesis.

Results

A. CSF *cis* pQTL mapping and fine mapping

We profiled CSF proteomes of 1,005 ADNI participants of European ancestry, among whom 45% were female and the mean age was 73.1 years (Table S1). A total of 2,961 CSF proteins were considered for *cis* pQTL mapping after quality control and regressing out effects of proteomic sequencing batch, age, clinical diagnosis, and surrogate variables. We defined *cis* pQTLs as SNPs located within 500 Kb up or downstream of the corresponding protein-coding genes that have SNP-protein pair association at $p < 5 \times 10^{-8}$. We identified 67,296 *cis* pQTLs, and, after clumping for linkage disequilibrium (LD) r^2<0.1 to keep approximately independent sites, there were 1,417 index pQTLs for 654 unique genes, with a median of 1 SNP per gene (range 1-13, interquartile range [IQR] of 2; Table S2-3).

To pinpoint putative causal pQTLs underlying the expression of CSF proteins, we performed pQTL fine mapping with SuSiE[15] to identify credible sets, which are the smallest sets of variants that have 95% probability of including causal pQTLs for these CSF proteins. We identified 1,092 credible sets for 739 proteins, with a median number of 6 SNPs per credible set. The higher number of proteins with credible sets compared to those with index pQTLs is due to the conservative p-value threshold used to

define a *cis* pQTL. Imposing a posterior inclusion probability (PIP) of >50% on pQTLs in credible sets, we identified 457 putative causal pQTLs for 346 unique proteins (Table S4).

To determine whether sex has an influence on the genetic regulation of protein expression in CSF, we performed sex-stratified *cis* pQTL mapping. We detected 30,383 *cis* pQTLs in males and 38,058 *cis* pQTLs in females. Among the cis pQTLs from the joint and sex-stratified pQTL mapping, we examined whether they interact with sex. We identified 2 pQTLs with significant interaction with sex (i.e., $sex \times pQTL$ term had FDR <0.05; Table S5). These sex-biased CSF pQTLs were associated with expression of pregnancy zone protein (PZP).

B. CSF *trans* pQTL mapping

CSF *trans* pQTL were defined as pQTLs within ± 500 kb of the gene and associated with its protein abundance at $p < 1.70 \times 10^{-11}$ ($p < 5 \times 10^{-8} \div 2931\ genes$). With this threshold, there were 8,411 *trans* pQTLs, corresponding to 130 index *trans* pQTLs (after clumping at LD r^2 <0.1) for 94 unique genes (*trans* pGene; **Figure 2**; Table S6-8). There were 4 proteins (CPN2, CTSS, LAMA2, and LAMB1) that have *trans* pQTLs located on different chromosomes.

The *APOE* locus was the *trans* pQTL for SMOC1 and the 14-3-3 family of proteins (YWHAE, YWHAG, and YWHAZ), which have been implicated in early AD pathophysiology and neurodegeneration[16] (Table S8). Notably, *trans* pQTL mapping stratified by disease status (AD vs. MCI vs. control) revealed that the association between *APOE* and SMOC1 and 14-3-3 proteins, respectively, depended on AD disease state (Table S9). Further stratification by CSF beta-amyloid and tau levels using ADNI thresholds[17] (i.e., A+T+ vs. A-T-) eliminated the APOE association with all four proteins (Table S9). Together, results from these stratification analyses suggest that the trans effect of *APOE* on SMOC1 and the 14-3-3 family of proteins are likely confounded by AD disease status.

We compared our CSF trans pQTLs to those identified in a recent aptamer-based CSF pQTL study, which also identified APOE as a *trans* pQTL for SMOC1 and the 14-3-3 proteins[13]. Their findings were consistent with our unstratified results. By contrast, that study[13] reported 335 proteins with trans pQTLs within ± 1Mb of the APOE, a markedly larger number than SMOC1 and 14-3-3 proteins

suggesting differences in platforms, statistical power, or analytic design (e.g., our decision to regress out the effect of surrogate variables before pQTL mapping to account for potentially hidden confounders). Regarding concordance between ours and this published CSF pQTL study, among the 94 unique *trans* pGenes we identified, 55 were also trans pGenes in that study that could be used for comparison[13]. We found that 32.7% (18 of 55) of these trans pGenes have a *trans* pQTL in at least moderate LD ($r^2 > 0.3$) with each other, suggesting modest overall agreement.

C. Cross tissue and platform pQTL comparisons

We compared the genetic control of protein expression in different tissues – brain, CSF, and plasma - using our CSF pQTLs and published pQTLs. In addition to tissue difference, proteomic data in published studies were profiled using different technologies – TMT-MS (brain pQTLs)[10], SOMAscan (CSF pQTLs[13] and plasma pQTLs[5]), and Olink (plasma pQTLs)[6] (**Figure 3a**). Specifically, TMT-MS profiled 2961 proteins in CSF and 9,980 in brain[10] after quality control and removing proteins with missing data. The aptamer-based approach profiled 6,133 protein in CSF[13] and 4,656 proteins in plasma[5], and antibody-based profiling measured 2460 proteins in plasma[6] in the published studies used for comparison (**Figure 3a**). In CSF, between MS- and aptamer-based CSF proteomes, there were 1895 proteins measured in common, reflecting 64% (1895 of 2961) of the MS-based CSF proteome and 31% of the aptamer-based CSF proteome (**Figure 3b**). There were 1,066 proteins measured only in MS-based CSF proteomes, revealing the value of MS-based CSF profiling in providing novel biological insight (**Figure 3b**).

Regarding different tissues, between the CSF and brain MS-based proteomes, there were 2,200 proteins measured in common, reflecting 74% of the MS-based CSF proteome and 22% of the MS-based brain proteome (**Figure 3c;** Table S10). Between CSF and plasma, the MS-based CSF proteome measured 848 proteins not profiled in either CSF or plasma by other technologies (**Figure 3d**), again highlighting the additional insight into biology that MS-based CSF proteome can provide.

We examined π_1 rates for pQTL mapping between this study (discovery set) and other published pQTL studies (replication set). The π_1 statistic[18] estimates the rate of pQTLs identified in the discovery

dataset that are also pQTLs in the replication dataset among the proteins profiled in both sets (**Figure 3e**). For the 2,200 proteins profiled in both CSF and brain with MS-based proteomics, we found a moderately high π_1 rate of 0.83 for *cis* CSF pQTLs. Among the 1,895 proteins profiled in both MS- and aptamer-based CSF proteomes, we found a moderate π_1 rate of 0.73 for *cis* pQTLs and lower π_1 rate of 0.56 for trans pQTLs. For the 1,727 proteins profiled by MS-based CSF and aptamer-based plasma, there was a similar pattern of moderate π_1 rate of 0.78 for cis pQTLs and lower π_1 rate of 0.42 for trans pQTLs (Figure 3e). Interestingly, for the 1,057 proteins profiled by MS-based CSF and antibody-based plasma, we found comparatively high π_1 rate of 0.97 and 0.85 for cis and trans pQTLs, respectively (Figure 3e).

Furthermore, we tested which CSF *cis* pQTLs are also brain *cis* pQTLs by colocalization analysis[3]. Among the 654 CSF genes with at least one cis pQTLs, 445 genes were profiled in a published brain pQTL study[10]. Using COLOC[3], we found that 35% of the pGenes (159 of 445) showed evidence of sharing the same pQTLs between brain and CSF (Table S11).

D. Identifying AD candidate causal proteins in CSF

To identify proteins in CSF that are consistent with a causal role in AD, we integrated the largest AD GWAS results[19] with CSF proteomic and genetic data via a proteome-wide association study (PWAS) using FUSION[4]. The goal of the PWAS was to identify proteins whose genetically regulated expression levels in CSF are associated with AD. In the PWAS, we first delineated the proteins with significant genetic regulation of their CSF abundance level (i.e., heritability $p < 0.01$) and determined the effect size of their genetic regulation (also referred to as “protein weights”). We found 819 proteins with heritable abundance and integrated them with AD GWAS summary statistics, resulting in 41 proteins with genetic evidence for association with AD (Table S12). Among these 41 proteins, we performed Mendelian randomization using AD GWAS results and the CSF *cis* pQTLs following SMR[2]. The Mendelian randomization is complementary to the PWAS and provides orthogonal confirmation that these proteins causally mediate the AD GWAS signals. Moreover, the HEIDI test in SMR identifies associations likely driven by LD rather than a shared causal variant. Among these 41 PWAS-significant proteins, 24 were

significant in SMR and passed the HEIDI test, suggesting a causal role in AD (**Table 1**; Table S13). Among these 24 AD candidate causal proteins, 11 are novel and 13 have been identified in prior proteogenomic studies in brain (n=8) or CSF (n=10; **Figure 4a**; Table S14)[10,13]. When compared to plasma proteogenomic studies, only one common candidate causal protein was found – PILRB -- leaving 10 novel AD causal proteins identified in the current MS-based CSF proteogenomic study (**Figure S1**). The 24 candidate AD causal proteins identified in CSF are involved in modulating microglia and macrophage activation (PILRA, PILRB, CD300A, TREM2, SIRPB1, SIRPA, and CD33)[20–23], regulating neurovascular growth and remodeling (LGALS3, NTRK1, EGFR, ACE)[24–27], vitamin D metabolism (LRP2, GC), and proteolysis and lysosomal function (CTSH, CLN5, GRN, **Figure 4b-c**)[28–31]. In addition, IGHV3-72, IGHG2, and IGHV2-70 are notably involved in humoral immunity, with IGHV2-70 also previously identified as AD candidate causal protein in brain (**Figure 4a**), linking humoral immunity to AD risk.

