Distribution Agreement

Signature:

______________________________ 4/14/2025
Eseza Kironde Date

Proteome-Wide Association Analysis of Alzheimer's Disease Using Cerebrospinal Fluid Protein Data

By

Eseza Kironde
Master of Public Health

Department of Biostatistics and Bioinformatics

Michael P. Epstein
Committee Chair

Jingjing Yang
Committee Member

Proteome-Wide Association Analysis of Alzheimer's Disease Using Cerebrospinal Fluid Protein Data

By

Eseza Kironde

B.A., Smith College 2020

Thesis Committee Chair: Michael P. Epstein, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University in
partial fulfillment of the requirements for the degree of
Master of Public Health in Biostatistics and
Bioinformatics
2025

Abstract

Proteome-Wide Association Analysis of Alzheimer's Disease Using Cerebrospinal Fluid Protein Data

By Eseza Kironde

Alzheimer's disease (AD) is a progressive neurodegenerative disorder with high heritability (60-80%). While genome-wide association studies (GWAS) have identified over 75 genetic risk loci, determining causal genes and biological mechanisms remains challenging. This study addresses these limitations by adapting OTTERS, a framework originally designed for transcriptome-wide association studies (TWAS), to perform proteome-wide association studies (PWAS) using cerebrospinal fluid (CSF) protein data.

Unlike traditional PWAS that rely on plasma samples or require individual-level reference data, our approach leverages summary-level CSF proteomics, which more directly reflects brain pathology in neurological disorders. We analyzed 741 CSF proteins using four different polygenic risk score (PRS) methods to account for diverse genetic architectures and combined results through an omnibus test.

Our analysis identified 13 proteins significantly associated with Alzheimer's disease after Bonferroni correction ($p < 6.75E-5$). Clusterin (*CLU*) showed the strongest association ($p = 3.30E-35$), followed by Interleukin-34 (IL34) and Fructose-bisphosphate aldolase A (*ALDOA*). While ten proteins confirmed previous AD associations, three novel proteins (*ALDOA*, *RNF43*, and *SIGLEC9*) represent potential new biomarkers or therapeutic targets.

This study demonstrates that CSF proteogenomics offers valuable insights into AD pathogenesis and that summary-level data approaches can maintain statistical power while increasing data accessibility. Our findings bridge the gap between genetic associations and protein-level changes, providing a framework for understanding the biological mechanisms underlying Alzheimer's disease.

Proteome-Wide Association Analysis of Alzheimer's Disease Using Cerebrospinal Fluid Protein Data

By

Eseza Kironde

BA., Smith College 2020

Thesis Committee Chair: Michael P. Epstein, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree
of Master of Public Health in Biostatistics
2025

Acknowledgements

I am deeply grateful to my thesis advisor, Dr. Michael Epstein, for his invaluable guidance, mentorship, and insight throughout this journey. I also extend sincere thanks to Dr. Jingjing Yang, whose thoughtful feedback and expertise significantly strengthened this work.
Special appreciation goes to Qile Dai—your technical support and willingness to help were truly instrumental to this analysis.

To my sisters, Malaika and Erina, and my dear friends, Ann, Sakaiza, and Josephine—your unwavering support, encouragement, and presence made this achievement possible
.
Finally, to my parents, Eric Kironde and Elizabeth Gowa Kironde—thank you for believing in me, especially in moments when I doubted myself. Your love has been my greatest strength.

Table of Contents

CHAPTER 1: INTRODUCTION

Alzheimer's disease (AD) is a disorder that causes progressive neurodegeneration of cells in the brain, leading to a reduction in cognitive function, memory, thinking and behavior [1]. AD is characterized by an accumulation of amyloid plaques and a massive loss of neurons [2, 3]. AD is the most common type of dementia and has a strong genetic component. Heritability, defined as the proportion of observed phenotypic variance that can be attributed to genetic factors [4], is estimated within families to be approximately 60-80% for AD, which indicates that genetic factors play a strong role in disease risk [5-7].

To identify genetic factors involved in AD, many studies perform testing using Genome Wide Association studies (GWAS). A GWAS is an observational study that examines genetic variants (Single Nucleotide Polymorphisms or SNPs) across an organism's entire genome, where researchers collect DNA samples from individuals with and without the condition, analyze the SNPs simultaneously and perform statistical analyses to identify variants that occur more frequently in affected individuals. [8, 9]. Although previous AD GWAS studies have found over 75 genetic risk loci [10-13], they cannot on their own specify which genes are causal or explain the biological mechanisms that these variants increase AD risk.

