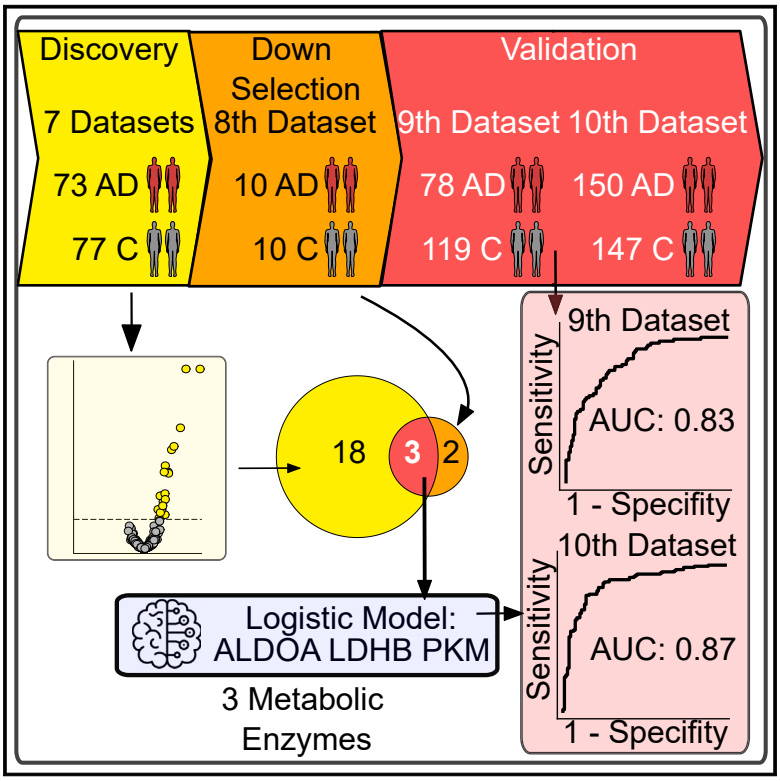
Cell Reports
Medicine

Article

Meta-analysis of published cerebrospinal fluid proteomics data identifies and validates metabolic enzyme panel as Alzheimer's disease biomarkers

Graphical abstract



Highlights

- We provide a generic strategy for unbiased meta-analysis of proteomics data
- We apply our strategy to 664 CSF samples from ten Alzheimer's disease cohorts
- Mostly metabolic enzymes are found to be altered in CSF samples
- Discovery and validation of three CSF biomarkers in 170 and 494 samples, respectively

Authors

Patrick W. van Zalm, Saima Ahmed, Benoit Fatou, ..., Henrik Zetterberg, Judith A. Steen, Hanno Steen

Correspondence

hanno.steen@childrens.harvard.edu

In brief

Van Zalm and colleagues present a strategy for meta-analysis of proteomics data. Its application is shown in CSF proteomics datasets from ten Alzheimer's cohorts. Three metabolic enzyme biomarkers are found to be significantly altered in Alzheimer's disease in 170 samples from eight cohorts and further validated in 494 samples.

van Zalm et al., 2023, Cell Reports Medicine *4*, 101005
April 18, 2023
https://doi.org/10.1016/j.xcrm.2023.101005

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Meta-analysis of published cerebrospinal fluid proteomics data identifies and validates metabolic enzyme panel as Alzheimer's disease biomarkers

Patrick W. van Zalm,[1,2] Saima Ahmed,[1] Benoit Fatou,[1] Rudy Schreiber,[2] Omar Barnaby,[1] Adam Boxer,[3] Henrik Zetterberg,[4,5,6,7] Judith A. Steen,[8,9,10] and Hanno Steen[1,9,10,11,*]

[1]Department of Pathology, Boston Children's Hospital, and Department of Pathology, Harvard Medical School, Boston, MA, USA
[2]Department of Neuropsychology and Psychopharmacology, EURON, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, the Netherlands
[3]Memory and Aging Center, Department of Neurology, Weill Institute for Neuroscience, University of California, San Francisco, CA, USA
[4]Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, the Sahlgrenska Academy at the University of Gothenburg, Mölndal, Sweden
[5]Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden
[6]UK Dementia Research Institute at UCL, London, UK
[7]Department of Neurodegenerative Disease, UCL Institute of Neurology, London, UK
[8]F.M. Kirby Neurobiology Center, Boston Children's Hospital, and Department of Neurology, Harvard Medical School, Boston, MA, USA
[9]Neuroiology Program, Boston Children's Hospital, Boston, MA, USA
[10]These authors contributed equally
[11]Lead contact
*Correspondence: hanno.steen@childrens.harvard.edu
https://doi.org/10.1016/j.xcrm.2023.101005

SUMMARY

To develop therapies for Alzheimer's disease, we need accurate *in vivo* diagnostics. Multiple proteomic studies mapping biomarker candidates in cerebrospinal fluid (CSF) resulted in little overlap. To overcome this shortcoming, we apply the rarely used concept of proteomics meta-analysis to identify an effective biomarker panel. We combine ten independent datasets for biomarker identification: seven datasets from 150 patients/controls for discovery, one dataset with 20 patients/controls for down-selection, and two datasets with 494 patients/controls for validation. The discovery results in 21 biomarker candidates and down-selection in three, to be validated in the two additional large-scale proteomics datasets with 228 diseased and 266 control samples. This resulting 3-protein biomarker panel differentiates Alzheimer's disease (AD) from controls in the two validation cohorts with areas under the receiver operating characteristic curve (AUROCs) of 0.83 and 0.87, respectively. This study highlights the value of systematically re-analyzing previously published proteomics data and the need for more stringent data deposition.

INTRODUCTION

The most prevalent form of dementia, Alzheimer's disease (AD), is characterized by gradual deterioration of memory, thinking, and reasoning as well as by depression.[1] Differentiation of AD from other types of dementia, and hence targeted therapeutic development, has been impeded by the variability of clinical symptoms both within and between dementias[2] but could be facilitated by discovery of biomarkers in body fluids, especially cerebrospinal fluid (CSF) due to its proximity to and interactions with the brain.[3,4] The literature contains numerous reports of molecules selectively enriched in AD CSF. Unfortunately, few of these so-called biomarkers have been validated in independent cohorts. Many factors affecting variability such as small cohort sizes, inconsistent diagnostic criteria, and differences in sample handling, data acquisition, and analysis could underlie lack of irreproducibility of proteomics data. We therefore hypothesized that we could identify robust AD biomarkers by performing an unbiased meta-analysis of publicly available proteomics datasets from studies aimed at identifying AD biomarkers.

Although meta-analyses are common in other omics fields, leveraging of multiple published datasets to overcome statistical power limitations and lack of reproducibility has been neglected in the proteomics field. Only very few proteomics meta-analysis papers have recently been published.[5–7] Even so, that meta-analysis used the published result tables at face value, not addressing the heterogeneity introduced by uncoordinated databases and data search strategies. In contrast, our meta-analysis strategy attempted to minimize such heterogeneities.

In our study, we retrieved and re-analyzed the raw liquid chromatography mass spectrometry (LC/MS) data from six vetted published studies and one in-house dataset on quantitative CSF proteomics in the context of AD. The identified biomarker candidates, which were highly enriched in proteins associated

with metabolic processes, were validated using a two-pronged approach: firstly, by analyzing CSF specimens from an additional independent cohort to down-select the list of biomarker candidates and, secondly, by querying two recently published large-scale CSF proteomics studies aimed at identifying biomarkers for AD.[8,9] To minimize risk of overfitting, each dataset was only used once during the meta-analysis. This strategy resulted in identification of a set of three metabolic enzymes as CSF protein biomarkers, which were discovered and validated in arguably the largest (n = 664) and most diverse set (ten different cohorts) of patients with AD and age-matched healthy controls reported to date. All three members of the discovered biomarker panel are involved in glycolysis, i.e., a pathway well known to be dysregulated in AD.[10,11]

RESULTS AND DISCUSSION

Data pre-processing

We first identified potentially usable CSF proteomics datasets by performing an extensive literature review in PubMed. We retrieved 394 proteomic AD-focused biomarker discovery studies (Figure 1A) and chose for analysis those that (1) were published between January 1, 2010, and January 31, 2019; (2) used CSF collected *ante mortem* by lumbar puncture; (3) used LC-tandem MS (MS/MS) operated in data-dependent acquisition (DDA) mode for proteomic profiling; (4) used high-resolution/high-accuracy instrumentation; (5) made raw high-resolution/high-accuracy LC/MS data available for re-analysis; (6) provided discernable and appropriate naming of the data files; and (7) shared the relevant meta-information. Six studies and their respective datasets met our criteria (Table 1; see also STAR Methods). We included one additional in-house dataset (AD 4 and five control samples, tandem mass tagging [TMT] study with samples provided by UCSF, San Francisco, CA, USA) resulted in a total of 73 AD cases and 77 controls from seven independent cohorts assembled in five countries on both sides of the Atlantic Ocean (Figure 1A).[12–17] Next, we retrieved raw LC/MS data from these seven cohorts and re-searched against the UniProt human protein canonical sequence database (downloaded on January 17, 2019: 20,320 entries) using MSFragger/Fragpipe.[18–20] This method minimizes variability due to varying protein sequence databases (e.g., with or without isoforms, or different database versions) and differences in the search, scoring, quantification, and grouping algorithms. Moreover, by concentrating on the canonical protein forms, we avoided the problems of artifactual isoform calling within a dataset during the protein grouping caused by the presence or absence of spurious peptide spectral matches. To normalize data, we used the average of the median intensities of reference sample. As alternative, if no reference samples were available, the median of summed intensities of all samples was used (Figures 1B and1C).