To determine whether these CSF AD candidate causal proteins showed altered abundance in AD in CSF or brain, we first performed proteome-wide differential abundance analysis for AD in CSF (using ADNI CSF proteomes, **Figure S2a**) and in brain (using ROS/MAP brain proteomes, **Figure S2b**). In CSF, 297 proteins were differentially abundant in AD compared to controls (FDR < 0.05; n = 491 participants; Table S15), while in brain, 3,038 proteins were differentially abundant in AD (FDR <0.05; n = 553 donors; Table S16). The proteins with altered abundance in AD in CSF were enriched for processes involving axon and neuron generation and development, cell growth, synapse organization, cell-cell signaling, and synaptic signaling (Table S17). There were 117 proteins that had altered abundance in AD in both CSF and brain (**Figure S2c**; Table S18) and they were enriched in processes involving synapse structure and organization, cell junction organization, synaptic signaling, cell-cell signaling, neuron development, and regulation of cellular response to stress based on gene set enrichment analysis (Table S19).

We then examined whether the 24 AD causal proteins identified in CSF are also differentially expressed in AD in CSF or brain. We found that none were differentially abundant in CSF, but two – CRIP1 and TXNDC16 – were differentially abundant in AD in brain tissue. Conversely, among the 64 AD candidate causal proteins previously identified in brain[32], none were differentially abundant in CSF, but 22 were differentially abundant in brain tissue: CCDC6, COG7, CPD, CPLX1, DENR, DST, ELOC,

GPRC5B, ICA1L, MAP2K1, NDUFS2, PPP1R14B, RCSD1, RETREG3, RIPK2, SNX1, STARD3, STIP1, STX6, TMEM106B, TPCN1, and VPS37B.

Discussion

In this study, we conducted deep MS-based proteomic profiling of CSF, including 1066 proteins not measured in prior affinity-based CSF studies, to perform cis and trans pQTL mapping. We found 1,417 index cis pQTLs for 654 unique proteins and 457 putatively causal pQTLs (at PIP > 50%) for 346 unique proteins via fine mapping. We also identified sex-specific CSF pQTLs for PZP. Furthermore, *trans* pQTL mapping identified 130 index *trans* pQTLs for 94 unique proteins. Cross-tissue comparisons highlighted relative consistency of the CSF pQTLs across different tissues and proteomic platforms. Finally, through PWAS, we sought to resolve AD dementia GWAS signals to CSF proteins and identified 41 CSF proteins genetically linked to AD, of which 24 were confirmed by SMR/HEIDI testing as causally mediating AD GWAS signals. Of the 24, 10 reveal novel insights into AD pathophysiology not found in published proteogenomic studies in CSF, plasma, or brain. Notably, IG heavy chain proteins (IGHG2, IGHV2-70, IGHV3-72) are among these AD candidate causal proteins, expanding the role of the immune system in AD to include humoral immunity.

The 24 candidate AD causal proteins in CSF from this study play important roles in immune response, lysosomal function, and neurovascular growth and remodeling. Strikingly, 10 of these 24 proteins are components of cell-surface immune recognition - CD300A, CD33, IGHG2, IGHV2-70, IGHV3-72, PILRA, PILRB, SIRPA, SIRPB1, and TREM2. Among these are four distinct proteins for two different immune receptor families - PILRA/PILRB and SIRPA/SIRPB1 - with notable role in fine tuning immune response among myeloid cells[20,21]. Moreover, we found two levels of immunoglobulin function implicated by IGHV3-72, IGHG2, and IGHV2-70. IGHG2 defines one of the four major IgG isotypes – IgG2 – which preferentially recognizes polysaccharide antigens[33]. Both IGHV3-72 and IGHV2-70 are germline heavy chain variable segments whose allelic diversity can contribute to population level differences in humoral response[34]. TREM2 is a well-known microglial receptor with a rare missense variant associated with AD[35], and recent work suggests CD33 activation can inhibit TREM2-mediated phagocytosis in microglia[22,23].

Six of the 24 CSF AD candidate causal proteins are linked with lysosomal function. Notable among these is LGALS3, which encodes Galectin-3 and has diverse roles in cell growth, cell-cell adhesion, and phagocytosis by macrophages[25]. GRN encodes progranulin (PGRN), which has critical roles in lysosomal function and autophagy regulation[30]. GRN mutations cause frontotemporal dementia[36], and reduced PGRN levels lead to hyperactivated microglia and are linked to TDP-43 pathology.[30]. Other genes related to lysosomes include CTSH, encoding cathepsins H protease, and CLN5, recently identified as bis(monoacylglycero)phosphate (BMP) synthetase that is essential for normal lysosomal function[31]. CTSH is a cysteine protease found in microglia, with activity positively linked to neuroinflammation and hippocampal atrophy in preclinical models[28,29]. Keeping with receptors and intracellular vesicles, LRP2 encodes low-density lipoprotein (LDL) receptor-related protein 2 and plays an important role in endocytosis across a wide range of ligands[37]. EGFR and NTRK1 are notable for their classically described roles in endothelial and neuronal growth and survival but also contribute to inflammatory response[38–40]. EGFR is a receptor tyrosine kinase traditionally associated with the AKT/PI3K pathway. Its upregulation occurs during neural development, chronic neurodegeneration, and acute neural insult. In glia, EGFR upregulation induces inflammatory cytokines such as IL1β, IL6, and IL18, triggering microglial activation and a neuroinflammatory phenotype. EGFR upregulation and activation have been observed to increase plaque accumulation and enhance the toxicity of existing Aβ42 by hyperphosphorylating tau and promoting additional Aβ42 production[41].

Eight AD candidate causal proteins in CSF were also found using brain tissue in an independent brain proteogenomic study and dataset[32], suggesting these are among the most reliable findings. These 8 proteins (Table 1) function along two major axes: neuroimmune activation (SIRPA, SIRPB1, CTSH, GRN, EGFR, GC) and impaired proteostasis (ACE, GRN, and EGFR), with several proteins contributing to both. Signal Regulator Protein Alpha (SIRPA) and Signal Regulator Protein β1 (SIRPB1) are microglial surface proteins that regulate neuroimmune activation and clearance. Specifically, SIRPA is involved in the "don't eat me" signal, which prevents overactive phagocytosis under homeostatic conditions, while the upregulation of SIRPB1 is associated with increased microglial phagocytosis[42,43]. Vitamin D binding protein (GC) regulates vitamin D bioavailability, which modulates neuroinflammation through multiple

mechanisms including neurotransmitter regulation and immune cell phenotype control[44,45]. Regarding proteostasis, altered protein expression of ACE, GRN, and EGFR together can affect protein aggregation and clearance[26,30,39,46]. The strongest negative association in our AD analysis was with angiotensin-converting enzyme (ACE), a regulator of blood pressure via the renin-angiotensin system and its dysfunction is an established risk factor for hypertension, but ACE is also described to convert Aβ42 into less pathological forms[26,46].

Interpretation of our findings should take into consideration potential limitations of the study. First, the sample size of the CSF samples may limit our ability to detect lower abundant proteins and weaker pQTL signals or sex effects. This is somewhat mitigated by finding of moderate to high π_1 rates between pQTLs for genes measured in other tissues and across platforms. Curiously, we found less evidence for sex-biased pQTLs in CSF compared to earlier studies of brain[11], which might be related to our sample size and breadth of proteins detected. The sex-specific finding of PZP presents an intriguing avenue for future work on the role of biological sex on AD dementia given its roles in protease inhibition and immune regulation[47,48] that is consistent with the major biologic themes identified. Second, the APOE locus had *trans* pQTL for SMOC1 and 14-3-3 proteins, which is consistent with a recent aptamer-based study[13], but our work identified only four trans pQTLs compared to 338 trans pQTLs ±1 Mb of APOE found in that study. This discrepancy is likely related to our use of surrogate variable analysis and regression of disease status from proteomics data, as well as differences in proteomics technology and sample size. Additional CSF datasets with different technologies will help determine whether APOE exerts a broad pleiotropic effect through direct changes in CSF proteins or indirectly through its influence on AD status. Third, participants in our study are of European ancestry with high degrees of educational attainment. Our results ought to be tested for generalizability and future work would benefit from sampling across broader ranges of ancestries and life experiences. Fourth, we acknowledge that the 24 proteins we identified as putatively causal in AD could be influenced by genetic pleiotropy and experimental testing is suggested.

Our study has several strengths. TMT mass spectrometry is an untargeted, discovery-driven approach that does not require prior knowledge or the development of aptamers or antibodies for the

proteins being profiled, enabling the identification of novel proteins. For example, TMT-MS quantified 1066 CSF proteins that were not measured in prior affinity-based CSF proteomic studies. We also systematically compared the MS-based CSF proteome with those generated from CSF, brain, and plasma using a range of proteomic technologies. Our analytical strategy is rigorous, incorporating surrogate variable analysis to minimize potentially hidden confounding variables on pQTL mapping. Furthermore, we performed PWAS followed by Mendelian randomization with SMR/HEIDI to minimize the influence of LD on our findings and to find the most rigorously supported AD causal proteins.