Approaches that combine multi-omic data with GWAS have improved our ability to identify causal genes for traits and diseases. A standard -omic integration method is a transcriptome-wide association study (TWAS) [14, 15], which involves building predictive models of gene expression using reference panels with both genetic and expression data and then using these models to predict gene expression in GWAS samples based on genotypes. While TWAS has proven useful, the approach is limited because it captures primarily transcriptional effects and potentially misses post-transcriptional and post-translational regulatory mechanisms that influence disease. This has led to the advancement of Proteome-Wide Association Studies (PWAS), which studies the role of genetically-regulated protein abundance, which provides clearer links to biological outcomes [16]. These studies use protein Quantitative Trait Loci (PQTLs), which are genetic variants affecting protein levels from reference panels, to predict protein abundance in GWAS samples.

However, most PWAS studies narrowly focus on plasma proteins [17]. This is limiting for neurological disease research because plasma proteogenomics shares little overlap with brain proteogenomics, with proteomic studies revealing only approximately 25-30% of pQTL shared between CSF and plasma and an even smaller overlap of about 13% between brain frontal cortex and plasma [18, 19]. For AD research, brain specific protein changes are critical as they reflect the neurodegenerative processes happening in affected tissues. In contrast to plasma-based studies, cerebrospinal fluid (CSF) proteogenomic studies have successfully identified causal genes for several disease-associated GWAS loci in AD, Parkinson's disease and other neurodegenerative disorders[18, 20-23]. CSF is studied because it is accessible from living adults through lumbar puncture, making it more practical than obtaining brain tissue while still providing valuable neurological biomarkers that share important molecular signatures with the brain. Although plasma is more easily accessible than CSF, CSF directly interacts with the brain, which makes it vital for studying AD-related protein changes. This tissue specificity emphasizes the importance of analyzing CSF rather than plasma when investigating neurological diseases, as CSF provides a more accurate reflection of protein regulation in the central nervous system that cannot be adequately captured through peripheral blood measurements.

Traditional PWAS assumes access to individual-level reference data, which has limitations. Individual-level data is often restricted due to privacy concerns, requires complex data usage agreements, and typically has smaller sample sizes compared to summary statistics. Tools that process summary-level reference data are preferred to individual level data because summary data is more widely shared, has a larger sample size and has fewer privacy concerns. To enable TWAS using summary-level reference data, Dai et al. (2022) developed a statistical approach called OTTERS to train prediction models through multiple Polygenic Risk Score (PRS) methods [24]. PRS methods are statistical approaches that combine the effects of many genetic variants (SNPs) that were estimated by single variant GWAS to estimate a set of SNP effects for predicting a single score representing a predicted phenotype from genetics data. The PRS methods are used by OTTERS because both quantitative molecular traits such as gene expression and protein abundances can be considered as quantitative phenotypes.

To leverage the advantages of summary-level data, OTTERS uses 4 methodological approaches to model the genetic architecture of molecular traits: (1) P-value Thresholding with linkage disequilibrium clumping (P+T) [25, 26], (2) the frequentist LASSO regression-based lassosum [27], (3) Bayesian multivariable regression with continuous shrinkage (PRS-CS)[28], and (4) nonparametric Bayesian Dirichlet Regression [29] method (SDPR)[30]. These diverse methods account for different possible genetic architectures underlying the origin of gene expression or protein abundances. OTTERS culminates in an omnibus test that combines all the p-values for improved performance while protecting against inflated type 1 error. OTTERS is suitable for integrating summary-level xQTL data of quantitative molecular traits, such as gene expression and protein abundances, with GWAS summary data of complex phenotypes, for the purpose of detecting risk genes with their genetic effects mediated through the genetically regulated molecular traits [24, 31].

In this study, we've adapted OTTERS for protein data analysis (PWAS instead of TWAS), using CSF pQTL data from[19] to identify genes that influence Alzheimer's disease risk through related protein abundance. This is valuable because proteins are the functional units in cells and directly mediate biological processes. Proteins often correlate better with disease states than gene expressions, as gene expressions don't always translate to functional protein activity. Additionally, identifying causal proteins may provide more actionable therapeutic targets, and the CSF proteome is especially relevant for neurological disorders like Alzheimer's disease.

By extending OTTERS from gene expression to protein abundance, we bridge the gap between genetic associations and biological mechanisms in Alzheimer's disease research. When we find genetic variants that influence proteins in CSF, we can identify proteins that are strong predictors of AD by mapping back to the genetic variants that predict their abundances and determining if these variants are associated with AD status in GWAS data. This approach has the potential to reveal novel protein targets for further investigation and potential therapeutic development.

CHAPTER 2: METHODS

Data sources

The CSF pQTL summary statistics came from Western et al. (2024)[19]. This dataset comprised 3,506 individuals including 1,243 cognitively typical controls, 1,021 patients with late-onset Alzheimer's disease and 1,242 individuals with other neurodegenerative disorders.

The Alzheimer's disease GWAS summary statistics came from Bellenguez et al. (2022)[13], which contains genome-wide significant associations from the meta-analysis of 111,326 AD cases (including both clinically diagnosed and proxy cases) and 677,663 controls.