We applied an in-house-developed outlier detection method to each dataset based on the assumption that most proteins should correlate between samples and that the lack thereof indicates problems with sample collection, processing, or data acquisition. This method correlates all protein intensities from each sample with the dataset-specific median intensities. Samples with correlation values four standard deviations from the mean correlation value were classified as outliers and excluded from further analysis (Figure 1D). The outlier selection method led to the removal of five samples from the seven datasets, resulting in 69 AD cases and 76 controls. Manual inspection of the outlier detection method in principal-component analysis (PCA) provided evidence that our objective outlier detection method was effective in removing outliers (Figure S1). This outlier detection method has the advantage of allowing for quickly and easily adjusting the desired stringency. For example, if one can test many potential biomarkers in future studies, less stringent criteria may be used. On the other hand, if more elaborate functional or antibody-based assays such as ligand-binding assays or ELISAs are used for validation, then higher stringency might be more appropriate.

Following outlier removal, protein intensities were dataset-wise normalized using the Z score calculated from the control samples within each dataset. Differences in instrumentation and methodology used for each of the datasets precluded a direct comparison of the reported intensity values as evidenced in the PCA plot, showing that the raw intensity values from the different datasets were completely separated (Figure 1E). PCAs of the data before and after the Z score normalization (Figures 1E and 1F) clearly show the need for and suitability of the applied Z score normalization procedure.

Discovery of biomarker candidates

The curated combined dataset after removal of outliers included 145 samples with a total of 2,808 identified proteins (Figure S2A). We first performed a Benjamini-Hochberg-corrected Fisher's exact test to query whether any protein was specifically detected in only AD cases or controls. This analysis did not result in any significant proteins when using an adjusted p value of <0.05 as a cutoff value. Therefore, we decided to keep only those proteins observed in at least 30% of the samples, leaving 434 proteins (Figure S2B). Next, we used the non-parametric Mann-Whitney U test followed by Benjamini-Hochberg multiple testing correction. The Mann-Whitney U test yielded 21 proteins (all upregulated) with statistically significant abundance differences between AD and controls (Figures 2A, S3A, and S3B).

Interestingly, some of the proteins that we found to be significant in the meta-analysis were not found to be significant in some of the individual datasets (Figure S3C), which supports our hypothesis that the meta-analysis of multiple independent datasets can reveal novel biomarker candidates.

Importantly, some of our biomarker candidates have previously been linked to AD, suggesting that our meta-analysis strategy yields meaningful data. For example, pyruvate kinase (PKM) and fructose-bisphosphate aldolase A (ALDOA) were previously reported to be significantly enriched in AD mutation carriers.[21] Another targeted proteomics study found osteopontin (SPP1), malate dehydrogenase (MDH1), and insulin-like growth factor-binding protein 2 (IGFBP2) to be significantly enriched in the CSF of AD cases compared with controls.[22] In contrast, some proteins that have been repeatedly considered promising candidates such as Chitinase-3-like protein 1 (YKL-40), neurofilament light chain, or neurosecretory protein VGF were not found to be significant in our meta-analysis.[23–25] The inconsistencies between the different proteomics studies as well as currently pursued biomarkers highlights the need for testing and validation

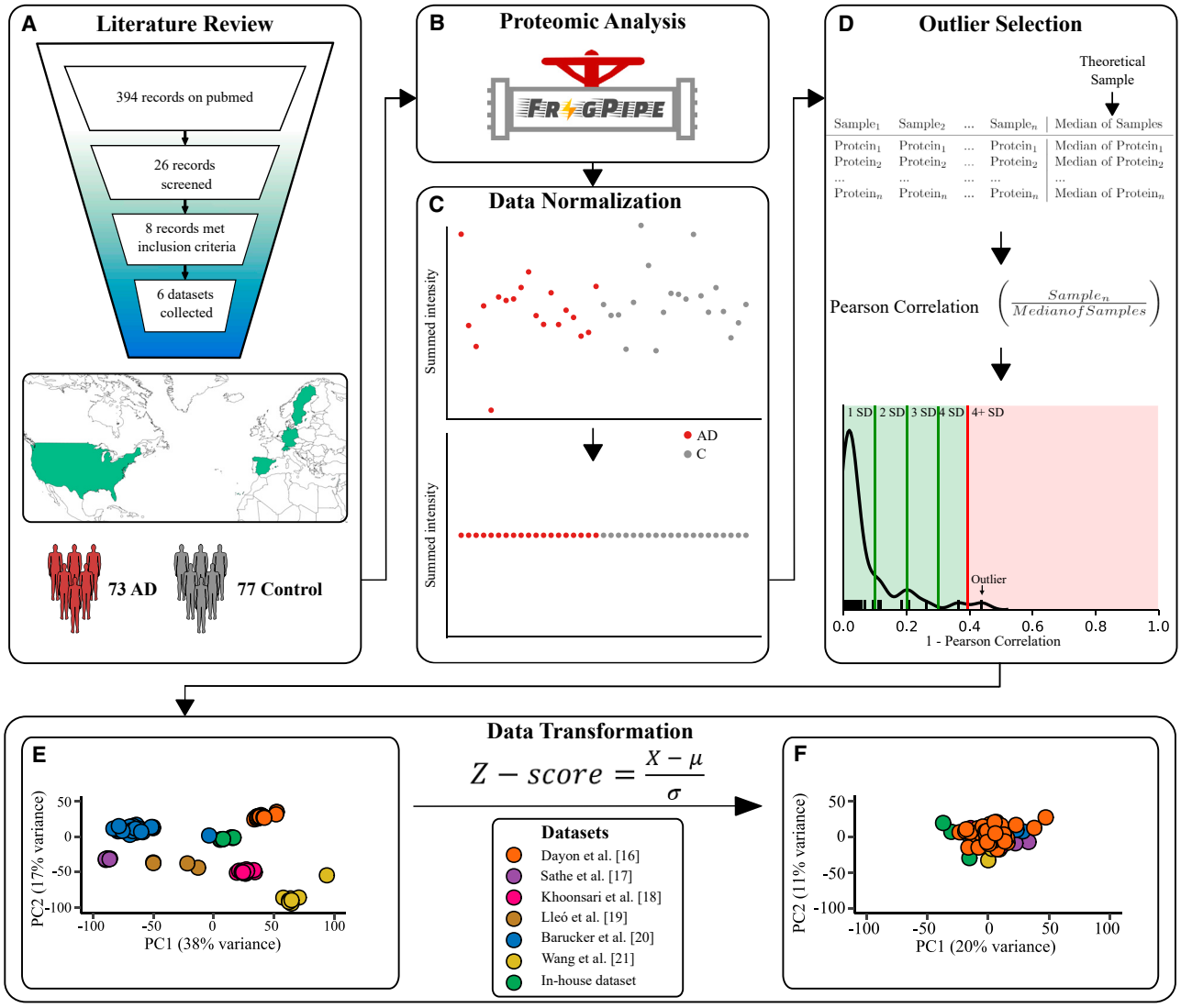


Figure 1. Data collection and pre-processing

(A) The literature was searched on PubMed, resulting in 394 records, of which six did result in the collection of raw MS data after the exclusion and inclusion criteria were taken into account. With the addition of one in-house dataset, a total of 150 samples were collected, derived from seven cohorts from five countries. (B–F) All data was systematically re-analyzed in MSFragger/Fragpipe (B) followed by data normalization (C). For the systematic removal of outliers, we developed a method where a theoretical sample was created by calculating the median intensity of each protein in a dataset (D). Next, Pearson correlation between each of the individual samples and the theoretical sample was calculated, and any sample that was more than three standard deviations removed from the theoretical sample was considered an outlier and removed from downstream analysis. The process described in (B)–(D) was repeated for each of the seven cohorts, whereafter data were combined, resulting in a PCA plot, as shown in (E). To overcome the variability between datasets, a *Z* score transformation was used, which resulted in a homogeneous dataset (F). *Z* score transformation is a promising approach for comparing heterogeneous datasets.

in multiple independent studies in order to obtain robust and unbiased results and thereby identify proteins most likely to be associated with the pathophysiology of AD pathology in CSF.

Bioinformatic analysis of biomarker candidates

We used two complementary methods to identify physical and functional interactions of potential biomarkers. We exported the 21 significant proteins to the CytoScape ClueGO tool to first enrich for GO biological processes followed by the clustering of non-redundant terms (Figure 2B). Next, we enriched the 21 biomarker candidates against the Molecular Signatures Database (MSigDB) Hallmark dataset (Figure 2C).[26] Remarkably, the ClueGO tool enriched for four clusters of GO annotations, of whom three were related to energy and metabolism. Next, the MSigDB Hallmark revealed glycolysis-related protein enrichment with the highest statistical confidence, pointing to an important role of metabolism in AD.

