Why Inclusion Matters for Alzheimer's Disease Biomarker Discovery in Plasma

Mostafa J. Khan[a], **Heather Desaire**[b], **Oscar L. Lopez**[c,d], **M. Ilyas Kamboh**[d,e,f], **Renã A.S. Robinson**[a,g,h,i,j,*]

[a]Department of Chemistry, Vanderbilt University, Nashville, TN, USA

[b]Department of Chemistry, University of Kansas, Lawrence, KS, USA

[c]Department of Neurology, University of Pittsburgh, Pittsburgh, PA, USA

[d]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA

[e]Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA, USA

[f]Department of Epidemiology, University of Pittsburgh, Pittsburgh, PA, USA

[g]Vanderbilt Memory and Alzheimer's Center, Vanderbilt University Medical Center, Nashville, TN, USA

[h]Vanderbilt Institute of Chemical Biology, Vanderbilt University, Nashville, TN, USA

[i]Vanderbilt Brain Institute, Vanderbilt University Medical Center, Nashville, TN, USA

[j]Department of Neurology, Vanderbilt University Medical Center, Nashville, TN, USA

Abstract

Background: African American/Black adults have a disproportionate incidence of Alzheimer's disease (AD) and are underrepresented in biomarker discovery efforts.

Objective: This study aimed to identify potential diagnostic biomarkers for AD using a combination of proteomics and machine learning approaches in a cohort that included African American/Black adults.

Methods: We conducted a discovery-based plasma proteomics study on plasma samples ($N = 113$) obtained from clinically diagnosed AD and cognitively normal adults that were self-reported African American/Black or non-Hispanic White. Sets of differentially-expressed proteins were then classified using a support vector machine (SVM) to identify biomarker candidates.

Results: In total, 740 proteins were identified of which, 25 differentially-expressed proteins in AD came from comparisons within a single racial and ethnic background group. Six proteins were differentially-expressed in AD regardless of racial and ethnic background. Supervised classification by SVM yielded an area under the curve (AUC) of 0.91 and accuracy of 86% for

*Correspondence to: Prof. Renã A. S. Robinson, Department of Chemistry, Vanderbilt University, 5423 Stevenson Center, Nashville, TN 37235, USA. Tel.:+1 615 343 0129; rena.as.robinson@vanderbilt.edu.

differentiating AD in samples from non-Hispanic White adults when trained with differentially-expressed proteins unique to that group. However, the same model yielded an AUC of 0.49 and accuracy of 47% for differentiating AD in samples from African American/Black adults. Other covariates such as age, *APOE4* status, sex, and years of education were found to improve the model mostly in the samples from non-Hispanic White adults for classifying AD.

Conclusion: These results demonstrate the importance of study designs in AD biomarker discovery, which must include diverse racial and ethnic groups such as African American/Black adults to develop effective biomarkers.

Keywords

African American; Alzheimer's disease; biomarker; Black; discovery; disparities; machine learning; plasma; proteomics; race

INTRODUCTION

One of the fastest growing populations in the United States are ethnic minorities, such that Hispanics and African American/Blacks are estimated to account for 40% of older adults in 2050 [1–3]. This is critical for the field of Alzheimer's disease (AD) which disproportionately impacts these populations and accounts for higher incidence rates of 1.5 to 2x for African American/Black adults [4–6]. In addition to disparities in disease incidence, there is a disparate economic burden of AD costs and caregiving for African American/Black families [7, 8]. Racial and ethnic disparities in AD are multi-factorial and can include contributions from socioeconomic status [1, 5, 9, 10], quality and level of education [11–14], comorbidities [1, 15], genetic risk factors [15], environmental stressors [16, 17], healthcare access [1, 5], and systemic racism [18]. Disease pathology is similar in African American/Black and Hispanic adults with regards to what we have learned for decades about AD in non-Hispanic White populations [5, 19–23]. Although, understanding disease pathogenesis is complicated by higher frequency of mixed dementia cases and vascular comorbidities that are often found for example in African American/Black adults with AD [19, 21, 22].

Importantly, for better diagnosing AD in all communities and for developing effective therapies, better strategies are needed to increase research participation of African American/Black adults into AD and related dementia research [24]. Currently, African American/Black adults participate in clinical trials at a rate of <~5% and in AD research assessments at <~13% [25–27]. Recently, there have been reports that demonstrate that biomarker discovery efforts for AD need to be inclusive of African American/Black adults as the standard cerebrospinal fluid (CSF) biomarkers, total-tau and phosphorylated (p)-tau, have different concentration levels in a cohort of African American/Black adults when compared to non-Hispanic White adults [28–33]. Other circulatory proteins in plasma also have been reported to have differences in African American/Black AD patients such as interleukin-9 [34]. Examples of plasma biomarker studies that include or solely focus on African American/Black adults are few [35–37]. This is critical as plasma is a less-invasive biofluid to obtain than CSF, and more importantly, as diagnostic or prognostic biomarkers for AD need to be effective for all.