Protein Selection and Processing

For our analysis, we focused on 831 CSF proteins measured in the Western et al. study. We created an annotation file mapping each protein to its corresponding gene and genomic coordinates through a multi-step process. First, we matched protein identifiers (SOMASeqID format e.g. X10000.28) to their respective gene symbols. For proteins present in the ROSMAP (Religious Orders Study and Memory and Aging Project) dataset, we used their existing gene annotation file to obtain genomic coordinates. For the remaining proteins, we retrieved coordinates from the hg38 human reference genome annotation. We then filtered out genes on sex chromosomes to focus solely on autosomal chromosomes and created chromosome-specific annotation files for further analysis.

Of the 831 proteins, we successfully processed 765 through our initial data extraction pipeline. The remaining 66 proteins were excluded because they didn't exist in the Western et al. dataset. During the OTTERS analysis, an additional 24 proteins were excluded due to insufficient cis-SNPs for model convergence, resulting in 741 with valid statistical results.

Data Extraction and Preparation

While OTTERS was originally designed for TWAS using eQTLs, we adapted the framework to perform a proteome-wide association test using pQTL data. From the original pQTL summary statistics, we extracted data for the 765 CSF proteins. For each protein, we identified cis-SNPs located within ±1 Mb of the annotated gene boundaries and formatted their effect sizes and sample sizes to match OTTERS file structures. We also extracted the corresponding SNPs from the Alzheimer's disease GWAS summary statistics reported by Bellenquez et al. (2024) [13] and formatted the data accordingly. Additionally, we obtained European ancestry linkage disequilibrium (LD) reference panels. These LD panels helped us account for correlation structures between genetic variants, which is important when combining multiple SNPs into a predictive score. Without accounting for the correlations, we would risk conflating the effects of variants with high LD, leading to biased results.

OTTERS Two-Stage Analysis

Our PWAS framework followed OTTERS' two-stage approach outlined in Figure 1:

In stage 1 (Estimating pQTL weights), we applied four different polygenic risk score methods to estimate cis-pQTL weights using our formatted pQTL summary data from CSF and the LD reference panel: (P-value Thresholding with LD clumping (P+T,) Frequentist lassosum, Nonparametric Bayesian SDPR, Bayesian PRS-CS). Each of these methods produced a separate set of weights that reflect how strongly each SNP within a gene influences corresponding protein abundance levels within the reference CSF.

In stage 2 (association testing with AD) we used the pQTL weights from stage 1 to perform association testing with AD by combining the weights with the AD GWAS summary statistics. For each protein and each PRS method, stage 2 produced a PWAS Z-score and corresponding p-value indicating the strength of association between the genetically regulated protein levels and AD risk.

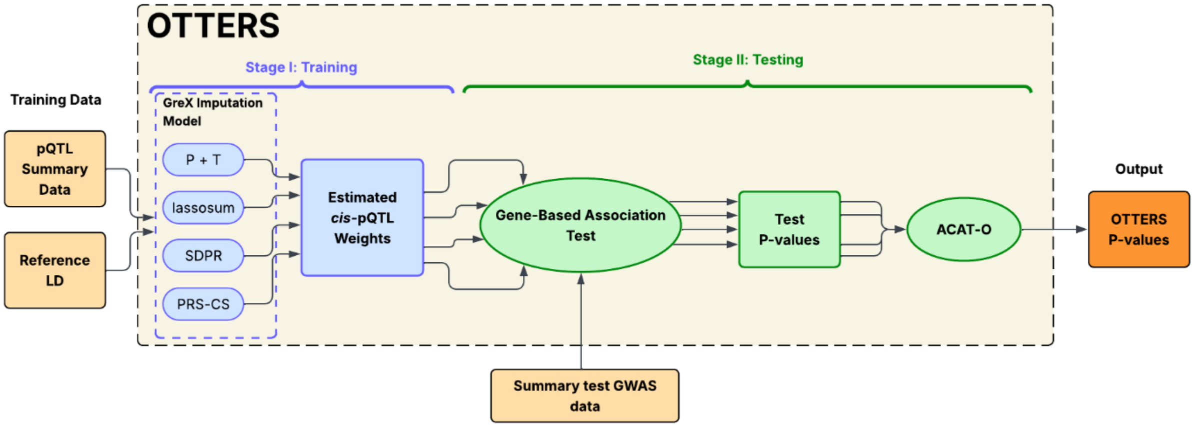


Figure 1. Overview of PWAS analysis using OTTERS framework [24]. Stage I: OTTERS uses 4 PRS models to estimate pQTL weights based on cis-SNPs (±1 Mb) and CSF pQTL summary statistics. Stage II: OTTERS conducts association testing between genetically predicted protein abundance and AD risk using GWAS summary statistics. The method concludes with an omnibus test that combines p-values from all 4 PRS methods.