Our findings are supported by a plethora of published studies describing the dysregulation of redox processes and mitochondrial function in AD-diseased brain tissue.[8,9,27–33] Furthermore,

Table 1. Information about the six studies with associated AD CSF MS data selected for re-analysis

Reference	Country of sample origin	Number of AD samples	Number of control samples	Disease definition	Control definition	Type of MS	Peptide labeling	Enrichment or depletion
Dayon et al.[12]	Switzerland	30	30	CSF P-tau181/Aβ1–42 ratio >0.0779	healthy controls, CDR = 0	Orbitrap Elite	TMT 6 plex	MARS14
Sathe et al.[13]	USA	5	5	in-house classification; similar to NINCDS-ADRDA	healthy controls, CDR = 0	Orbitrap Fusion Lumos	TMT 10 plex	MARS14
Khoonsari et al.[14]	Sweden	10	10	brain imaging, laboratory testing, neurological and cognitive examinations	control had normal cognition according to MMSE performance	7T LTQ-FT	None	MARS-Hu7
Lleó et al.[15]	Spain	1[a]	3[a]	NINCDS-ADRDA	assessed by neurologist	Orbitrap Velos Pro	none	albumin and IgG
Barucker et al.[16]	Germany	19	20	MMSE/MRI classification	age matched	Orbitrap Velos	none	no information
Wang et al.[17]	USA	4	4	NINCDS-ADRDA	controls had normal cognition according to MMSE performance	Q Exactive	none	glycoproteomics

Information about the origin of the samples, the number of samples for AD and control, the type of MS instruments used, quantification methods, and information about enrichment of the samples. NINCDS-ADRDA, National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association; CDR, clinical dementia rating; MMSE, mini-mental state exam (MMSE); IgG, immunoglobulin G.
[a]The samples that were used by Lleó et al. are pooled samples comprising 10 CSF samples per pool.

hypometabolism in patients with mild cognitive impairment (MCI) and AD has been reported using Fluoro-2-deoxy-D-glucose positron emission tomography (FDG-PET) imaging in affected brain regions.[34] Redox proteomics on AD postmortem brain tissue has shown that multiple metabolic pathways are affected, including pyruvate kinases, fructose bisphosphate aldolases, malate dehydrogenases, and enolases, all proteins that we observe to be significantly upregulated in the CSF of patients with AD.[11,35] More specifically, these proteins were found to be inactivated due to oxidation, leading to reduced activity of adenosine triphosphate (ATP) synthesis and subsequent reduction of glucose metabolism in the brain. The observed upregulation of such proteins in CSF might be a direct representation of the diminished activity of the brain hypometabolism.

Down-selection of biomarker candidates

To further validate the 21 biomarker candidates in a data-driven unbiased fashion, we used a two-step approach: firstly, we down-selected the 21 biomarker candidates. To this end, an independent second in-house dataset was generated comprising AD 10 (mean age 73.2 [±9.4], 50% females) and ten non-AD (mean age 71.6 [±7], 50% females) control samples—the samples had been collected in Gothenburg, Sweden. The analysis by unbiased discovery LC-MS without any depletion and/or fractionation resulted in the FragPipe-based identification and quantification of 488 CSF proteins in this down-selection dataset. The same data analysis workflow that we used for the meta-analysis was followed, i.e., data normalization, test for outliers (which removed one sample), Fisher's exact test (which did not identify any AD or control-specific proteins), and filtering based on completeness (70% cutoff).

The statistical analysis of the validation cohort using Benjamini-Hochberg-corrected Mann-Whitney U test resulted in a set of five proteins that were significantly different between AD and control CSF (adjusted $p < 0.05$) (Figure 3A). Of these five proteins, three overlapped with the biomarker candidates identified in our meta-analysis, namely ALDOA, L-lactate dehydrogenase B chain (LDHB), and PKM (Figures 3B and 3C).

These three markers have been described in multitude for AD: ALDOA has been found to be on the protein level in the cortex and substantia nigra and also affected gene expression the entorhinal cortex.[36,37] In the hippocampal proteome of patients with AD, PKM was found to be altered, and in a genome-wide function screen study, it was found that PKM is a regulator of amyloid β production, which was validated in a mouse study.[38,39] Next, LDHB together with other mitochondrial proteins was affected in an AD brain tissue transcriptomics study.[40] Querying the Human Protein Atlas for these three proteins showed that all three are indeed cytosolic proteins, with an even expression across all brain regions.[41,42] This observation raised the question how these bona fide cytosolic proteins were observed as being dysregulated in the CSF: were they actively secreted, or are they the result of neuronal death, which results in the spillage of these proteins into the CSF? As we did not observe any enrichment of other abundant intracellular proteins such as ribosomal or proteasomal proteins or histone, which would be associated with cell death, we hypothesize that those glycolytic proteins in the CSF are indeed the result of a secretion process.

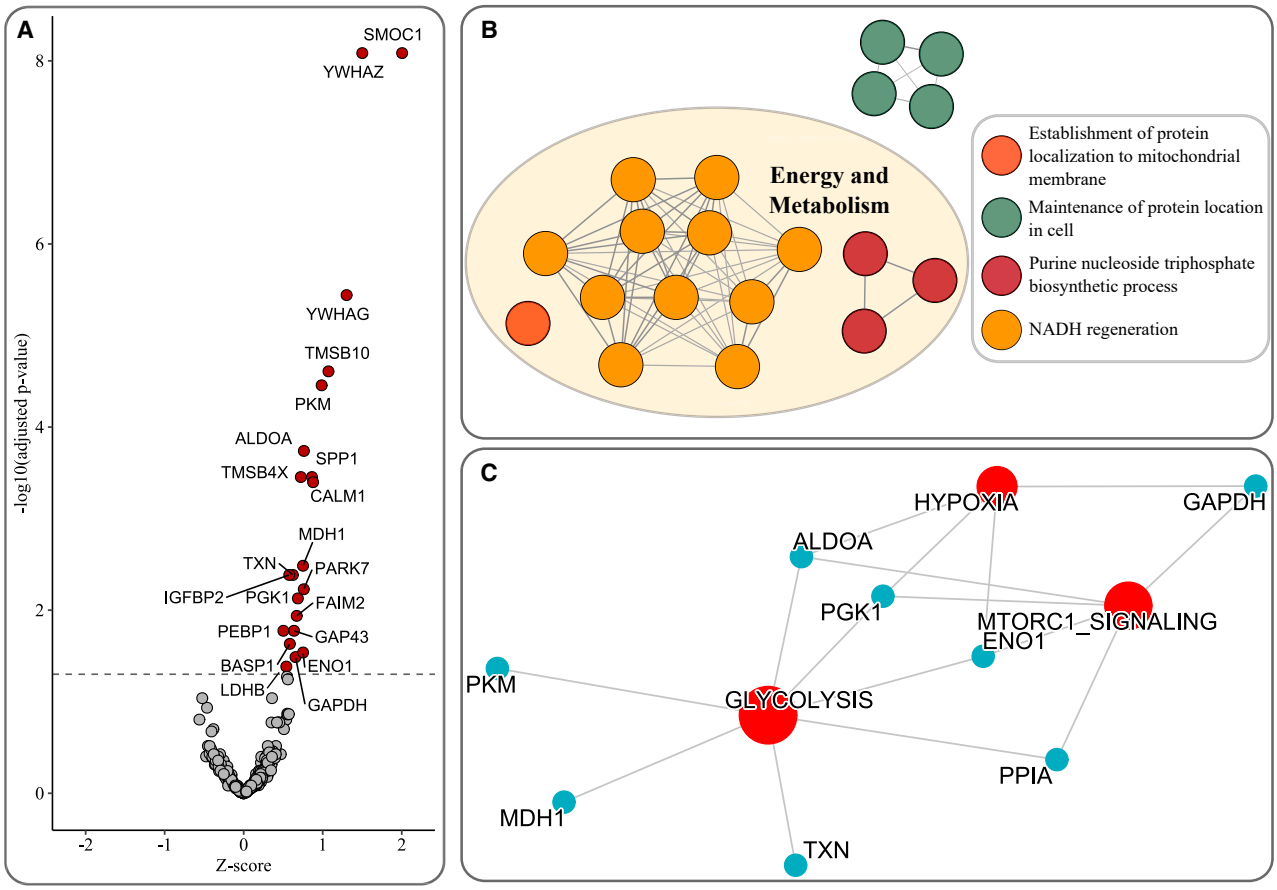


Figure 2. Discovery cohort: Statistics and enrichment

(A) The *Z* scored data was statistically analyzed with the Mann-Whitney U test and Benjamini-Hochberg multiple comparison correction, resulting in the discovery of 21 biomarker candidates across seven independent datasets. The 21 biomarker candidates were held against the Gene Ontology (GO) Biological Process database and visualized using the Cytoscape ClueGO tool, which showed that three out of the four clusters of GO annotations are linked to energy and metabolism.

(B) Biomarker candidates were held against the MSigDB Hallmark database, which particularly enriched for glycolysis (C).