In summary, among 1,005 ADNI participants, we profiled 2,961 CSF proteins, including 1,066 previously unmeasured proteins, using MS-based proteomics. We identified *cis* and *trans* pQTLs and uncovered 24 candidate causal proteins for AD in CSF, including 8 that were previously identified in brain in an independent study. These CSF candidate AD causal proteins are predominantly linked to humoral and innate immune functions and lysosomal functions, presenting viable and promising targets for future therapeutic development.

Methods

Study participants

ADNI: Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). We used phenotypic, genetic, and proteomic data of participants recruited by ADNI. It enrolls participants aged 55 or older, who undergo a series of initial tests that are repeated at intervals over subsequent years, including a clinical evaluation, neuropsychological tests, genetic testing, lumbar puncture, and MRI and PET scans. ADNI has been approved by an Institutional Review Board.

ROS/MAP: Participants were recruited by the Religious Orders Study (ROS) or the Rush Memory and Aging Project (MAP), two ongoing longitudinal cohort studies of aging and dementia approved by an Institutional Review Board of Rush University Medical Center[49,50]. Both studies recruit older adults who enroll without known dementia and agree to annual clinical evaluation and brain donation at death. All participants signed an informed consent, the Anatomical Gift Act, and a repository consent to allow their study data to be repurposed. This study included 770 participants who had tandem mass tag (TMT) proteomics generated from the dorsal lateral prefrontal cortex (DLPFC), as described[51], to compare with the ADNI MS-based CSF proteomics.

Genetic data

ADNI genetic data was from array-based genotyping or whole genome sequencing data (WGS). For array-based genotyping, we performed quality control of the data using PLINK 1.9[52] before performing imputation. We excluded subjects with an overall genotyping missingness > 10% or having mismatch between phenotypic and genotypic sex. For WGS data, we used Bystro[53] to perform initial quality control, including determining 5 standard deviation outliers based on heterozygosity, silent replacement ratio, and genotyping completeness.

We restricted our genetic analysis to individuals with genetic similarity to European reference populations, as this was the only group with sufficient sample size for meaningful results in the ADNI dataset. To determine genetic similarity, we performed multidimensional scaling (MDS) analysis using PLINK 1.9, comparing individual-level genetic data to Phase 3 1000 Genomes European reference samples. For both array-based and WGS data, we included SNPs with MAF > 0.05 that passed LD filtering (100 Kb windows, 5 Kb steps, $r^2 < 0.15$ for array data, $r^2 < 0.05$ for WGS data). We excluded ambiguous variants (AT/TA/GC/CG) and retained only participants clustering within 6 standard deviations of the 1000 Genomes European reference populations.

We imputed array-based genotyping data using the TOPMed[54] reference panel with hg38 coordinates following recommended quality control procedures: updating strand orientation, genomic positions, and reference/alternative allele assignments; removing A/T and G/C SNPs with MAF > 0.40;

and excluding SNPs with differing alleles, allele frequency differences > 0.2, or absence from the reference panel. After imputation using TOPMed default parameters[54,55], we retained SNPs with imputation quality $R^2 > 0.3$. We then merged the imputed array data with WGS data from participants with genetic similarity to European reference populations, prioritizing WGS when both data types were available for the same individual. Post-merger quality control included removing variants with Hardy-Weinberg equilibrium deviation ($p < 1\times10^{-7}$), missing genotype rate > 10%, or MAF < 0.05. Finally, we excluded one individual from each pair of second-degree or closer relatives, retaining the individual with less genotyping missingness.

Proteomic data

Cerebrospinal Fluid Proteomics

Protein Digestion and Isobaric Tandem Mass Tag (TMT) Peptide Labeling: Baseline CSF samples from 1105 ADNI participants were digested as described before[56,57]. To normalize protein quantification across batches, a global internal standard (GIS)[58] was created by aliquoting an equal volume (100 µL) of each CSF sample, post-digestion and HLB clean-up. All individual samples and GISs were dried using a speed vacuum concentrator (Labconco) and re-suspended in 100 mM TEAB buffer (50 µL). The TMTpro labeling reagents (5mg) were equilibrated to room temperature, and anhydrous ACN (200 µL) was added to each reagent channel (TMT reagent 16plex lot# YA357799; 134C and 135N lot# YB370079). Each channel was gently vortexed for 5 min, and then 10 µL from each TMT channel was transferred to the peptide solutions and allowed to incubate for 1 h at room temperature. The reaction was quenched with 5% (vol/vol) hydroxylamine (5 µL) (Pierce). All channels were then combined and dried by SpeedVac (LabConco). The combined sample was then resuspended in 500 µL of 0.1% TFA and then diluted 1:1 with 4% H3PO4 and desalted with a 30 mg MCX column (Waters). Samples were loaded onto the MCX column and then washed with 100 mM ammonium formate in 2% formic acid followed by a methanol and eluted with 300uL of 5% ammonium hydroxide in methanol. The eluates were then dried to completeness using a SpeedVac (LabConco).

High-pH Off-line Fractionation: Dried samples were re-suspended in high pH loading buffer (0.07% vol/vol NH4OH, 0.045% vol/vol FA, 2% vol/vol ACN) and loaded onto a Water's BEH 1.7 um 2.1mm by 150mm. A Thermo Vanquish was used to carry out the fractionation. Solvent A consisted of 0.0175% (vol/vol) NH4OH, 0.01125% (vol/vol) FA, and 2% (vol/vol) ACN; solvent B consisted of 0.0175% (vol/vol) NH4OH, 0.01125% (vol/vol) FA, and 90% (vol/vol) ACN. The sample elution was performed over a 25 min gradient with a flow rate of 0.6 mL/min. A total of 192 individual equal volume fractions were collected across the gradient and subsequently pooled by concatenation into 96 fractions and dried using a SpeedVac.

Liquid Chromatography Mass Spectrometry: All fractions were resuspended in an equal volume of loading buffer (0.1% FA, 0.03% TFA, 1% ACN) and analyzed by liquid chromatography coupled to tandem MS. Peptide eluents were separated on Water's CSH column (1.7um resin 150um by 15 cm) by a Vanquish Neo (ThermoFisher Scientific). Buffer A was water with 0.1% (vol/vol) formic acid, and buffer B was 99.9% (vol/vol) acetonitrile in water with 0.1% (vol/vol) formic acid. The gradient was from 3% to 35% solvent B over 17 mins followed by column wash and equilibration for a total of 23 mins. Peptides were monitored on an Orbitrap Astral spectrometer (ThermoFisher Scientific) fitted with a high-field asymmetric waveform ion mobility spectrometry (FAIMS Pro) ion mobility source (ThermoFisher Scientific). Two compensation voltages (CV) of -45 and -60 were chosen for the FAIMS. Each cycle consisted of one full scan acquisition (MS1) with an m/z range of 400-1500 at 120,000 resolution and standard settings and as many tandem (MS/MS) scans in 1.5 seconds. The Astral higher energy collision-induced dissociation (HCD) tandem scans were collected at 35% collision energy with an isolation of 0.5 m/z, an AGC setting of 100%, and a maximum injection time of 20ms. Dynamic exclusion was set to exclude previously sequenced peaks for 30 seconds within a 5-ppm tolerance window.

Database Search: The de-identified 6958 raw files (across 65 TMT-18 plexes) have been deposited at synapse.org with the SynID: syn59804727. These files were subsequently searched using FragPipe (version 22.0) as described[59–61]. The search was performed with a database of canonical Human

proteins downloaded from Uniprot (20,402; accessed 02/11/2019), with added specific peptides for APOE E4 and E2 alleles, ABETA40, and ABETA42 peptides (total of 20405 sequences)[62]. The workflow used in FragPipe followed default TMT-16 plex parameters, used for both TMT-16 and TMT-18 experimental design (TMTpro). Briefly, precursor mass tolerance was -20 to 20 ppm, fragment mass tolerance of 20 ppm, mass calibration and parameter optimization were selected, and isotope error was set to -1/0/1/2/3. Enzyme specificity was set to semi-tryptic and up to two missing trypsin cleavages were allowed. Peptide length was allowed to range from 7 to 42 and peptide mass from either 200 to 5,000 Da. Variable modifications that were allowed in our search included: oxidation on methionine, N-terminal acetylation on protein, N-terminal acetylation on peptide along with off-target TMT tag modification on Serine, Threonine and Histidine, with a maximum of 3 variable modifications per peptide. Peptide Spectral Matches were validated using Percolator[63]. The false discovery rate (FDR) threshold was set to 1% and protein and peptide abundances were quantified using Philosopher[64] for downstream analysis. The data from all batches were merged and protein levels were first scaled by dividing each protein intensity by intensity sum of all proteins in each sample followed by multiplying by the maximum protein intensity sum across all samples (i.e., column normalized). Instances where the intensity was ‘0’ were treated as ‘missing values’.

Quality Control: For CSF proteomic quality control, we kept proteins with data available in ≥ 200 subjects, selecting the highest abundance entry for proteins with multiple UniProt annotations. Protein abundances were normalized by sample-specific total protein content and $\log_2$-transformed. We checked for blood contamination by testing whether any sample was an outlier for hemoglobin subunits but found none. We excluded samples > 4 standard deviations from the mean of the first two principal components using iterative PCA. Linear regression was used to remove technical (batch) and biological (age, clinical diagnosis) effects from protein abundances. We then applied surrogate variable analysis (SVA package[65]) to identify and remove 67 additional confounding factors from the residual proteomic data. The final proteomic profiles were z-score standardized and limited to samples with available genetic data, resulting in 1006 subjects with 2949 proteins for pQTL mapping.