Omnibus Testing with ACAT-O

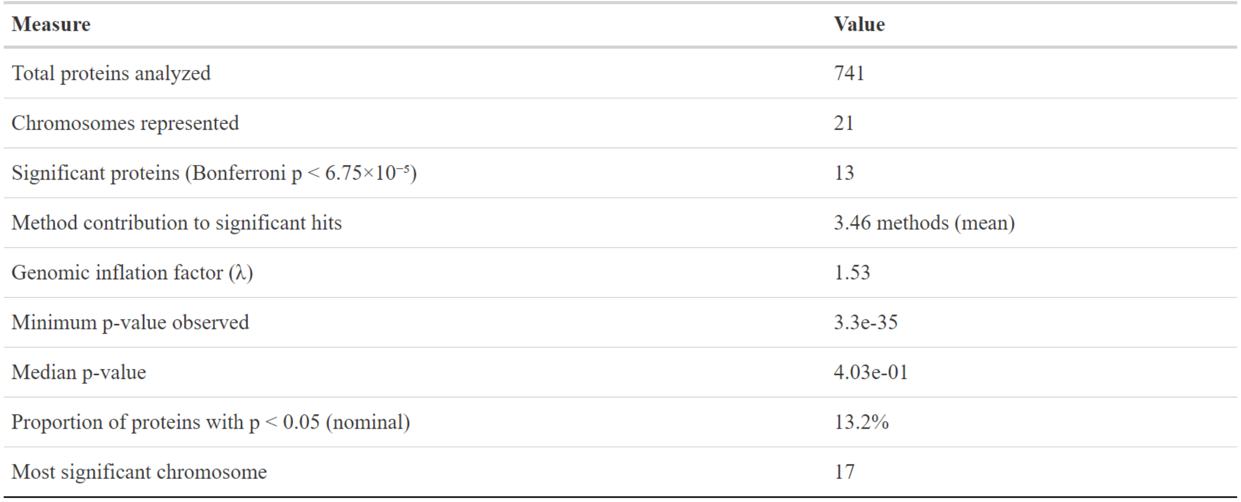
Since different PRS methods make different assumptions about genetic architecture, we used the Aggregated Cauchy Association Test (ACAT-O) to combine p-values from all four methods. Unlike methods that assume independence between tests, the ACAT-O accounts for potential correlations, using a Cauchy distribution for inference and borrowing strength across different models to increase statistical power while controlling type 1 error.

This gave us a single omnibus p-value for each protein, which represented its overall association with AD. The ACAT-O output included both the combined p-value and the number of PRS methods that contributed valid p-values to the test. A method might not contribute a valid p-value if there were insufficient SNPs within the gene region or if the algorithm failed to reach a stable solution due to issues like high correlation between genetic variants.

Of the 765 proteins initially processed, 741 (97%) had valid results from at least one PRS method and were included in the final analysis. We applied Bonferroni correction to account for multiple testing, using a significance threshold of p<6.75E-5 (0.05/741) to control for type 1 errors.

CHAPTER 3: RESULTS

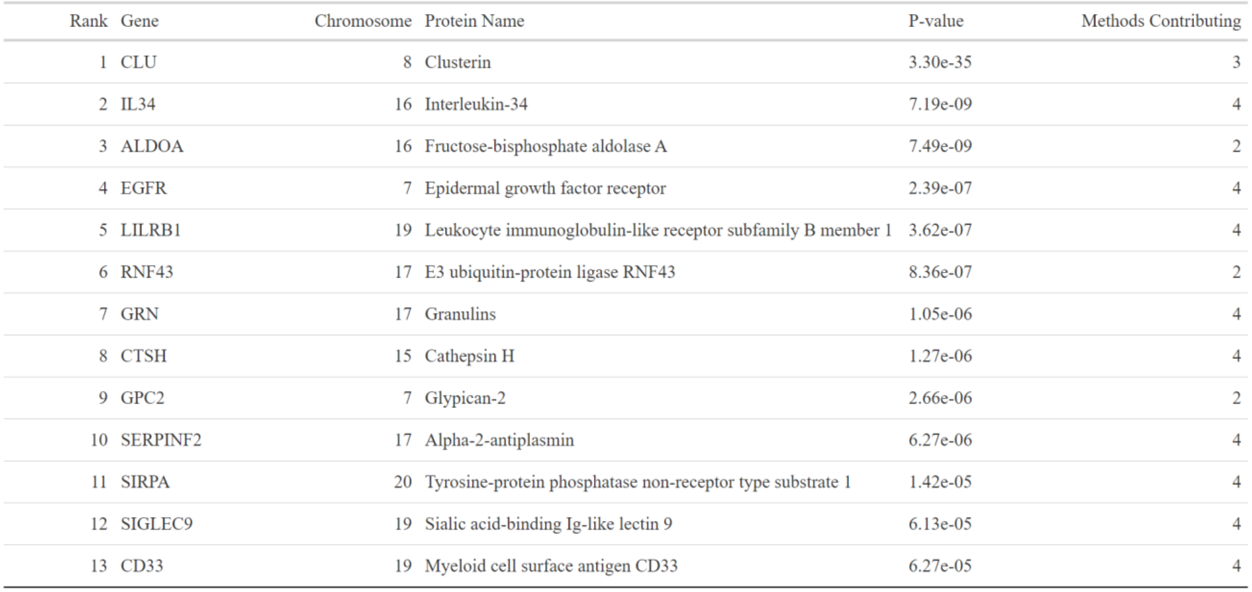
Of the 831 proteins in our annotation file, we successfully analyzed 741 using the OTTERS framework. From this analysis, we identified 13 proteins that were significantly associated with Alzheimer's disease after Bonferroni correction (p<6.75E-5) (Table 1). The most striking result was for Clusterin (*CLU*) (p = 3.30E-35), which showed a strong relationship with AD risk.