Validation of biomarker candidates

For a second validation of our panel of biomarkers candidates, we used two additional independent datasets. To this end, we leveraged two recently published large-scale CSF proteomics studies aimed at identifying AD biomarkers: Johnson et al. used a TMT strategy to analyze CSF samples from 150 AD cases (mean 68.2 [±8.3], 55% females) and 147 controls (mean 65.0 [±8.2], 72% females), while Bader et al. applied a label-free quantification (LFQ) data-independent acquisition strategy to the analysis of CSF samples from 88 AD cases (mean 71.7 [±8.1], 55% females) and 109 controls (mean 65.5 [±14.4], 46% females).[8,9] From each dataset, we extracted the intensities of the three biomarker candidates ALDOA, LDHB, and PKM, which were all considered significant in both datasets. We then tested these three proteins individually for their differentiating capabilities using receiver operating characteristic analysis. The areas under the receiver operating characteristic curve (AUROCs) for ALDOA, LDHB, and PKM were 0.85, 0.79, and 0.82, respectively, for the Johnson et al. dataset (Figure 4A) and 0.81, 0.78, and 0.73, respectively, for the Bader et al. dataset (Figure 4B). Next, we used logistic regression modeling to assess the differentiating capabilities of the biomarker panel. To minimize overfitting risk, we performed two separate analyses: firstly, we trained a model on the Johnson et al. dataset and validated on the Bader et al. data. Secondly, we trained a model on the Bader et al. dataset and validated on the Johnson et al. data. These analyses resulted in very similar AUROCs of 0.87 and 0.83 for Johnson et al. and Bader et al. datasets, respectively, confirming that there is no overfitting observable.

To better assess our purely data-driven three-step approach, we evaluated the list of 21 biomarker candidates we identified in the discovery phase of our proteomic meta-analysis. To this end, we investigated the following aspects of these 21 proteins in the two validation datasets: (1) which proteins were observed in both validation datasets, (2) how their individual performance is (as assessed by AUROC averaged across the two validation datasets), and (3) how combinations of three perform between the two datasets (as assessed by AUROC averaged across the two validation datasets). These criteria were compared with the values of the three proteins selected by our three-step

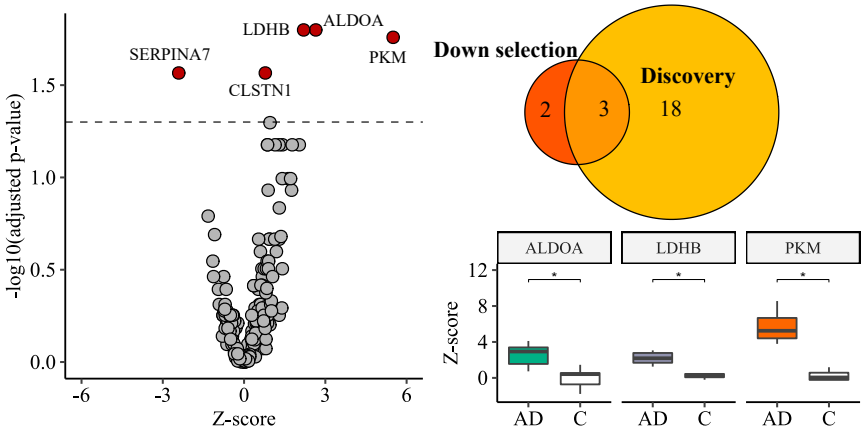


Figure 3. Biomarker down-selection cohort

(A and B) Statistical analysis (Mann-Whitney U test and Benjamini-Hochberg correction) of the pruning cohort did result in five significant proteins (A), of which three proteins did overlap with the results of meta-analysis (B).

(C) Boxplots of the three significant biomarker candidates (ALDOA, LDHB, PKM) are shown.

approach. Firstly, of the 21 proteins identified in our three-step approach, 15 were also observed in the two validation datasets. Secondly, those 15 proteins covered an AUROC range of 0.52–0.85, with YWHAZ showing the best individual performance. The three proteins selected by our method ranked second, third, and fourth based on their individual performances, with AUROCs of 0.74–0.84 (Figure S4). Thirdly, of the 455 possible three-protein combinations, our set of three proteins ranked 50th (11th percentile) with an AUROC of 0.85 (Table S1). In comparison, the best three-protein combination featuring ALDOA, GAPDH, and YWHAZ resulted in an AUROC of 0.89. As such, it can be stated that our purely data-driven objective approach, which can be easily implemented in a data analysis pipeline, resulted in the selection of three proteins with excellent, albeit not necessarily maximum, performance; maximum performance can be achieved by a more brute force approach. The decision whether to use a two-step approach with brute force or a three-step approach with an additional down-selection step will have to be made on a case-by-case basis depending on the number of proteins selected in the discovery phase.

Limitations of the study

We would like to point out that when exploring MS literature and data repositories for appropriate datasets problems with correct data labeling, completeness of data files and demographic/clinical information became apparent (Table 2). The incomplete demographic variables made in-depth analysis of the effects of, e.g., gender or ethnicity impossible. This highlights the need to ensure that publicly available datasets are complete and carefully labeled and that all relevant demographic/clinical information are available. The mere availability of raw data does not make them useful. Instead, for a dataset to be useful, standards such as those outlined by the FAIR principles have to be met and ensured as part of the manuscript review process.[43,44] Inherent limitations of meta-analyses include lack of control over sample collection and processing as well as the possibility of missing biomarker candidates that can only be found with specific methodology. However, any set of biomarkers discovered and validated using heterogeneous datasets is likely free of systematic biases and is truly robust.

Conclusion

In summary, we provide a generic strategy for an unbiased meta-analysis of proteomics data exemplified by application to studies of CSF samples from patients with AD. These findings highlight the power of analyzing larger numbers of patients from various independent cohorts that are increasingly available in public data repositories. Our results strongly support the original hypothesis that combining new and existing independent cohorts into three unrelated meta-cohorts for discovery, down-selection, and validation of biomarker candidates, respectively, leads to a superior biomarker panel by leveraging the existing analyses of a wide range of independent datasets collected with heterogeneous methodologies. Our exceedingly well-validated biomarker panel comprising three glycolytic enzymes is consistent with the well-described dysregulation of the redox metabolic pathways in AD and the concept that metabolic dysregulation is the strongest overarching feature of AD. These data suggest that a better understanding the cause-and-effect relationship of glycolysis with AD might be a key not only for diagnosing AD in living patients but also to development of alternatives to current therapeutic approaches, which primarily target amyloid β and/or tau. A logical next step is to design a prospective validation study that includes symptomatic as well as healthy controls, i.e., patients with non-AD dementias, to determine the specificity of our candidate biomarker panel for AD.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Biomarker discovery
  - Down-selection of biomarker candidates
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Biomarker discovery
  - Down-selection
  - Validation of biomarker candidates
- ADDITIONAL RESOURCES

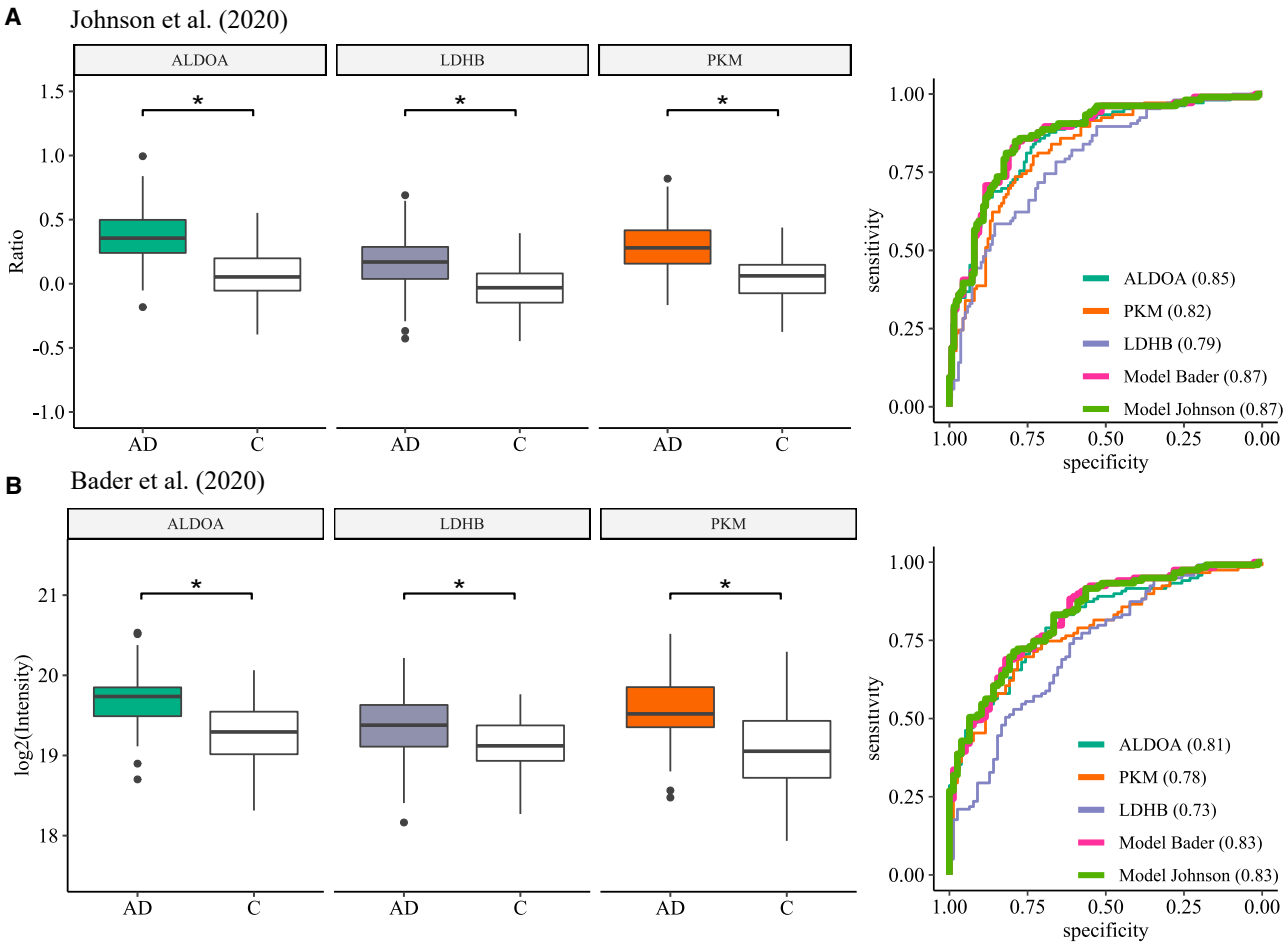


Figure 4. Biomarker validation cohorts

The postanalysis protein matrices of two large-scale AD CSF publications were downloaded to assess biomarker efficacy of the three biomarker candidates ALDOA, PKM, and LDHB. Each of the three proteins was found to be significant between AD and controls in both the Johnson et al. (A) and Bader et al. (B) datasets, further validating the efficacy of its differentiating capabilities. Next, a logistic regression model of the three biomarker candidates was trained for each of the validation datasets, followed by testing of both models on the two validation cohorts. AUROC of the single proteins as well as the two models is shown in the legend.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2023.101005.