CSF pQTL mapping: We performed pQTL mapping using SNPs with MAF ≥ 0.05, defining cis pQTLs as variants within 1 Mb of the corresponding protein-coding gene (coordinates from BioMart[66], accessed February 10, 2025) and trans pQTLs as variants outside this window. Linear models were fitted using PLINK[52] with SNP genotype as the predictor and normalized protein expression as the outcome, adjusting for sex, 10 principal components, and genetic data source (WGS vs. array). Significance thresholds were $p < 5x10^{-8}$ for cis pQTLs and $p < 1.706 \times 10^{-11}$ (5×10^{-8} / 2,931 proteins) for trans pQTLs. To address LD among significant pQTLs, we used PLINK's '--clump' function to identify independent index pQTLs with pairwise $r^2 < 0.1$ within distance of 5 Mb. We estimated the inflation factor lambda for the pQTL mapping and found the median lambda to be 1.002, range [0.975 – 1.028]

CSF pQTL fine-mapping: We performed pQTL fine-mapping with SuSiE[15] following its default parameters. SuSiE was designed to identify as many credible sets as the data support, each with as few variants as possible. For a given gene and its corresponding variants in the *cis* regulatory region, the output was the number of credible sets that had 95% probability of containing a variant with nonzero causal effect. We set the maximum number of credible sets for a gene to be 10 (default value).

CSF pQTL π_1 rate**:** We compared our CSF pQTL findings (2,961 proteins, TMT-MS) with pQTLs from four studies: brain tissue (Wingo et al., n=1,013, TMT-MS)[10], CSF (Western et al., n=3,506, SomaScan)[13], and plasma (Ferkingstad et al., n=35,559, SomaScan; Sun et al., n=34,557, Olink)[5,6]. To assess replication, we estimated π_1 (proportion of true associations) for each comparison by first calculating π_0 using the qvalue package, which takes the p-values from our index pQTLs when tested in each comparison study. Comparisons were made within each QTL type (i.e., cis to-cis or trans to-trans).

Integration of CSF proteomics and genetic data with AD GWAS results: We integrated CSF proteomic and genetic data with AD GWAS[19] results using two complementary approaches. First, we performed PWAS using FUSION[4], restricting genotypes to SNPs in the European LD reference panel.

We estimated SNP-based heritability for each protein and declared those with $p < 0.01$ heritable. For heritable proteins, we trained prediction models (BLUP, LASSO, elastic net, BSLMM) using SNPs within 500 kb of the gene, retaining weights from the best-performing model. These weights were integrated with AD GWAS summary statistics to identify cis regulated proteins associated with AD (FDR < 0.05). Second, we performed Summary data-based Mendelian Randomization (SMR)[2] on PWAS-significant proteins using pQTL and AD GWAS data to test whether CSF proteins mediate SNP-trait associations. We applied HEIDI[2] testing to exclude associations likely due to LD (HEIDI $p < 0.05$). Proteins were considered causally implicated if they met all criteria: PWAS FDR < 0.05, SMR $p < 0.05$, HEIDI $p > 0.05$, and consistent effect directions on AD between PWAS and SMR analyses. All coordinates used GRCh38.

Comparing AD causal proteins identified in CSF between this and a published study[13]: Since the published study performed the PWAS for both *cis and trans* pQTLs (and not gene-based) using FUSION pipeline, and their cis and trans pQTLs were operationally defined as within 1MB window up and downstream of the gene, we compared our results with their results from cis pQTLs. Additionally, for genes with multiple somaIDs, we kept the top PWAS entry. In the 38 prioritized genes in the published study, 27 were results from cis pQTLs. We compared our CSF causal proteins with these 27.

STRING database: We use the STRING database[67], which curates evidence from experimental data, computational predictions, text mining, and various databases, to examine protein-protein interactions (PPIs) encompassing physical PPI and functional associations. We input the 24 candidate AD causal proteins identified in CSF into STRING and examined their PPIs and biological pathways they are enriched in at FDR $p<0.05$.

Differentially expressed proteins in AD in both CSF and brain: We used available MS-based proteomic brain profiles from donors of the ROS/MAP studies as described[51]. To make the results comparable between brain and ADNI CSF proteomics data, we used clinically defined AD dementia as

case and cognitively normal controls in both CSF and brain tissue. Regression models were fit for protein abundance adjusted for sex, age, and education. Proteins with FDR < 0.05 were considered significantly differentially expressed. AD diagnosis in ADNI was based on neuropsychological testing and a clinical exam with a neurologist at the time CSF was drawn for proteomic profiling.

Functional enrichment of differentially expressed proteins was performed using the clusterProfiler R package (v4.13.4)[68]. All proteins that passed quality control (n = 2,961) were used as the background set. Overrepresentation analysis was conducted using a one-sided Fisher's exact test, and p-values were adjusted using the Benjamini-Hochberg false discovery rate (FDR) method. Gene Ontology (GO) terms with an adjusted p-value < 0.05 were considered statistically significant.

Examine expression of candidate AD causal proteins in CSF: Our prior work has identified 64 candidate causal brain protein in AD[10]. In the current study, we examined the overlap between these candidate causal proteins from brain and the proteins differentially expressed in AD in CSF.

References

1. The GTEx Consortium *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
2. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
3. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLOS Genet.* **16**, e1008720 (2020).
4. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
5. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712–1721 (2021).
6. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
7. Wingo, T. S. *et al.* Brain proteome-wide association study implicates novel proteins in depression pathogenesis. *Nat. Neurosci.* **24**, 810–817 (2021).
8. 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).
9. Yang, C. *et al.* Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders. *Nat. Neurosci.* **24**, 1302–1312 (2021).
10. Wingo, A. P. *et al.* Multiancestry brain pQTL fine-mapping and integration with genome-wide association studies of 21 neurologic and psychiatric conditions. *Nat. Genet.* 1–10 (2025) doi:10.1038/s41588-025-02291-2.
11. Wingo, A. P. *et al.* Sex differences in brain protein expression and disease. *Nat. Med.* **29**, 2224–2232 (2023).
12. Sasayama, D. *et al.* Genome-wide quantitative trait loci mapping of the human cerebrospinal fluid proteome. *Hum. Mol. Genet.* **26**, 44–51 (2017).
13. Western, D. *et al.* Proteogenomic analysis of human cerebrospinal fluid identifies neurologically relevant regulation and implicates causal proteins for Alzheimer's disease. *Nat. Genet.* **56**, 2672–2684 (2024).
14. Hansson, O. *et al.* The genetic regulation of protein expression in cerebrospinal fluid. *EMBO Mol. Med.* **15**, e16359 (2023).
15. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).
16. Johnson, E. C. B. *et al.* Cerebrospinal fluid proteomics define the natural history of autosomal dominant Alzheimer's disease. *Nat. Med.* **29**, 1979–1988 (2023).

17. Shaw, L. M. & Alzheimer's Disease Neuroimaging Initiative. Predictive Performance of CSF Ab1-42 and Tau for Cognitive Decline and Progression to Dementia. (2025).
18. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9440–9445 (2003).
19. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).
20. Shiratori, I., Ogasawara, K., Saito, T., Lanier, L. L. & Arase, H. Activation of Natural Killer Cells and Dendritic Cells upon Recognition of a Novel CD99-like Ligand by Paired Immunoglobulin-like Type 2 Receptor. *J. Exp. Med.* **199**, 525–533 (2004).
21. Barclay, A. N. & Brown, M. H. The SIRP family of receptors and immune regulation. *Nat. Rev. Immunol.* **6**, 457–464 (2006).
22. Griciuc, A. *et al.* TREM2 Acts Downstream of CD33 in Modulating Microglial Pathology in Alzheimer's Disease. *Neuron* **103**, 820-835.e7 (2019).
23. Wißfeld, J. *et al.* Reporter cell assay for human CD33 validated by specific antibodies and human iPSC-derived microglia. *Sci. Rep.* **11**, 13462 (2021).
24. Leung, Y. Y. *et al.* Integrated genomic analysis and CRISPRi implicates EGFR in Alzheimer's disease risk. 2025.06.25.25328705 Preprint at https://doi.org/10.1101/2025.06.25.25328705 (2025).
25. Sano, H. *et al.* Critical role of galectin-3 in phagocytosis by macrophages. *J. Clin. Invest.* **112**, 389–397 (2003).
26. Chou, C.-L. & Yeh, H.-I. The Role of the Renin-Angiotensin System in Amyloid Metabolism of Alzheimer's Disease. *Acta Cardiol. Sin.* **30**, 114–118 (2014).
27. Hamano, T. *et al.* Abnormal intracellular trafficking of high affinity nerve growth factor receptor, Trk, in stable transfectants expressing presenilin 1 protein. *Brain Res. Mol. Brain Res.* **137**, 70–76 (2005).
28. Fan, K. *et al.* The induction of neuronal death by up-regulated microglial cathepsin H in LPS-induced neuroinflammation. *J. Neuroinflammation* **12**, 54 (2015).
29. Wang, Y. *et al.* Cathepsin H: Molecular characteristics and clues to function and mechanism. *Biochem. Pharmacol.* **212**, 115585 (2023).
30. Vardarajan, B. N. *et al.* Progranulin mutations in clinical and neuropathological Alzheimer's disease. *Alzheimers Dement.* **18**, 2458–2467 (2022).
31. Medoh, U. N. *et al.* The Batten disease gene product CLN5 is the lysosomal bis(monoacylglycero)phosphate synthase. *Science* **381**, 1182–1189 (2023).
32. Wingo, A., Liu, Y. & Wingo, T. Multi-ancestry brain pQTL fine-mapping and integration with genome-wide association studies of 21 neurologic and psychiatric traits. *Nat. Genet. 2025 Press*.
33. Vidarsson, G., Dekkers, G. & Rispens, T. IgG Subclasses and Allotypes: From Structure to Effector Functions. *Front. Immunol.* **5**, 520 (2014).