Measure	Value
Total proteins analyzed	741
Chromosomes represented	21
Significant proteins (Bonferroni $p < 6.75\times10^{-5}$)	13
Method contribution to significant hits	3.46 methods (mean)
Genomic inflation factor (λ)	1.53
Minimum p-value observed	3.3e-35
Median p-value	4.03e-01
Proportion of proteins with $p < 0.05$ (nominal)	13.2%
Most significant chromosome	17

Table 1. Summary statistics of ACAT-O p-values in CSF PWAS of Alzheimer's Disease.

Out of the 13 significant proteins identified, the number of contributing PRS methods varied, as shown in Table 2. The most significant protein, Clusterin (*CLU*) (p = 3.30 E-35), had contributions from only 3 of the 4 possible methods, yet still achieved an extremely significant p-value, which highlights the strength of its association with AD. In contrast, proteins such as Interleukin-34 (*IL34*), Leukocyte immunoglobulin-like receptor subfamily B member 1 (*LILRB1*), Granulins (*GRN*), Cathepsin H (*CTSH*), and Alpha-2-antiplasmin (SERPINF2) had contributions from 4 methods, demonstrating more consistency. Proteins with contributions from fewer methods, such as Fructose-bisphosphate aldolase A (*ALDOA*), E3 ubiquitin-protein ligase RNF43 (*RNF43*), and Glypican-2 (*GPC2*) (each with 2 contributing methods), could represent cases where the genetic architecture was unique enough to only be captured by the training method specifically tailored to detect such architecture. These differences in method contribution highlight the value of the omnibus approach, which can detect significant associations regardless of which specific methods best capture the underlying genetic architecture.



Rank	Gene	Chromosome	Protein Name	P-value	Methods Contributing
1	CLU	8	Clusterin	3.30e-35	3
2	IL34	16	Interleukin-34	7.19e-09	4
3	ALDOA	16	Fructose-bisphosphate aldolase A	7.49e-09	2
4	EGFR	7	Epidermal growth factor receptor	2.39e-07	4
5	LILRB1	19	Leukocyte immunoglobulin-like receptor subfamily B member 1	3.62e-07	4
6	RNF43	17	E3 ubiquitin-protein ligase RNF43	8.36e-07	2
7	GRN	17	Granulins	1.05e-06	4
8	CTSH	15	Cathepsin H	1.27e-06	4
9	GPC2	7	Glypican-2	2.66e-06	2
10	SERPINF2	17	Alpha-2-antiplasmin	6.27e-06	4
11	SIRPA	20	Tyrosine-protein phosphatase non-receptor type substrate 1	1.42e-05	4
12	SIGLEC9	19	Sialic acid-binding Ig-like lectin 9	6.13e-05	4
13	CD33	19	Myeloid cell surface antigen CD33	6.27e-05	4

Table 2. Proteins significantly associated with Alzheimer's disease after Bonferroni corrections. 13 proteins were identified using OTTERS framework with 4 PRS methods on CSF protein data. The rank is based on statistical significance, with p-value indicating the strength of association. "methods contributing" column shows how many PRS methods detected the protein's association with AD.

The Manhattan plot (Figure 2) illustrates the distribution of protein association across all chromosomes. There were clusters of significant proteins on chromosomes 16 (*IL34, ALDOA*), 17 (*RNF43, GRN, SERPINF2*), and 19 (*LILRB1, SIGLEC9, CD33*), which contain known AD risk loci [13, 32-34]. The extremely strong signal from Clusterin (*CLU*) on chromosome 8 also stands out.