ACKNOWLEDGMENTS

We are grateful to all researchers that made their LC/MS data and the relevant meta-data publicly available—either by uploading them to well-established data depositories or by responding to our request. Without their willingness to share their old data, this study would not have been possible. In particular, we would like to acknowledge Drs. Murat Eravci and Christoph Weise, who went out of their way to locate the raw LC/MS data and the associated meta information. We would like to acknowledge Dr. Nancy Chamberlin for her insightful editing of the manuscript. Finally, we would like to acknowledge Drs. Al Ozonoff and Manja Koch for discussions about the validity of envisioned statistical approaches and/or statistical interpretation of the findings. The authors acknowledge the following funding for the research described in the manuscript: R.S. received funding from Research Foundations Flanders. H.Z. is supported by grants from the Swedish Research Council (#2018-02532); the European Research Council (#681712); the Swedish State Support for Clinical Research (#ALFGBG-720931); the Alzheimer Drug Discovery Foundation (ADDF), USA (#201809–2016862); the AD Strategic Fund and the Alzheimer's Association (#ADSF-21-831376-C, #ADSF-21-831381-C, and #ADSF-21-831377-C); the Olav Thon Foundation; the Erling-Persson Family Foundation, Stiftelsen för Gamla Tjänarinnor, Hjärnfonden, Sweden (#FO2019-0228); the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 860197 (MIRIADE); and the UK Dementia Research Institute at UCL. A.B. received funding from the National Institutes of Health/National Institute of Aging: UCSF FTLD PPG (P01AG019724) and ADRC (P30AG062422). J.A.S. received grants from the National Institutes of Health/National Institute of Aging (R01AG071858-01), the Alzheimer's Drug Discovery Foundation, and the Tau Consortium. H.S. has received grants from the National Institute of Health/National Institute of Health & National Institute of Allergy and Infectious Diseases (U24AI152179-02S1). H.S. and J.A.S. received funding from The Diagnostics Accelerator at the Alzheimer's Drug Discovery Foundation.

AUTHOR CONTRIBUTIONS

H.S. and J.A.S. were the overall project leaders and devised the concept and the analytical strategies. Literature review, data collection, data analysis, and

Table 2. Information on the seven datasets that were included for the meta-analysis

	Unique proteins	Cutoff (70%)	Significant (p adjusted < 0.05)	Outliers AD	Outliers control	Available clinical information
Dayon et al.[12]	1,084	245	45	1	0	CSF AB42, tau, ptau, serum albumin
Sathe et al.[13]	2,533	2,522	0	1	0	AD vs. control
Khoonsari et al.[14]	522	409	0	1	0	AD vs. control, age, gender (all male), age of onset (only AD), AB42 (only AD), tau (only AD), ptau (only AD).
Lleó o et al.[15]	1,418	N/A	N/A	0	0	AD vs. control (pooled of 10 samples)
Barucket et al.[16]	914	461	3	1	0	AD vs. control
Wang et al.[17]	433	252	0	0	0	AD vs. control
In-house	1,148	863	0	0	1	AD vs. control

Information includes proteins identified before and after 70% cutoff filter, number of significant proteins following a Mann-Whitney U test and Benjamini-Hochberg correction, number of outliers between AD and control, and the available clinical information for each of the datasets. Cutoff filter information and statistical analysis in the Lleó et al. dataset was not possible due to the small sample size.

data interpretation were done by P.W.v.Z. with the help of B.F., M.K., and H.S. O.B. processed and analyzed the in-house dataset samples used for the biomarker discovery part. S.A. processed the validation cohort samples and contributed to the data analysis. H.Z. provided the validation sample set and contributed to the data interpretation and the manuscript. R.S. gave insightful comments and contributed to the data interpretation and the manuscript. After the initial drafting by P.W.v.Z., the manuscript was written by P.W.v.Z., H.S., and J.S. All authors contributed to the article and approved the submitted version.

DECLARATION OF INTERESTS

R.S. is member of the European Behavioral Pharmacology Society. H.Z. is chair of the Alzheimer's Association Global Biomarker Standardization Consortium and the Alzheimer's Association Biofluid-Based Biomarker PIA. J.A.S. reports patents for tau therapeutics and biomarkers. H.S., J.A.S., and P.W.v.Z. have submitted a patent application for markers described in this manuscript. H.S. and J.A.S. report additional patent applications (unrelated to the work described in the manuscript) around tau-PTM-based biomarkers for AD and other tauopathies.

Received: June 15, 2022
Revised: October 10, 2022
Accepted: March 17, 2023
Published: April 18, 2023

REFERENCES

1. Sancesario, G.M., and Bernardini, S. (2018). Diagnosis of neurodegenerative dementia: where do we stand, now? Ann. Transl. Med. *6*, 340. https://doi.org/10.21037/21001.
2. Karantzoulis, S., and Galvin, J.E. (2011). Distinguishing Alzheimer's disease from other major forms of dementia. Expert Rev. Neurother. *11*, 1579–1591. https://doi.org/10.1586/ern.11.155.
3. Blennow, K., Dubois, B., Fagan, A.M., Lewczuk, P., de Leon, M.J., and Hampel, H. (2015). Clinical utility of cerebrospinal fluid biomarkers in the diagnosis of early Alzheimer's disease. Alzheimers Dement. *11*, 58–69.
4. Teunissen, C., Verheul, C., and Willemse, E. (2018). The use of cerebrospinal fluid in biomarker studies. In Handbook of clinical neurology (Elsevier), pp. 3–20.
5. Haytural, H., Benfeitas, R., Schedin-Weiss, S., Bereczki, E., Rezeli, M., Unwin, R.D., Wang, X., Dammer, E.B., Johnson, E.C.B., Seyfried, N.T., et al. (2021). Insights into the changes in the proteome of Alzheimer disease elucidated by a meta-analysis. Sci. Data *8*, 312. https://doi.org/10.1038/s41597-021-01090-8.
6. Bai, B., Vanderwall, D., Li, Y., Wang, X., Poudel, S., Wang, H., Dey, K.K., Chen, P.-C., Yang, K., and Peng, J. (2021). Proteomic landscape of Alzheimer's Disease: novel insights into pathogenesis and biomarker discovery. Mol. Neurodegener. *16*, 55. https://doi.org/10.1186/s13024-021-00474-z.
7. Pedrero-Prieto, C.M., García-Carpintero, S., Frontiñán-Rubio, J., Llanos-González, E., Aguilera García, C., Alcaín, F.J., Lindberg, I., Durán-Prado, M., Peinado, J.R., and Rabanal-Ruiz, Y. (2020). A comprehensive systematic review of CSF proteins and peptides that define Alzheimer's disease. ClinClin. Proteomicsomics *17*, 21. https://doi.org/10.1186/s12014-020-09276-9.
8. Bader, J.M., Geyer, P.E., Müller, J.B., Strauss, M.T., Koch, M., Leypoldt, F., Koertvelyessy, P., Bittner, D., Schipke, C.G., Incesoy, E.I., et al. (2020). Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease. Mol. Syst. Biol. *16*, e9356. https://doi.org/10.15252/msb.20199356.
9. Johnson, E.C.B., Dammer, E.B., Duong, D.M., Ping, L., Zhou, M., Yin, L., Higginbotham, L.A., Guajardo, A., White, B., Troncoso, J.C., et al. (2020). Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. Nat. Med. https://doi.org/10.1038/s41591-020-0815-6.
10. Saito, E.R., Miller, J.B., Harari, O., Cruchaga, C., Mihindukulasuriya, K.A., Kauwe, J.S.K., and Bikman, B.T. Alzheimer's Disease Alters Oligodendrocytic Glycolytic and Ketolytic Gene Expression. Alzheimer's & Dementia *N/a*. https://doi.org/10.1002/alz.12310.
11. Tramutola, A., Lanzillotta, C., Perluigi, M., and Butterfield, D.A. (2017). Oxidative stress, protein modification and Alzheimer disease. Brain Res. Bull. *133*, 88–96. https://doi.org/10.1016/j.brainresbull.2016.06.005.
12. Dayon, L., Cominetti, O., Wojcik, J., Galindo, A.N., Oikonomidi, A., Henry, H., Migliavacca, E., Kussmann, M., Bowman, G.L., and Popp, J. (2019). Proteomes of paired human cerebrospinal fluid and plasma: relation to blood-brain barrier permeability in older adults. J. Proteome Res. *18*, 1162–1174.
13. Sathe, G., Na, C.H., Renuse, S., Madugundu, A.K., Albert, M., Moghekar, A., and Pandey, A. (2018). Quantitative Proteomic Profiling of Cerebrospinal Fluid to Identify Candidate Biomarkers for Alzheimer's Disease (PROTEOMICS–Clinical Applications), p. 1800105.
14. Khoonsari, P.E., Häggmark, A., Lönnberg, M., Mikus, M., Kilander, L., Lannfelt, L., Bergquist, J., Ingelsson, M., Nilsson, P., Kultima, K., and Shevchenko, G. (2016). Analysis of the cerebrospinal fluid proteome in Alzheimer's disease. PLoS One *11*, e0150672.
15. Lleó, A., Núñez-Llaves, R., Alcolea, D., Chiva, C., Balateu-Paños, D., Colom-Cadena, M., Gomez-Giro, G., Muñoz, L., Querol-Vilaseca, M.,

Pegueroles, J., et al. (2019). Changes in synaptic proteins precede neurodegeneration markers in preclinical Alzheimer's disease cerebrospinal fluid. Mol. Cell. Proteomics. *18*, 546–560.