34. Mikocziova, I., Greiff, V. & Sollid, L. M. Immunoglobulin germline gene variation and its impact on human disease. *Genes Immun.* **22**, 205–217 (2021).
35. Jonsson, T. *et al.* Variant of TREM2 associated with the risk of Alzheimer's disease. *N. Engl. J. Med.* **368**, 107–116 (2013).
36. Cruts, M. *et al.* Null mutations in progranulin cause ubiquitin-positive frontotemporal dementia linked to chromosome 17q21. *Nature* **442**, 920–924 (2006).
37. Beenken, A. *et al.* Structures of LRP2 reveal a molecular machine for endocytosis. *Cell* **186**, 821-836.e13 (2023).
38. Minnone, G., De Benedetti, F. & Bracci-Laudiero, L. NGF and Its Receptors in the Regulation of Inflammatory Response. *Int. J. Mol. Sci.* **18**, 1028 (2017).
39. Choi, H.-J., Jeong, Y. J., Kim, J. & Hoe, H.-S. EGFR is a potential dual molecular target for cancer and Alzheimer's disease. *Front. Pharmacol.* **14**, 1238639 (2023).
40. Kim, J. *et al.* Inhibiting EGFR/HER-2 ameliorates neuroinflammatory responses and the early stage of tau pathology through DYRK1A. *Front. Immunol.* **13**, (2022).
41. Jayaswamy, P. K. *et al.* Implicative role of epidermal growth factor receptor and its associated signaling partners in the pathogenesis of Alzheimer's disease. *Ageing Res. Rev.* **83**, 101791 (2023).
42. Gaikwad, S. *et al.* Signal regulatory protein-beta1: a microglial modulator of phagocytosis in Alzheimer's disease. *Am. J. Pathol.* **175**, 2528–2539 (2009).
43. PubChem. SIRPB1 - signal regulatory protein beta 1 (human). https://pubchem.ncbi.nlm.nih.gov/gene/SIRPB1/human.
44. Li, L., Han, B., Kong, Y., Zhang, G. & Zhang, Z. Vitamin D binding protein in psychiatric and neurological disorders: Implications for diagnosis and treatment. *Genes Dis.* **11**, 101309 (2024).
45. Groves, N. J., McGrath, J. J. & Burne, T. H. J. Vitamin D as a neurosteroid affecting the developing and adult brain. *Annu. Rev. Nutr.* **34**, 117–141 (2014).
46. Zou, K. *et al.* Angiotensin-Converting Enzyme Converts Amyloid β-Protein 1–42 (Aβ1–42) to Aβ1–40, and Its Inhibition Enhances Brain Aβ Deposition. *J. Neurosci.* **27**, 8628–8635 (2007).
47. Cater, J. H. *et al.* Human pregnancy zone protein stabilizes misfolded proteins including preeclampsia- and Alzheimer's-associated amyloid beta peptide. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 6101–6110 (2019).
48. Wu, Y., Zhao, Z., Deng, X., Jia, J. & Yuan, G. Pregnancy zone protein, a potential research target in multiple diseases. *Gene* **935**, 149013 (2025).
49. Bennett, D. A., Schneider, J. A., Arvanitakis, Z. & Wilson, R. S. Overview and findings from the religious orders study. *Curr. Alzheimer Res.* **9**, 628–645 (2012).
50. Bennett, D. A. *et al.* Overview and findings from the rush Memory and Aging Project. *Curr. Alzheimer Res.* **9**, 646–663 (2012).

51. Mei, Z. *et al.* Unraveling sex differences in Alzheimer's disease and related endophenotypes with brain proteomes. *Alzheimers Dement. J. Alzheimers Assoc.* **21**, e70206 (2025).
52. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
53. Kotlar, A. V., Trevino, C. E., Zwick, M. E., Cutler, D. J. & Wingo, T. S. Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale. *Genome Biol.* **19**, 14 (2018).
54. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
55. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
56. Dammer, E. B. *et al.* Proteomic analysis of Alzheimer's disease cerebrospinal fluid reveals alterations associated with APOE ε4 and atomoxetine treatment. *Sci. Transl. Med.* **16**, eadn3504 (2024).
57. Modeste, E. S. *et al.* Quantitative proteomics of cerebrospinal fluid from African Americans and Caucasians reveals shared and divergent changes in Alzheimer's disease. *Mol. Neurodegener.* **18**, 48 (2023).
58. Ping, L. *et al.* Global quantitative analysis of the human brain proteome in Alzheimer's and Parkinson's Disease. *Sci. Data* **5**, 180036 (2018).
59. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).
60. Bangs, M. C. *et al.* Proteomic Subtyping of Alzheimer's Disease CSF links Blood-Brain Barrier Dysfunction to Reduced levels of Tau and Synaptic Biomarkers. *BioRxiv Prepr. Serv. Biol.* 2025.03.14.643332 (2025) doi:10.1101/2025.03.14.643332.
61. Trautwig, A. N. *et al.* Molecular Subtyping Based on Hippocampal Cryptic Exon Burden Reveals Proteome-wide Changes Associated with TDP-43 Pathology across the Spectrum of LATE and Alzheimer's Disease. 2025.05.30.656396 Preprint at https://doi.org/10.1101/2025.05.30.656396 (2025).
62. Seifar, F. *et al.* Large-scale deep proteomic analysis in Alzheimer's disease brain regions across race and ethnicity. *Alzheimers Dement. J. Alzheimers Assoc.* (2024) doi:10.1002/alz.14360.
63. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).
64. da Veiga Leprevost, F. *et al.* Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat. Methods* **17**, 869–870 (2020).
65. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinforma. Oxf. Engl.* **28**, 882–883 (2012).

66. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
67. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
68. G, Y., Lg, W., Y, H. & Qy, H. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics J. Integr. Biol.* **16**, (2012).

Acknowledgments

We gratefully acknowledge the generosity of the participants and hard work of the ADNI study staff. We gratefully acknowledge the research volunteers and staff of ROS/MAP for their participation and contributions. We thank Jiaqi Liu for assistance in creating Figure 2.

Conflict of Interest Statement

A.I.L. serves as a consultant to Cognito, Asha Therapeutics, NextSense and Cognition Therapeutics. D.M.D., A.I.L., and N.T.S. are co-founders, employees, consultants, and/or shareholders of EmTheraPro. D.M.D. and N.T.S. are co-founders of Arc proteomics. N.T.S. is a co-founder of Stitch-Rx. T.S.W. is a co-founder of revXon. The remaining authors declare no competing interests.

Resources

Data: Raw mass spectrometry data and pre- and post-processed plasma protein abundance data and case traits related to this manuscript are available at https://www.synapse.org/Synapse:syn59804727 on the AMP-AD Knowledge Portal, which is a platform for accessing data, analyses and tools generated by the AMP-AD Target Discovery Program and other programs supported by the National Institute on Aging to enable open-science practices and accelerate translational learning. The data, analyses and tools are shared early in the research cycle without a publication embargo on secondary use. ADNI data are available on https://adni.loni.usc.edu/.

Software: FragPipe (v22.0), MSFragger (v4.1), IonQuant (v1.10.27), diaTracer (v1.1.5), Philosopher (v5.1.1), Percolator (v3.6.5), R (v4.4.2), FUSION (commit e1ba5f7; https://github.com/gusevlab/fusion_twas), SMR (v1.02; https://yanglab.westlake.edu.cn/software/smr), susieR (v0.12.35; https://stephenslab.github.io/susieR/), SVA (v3.20.0; https://bioconductor.org/packages/sva/), and plink (v2.0; https://www.cog-genomics.org/plink/2.0/).

Funding

I01 BX005686 (A.P.W.), IK4 BX005219 (A.P.W.), P30 AG066511 (A.I.L.), R01 AG072120 (A.P.W., T.S.W.), R01 AG075827 (A.P.W., T.S.W.), R01 AG079170 (T.S.W.), RF1 AG062181 (N.T.S.), U01 AG046161 (A.I.L.), U01 AG061356 (A.I.L.), U01 AG061357 (A.I.L., N.T.S.), U01 NS128433 (A.I.L., N.T.S.), as well as the Foundation for the National Institutes of Health AMP-AD 2.0 grant. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2- 0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech,

Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

Contributions

APW, NTS, and TSW Conceptualization. AIL, APW, AS, DAB, DMD, EJF, FW, NTS, TSW, and ZM. Data Curation. APW, ESG, MW, SMV, TSW, YL, and ZM. Formal analysis. APW, MW, TSW, and ZM. Initial Draft. APW and TSW Supervision. All authors revised and provided critical feedback on the manuscript and approved the manuscript.