Plasma proteomics is a growing field within AD biomarker discovery [38–45]. Many studies have conducted plasma proteomics analyses across the AD spectrum from cognitively normal (CN) to mild cognitively impaired to confirmed early and late-onset AD [46–64]. The most widely used biomarkers in plasma are amyloid-beta 40 ($A\beta_{40}$) and 42 ($A\beta_{42}$) peptides [38, 45, 65–67]. Other potential plasma protein biomarkers, a few of which have been validated within the same cohort [68] and few in independent cohorts [46, 48, 59, 69, 70], also exist. A recent review of plasma biomarkers for AD [40] analyzed findings from 22 previous discovery-based proteomics studies and found a lack of reproducibility across those studies. For example, the review reported alpha-2 macroglobulin as a biomarker candidate in six different studies, while pancreatic polypeptide was reported in five studies [40]. The remaining proteins reported in more than two studies included apolipoprotein A1, afamin, fibronectin, apolipoprotein A4, alpha –1- antitrypsin, fibrinogen-γ chain, insulin like growth factor binding protein-2, macrophage inflammatory protein 1-α, beta-2 glycoprotein, and complement factor B [40].

Few plasma biomarker studies utilized diverse cohorts that incorporated African American/ Black adults into the study design [35, 37, 71]. Plasma biomarkers for an amnestic MCI cohort of females from African American/Black, Hispanic, and non-Hispanic White backgrounds were identified as having levels unique to a given racial and ethnic group and highlight the importance of including diverse groups in biomarker studies [37]. Age was deemed a critical factor in mid-life plasma $A\beta$ concentrations in a large cohort of European and African American/Black adults, whereby there were also different genes that had race-specific changes [35]. For example, cystathionine beta-synthase gene had genome-wide significant association with plasma homocysteine levels in African American/ Blacks and Yoruba cohorts, that are associated with African ancestry [71]. Inclusion of African American/Black adults in study designs is critical to ensure that specific biomarkers, combinations of biomarkers in a panel, or biomarker levels are able to accurately distinguish and diagnose disease in all groups.

Herein, we conducted a pilot study with plasma samples available from the University of Pittsburgh Alzheimer's Disease Research Center (ADRC) that included self-reported African American/Black and non-Hispanic White adults that participated in ADRC research. Individual participants were either cognitively normal or had a clinically confirmed diagnosis of AD, and plasma was collected from a time point away from baseline that was coincident with clear disease pathology and clinical diagnosis. The samples ($N = 113$) were randomly divided into two sets in order to accommodate a study design that would allow us to 1) conduct discovery-based proteomics to identify differentially-expressed proteins in a cohort that included African American/Black adults; 2) use the data generated from the discovery-based proteomics studies to determine the utility of the selected panel of proteins as diagnostic AD biomarkers using machine learning; and 3) determine the extent to which the model performance depended on the racial and ethnic background of the training set samples. Comprehensive plasma proteomics biomarker discovery was conducted using quantitative tandem mass tags of peptides from plasma immunodepletion, liquid chromatography (LC), and mass spectrometry (MS) workflows. Processed proteomics data along with machine learning was used to test whether the racial and ethnic background of the training set samples impacts the accuracy of the biomarker panels for AD. In this

approach, we divided the cohort into two sets: Set 1 ($N = 73$) was used to select the optimal protein features and Set 2 ($N = 40$) was used to validate the results.

METHODS

Plasma sample collection

Human plasma samples ($N = 113$) from African American/Black and non-Hispanic White individuals were obtained from the University of Pittsburgh Alzheimer's Disease Research Center (ADRC). Detailed characteristics of the individuals are provided in Table 1. Approval for the participation of human subjects were obtained by the Institutional Review Boards of the University of Pittsburgh and Vanderbilt University. The Mini-Mental State Examination (MMSE) was performed, and disease individuals were clinically diagnosed with mild to moderate dementia at the time of blood draw according to the National Institute on Aging-Alzheimer's Association and National Alzheimer's Coordinating Center criteria. The samples were divided into two separate sample sets of $N = 73$ and $N = 40$, using a blinded study design. Plasma samples were obtained from participants at the most recent blood draw at the start of this study in December 2016, in order to analyze plasma that was close in chronological year and storage conditions and that corresponded to clear disease diagnosis.