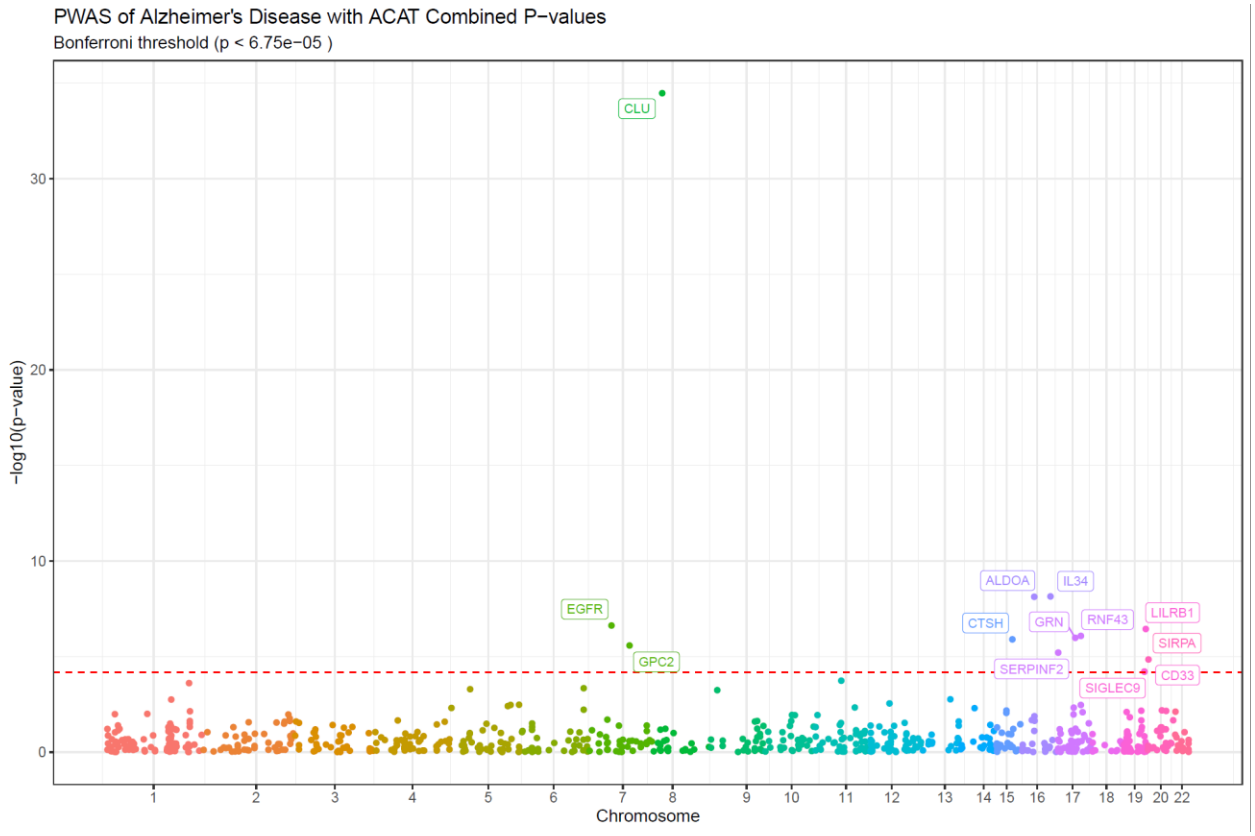


Figure 2. Manhattan plot of -log10 PWAS p-values across 741 CSF proteins. Each point represents the omnibus p-value for a protein. The red dashed line indicates the Bonferroni significance threshold (p < 6.75E-5). Thirteen proteins exceed this threshold and are labeled with their gene names. Clusterin (*CLU*) on chromosome 8 shows the strongest association (p = 3.30E-35).

Several other proteins showed significant associations, including Interleukin-34 (*IL34*) (p = 7.19E-09) and Fructose-bisphosphate aldolase A (*ALDOA*) (p = 7.49 E-09) on chromosome 16, and Epidermal growth factor receptor (*EGFR*) (p = 2.39E-07) on chromosome 7. The presence of multiple significant proteins on chromosome 19, where the *APOE* gene (strongest genetic risk factor for AD) is located, is also noteworthy [35-38].

Of the 13 significant genes we identified, 10 have been previously associated with AD in previous studies: *CLU, IL34, EGFR, GRN, CTSH, SIRPA, CD33, LILRB1, SERPINF2*, and *GPC2.* [13, 32-34]. Our findings further validate these genes' association with AD risk. Our analysis also revealed three genes not previously highlighted in AD research *ALDOA, RNF43,* and *SIGLEC9*. These represent potential biomarkers or therapeutic targets for AD. Particularly, *ALDOA* (p = 7.49E-09) on chromosome 16 showed a strong association and may represent a new pathway involved in AD pathology.

The quantile-quantile (QQ) plot (Figure 3) shows some deviation of p-values from the expected distribution under the null hypothesis, which could result from polygenicity of AD [39]. We will explore the use of variance-control methods to deal with this bias in future work.