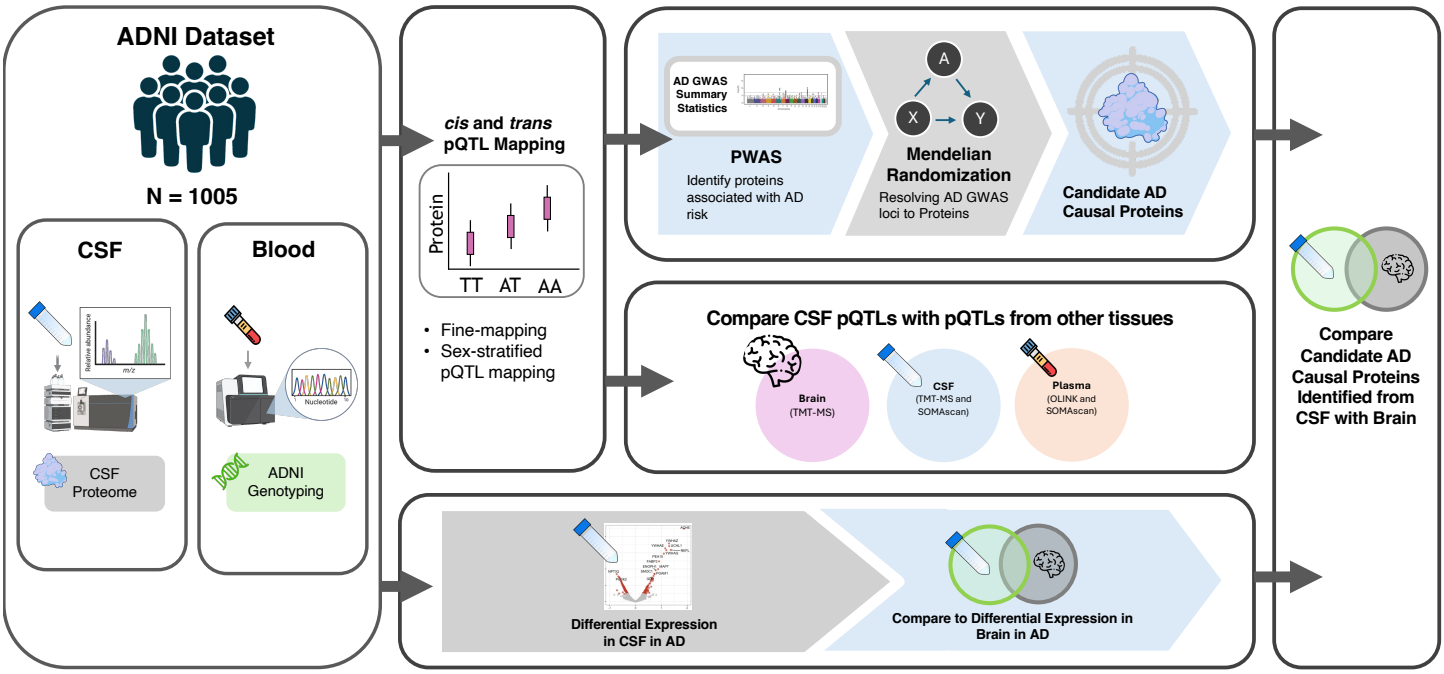


Figure 1. Study Overview.

Study participants were individuals from ADNI who donated cerebrospinal fluid (CSF) and had available genotyping. Mass spectrometry (MS) was used to generate CSF proteomes and those were used to perform *cis* and *trans* protein quantitative trait locus (pQTL) mapping. Fine mapping was performed on the *cis* pQTLs. Sex-specific pQTLs were also identified. MS-based proteomic pQTLs were compared with published studies using CSF, brain, and plasma tissue and MS and affinity-based approaches for measuring the proteome. The CSF proteome was compared with brain proteome in differential analysis of AD dementia. Finally, the CSF pQTLs were used to infer AD causal proteins that were compared with results of brain illustrating CSF-specific findings.

Figure 2. Two-dimensional Manhattan plot for MS-based CSF pQTLs.

There are 1,417 index cis pQTLs and 130 index trans pQTLs plotted. The x-axis is the genomic position of the pQTL SNP, and the y-axis is the gene position for the tested protein. The *cis* pQTLs are shown in orange while *trans* pQTLs are shown in blue with their corresponding protein labels. Boxes around gene labels indicate proteins with more than one pQTL on separate chromosomes.

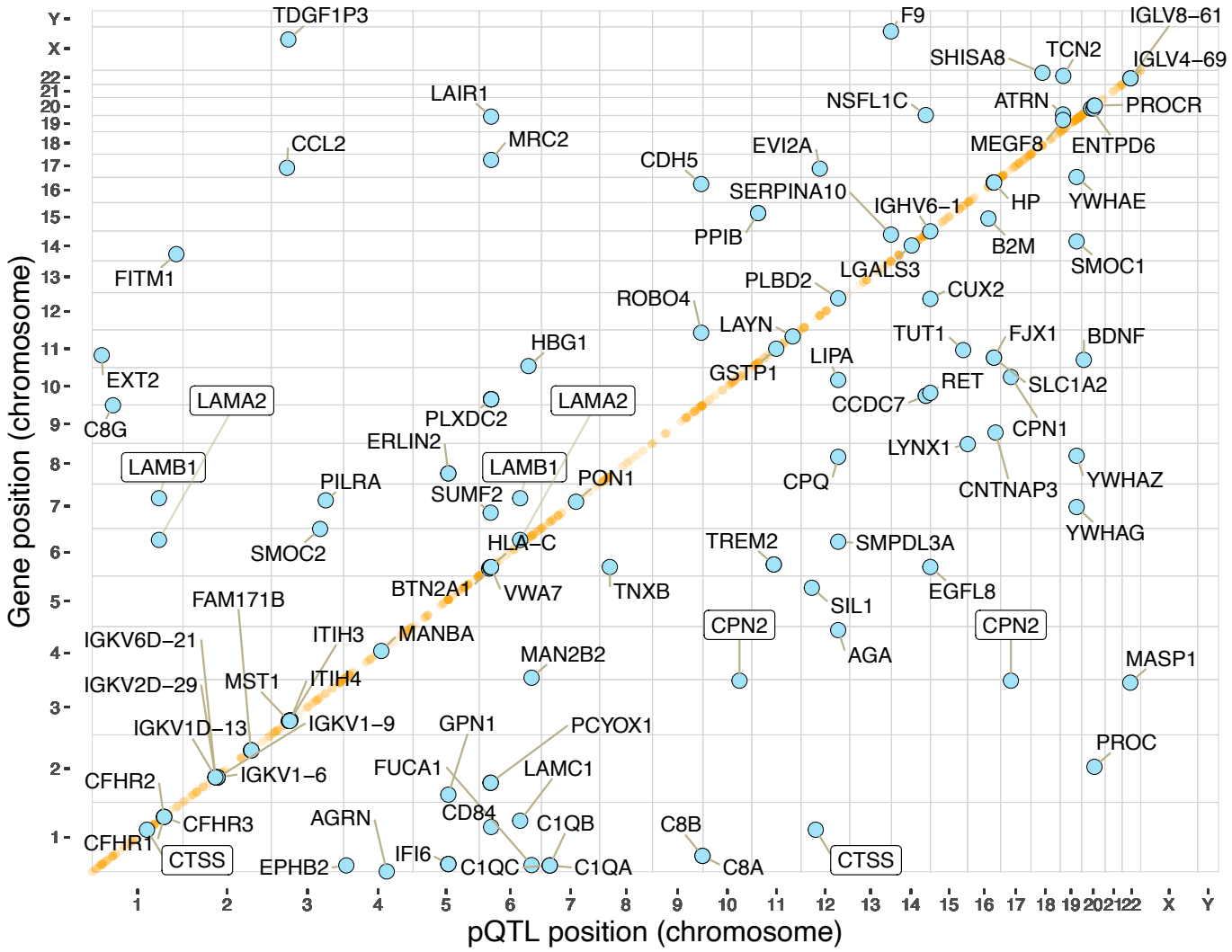


Figure 3. Comparison of MS-based CSF proteomic profiles and pQTLs across different tissues and technologies.