Plasma depletion

Plasma samples were depleted of the top six most abundant proteins (albumin, IgG, IgA, α1-antitrypsin, transferrin, and haptoglobin) using the Multiple Affinity Removal System (MARS) Column Human 6 (Agilent, Santa Clara) according to manufacturer's instructions. In brief, 30 μL of crude plasma sample was diluted 4 times using buffer A (Agilent, Santa Clara) and centrifuged at 16000 g for 1 min through a 0.22 μm spin filter to remove particulates. The sample was injected onto the MARS 6 column using a Waters Alliance 2695 Separation module LC system and the resulting fractions were collected. The flow through fractions were concentrated using a 5kDa molecular weight cutoff concentrator at 4,695 g for 1.5 h followed by a bicinchoninic acid (BCA) assay to determine protein concentration. A pooled sample containing equal amounts of protein from each of the plasma samples was generated and used as quality control (QC) sample

Digestion

Samples were randomized into eight and four batches respectively, for Set 1 and Set 2 with corresponding QC sample in each batch. In Set 1, in solution digestion was performed in 100 mM ammonium bicarbonate buffer. Proteins (100 μg) were reduced using 200 mM dithiothreitol (DTT) for 45min at 55°C, while alkylation was performed using 200 mM iodoacetamide (IAM) in the dark for 30 min. Finally, proteins were digested using trypsin/Lys-C mix (Promega, Madison) overnight at 37°C (1:50 enzyme:protein ratio). The digested samples were acidified with formic acid and desalted using HLB cartridges (Waters Corporation, Milford) per manufacturer's instructions. For Set 2, the digestion was carried out using the filter assisted sample preparation (FASP) protocol [72]. In brief, proteins (100 μg) were transferred onto a 10 kDa molecular weight cutoff filter (Sartorius, Gloucestershire, UK) and reduced for 15 min with 20 mM DTT in 100 mM Tris with 8 M urea. This was followed by centrifugation at 14000 rpm for 30 min and the resulting

filtrate was discarded. Proteins were then alkylated with 20 mM IAM in the dark for 15 min followed by centrifugation to remove the excess reagents. Samples were washed using 100 mM Tris with 1 M urea in a centrifuge at 14000 rpm and trypsin/Lys-C mix (Promega, Madison) was added and digested for 8 h at 37°C (1:50 enzyme:protein ratio). After digestion, the peptides were acidified with formic acid and desalted using HLB cartridges. TMT 10-plex or 11-plex labeling (ThermoFisher Scientific, Waltham) was performed following manufacturer's instruction with 25 μg of peptides. Labeled peptides were desalted and separated into 12 fraction using high pH (pH = 10) reversed-phase fractionation with acetonitrile (ACN(%)– 3, 5, 8, 10, 13, 18, 22, 30, 45, 60, 80, 95) on an HLB cartridge. Fractions were dried down and reconstituted in water with 0.1% formic acid.

LC-MS/MS and MS^3 parameters

Peptides were analyzed using an Orbitrap Fusion Lumos (ThermoFisher Scientific, Waltham) in positive ionization mode. The samples were loaded onto a self-packed C18 (5 μm, 200Å, MICHROM Biore-sources Inc.) trap column (100 μM ID × 2.5 cm, IntegraFrit Capillary), and separation was performed on an in-house packed C18 (2.5 μm, 100Å, XBridge BEH from Waters) capillary column (100 μM ID × 25 cm, Polymicro Technologies) at 300 nL/min using solvent A (water with 0.1 % formic acid) and solvent B (acetonitrile with 0.1% formic acid). The gradient was as follows: 0–7min, 10% B; 7–67min, 10–30% B; 67–75min, 30–60% B; 75–77 min, 60–90% B; 77–82min, 90% B; 82–83min, 90–10% B; and 83–100min, 10% B. Full MS scans were acquired over a mass range of m/z 375–1500 at a resolution of 120,000 with the automatic gain control (AGC) target set at 4×10^5 ions and maximum ion injection (IT) time of 50 ms. Data dependent acquisition (DDA) was used to acquire MS/MS spectra with a cycle time of 3 s. MS/MS fragmentation was performed using collision-induced dissociation (CID) with an NCE = 35%. The AGC was set at 1×10^4 using an isolation width of 0.7 m/z, maximum injection time of 100 ms, and a dynamic exclusion of 20 s. Synchronous precursor selection (SPS) mode was used for collecting MS^3 spectra of the top 10 most intense ions from the MS/MS fragments. Higher-energy collisional dissociation (HCD) was used for MS^3 with the following Orbitrap parameters: NCE = 55%, scan range = 100–400 m/z, resolution = 60,000, AGC = 5×10^4, maximum injection time = 18 ms and isolation width = 2 m/z. Each fraction was injected in duplicate and the injection order was randomized for each batch.

Data analysis

Raw files were analyzed using Proteome Discoverer software (version 2.2) and searched against the Uniprot human reviewed protein database (07/17/2018, 20289 sequences) using SEQUEST-HT. The following parameters were used: maximum two trypsin miscleavages, precursor mass tolerance 10 ppm, fragment mass tolerance 0.6 Da; dynamic modification of methionine oxidation (+15.995 Da), protein N-termini acetyl (42.011 Da), TMT 10 (229.163 Da)/11 plex (229.169 Da) on peptide N-termini and lysine residue, static modification of cysteine carbamidomethyl (+57.02 Da). Decoy database searching was employed to generate high confidence peptides (FDR < 1%). TMT reporter ions (i.e., m/z 126 – 131) were identified with the following parameters: most confident centroid and 20 ppm reporter ion mass tolerance. Technical replicates and fractions from each batch were combined into one

result file. Finally, the individual batch data were combined together for further processing using an in-house Python script.