16. Barucker, C., Sommer, A., Beckmann, G., Eravci, M., Harmeier, A., Schipke, C.G., Brockschnieder, D., Dyrks, T., Althoff, V., Fraser, P.E., et al. (2015). Alzheimer amyloid peptide Aβ 42 regulates gene expression of transcription and growth factors. J. Alzheimers Dis. *44*, 613–624.
17. Wang, J., Cunningham, R., Zetterberg, H., Asthana, S., Carlsson, C., Okonkwo, O., and Li, L. (2016). Label-free quantitative comparison of cerebrospinal fluid glycoproteins and endogenous peptides in subjects with Alzheimer's disease, mild cognitive impairment, and healthy individuals. Proteomics. Clin. Appl. *10*, 1225–1241.
18. Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D., and Nesvizhskii, A.I. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. Nat. Methods *14*, 513–520. https://doi.org/10.1038/nmeth.4256.
19. da Veiga Leprevost, F., Haynes, S.E., Avtonomov, D.M., Chang, H.-Y., Shanmugam, A.K., Mellacheruvu, D., Kong, A.T., and Nesvizhskii, A.I. (2020). Philosopher: a versatile toolkit for shotgun proteomics data analysis. Nat. Methods *17*, 869–870. https://doi.org/10.1038/s41592-020-0912-y.
20. Yu, F., Haynes, S.E., and Nesvizhskii, A.I. (2021). IonQuant enables accurate and sensitive label-free quantification with FDR-controlled match-between-runs. Mol. Cell. Proteomics. *20*, 100077. https://doi.org/10.1016/j.mcpro.2021.100077.
21. Ringman, J.M., Schulman, H., Becker, C., Jones, T., Bai, Y., Immermann, F., Cole, G., Sokolow, S., Gylys, K., Geschwind, D.H., et al. (2012). Proteomic changes in cerebrospinal fluid of presymptomatic and affected persons carrying familial Alzheimer disease mutations. Arch. Neurol. *69*, 96–104.
22. Heywood, W.E., Galimberti, D., Bliss, E., Sirka, E., Paterson, R.W., Magdalinou, N.K., Carecchio, M., Reid, E., Heslegrave, A., Fenoglio, C., et al. (2015). Identification of novel CSF biomarkers for neurodegeneration and their validation by a high-throughput multiplexed targeted proteomic assay. Mol. Neurodegener. *10*, 64.
23. Emami Khoonsari, P., Shevchenko, G., Herman, S., Musunuri, S., Remnestål, J., Brundin, R., Degerman Gunnarsson, M., Kilander, L., Zetterberg, H., and Nilsson, P. (2017). Chitinase-3-like Protein 1 (CH3L1) and Neurosecretory Protein VGF (VGF) as Two Novel CSF Biomarker Candidates for Improved Diagnostics in Alzheimer's Disease.
24. Laterza, O.F., Modur, V.R., Crimmins, D.L., Olander, J.V., Landt, Y., Lee, J.-M., and Ladenson, J.H. (2006). Identification of novel brain biomarkers. Clinical chemistry *52*, 1713–1721.
25. Selle, H., Lamerz, J., Buerger, K., Dessauer, A., Hager, K., Hampel, H., Karl, J., Kellmann, M., Lannfelt, L., and Louhija, J. (2005). Identification of novel biomarker candidates by differential peptidomics analysis of cerebrospinal fluid in Alzheimer's disease. Combinatorial Chemistry & High Throughput Screening *8*, 801–806.
26. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.
27. An, Y., Varma, V.R., Varma, S., Casanova, R., Dammer, E., Pletnikova, O., Chia, C.W., Egan, J.M., Ferrucci, L., Troncoso, J., et al. (2018). Evidence for brain glucose dysregulation in Alzheimer's disease. Alzheimers Dement *14*, 318–329. https://doi.org/10.1016/j.jalz.2017.09.011.
28. Trushina, E., Dutta, T., Persson, X.-M.T., Mielke, M.M., and Petersen, R.C. (2013). Identification of altered metabolic pathways in plasma and CSF in mild cognitive impairment and Alzheimer's disease using metabolomics. PLoS One *8*, e63644.
29. Reddy, P.H., Tripathi, R., Troung, Q., Tirumala, K., Reddy, T.P., Anekonda, V., Shirendeb, U.P., Calkins, M.J., Reddy, A.P., and Mao, P. (2012). Abnormal mitochondrial dynamics and synaptic degeneration as early events in Alzheimer's disease: implications to mitochondria-targeted antioxidant therapeutics. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease *1822*, 639–649.
30. Schmitt, K., Grimm, A., Kazmierczak, A., Strosznajder, J.B., Götz, J., and Eckert, A. (2012). Insights into mitochondrial dysfunction: aging, amyloid-β, and tau–a deleterious trio. Antioxid. Redox Signal. *16*, 1456–1466.
31. Selfridge, J.E., Lezi, E., Lu, J., and Swerdlow, R.H. (2013). Role of mitochondrial homeostasis and dynamics in Alzheimer's disease. Neurobiol. Dis. *51*, 3–12.
32. Yao, J., Rettberg, J.R., Klosinski, L.P., Cadenas, E., and Brinton, R.D. (2011). Shift in brain metabolism in late onset Alzheimer's disease: implications for biomarkers and therapeutic interventions. Mol. Aspects Med. *32*, 247–257.
33. Zhou, M., Haque, R.U., Dammer, E.B., Duong, D.M., Ping, L., Johnson, E.C.B., Lah, J.J., Levey, A.I., and Seyfried, N.T. (2020). Targeted mass spectrometry to quantify brain-derived cerebrospinal fluid biomarkers in Alzheimer's disease. ClinClin. Proteomicsomics *17*, 19. https://doi.org/10.1186/s12014-020-09285-8.
34. Moretti, D.V., Pievani, M., Pini, L., Guerra, U.P., Paghera, B., and Frisoni, G.B. (2017). Cerebral PET glucose hypometabolism in subjects with mild cognitive impairment and higher EEG high-alpha/low-alpha frequency power ratio. Neurobiol. Aging *58*, 213–224. https://doi.org/10.1016/j.neurobiolaging.2017.06.009.
35. Perluigi, M., Sultana, R., Cenini, G., Di Domenico, F., Memo, M., Pierce, W.M., Coccia, R., and Butterfield, D.A. (2009). Redox proteomics identification of 4-hydroxynonenal-modified brain proteins in Alzheimer's disease: role of lipid peroxidation in Alzheimer's disease pathogenesis. Proteomics. Clin. Appl. *3*, 682–693.
36. Zahid, S., Oellerich, M., Asif, A.R., and Ahmed, N. (2012). Phosphoproteome profiling of substantia nigra and cortex regions of Alzheimer's disease patients. J. Neurochem. *121*, 954–963. https://doi.org/10.1111/j.1471-4159.2012.07737.x.
37. Ding, B., Xi, Y., Gao, M., Li, Z., Xu, C., Fan, S., and He, W. (2014). Gene expression profiles of entorhinal cortex in Alzheimer's disease. Am J Alzheimers Dis Other Demen *29*, 526–532. https://doi.org/10.1177/1533317514523487.
38. Han, J., Hyun, J., Park, J., Jung, S., Oh, Y., Kim, Y., Ryu, S.-H., Kim, S.-H., Jeong, E.I., Jo, D.-G., et al. (2021). Aberrant role of pyruvate kinase M2 in the regulation of gamma-secretase and memory deficits in Alzheimer's disease. Cell Rep. *37*, 110102. https://doi.org/10.1016/j.celrep.2021.110102.
39. Hondius, D.C., van Nierop, P., Li, K.W., Hoozemans, J.J.M., van der Schors, R.C., van Haastert, E.S., van der Vies, S.M., Rozemuller, A.J.M., and Smit, A.B. (2016). Profiling the human hippocampal proteome at all pathologic stages of Alzheimer's disease. Alzheimers Dement. *12*, 654–668. https://doi.org/10.1016/j.jalz.2015.11.002.
40. Galea, E., Weinstock, L.D., Larramona-Arcas, R., Pybus, A.F., Giménez-Llort, L., Escartin, C., and Wood, L.B. (2022). Multi-transcriptomic analysis points to early organelle dysfunction in human astrocytes in Alzheimer's disease. Neurobiol. Dis. *166*, 105655. https://doi.org/10.1016/j.nbd.2022.105655.
41. Sjöstedt, E., Zhong, W., Fagerberg, L., Karlsson, M., Mitsios, N., Adori, C., Oksvold, P., Edfors, F., Limiszewska, A., Hikmet, F., et al. (2020). An atlas of the protein-coding genes in the human, pig, and mouse brain. Science *367*, eaay5947. https://doi.org/10.1126/science.aay5947.
42. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419. https://doi.org/10.1126/science.1260419.
43. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data

management and stewardship. Sci. Data *3*, 160018. https://doi.org/10.1038/sdata.2016.18.