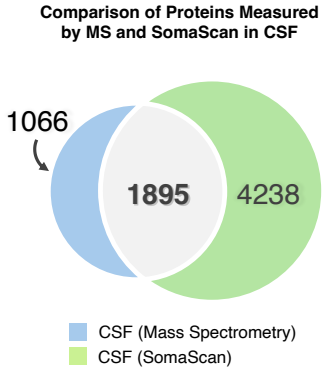
a. Overview of current and prior pQTL studies included in the comparative analysis, showing tissue type, sample size, proteomic technology, and number of genes analyzed. **b.** Comparison of CSF gene products measured by MS versus aptamer-based technology (in Western et al 2024) shows 1,895 proteins measured in common. **c.** Comparison of gene products measured by MS in CSF (current study) versus brain tissue (Wingo et al. 2025) revealed 2,200 shared genes. **(d)** Comparison of proteins profiled across CSF and plasma in MS- and affinity-based proteomic studies. Bars represent the number of genes detected in each unique combination of studies, with filled circles indicating which studies contribute to each intersection. The blue color indicates proteins measured by CSF TMT-MS and other platforms and purple indicates proteins only profiled by CSF TMT-MS alone. **e.** Replication rates (π_1) between the current MS-based CSF pQTLs and published pQTL studies for proteins profiled between the two technologies. Results are stratified by *cis* and *trans* pQTLs.

a

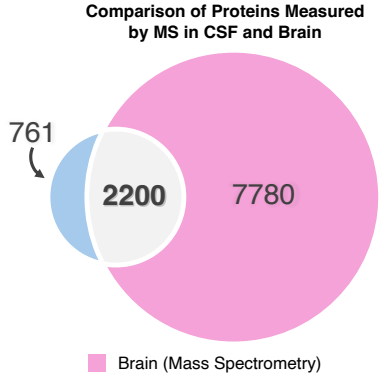
Current and Prior pQTL Studies

Tissue	Study	Sample Size	Technology	Proteins Analyzed
CSF	This study	1,006	MS	2,961
Brain	Wingo et al. 2025	1,013	MS	9,980
CSF	Western et al., 2024	3,506	SOMAScan	6,133
Plasma	Ferkingstad et al 2021	35,559	SOMAScan	4,656
Plasma	Sun et al., 2023	34,557	Olink	2,460

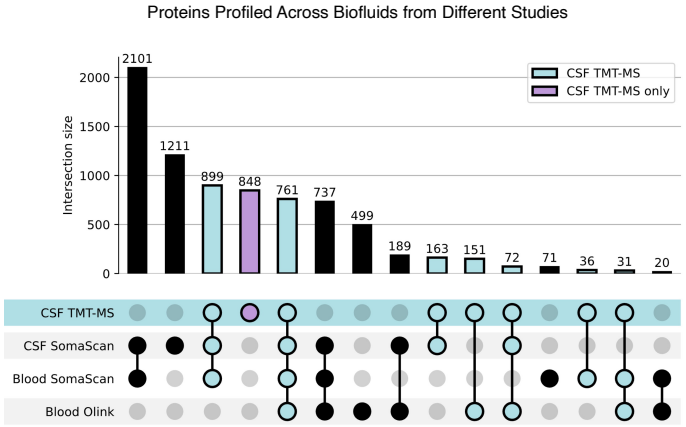
b



c



d



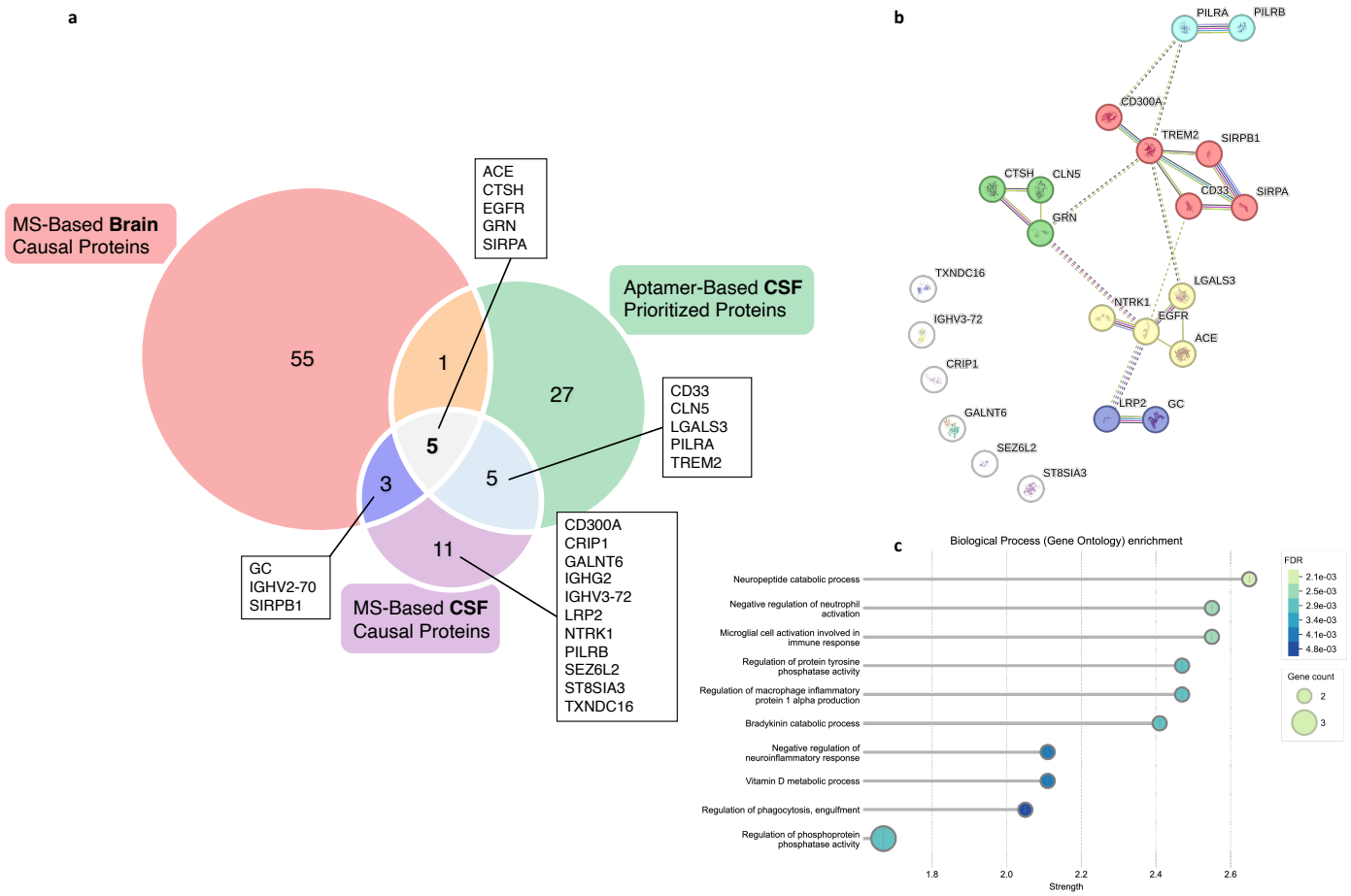
e

π_1 **rate** between the current MS-based CSF pQTLs (discovery) and published pQTL studies (as replication)

Comparison Tissue	*cis*	*trans*
Brain (MS)	0.82	-
CSF (SOMAScan)	0.73	0.56
Plasma (SOMAScan)	0.78	0.42
Plasma (Olink)	0.97	0.85

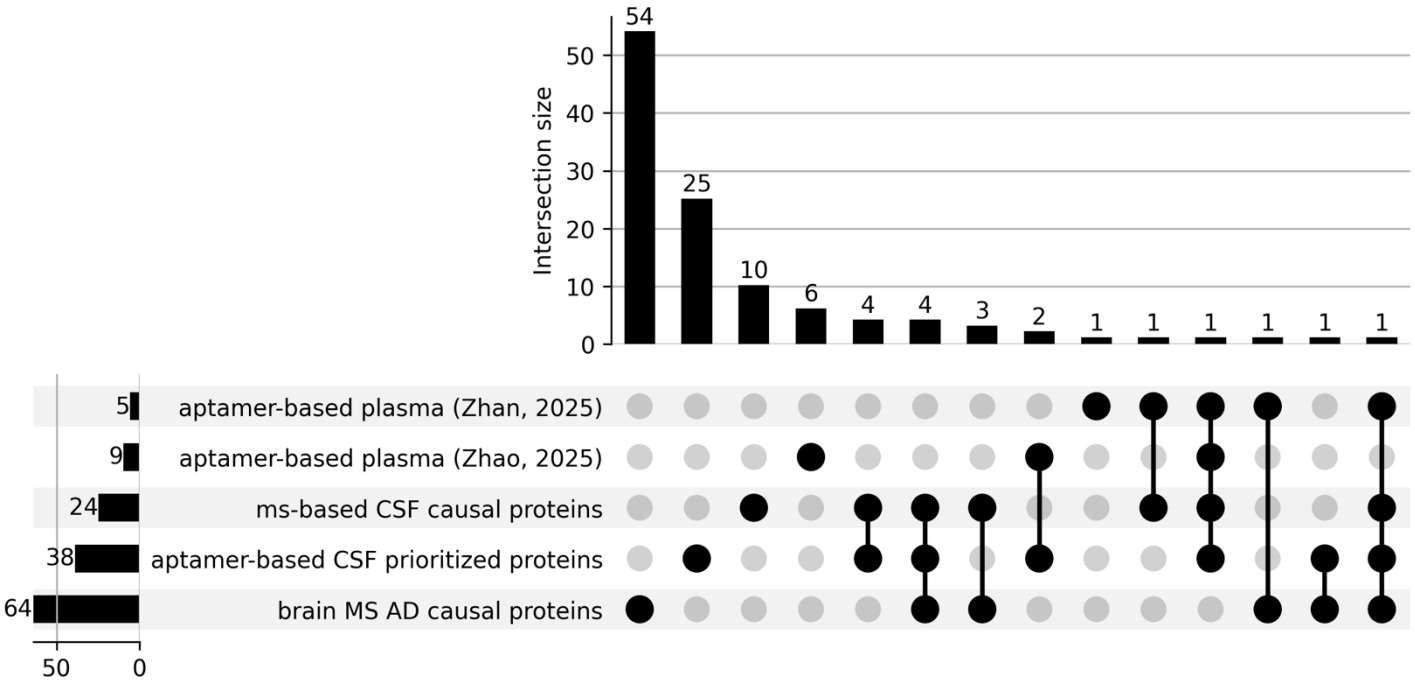
Figure 4. AD causal proteins and cross-tissue comparisons.