Protein filtering criteria included peptide spectral matches (PSMs) 2, and reporter ion intensity values above the minimum threshold in at least 75% of the TMT channels (i.e., present in any 54 of 73 samples or 30 of 40 samples). The final list of proteins was normalized using a two-step internal reference scaling (IRS) method which has been described previously [73]. Briefly, in-batch normalization was performed by calculating a scaling factor (SF), which is the ratio of the sum of intensity of the pooled channel to the sum of each individual TMT channel, followed by multiplying the intensities by the SF for each individual batch. Next, across-batch normalization was applied by the use of SF from the geometric mean of the TMT intensity of pooled samples. Once normalized, differentially-expressed proteins ($p < 0.05$) were determined by student's t-test's between AD and CN sample groups within each sample set. We did not use Bonferroni or other multiple hypothesis testing to keep a less stringent initial set of data to feed into the machine learning algorithm [74, 75]. A fold change cut-off of 1.23 (Set 1) and 1.33 (Set 2) was established based on biological and technical variation in the data [76]. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org/) via the PRIDE [77] partner repository with the dataset identifier PXD022265.

Machine learning

Data preparation—Supervised classification was performed using RStudio, R version 3.5.1. The data sets with at least 75% of the TMT channels filled were selected for further analysis. Models were built with two different approaches: 1) using only the protein data, and 2) using the protein data along with the variables of sex, age, years of education, and *APOE* status. In the models with combined data types, sex and *APOE* genotype status were converted to numeric variables. For *APOE* status, any patient with a 4/4 genotype was coded as a 4. Genotypes of 3/4 and 3/3 were coded as 3 and 2, respectively. All remaining patients, who had at least one *APOE2* allele (2/3, 2/4 genotypes), were coded as 1. Prior to classification, differentially-expressed proteins were selected as the protein feature set. These were either combined or not combined with the clinical variables and then the matrix was scaled using the embedded scale function in R. After scaling, any missing values were replaced with the average value for the given feature in the dataset under consideration. Since the data had been scaled, the average value was zero in each case.

Classification—All supervised classification was performed with a support vector machine (SVM) using the R package, e1071 [78]. All area under the receiver-operator curve (AUC) values reported herein were calculated using the R package, pROC [79]. In order to ensure comparability from one experiment and one data set to the next, all of the SVM classifications were performed using identical parameters (i.e., leave-one-out cross-validation was performed in every case, and hyperparameters were not optimized). The reported accuracy was based on comparing the true class values to the probabilities generated in the model. The probabilities were also used in calculating the AUC.

To account for class imbalance and the fact that the data sets had different proportions of cases versus controls, a random undersampling and aggregation technique was employed. During undersampling, a random set of training samples was selected for model building, with the number for each class being equal and determined by subtracting one from the smallest class of samples. (For example, if there were 39 AD patients and 34 controls, the number of samples in each class in the training set would be 33.) The test sample was always excluded from consideration, and then training samples from each class were randomly selected, based on the pre-set number of samples to be used in the model. The SVM classification commenced, and the probability of the sample being assigned to Group 1 was recorded. This process was repeated 300 times for each sample. The 300 probabilities that resulted from the 300 classifications for each sample were averaged in order to determine a single probability for each sample. This algorithm maximally leverages all available data while not imparting a bias in the results based on the relative sizes of the two classes in the training set. Example code is provided in the Supplementary Material.

RESULTS

Plasma samples (N = 113) from four study groups, African American/Black cognitively normal (African American/Black CN, N = 26) and AD (African American/Black AD, N = 30), non-Hispanic White cognitively normal (non-Hispanic White CN, N = 28) and AD (non-Hispanic White AD, N = 29), were obtained from the University of Pittsburgh ADRC. Generally, there were twice as many females in each group and no significant differences were found in CN compared to AD groups with regards to sex, age, year of blood draw, and presence of other comorbidities (diabetes, hypercholesterolemia, hypertension). Each of the groups had an average 13 years of education, except the non-Hispanic White CN group which had ~16 years of education (Table 1). The average MMSE scores for the CN samples were above 27 while for the AD samples they were below 15. Each patient sample also had information about *APOE* genotypes, and we note that the African American/Black AD group had a higher percentage (i.e., 23%) of the 4/4 genotype compared to non-Hispanic White AD group. AD groups overall had higher percentages of individuals who carried a homozygous and/or heterozygous 3/4 or 4/4 genotypes compared to the CN group.