44. Perez-Riverol, Y.; European Bioinformatics Community for Mass Spectrometry (2020). Toward a sample metadata standard in public proteomics repositories. J. Proteome Res. *19*, 3906–3909. https://doi.org/10.1021/acs.jproteome.0c00376.
45. Hansson, O., Zetterberg, H., Buchhave, P., Londos, E., Blennow, K., and Minthon, L. (2006). Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study. Lancet. Neurol. *5*, 228–234. https://doi.org/10.1016/S1474-4422(06)70355-6.
46. Vizcaíno, J.A., Csordas, A., Del-Toro, N., Dianes, J.A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., and Ternent, T. (2015). 2016 update of the PRIDE database and its related tools. Nucleic Acids Research *44*, D447–D456.
47. Bennike, T.B., Bellin, M.D., Xuan, Y., Stensballe, A., Møller, F.T., Beilman, G.J., Levy, O., Cruz-Monserrate, Z., Andersen, V., Steen, J., et al. (2018). A cost-effective high-throughput plasma and serum proteomics workflow enables mapping of the molecular impact of total pancreatectomy with islet autotransplantation. J. Proteome Res. *17*, 1983–1992.
48. Berger, S.T., Ahmed, S., Muntel, J., Cuevas Polo, N., Bachur, R., Kentsis, A., Steen, J., and Steen, H. (2015). MStern blotting–high throughput polyvinylidene fluoride (PVDF) membrane-based proteomic sample preparation for 96-well plates. Mol. Cell. Proteomics. *14*, 2814–2823.
49. Spellman, D.S., Wildsmith, K.R., Honigberg, L.A., Tuefferd, M., Baker, D., Raghavan, N., Nairn, A.C., Croteau, P., Schirm, M., Allard, R., et al. (2015). Development and evaluation of a multiplexed mass spectrometry based assay for measuring candidate peptide biomarkers in Alzheimer's Disease Neuroimaging Initiative (ADNI) CSF. Proteomics. Clin. Appl. *9*, 715–731.
50. RStudio-Team (2015). RStudio (Integrated Development for R).
51. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B *57*, 289–300.
52. Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pagès, F., Trajanoski, Z., and Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics *25*, 1091–1093. https://doi.org/10.1093/bioinformatics/btp101.
53. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for comparing biological themes among gene clusters. OMICS A J. Integr. Biol. *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.
54. Molecular Signatures Database (MSigDB) 3.0 | Bioinformatics | Oxford Academic https://academic.oup.com/bioinformatics/article/27/12/1739/257711?login=true.
55. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 77.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
PVDF 96 well membrane plate		
Biological samples		
cerebrospinal fluid	Zetterberg Lab (University of Gothenburg, Mölndal, Sweden)	
Chemicals, peptides, and recombinant proteins		
Urea	Sigma-aldrich	#SLCG4742
Ammonium Bicarbonate	Sigma-aldrich	#BCBS2121V
Formic acid	Sigma-aldrich	#STBJ9437
Dithiothreitol	Sigma-aldrich	#SLCD8671
Iodoacetamide	Sigma-aldrich	SLCC6164
LC-MS Water	Fisher Chemical	216530
Acetonitrile	Fisher Chemical	201791
Trypsin	Promega	V5111
Critical commercial assays		
INNOTEST hTAU	Fujirebio	81579
INNOTEST PHOSPHO-TAU	Fujirebio	81581
INNOTEST BETA-AMYLOID(1–42)	Fujirebio	81583
Deposited data		
LC-MS raw Proteomics Data	ProteomeXchange PRIDE	PXD022649
Data matrices and code	https://github.com/SteenOmicsLab/CSF-AD-meta-analysis	
Software and algorithms		
Fragpipe 17.1	https://fragpipe.nesvilab.org/	
R & Rstudio	https://www.rstudio.com/	
Cytoscape	https://cytoscape.org/	
ClueGO	https://apps.cytoscape.org/apps/cluego	
UniProt	https://www.uniprot.org/	

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Dr. Hanno Steen (Hanno.steen@childrens.harvard.edu).

Materials availability

No new reagents were generated in this study.

Data and code availability

The mass spectrometry proteomics data of the validation cohort has been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD022649. The 10 datasets and the R-code for analysis and visualisation including an easy-accessible R-markdown file can be found on the following Github page: https://github.com/SteenOmicsLab/CSF-AD-meta-analysis. All matrices with protein quantification for the discovery, down-selection and the two large scale proteomic studies (and their clinical files if required) can be found in the Github repository. Any additional information required to reanalyze the data reported in this work paper is available from the Lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

For the subsequent validation, we used 20 CSF samples provided by the Zetterberg lab (University of Gothenburg, Mölndal, Sweden). The samples were from patients who sought medical advice because of cognitive impairment. Patients were designated as normal (n = 10) or AD (n = 10) according to CSF biomarker levels, measured using INNOTEST assays (Fujirebio, Ghent, Belgium), using cut-offs that are >90% specific for AD: total tau (T-tau) >350 ng/L and Aβ42 < 530 ng/L.[45] None of the biochemically normal subjects fulfilled these criteria. To collect the CSF, lumbar punctures were performed in the morning. CSF was stored in polypropylene tubes and centrifuged to pellet any cell debris. After centrifugation, all CSF samples were frozen and stored at −80°C without thawing until the experiment. The Regional Ethics Committee at the University of Gothenburg approved the study.

METHOD DETAILS

Biomarker discovery

Literature review

To retrieve AD-related CSF proteomics datasets, PubMed was searched with multiple combinations of the following keywords: 'Alzheimer's disease', 'biomarker discovery', 'cerebrospinal fluid', 'proteomics', 'dementia', 'mass spectrometry', 'discovery proteomics' and 'neurodegeneration'. The resulting PubMed search results were downloaded and reviewed to determine if a paper described a study of CSF proteomes from AD patients using data-dependent acquisition (DDA) mode.

Exclusion criteria were: published before 2010; reviews (systematic or literature); written in other languages than English; no CSF proteomics; non-human CSF; no AD-related samples; only *post mortem* CSF; CSF not collected by a lumbar puncture; no information about the origin of samples; no description of sample preparation and/or MS techniques; use of non-DDA methods (e.g. SRM, MRM, PRM, Western Blot, 2D gel electrophoresis); no information about AD diagnosis criteria or CSF collection; peptidomics for biomarker discovery, isobaric labeling at the protein level or use of low resolution/low accuracy MS instrumentation.

Inclusion criteria were: published between Jan 1st, 2010 and Jan 31st, 2019; proteomic analyses of CSF from AD patients and in controls; CSF collection *ante mortem* by lumbar puncture; proteomic profiling using LC-MS/MS operated in DDA mode; use of high resolution/high accuracy instrumentation; well defined and described AD diagnosis and a clear definition of controls.

The selected papers were searched to determine if the paper describes the availability of the raw MS data in repositories such as PRIDE or MassIVE (massive.ucsd.edu).[46] If data were not available on repositories, authors were contacted directly requesting their raw LC-MS data. In some cases, LC-MS data were available, but the relevant meta-data was not. Some of these complications could be resolved by directly contacting the corresponding authors. If several methods were described in a single publication only the dataset with the largest number of identified proteins was used for further analysis.

One (unpublished) additional in-house dataset was used in this study. Quantitative proteomic mapping of these samples was performed using tandem mass tags (TMT) and analyzed on a Q Exactive Hybrid Quadrupole-Orbitrap mass spectrometer (ThermoFisher, Waltham, MA, USA). Similar to the publications selected in the literature review the CSF of the in-house dataset was collected ante mortem.

Down-selection of biomarker candidates

Sample processing, digestion and clean-up

CSF samples were prepared for proteomic analysis using an in-house-developed MStern Blotting protocol which was adapted for CSF samples.[47,48] Briefly, 100 μL of CSF samples were processed using a PVDF 96-well membrane plate (Merck-Millipore, MA, USA). Initially, the 100 μL of CSF was mixed with 100 μL Urea buffer (8M in 50mM ammonium bicarbonate (ABC). To further reduce the disulphide bonds on the proteins 30 μL Dithiothreitol (DTT) (0.05 M in water) was added and incubated in a thermomixer (300rpm) for 20 min at room temperature. To prevent the re-formation of disulphide bonds, 30 μL Iodoacetamide (IAA) (0.25 M in water) was added and incubated in a thermomixer (300rpm) for 20 min at room temperature in the dark.