a. Comparison of AD causal proteins identified in brain and CSF using MS- and aptamer-based proteogenomics. The Venn diagram depicts proteins identified in published work (Wingo et al., 2025; Western et al., 2024) and in this MS-based CSF study. **b.** STRINGdb diagram for 24 AD causal proteins identified using MS-based CSF proteogenomics (two IGH proteins contained insufficient information and not depicted here – IGHV2-70 and IGHG2). Markov clustering reveals four functional modules comprised of a total 5 clusters known to contribute to distinct aspects of AD pathology. Clusters 1 and 2 (denoted by light blue and red) contain proteins that modulate microglia and macrophage activation[1–5]. Cluster 3 (yellow) contains proteins that regulate neurovascular growth and remodeling[3,6–9]. Cluster 4 (dark blue) are proteins governing vitamin D metabolism, which is implicated in many neuroinflammatory and neurodegenerative disorders. Cluster 5 (green) are proteins involved in proteolysis and lysosomal function[10–13]. The isolated nodes, include IGHV3-72 and the two additional immunoglobulin heavy chain components, suggesting a role for immunoglobulins in AD, while the remaining isolated nodes are best characterized in cell growth and carcinogenesis. **c.** Result of biological process enrichments highlight leukocyte activation, regulation of phagocytosis, and neuropeptide catabolic processes as the top 3 enriched FDR significant processes among the 24 causal AD proteins in CSF. The top three biological process enrichments by FDR highlight leukocyte activation, phagocytosis, and neuropeptide catabolism. In the context of our study, “leukocyte activation” and “regulation of phagocytosis” most strongly point to microglial activity, which is consistent with observed microglial dysfunction in AD[4,14,15].



Supplementary Figure 1. Comparison of AD causal proteins identified in CSF, brain, and plasma.

Overlap between AD causal proteins identified in either CSF, brain, or plasma.



Aptamer-based plasma study was from Zhan et al 2025 or Zhao et al 2025. MS-based CSF causal proteins are from the current study. Aptamer-based CSF prioritized proteins are from Western et al, 2024. Brain MS AD causal proteins are from Wingo et al, 2025.

Supplementary Figure 2. Comparison between CSF and brain proteins differentially expressed in AD dementia.

a. Differential expression analysis of MS-based CSF proteome for AD dementia status in ADNI. **b.** Differential expression of MS-based brain proteome for AD dementia in ROS/MAP. **c.** Comparison between CSF and brain for proteins showing differential abundance in AD dementia. There were 2,961 proteins profiled in both CSF and brain, among those 117 were differentially expressed in AD dementia in both tissues. For the volcano plots in (a) and (b), red indicates differential expression at FDR <0.05 (significant) and black conveys not differentially expressed (FDR >0.05, non-significant). Beta-value was plotted on the x-axis and -log10 p-value on the y-axis.

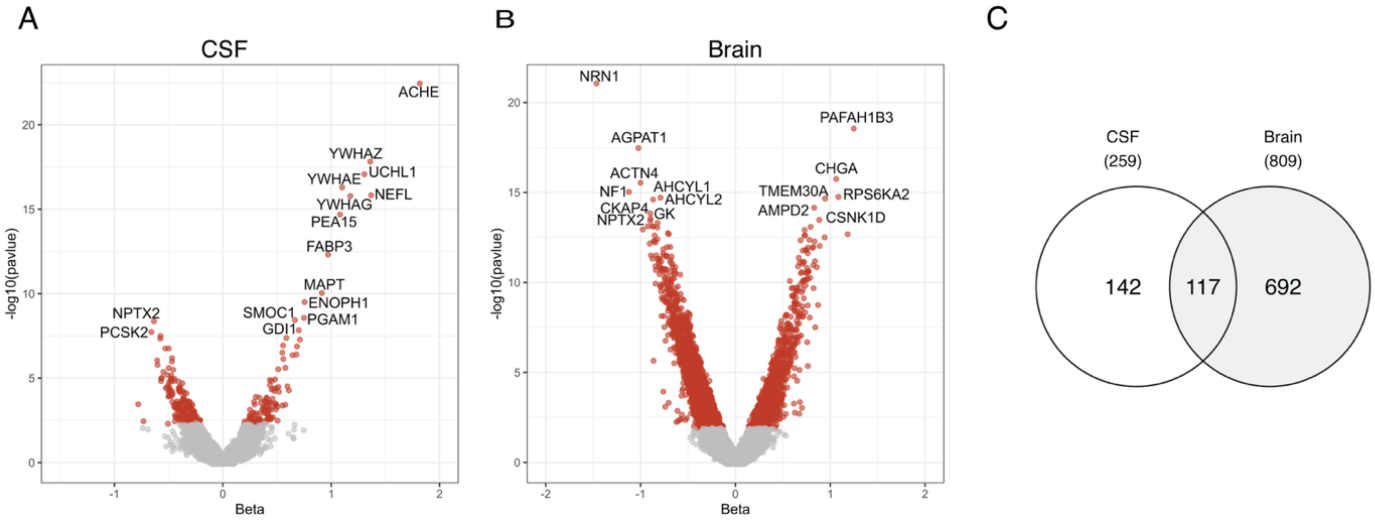


Table 1. **Candidate AD causal proteins identified in CSF (n = 24)**
Listed are proteins found to be significant in the PWAS of AD and SMR/HEIDI of AD using proteogenomic data from CSF (n = 24). Significance was defined as meeting all the following criteria: PWAS FDR p<0.05, SMR p<0.05; HEIDI p >0.05; SMR beta and PWAS Z have consistent direction of association. Candidate causal proteins in brain were obtained from integrating the same AD GWAS with brain proteogenomic data via PWAS of AD and SMR/HEIDI of AD based on similar criteria. These analyses were previously performed and published by Wingo et al, 2025[7]. Rows were sorted by whether the candidate causal proteins identified in CSF are also candidate causal proteins identified in brain and by PWAS p-values.

	Protein	CHR	PWAS Z	PWAS p	PWAS FDR	SMR beta	SMR p	HEIDI p	Causal protein in brain
1	ACE	17	-8.27	1.29E-16	2.11E-14	-0.12	1.22E-11	0.40	**yes**
2	GRN	17	-7.02	2.19E-12	2.99E-10	-0.19	5.76E-07	0.19	**yes**
3	IGHV2-70	14	4.87	1.12E-06	9.17E-05	0.06	3.55E-06	0.17	**yes**
4	CTSH	15	4.51	6.41E-06	4.04E-04	0.05	3.25E-06	0.25	**yes**
5	EGFR	7	4.48	7.48E-06	4.38E-04	0.12	8.05E-04	0.13	**yes**
6	SIRPB1	20	-4.30	1.70E-05	9.28E-04	-0.04	3.60E-05	0.44	**yes**
7	SIRPA	20	4.15	3.29E-05	1.52E-03	0.03	2.61E-05	0.41	**yes**
8	GC	4	-3.73	1.92E-04	6.55E-03	-0.04	3.91E-04	0.67	**yes**
9	PILRA	7	8.60	7.77E-18	2.12E-15	0.12	2.39E-12	0.34	no
10	PILRB	7	8.46	2.75E-17	5.63E-15	0.08	3.08E-13	0.95	no
11	TREM2	6	-5.75	9.08E-09	9.30E-07	-0.22	3.73E-05	0.22	no
12	SEZ6L2	16	-5.31	1.12E-07	1.02E-05	-0.14	1.85E-03	0.07	no
13	IGHV3-72	14	-4.61	4.02E-06	2.99E-04	-0.21	3.15E-03	0.30	no
14	CD33	19	4.24	2.25E-05	1.15E-03	0.04	3.58E-04	0.71	no
15	IGHG2	14	-4.07	4.74E-05	1.99E-03	-0.18	2.78E-03	0.30	no
16	GALNT6	12	-4.06	4.99E-05	1.99E-03	-0.04	1.29E-05	0.67	no
17	CLN5	13	-4.05	5.10E-05	1.99E-03	-0.05	1.27E-04	0.13	no
18	NTRK1	1	3.76	1.73E-04	6.16E-03	0.12	1.69E-03	0.65	no
19	ST8SIA3	18	3.61	3.07E-04	8.98E-03	0.05	4.95E-04	0.94	no
20	CRIP1	14	3.60	3.18E-04	8.98E-03	0.05	2.64E-02	0.06	no
21	LRP2	2	3.34	8.32E-04	2.00E-02	0.10	8.84E-03	0.97	no
22	CD300A	17	3.22	1.30E-03	3.04E-02	0.08	4.23E-03	0.23	no
23	LGALS3	14	-3.08	2.04E-03	4.40E-02	-0.03	9.54E-04	0.19	no
24	TXNDC16	14	-3.05	2.27E-03	4.77E-02	-0.09	6.37E-03	0.64	no