A general overview of the plasma proteomics workflow employed for both Set 1 and Set 2 is shown in Fig. 1. For both experiments, all steps, except the digestion process were similar in order to simulate known variations in sample preparation study designs across proteomics laboratories. A total of 538 high confidence proteins (1% FDR, PSMs 2) were identified in Set 1, and 596 proteins were identified in Set 2 yielding a total of 740 proteins from both sets (Fig. 2). Between Set 1 and Set 2, 395 proteins were identified in both sets. While this level of coverage is desirable, many of the proteins were identified in only a sub-population of the samples, so different filtering criteria were assessed to choose a data set size that balanced the competing needs of retaining many proteins in the data set while having quantitative data for as many samples as possible for each of the proteins. When considering only proteins with 50% of the TMT channels present, 285 and 380 proteins remained in Set 1 and Set 2, respectively. When considering only proteins with 75% of the TMT channels present, 249 and 314 proteins, respectively, were present in Set 1 and Set 2, with > 95% of the proteins in Set 1 also appearing in Set 2. This filtering option results

in significantly better coverage for each protein while only minimally shrinking the size of the data set. We also considered filtering the data to include only proteins with 100% of the TMT channels present, which resulted in 189 proteins in Set 1, 257 proteins in Set 2, and 183 proteins in common between the two sets. This filtering option would likely remove too many important but low-abundant proteins. We selected those proteins present in 75% of the patient samples for further analysis, as this data set best balanced the needs of retaining as many proteins as possible while providing quantitative data for as many samples as possible for each protein.

Differentially-expressed proteins in AD

Since only a small fraction of plasma proteins were expected to be differentially expressed between AD and CN groups, a robust process to select the optimal proteins for machine learning was needed. Thus, we focused on identifying differentially-expressed proteins in Set 1 (N = 34 CN, N = 39 AD) which had a larger number of samples compared to Set 2 (N = 20 CN, N = 20 AD). For this initial analysis, we combined data from all of the CN individuals into one group and all with AD into a second group. Figure 3a displays a volcano plot distribution of protein TMT ratios of the AD compared to CN groups. Of the proteins with significant p-values in Set 1 ($p < 0.05$), four proteins had fold-changes that were > 1.18: beta-ala-his dipeptidase (FC = 0.73, p 0.0001), keratin type I cytoskeletal 9 (FC = 0.71, p 0.049), apolipoprotein L1 (FC = 0.84, p 0.03), and adiponectin (FC = 1.40, p 0.02). Beta-Ala-His dipeptidase, keratin type I cytoskeletal 9, and apolipoprotein L1 were all lower in AD compared to CN, while adiponectin was higher in AD. These changes are consistent with literature reports [80–82]. Thus, these four proteins from Set 1 were selected as the protein feature to use in subsequent machine learning studies, and their utility for confirming AD was tested in both Set 1 and Set 2.

Using an SVM classifier and leave-one-out cross-validation, we determined the utility of the four selected proteins for confirming AD. Two models were tested: one included only the four proteins (beta-ala-his dipeptidase, keratin type I cytoskeletal 9, apolipoprotein L1, and adiponectin), and the second contained these proteins along with four additional variables: age, sex, years of education, and *APOE* status. These models were tested twice, using samples from either Set 1 or Set 2. The classification accuracies of the two models in Set 1 and Set 2, are shown in Fig. 4.

Several significant outcomes are noted based on results in Fig. 4. In all four classifications, the non-Hispanic White adult samples had higher accuracy than the African American/ Blacks adult samples.

Additionally, in both Set 1 and Set 2, a higher overall accuracy was obtained when the other variables were included. However, the variables of age, sex, years of education, and *APOE* status, provided a bigger boost to the accuracy of the samples from non-Hispanic White adults compared to African American/Black adults. The fact that the two tested models were not as effective for the samples from African American/Black adults as they were for the samples from non-Hispanic White adults lead us to consider race-stratification of proteomics data prior to machine learning.

Differentially-expressed proteins in AD in race-stratified groups

Differentially-expressed proteins between AD and CN groups for the African American/ Black and non-Hispanic White samples in Set 1, are shown using volcano plots in Fig. 3b and 3c. Nine proteins were differentially expressed for the non-Hispanic White samples. Most of these proteins were decreased in AD with serum amyloid A-1 protein having the largest change (FC = 0.5, p 0.02). The remaining proteins included: beta-ala-his dipeptidase (FC = 0.69, p 0.0009), dopamine beta-hydroxylase (FC = 0.69, p 0.03), apolipoprotein C3 (FC = 0.74, p 0.001), serum amyloid A-4 protein (FC = 0.8, p 0.02), multimerin-2 (FC = 0.8, p 0.01), apolipoprotein E (FC = 0.81, p 0.045), and afamin (FC = 0.81, p 0.01). Adiponectin (FC = 1.61, *p*-value 0.02) was the only proteins that increased in non-Hispanic White AD samples.

Two proteins were differentially-expressed in samples from African American/Black adults in Set 1 (Fig. 3c): beta-ala-his dipeptidase (FC = 0.78, p 0.04) and keratin type II cytoskeletal 1(FC = 0.59, p 0.048). Only beta-ala-his dipeptidase was differentially-expressed in AD for both African American/Black and non-Hispanic White samples. A list of differentially-expressed proteins with corresponding p- and fold-change values for Set 1 and Set 2 are provided in Table 2 and Supplementary Table 1 respectively.