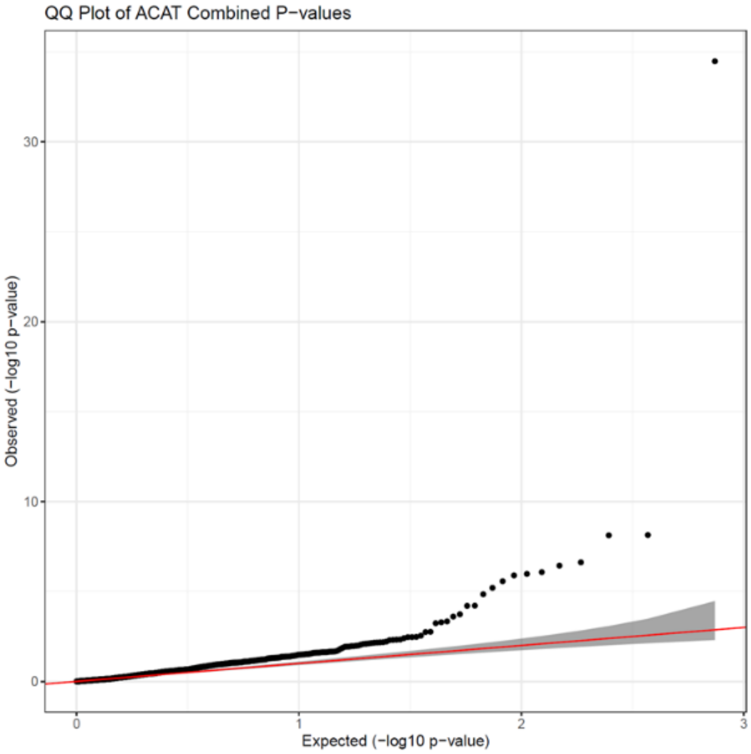


Figure 3. QQ plot of observed vs. expected p-values for protein-Alzheimer's Disease associations. Significant proteins deviate above the expected distribution line with Clusterin (*CLU* showing the most extreme significance (-log10(p) ≈ 35). This pattern confirms true biological signals rather than systematic bias. Genomic inflation factor (λ) = 1.53.

The strength and number of associations identified through the ACAT-O shows the advantage of integrating multiple statistical methods to capture the diverse genetic architectures underlying protein regulation. By using 4 PRS methods through OTTERS, we were able to identify strong associations that might have been missed by using a single approach.

More research can be done to characterize these proteins and understand their relationships to AD pathways. It would be useful to determine whether these proteins represent potential biomarkers or therapeutic targets for AD drugs in the future.

CHAPTER 4: DISCUSSION

This study demonstrates the successful adaptation of OTTERS, a framework that was originally designed for TWAS studies, to perform PWAS studies using summary-level data. This study shows that OTTERS can be used to identify disease-associated proteins. This is valuable for studying diseases like Alzheimer's, where protein changes may directly reflect disease processes more than gene expression changes. Our successful identification of 13 significant proteins associated with AD validates the cross-platform application of OTTERS.

An important advantage of OTTERS is the ability to use summary level data to find these associations. Summary-level data is much more widely accessible compared to individual-level data, due to privacy concerns. OTTERS allowed us to use pQTL data from Western et al. study that had 3506 individuals and GWAS data from the comprehensive AD GWAS by [13] (111,326 cases and 677,663 controls). These large sample sizes increase statistical power and improve the reliability of our findings. Additionally, summary level methods like OTTERS are more computationally efficient than methods that require individual-level data, as they avoid the need to store and process sensitive health information, while maintaining statistical power of large studies. However, a limitation to consider is that individual-level data allows for more flexibility with modeling approaches and allows exploration of trends that might be lost in summary statistics.

Some other limitations to consider is that the SOMAscan platform used in the Western et al. study, while comprehensive, was not able to interrogate all the relevant CSF proteins for our analysis. Secondly, because our analysis focuses on genetic components of protein regulation, it might miss environmentally influenced protein changes that might be relevant to AD. Lastly, the use of the European ancestry population as the LD reference and for the GWAS data limits the ability to generalize our findings to other ancestral groups.

Our work opens the door to continuing research on the 13 identified proteins, which can produce insights into AD pathology. OTTERS can be applied to other neurodegenerative diseases or extend to different tissue/ protein types. Future improvements could include incorporating additional PRS methods to capture even more diverse genetic architectures and expanding into more diverse ancestral populations. Longitudinal studies examining how these protein associations relate to disease progression could also provide valuable insights. Finally, developing drug repurposing strategies targeting the identified proteins is a translational direction that could accelerate therapeutic development for Alzheimer's disease. By bridging the gap between genetic associations and protein-level changes, our approach provides a framework to understand the biological structure of Alzheimer's disease and potentially other neurological diseases.

REFERENCES

1. Breijyeh Z, Karaman R. Comprehensive Review on Alzheimer's Disease: Causes and Treatment. Molecules. 2020;25(24).

2. De-Paula VJ, Radanovic M, Diniz BS, Forlenza OV. Alzheimer's disease. Subcell Biochem. 2012;65:329-52.

3. Cipriani G, Dolciotti C, Picchi L, Bonuccelli U. Alzheimer and his disease: a brief history. Neurol Sci. 2011;32(2):275-9.

4. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. Nature Reviews Genetics. 2013;14(2):139-49.

5. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, et al. Role of Genes and Environments for Explaining Alzheimer Disease. Archives of General Psychiatry. 2006;63(2):168-74.