Reduced and alkylated CSF protein suspension was transferred to a 96 well polyvinylidene fluoride (PVDF) membrane (MSIPS4510, Millipore, MA, USA), which had been activated with 150 μL 70% ethanol and subsequently primed with 200 μL of urea buffer. To facilitate the transfer of the solution through the PVDF membrane a vacuum manifold was used. CSF proteins are captured on the PVDF membrane and were washed with 200 μL 50 mM ABC before applying 100 μL digestion buffer (0.4 μg Trypsin (V5111, Promega, WI, USA) in 50 mM ABC) to the 96-wells plate. The 96-wells plate was wrapped in parafilm and put in a 37°C dark humidified incubator for 2 h to facilitate digestion of the proteins. After incubation, the remaining digestion buffer was evacuated from the 96-wells PVDF membrane plate using a vacuum manifold. Proteins, now peptides, were eluted twice with 150 μL of 40% acetonitrile (ACN), 0.1% formic acid (FA). The flow-through was pooled in a 96-wells plate which was centrifuged to dryness in a vacuum centrifuge.

For sample desalting, peptides were resuspended in 100 μL of 0.1% FA and transferred to a 96 wells MACROSPIN C18 plate (Targa, Nest Group, MA, USA) which had previously been activated with 100 μL of 70% ACN, 0.1% FA followed conditioning with 100 μL 0.1% FA. To transfer the solutions through the MACROSPIN C18 plate, the plates were centrifuged at 2000g for 2 min. After capturing the peptides on the C18 beads the plate was washed with 100 μL of 0.1% FA followed by eluting the peptides with 100 μL 40% ACN, 0.1% FA and 100 μL 70% ACN, 0.1% FA. The captured eluents were dried down in a vacuum centrifuge and stored at −20°C until analysis.

LC-MS/MS analysis

To validate the biomarker candidates the prepared CSF samples were analyzed on an Orbitrap Q Exactive mass spectrometer (Thermo Scientific, Bremen, Germany). First, the tryptic digests were resuspended in 20 μL resuspension buffer (5% ACN, 5% FA) and placed into a nanoflow HPLC pump module LC autosampler (Eksigent/Sciex, Framingham, MA, USA) where 4 μL of the sample was loaded onto a PicoChip column (150 μm × 10 cm Acquity BEH C18 1.7 μm 130 Å, New Objective, Woburn, MA) which was kept at 50°C. The peptides were eluted off the PicoChip column using 2% of solvent B (0.1% FA in ACN) in solvent A (0.1% FA), which was increased from 2 to 30% in a 40 min ramp gradient and back to 35% on a 5 min ramp gradient with a flow rate of 1000 nL/min. The Orbitrap settings were the following: positive DDA top 12 mode. MS1 scan settings: *m/z* range: 375–1400, resolution 70000 @ *m/z* 200, AGC target 3e6, max IT 60 ms. MS scan settings: resolution 17500 @ *m/z* 200, AGC target 1e5, max IT 100 ms, isolation window *m/z* 1.6, NCE 27, underfill ratio 1% (intensity threshold 1e4), charge state exclusion unassigned, 1, >6, peptide match preferred, exclude isotopes on, dynamic exclusion 40 s.

QUANTIFICATION AND STATISTICAL ANALYSIS

Biomarker discovery

Proteomic analysis

All downloaded raw LC/MS data were analyzed in Fragpipe 17.1 using the human UniprotKB/Swiss-prot protein sequence (without isoforms) database which was downloaded on January 17th, 2019 with a total of 20623 entries.[18–20] A maximum of two missed tryptic cleavages were allowed and we set the peptide length between 7 and 50 amino acids long. We included cysteine residues (fixed), acetylation of the N-terminal of proteins (variable), and oxidation of methionine (variable) as modifications with a maximum of three modifications on a peptide. For the TMT studies a (+229.163 Da) modification at the N terminus of the peptide as well as at lysine was set as fixed modifications. 1% FDR was allowed for percolator and proteinProphet. IonQuant included the match-between-runs settings where we required at least one ion per peptide for quantification.

Data analysis

Protein identification and quantification outputs were exported from Fragpipe which was loaded into R-studio where all the described analysis was executed unless described differently. Each step of the data preparation, statistical analysis and data visualization is available as an R-markdown file on the Github Repository.

Because some of the datasets included a reference node, two normalization methods were scripted in R. If there was no reference node in the data, the summed intensity for each sample was calculated followed by the calculation of the median of all summed intensities for each study individually. Next, for each sample in the dataset the normalization factor (NF) was calculated by dividing the median of all summed intensities by the summed intensity of a given sample. Subsequently, this factor was used to normalize the protein intensities of the corresponding sample. If a dataset did include one or more reference sample the median intensity of the reference sample was calculated and used to determine the NF. In the case of multiple reference nodes, the average of all median intensities was calculated and used to determine the NF. The NF was calculated by dividing the median intensity of the reference node (or average if multiple reference nodes) by the median intensity of a given sample. The NF was subsequently used to normalize the protein intensities of the corresponding sample.

In shotgun MS proteomic studies there is no standardized manner to select and remove samples that should be considered outliers. Usually, principal component analysis (PCA) plots of the samples are created, and samples are considered outlier based on their location relative to the other samples. This subjective approach can then lead to the removal of the outlier from datasets but lack reproducibility. For the re-analysis of datasets, we avoided this approach and developed a standardized approach for outlier identification instead. The normalized datasets were loaded into R where a theoretical sample was created based on the median intensities calculated for each protein identified in the respective dataset. Next, the Pearson correlation between each sample and this theoretical median sample was calculated in each dataset. Next, the standard deviations for the correlation coefficients were calculated and any sample with a correlation coefficient more than three standard deviations away from one was considered an outlier and removed from the dataset. A similar approach using three standard deviations as a criterion for being an outlier can be found in the paper by Spellman et al.[49]

The collected datasets from different labs used different methodologies and hardware set-ups for their proteomic profiling of the CSF samples. Therefore, the intensities of the different datasets were not directly comparable; instead, the Z-scores were calculated to allow for comparisons. For the *Z* score transformation, all data were first loaded into R where all zero intensities produced by Fragpipe were treated as missing values, i.e. these values were removed from the intensity matrix. Subsequently, all remaining intensity values were log2-and then *Z* score transformed: For a given protein in each dataset, the mean and standard deviation were calculated based on the intensities of the control samples in a dataset. These values were then used to calculate the Z-scores of the corresponding protein for all samples (*Z* score = (intensity – mean)/standard deviation). This process was then repeated for all protein across all datasets. Finally, all datasets were combined into one large dataset. The effectiveness of this procedure was confirmed by comparing the PCA plots before and after *Z* score transformation.

Due to high variability and/or incompleteness of meta data such as age, gender, or Braak stages we were only able to test for differences in proteins between AD and controls without control for potential confounders. All statistical analysis was executed in R-studio.[50] First, a Fisher exact test was used to identify proteins that might show statistical significance due to the percent

presence/absence in the AD vs. control group. Next, proteins with more than 70% missing values were removed from the dataset. For statistical analysis, the non-parametric Mann-Whitney U test was used. The non-parametric test was chosen as with such variable data assumptions for the parametric equivalent could not be assured. The resulting p values were corrected for multiple comparison using the Benjamini-Hochberg procedure.[51] Last, since it was hypothesized that the meta-analysis would results in more biomarker candidates compared to analyzing each dataset separately, we analyzed each of the datasets used in this meta-analysis on its own with a Student's T-test and Benjamini-Hochberg correction. The results were used to create a heatmap where the p values of the found biomarker candidates of the meta-analysis were compared with the p values of the biomarker candidates when datasets were analyzed on their own.

The identified biomarker candidates were analyzed to determine their functional enrichment. We analyzed the biomarker candidates with the Cytoscape plug-in ClueGO.[52] ClueGO can compare and integrate clusters/groups of GO annotations based on kappa statistics to connect GO terms to one another. We also tested against the MSigDB Gene Sets Hallmark datasets using the msigdbr-package followed by visualization using the Clusterprofiler R package.[53,54]

Down-selection

Statistical analysis

Given our goal of mining and (re-)analyzing existing data using a standard and systematic data processing pipeline, all methods described above were also applied to this set of CSF samples, apart from Z-scoring the data. First, a two-sided Fisher's exact test with a Benjamini-Hochberg correction was used to determine if a protein was only identified in the AD or control group. Next, proteins with valid values less than 70% were removed from the dataset followed by analysis with a Mann-Whitney U-test and a Benjamini-Hochberg Correction. Non-parametric testing was chosen to mirror the analysis in the discovery step. Proteins found to be significant in the discovery and down-selection cohort were selected for final validation.

Validation of biomarker candidates

The Johnson et al. (2020) and Bader et al. (2020) datasets were used for final validation.[8,9] We extracted the tabular quantification data from their respective data repositories where we extracted clinical classification of samples and the quantitative values of the down-selected markers. These markers were then combined into a biomarker panel using logistic regression modeling. A model for each of the two datasets was created independently from one another. Finally, the single markers and both the models (i.e., model-Bader and model-Johnson) were tested on both datasets to assure model results were not due to overfitting. Testing was facilitated with the pROC R-package which visualized the results and extracted the area under the curve statistics of the receiver operating characteristic curve analysis.[55]

Next, we assessed the biomarker efficacy of all the biomarker candidates from the discovery cohort on the down-selection dataset and the two validation datasets through calculation of the area under the receiver operating characteristic curve (AUROC) for each protein. Last, we tested each possible 3-protein combination of the discovery cohort biomarker candidates in the two validation datasets through a brute-force method.

ADDITIONAL RESOURCES

Github repository which includes all datasets and code: https://github.com/SteenOmics/AD_CSF_Meta-Analysis.