Supervised classification of differentially-expressed proteins

Eight unique classifications were performed using the differentially-expressed proteins selected from volcano plots in Fig. 3b and 3c. The first SVM classification used the set of nine proteins that were differentially expressed in the non-Hispanic White group from Set 1 (Fig. 3b) as a feature set, and the non-Hispanic White group data from Set 1 for training. Similarly, a second classification was conducted in the same way with Set 2 samples. Two additional classifications were performed in which the nine proteins described above were combined with the four variables (age/sex/education/*APOE* status). In total, four classifications were performed using the proteins that were differentially expressed in the samples from non-Hispanic White AD adults in Set 1. A parallel set of four classifications was performed using the set of proteins that were differentially expressed in samples from African American/Black AD adults (Fig. 3c). These proteins were tested in Set 1 and Set 2, either on their own or with the four clinical variables. In every case where the feature set included proteins that had been differentially expressed in samples from African American/ Black adults, those samples from the set being tested, were used to train the model. Results for all eight classifications, separated by racial group, are shown in Table 3. It should be noted that Table 3 exclusively includes the accuracy for test data, not training data. Test samples were always left out when training the models (see Methods).

Overall, when differentially-expressed proteins from the non-Hispanic White group were used to classify AD in the samples, an effective model (i.e., AUC was 0.91 and the accuracy was 86%) was developed, but only for the samples from the non-Hispanic White adults. In contrast, the model performed noticeably worse (i.e., AUC of 0.49 and accuracy of 47%) when classifying the data from samples of African American/Black adults. When other (age, sex, education, *APOE* status) variables were also included in the model, the classification accuracy (i.e., AUC was 0.61 and the accuracy was 56%) of the samples

from African American/Black adults improved, yet not to the level of accuracy what would be required for an effective biomarker assay. Incorporating (age, sex, education, *APOE* status) variables improved the accuracy from 0.91 to 0.97 for the non-Hispanic White group, demonstrating that these variables could be combined with protein data to generate a model that can effectively predict AD in this demographic. This desirable outcome, which did not happen when those variables were included for the African American/Black group, suggests that other variables are needed to improve the accuracy of a given model for the African American/Black adult samples in this study.

When SVM classifications were performed on Set 2, the overall findings from Set 1 were generally replicated. Using data from samples in only the African American/Black adult group did not result in accurate classification (i.e., maximum AUC was 0.43 and the accuracy was 45%), either with or without (age, sex, education, *APOE* status) variables included. In Set 2, samples from non-Hispanic White adults were best classified by combining protein data and (age, sex, education, *APOE*) variables (Table 3).

Differences between the two racial groups are readily apparent when the classification model is built using differentially-expressed proteins ($N = 2$, Table 2) from the African American/Black group and only samples from this group to train the model. The best overall outcome for samples from non-Hispanic White adults occurred when Set 1 was classified

Using the clinical variables (sex/age/education/*APOE* status) and the protein features. In this case, an AUC of 0.91 and accuracy of 84% was obtained. Using similar data from samples in the African American/Black groups, resulted in worse classification performance (i.e., AUC of 0.47 and accuracy of 47%). The best overall outcome for samples from African American/Black adults was observed when no clinical variables were used and when testing data in Set 2. In that case, an improved classification was obtained (i.e., AUC of 0.84 and an accuracy of 70%) demonstrating how critical it was to evaluate various parameters and testing approaches in finding the best classification for both racial groups. However, it is clear that the performance outcomes of the machine learning model is also dependent on the samples and protein data obtained in training and test sets. In all the tests where samples from African American/Black adults were used for model training, more accurate classification without the age/sex/education/*APOE* status variables was obtained. By contrast, in three out of the four classifications where these variables were included for samples from non-Hispanic White adults, the samples from that group were more accurately classified.

DISCUSSION

In this study we performed two independent plasma proteomics experiments between AD and CN samples collected from African American/Black and non-Hispanic White participants from the University of Pittsburgh ADRC. At the time of this study, there were only 56 African American/Black participants that had banked plasma sample meeting our criterion. We used the demographics of those participants, primarily age, sex, and AD diagnosis, to match a similar size of non-Hispanic White participants. The AD diagnosis was clinically confirmed, and we used the most recent blood draw for these plasma analyses.

Most characteristics that we evaluated of the participants were generally similar between African American/Black and non-Hispanic White groups.