6. Andrade-Guerrero J, Santiago-Balmaseda A, Jeronimo-Aguilar P, Vargas-Rodríguez I, Cadena-Suárez AR, Sánchez-Garibay C, et al. Alzheimer's Disease: An Updated Overview of Its Genetics. Int J Mol Sci. 2023;24(4).

7. Bergem AL, Engedal K, Kringlen E. The role of heredity in late-onset Alzheimer disease and vascular dementia. A twin study. Arch Gen Psychiatry. 1997;54(3):264-70.

8. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010;363(2):166-76.

9. Pearson TA, Manolio TA. How to interpret a genome-wide association study. Jama. 2008;299(11):1335-44.

10. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. Nature. 2022;610(7933):704-12.

11. Fernandez-Rozadilla C, Timofeeva M, Chen Z, Law P, Thomas M, Schmit S, et al. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. Nat Genet. 2023;55(1):89-99.

12. Tcheandjieu C, Zhu X, Hilliard AT, Clarke SL, Napolioni V, Ma S, et al. Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. Nat Med. 2022;28(8):1679-92.

13. Bellenguez C, Küçükali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat Genet. 2022;54(4):412-36.

14. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nature Genetics. 2016;48(3):245-52.

15. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics. 2015;47(9):1091-8.

16. 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.

17. Ferkingstad E, Sulem P, Atlason BA, Sveinbjornsson G, Magnusson MI, Styrmisdottir EL, et al. Large-scale integration of the plasma proteome with genetics and disease. Nat Genet. 2021;53(12):1712-21.

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

19. Western D, Timsina J, Wang L, Wang C, Yang C, Phillips B, et al. Proteogenomic analysis of human cerebrospinal fluid identifies neurologically relevant regulation and implicates causal proteins for Alzheimer's disease. Nature Genetics. 2024;56(12):2672-84.

20. Sasayama D, Hattori K, Ogawa S, Yokota Y, Matsumura R, Teraishi T, et al. Genome-wide quantitative trait loci mapping of the human cerebrospinal fluid proteome. Hum Mol Genet. 2017;26(1):44-51.

21. Hansson O, Kumar A, Janelidze S, Stomrud E, Insel PS, Blennow K, et al. The genetic regulation of protein expression in cerebrospinal fluid. EMBO Mol Med. 2023;15(1):e16359.

22. Kaiser S, Zhang L, Mollenhauer B, Jacob J, Longerich S, Del-Aguila J, et al. A proteogenomic view of Parkinson's disease causality and heterogeneity. npj Parkinson's Disease. 2023;9(1):24.

23. Kauwe JS, Bailey MH, Ridge PG, Perry R, Wadsworth ME, Hoyt KL, et al. Genome-wide association study of CSF levels of 59 alzheimer's disease candidate proteins: significant associations with proteins involved in amyloid processing and inflammation. PLoS Genet. 2014;10(10):e1004758.

24. Dai Q, Zhou G, Zhao H, Võsa U, Franke L, Battle A, et al. OTTERS: a powerful TWAS framework leveraging summary-level reference data. Nat Commun. 2023;14(1):1271.

25. Privé F, Vilhjálmsson BJ, Aschard H, Blum MGB. Making the Most of Clumping and Thresholding for Polygenic Scores. Am J Hum Genet. 2019;105(6):1213-21.

26. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748-52.

27. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 2018;58(1):267-88.

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

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

30. Zhou G, Zhao H. A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics. PLoS Genet. 2021;17(7):e1009697.

31. Hu T, Liu Q, Dai Q, Parrish RL, Buchman AS, Tasaki S, et al. Proteome-wide association studies using summary pQTL data of three tissues identified 30 risk genes of Alzheimer's disease dementia. medRxiv. 2024.

32. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet. 2019;51(3):404-13.

33. Baker M, Mackenzie IR, Pickering-Brown SM, Gass J, Rademakers R, Lindholm C, et al. Mutations in progranulin cause tau-negative frontotemporal dementia linked to chromosome 17. Nature. 2006;442(7105):916-9.

34. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45(12):1452-8.

35. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science. 1993;261(5123):921-3.

36. Liu CC, Liu CC, Kanekiyo T, Xu H, Bu G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat Rev Neurol. 2013;9(2):106-18.

37. Lumsden AL, Mulugeta A, Zhou A, Hyppönen E. Apolipoprotein E (APOE) genotype-associated disease risks: a phenome-wide, registry-based, case-control study utilising the UK Biobank. EBioMedicine. 2020;59:102954.

38. Zhao J, Fu Y, Yamazaki Y, Ren Y, Davis MD, Liu C-C, et al. APOE4 exacerbates synapse loss and neurodegeneration in Alzheimer's disease patient iPSC-derived cerebral organoids. Nature Communications. 2020;11(1):5540.

39. Liang Y, Nyasimi F, Im HK. Pervasive polygenicity of complex traits inflates false positive rates in transcriptome-wide association studies. bioRxiv. 2024.