We identified 740 proteins in total using TMT-based quantitative proteomics on an Orbitrap Fusion Lumos MS instrument, which is on par with recent plasma proteomics publications [83–90]. However, with substantial increases in high pH reversed-phase fractionation in the sample processing, higher numbers of plasma protein identifications could have been identified [91]. Additionally, we note that our MS data acquisition method used MS^3 for TMT measurements in order to increase the quantitative accuracy of differentially-expressed proteins. However, using this approach is known to increase instrument duty cycle and result in lesser protein identifications. After applying stringent criteria for identification, presence of TMT reporter ion channels across samples, and fold-change and p-value cutoffs, we identified a total of 27 differentially-expressed proteins (see Methods) in AD. These proteins were either different in a specific racial group or in comparisons of both racial groups together. Several of these proteins had changes in AD that were previously reported [80, 81, 90]. Notable in this study was the inclusion of samples from African American/Black ADRC participants. Based on our analyses, there was only one differentially-expressed protein (beta-Ala-His dipeptidase) in AD that was significant in both African American/Black and non-Hispanic White groups. Lower expression of this protein in AD for both groups is consistent with previous reports [90].

One of the proteins that was only differentially expressed in non-Hispanic White samples included ApoE, which has been widely reported as potential plasma biomarker in AD. There have been contradictory reports of ApoE being higher in AD [92], while others have reported it as lower in AD [52]. In our study, in samples from non-Hispanic White adults, ApoE was lower in adults in AD, whereas it did not have a significant change in AD versus CN in samples from African American/Black adults. Other proteins that were lower in samples from non-Hispanic White AD included afamin, ApoC3, serum amyloid A1 protein, and serum amyloid A4 protein, all of which have been reported to change in previous studies. Direction of change in afamin, ApoC3, and adiponectin (Table 2) were also consistent with literature reports [47, 49, 50, 80, 90, 93–95]. Some of the differentially-expressed proteins (i.e., serum amyloid A1 and serum amyloid A4 proteins) that we identified were observed in other studies; however, the direction of change in AD versus CN was not consistent [90]. Two novel findings in this work were the differential expression of dopamine beta-hydroxylase and multimerin-2 which were only different in the non-Hispanic White group.

The main outcome from our machine learning analyses is that samples from the African American/Black and non-Hispanic White participants had notable differences in performance outcomes. Proteomics data from non-Hispanic White adults was classified substantially more accurately than data from African American/Black adults when both racial groups were combined in the training data (Fig. 4) and when only the non-Hispanic White group was used as training data (Table 3). The other main difference observed is that samples from the non-Hispanic White group were typically best classified when (age, sex, education, and *APOE* status) variables were included in the model, while the classification of the African American/Black group benefitted less from including these

variables. In fact, including these variables was detrimental in both cases where samples only from the African American/Black group were used to train the model and using proteins differentially expressed only in the African American/Black group. Even in the case where samples from both African American/Black and non-Hispanic White groups were used to identify differentially-expressed proteins and to train the model (Fig. 4), including (age, sex, education, and *APOE* status) variables only improved the classification of the samples from African American/Black adults in test data sets.

Overall, these studies show how critical it is for biomarker discovery efforts to be inclusive of individuals from various racial and ethnic backgrounds as it can have a huge impact on effectiveness of machine learning models. These studies found overall that samples from the non-Hispanic White group were more accurately classified with SVM based on changes in plasma proteins from CN and AD adults. Also, the addition of age, sex, years of education, and *APOE* status as variables in the model disproportionally improved the classification accuracy of the non-Hispanic White group compared to the African American/Black group. It is not clear from our findings that the machine learning performance outcomes are simply just tied to the self-reported race as many factors such as life experiences, stress, age, education, mixed-dementia pathology are intermixed in the construct of race and can contribute to protein levels in plasma. These factors should be considered for the African American/Black group; however, are not easy to evaluate in terms of contributions to differential protein expressions in AD, which ultimately was used to feed proteins into the machine learning model. Achieving satisfactory outcomes with machine learning for AD with samples from African American/Black adults was possible in this study for one set of plasma proteins and without the use of covariates of age, sex, years of education, and *APOE* status variables. However, these studies clearly point to a need for increased plasma proteomics studies and number of plasma samples from African American/Black adults that are cognitively normal and with clinical diagnoses of AD in order to facilitate training of machine learning models and importantly, ensure favorable outcomes for African American/ Black adults in plasma proteomics biomarker discovery efforts.

Study strengths and limitations

One of the major strengths of our study is the inclusion of plasma samples from both African American/Black and non-Hispanic White individuals. The number of samples from each group was evenly distributed, with no significant differences due to sex, age, or comorbidities. Also, the AD samples having higher percentage of *APOE4* allele present, as reported by previous studies [1] was maintained by the samples in this study. This is the first study of its kind to compare these two racial groups in plasma proteomics experiments and directly evaluate in machine learning models for AD biomarker discovery. The majority of the proteins we found to be differentially expressed have been previously reported to change in AD; yet a few findings, with regard to AD-related protein expression, are novel, particularly when considering the protein expression data specific to samples from African American/Black adults. These findings coupled with the fact that the number of studies involving samples from African American/Black adults are very limited, puts emphasis on the need for conducting more AD research including African American/Black patient

samples. This additional effort is necessary to both better understand disparities in disease incidence but also to ensure biomarker discovery efforts are effective for everyone.

We employed a randomized blinded study design and ensured samples from each study group were included in every TMT batch. This allowed us to minimize error due to sample preparation. Also, we performed MS^3 quantification, which provides highly accurate quantitative information and thus enhances confidence in the fold-changes observed for AD. This choice is potentially critical for facilitating biomarker discovery by focusing on robust protein changes [96]. Despite the use of MS^3, which requires a higher duty cycle, we were able to identify similar numbers of proteins compared to recent publications using MS/MS approaches [83–90].

Another strength is that we had enough plasma samples to establish two independent data sets so that the results obtained from a training set (Set 1) could be validated with a test set (Set 2). There was a high degree of overlap in terms of protein identifications in the training and test sets that could be considered for use in machine learning. Our study design allowed us to stratify our data post-analysis based on the self-reported racial and ethnic group of the participants. In this case, the machine learning model was able to differentiate AD samples with high accuracy for the non-Hispanic White group using a set of nine proteins that were selected based on volcano plot analysis. We believe it was a strength that our study design allowed us to test if self-reported race was a critical factor in the accuracy of potential biomarker candidates. The same protein set that produced high accuracy in the non-Hispanic White group performed poorly when applied to the African American/Black group in both data sets. This strongly indicates a need for more studies that have inclusive designs and for evaluation of whether self-reported race or other variables are critical for biomarker development.

A notable limitation of this study was that protein identifications were cut in half by filtering out proteins that were missing TMT values for at least 50% of the samples. We note that this loss is likely due to batch effects; for every additional TMT batch acquired, new proteins are observed while others can go undetected due to the stochastic nature of the data dependent acquisition [97]. Also, the use of MS^3 could result in proteins being missed, as they may not have been selected for fragmentation in one TMT batch but were in another. DIA or targeted MRM methods could avoid this issue by only focusing on known protein identifications throughout the entire run.

While our sample size was a total of $N = 113$, which was on par with our similar studies involving AD [58, 63, 64, 90, 95, 98], we note that for biomarker discovery efforts this sample size is still small. Obtaining samples from African American/Black participants is challenging with limited availability in ADRCs; however, we will need to work with other ADRCs and focused studies [99, 100] to generate large sample sizes especially from available African American/Black participants. Sample size is especially critical as any racial and ethnic group is not homogenous and existing knowledge in plasma proteomics for AD is based mostly on participants from non-Hispanic White and European backgrounds. We believe our moderate sample size also limited our ability to identify an adequate number and set of differentially-expressed proteins that could be effective for the machine learning

classification of African American/Black adults. This is recognized by our study design to split samples into two sets for training and validation which reduced sample size in each group. However, our sample size was substantially higher than in other reported AD biomarker studies of African American/Black adults [37].

Finally, we note another limitation to this study was not including additional variables other than age/sex/education/*APOE4* status that may have improved machine learning outcomes for samples from the African American/Black group. For example, it is recognized that biases in mini-mental state examination and other cognitive tests [101, 102] may result in inadequacies of generating similar types of participants and that quality of education may be a far more important factor than years of education to include in study designs [103]. Genetic ancestry could be included as an additional measure as well as the use of other types of genetic markers that have AD risk associated with African American/Black adults [104]. Stratification of groups based on self-reported race has limits also because it does not capture life-long experiences, such as discrimination and exposure to systemic racism [18], that have been shown to lead to inherent biases in healthcare and also impact plasma proteomic cytokine levels [105, 106].

CONCLUSIONS

Plasma proteomics analysis combined with classification by machine learning is a powerful strategy for identifying potential biomarker candidates that can be used for AD diagnosis. Plasma proteomics biomarker discovery efforts have largely excluded samples from African American/Black adults, and this study sought to include samples from this group to help facilitate biomarker discovery efforts for everyone. Our analyses demonstrated that potential biomarker candidates for AD diagnosis could be identified with high accuracy in plasma samples from non-Hispanic White adults that were cognitively normal or clinically diagnosed with AD, and that these same candidates were not effective in samples from African American/Black adults. Further improved machine learning outcomes for AD biomarker discovery were possible with the addition of variables such as age, sex, years of education, and *APOE4* status in the training model; however, these specific variables do not appear to be as effective for classifying samples from African American/Black adults with plasma proteomics data. Thus, the search is on for a better set of plasma proteins and/or combined use of clinical/demographic variables which can be used to ensure biomarker discovery efforts in AD are effective for everyone, including African American/Black adults.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

REFERENCES

[1]. Barnes LL, Bennett DA (2014) Alzheimer's disease in African Americans: Risk factors and challenges for the future. Health Aff (Millwood) 33, 580–586. [PubMed: 24711318]

[2]. Matthews KA, Xu W, Gaglioti AH, Holt JB, Croft JB, Mack D, McGuire LC (2019) Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015–2060) in adults aged 65 years. Alzheimers Dement 15, 17–24. [PubMed: 30243772]

[3]. Lines L, Sherif N, Wiener J (2014) Racial and ethnic disparities among individuals with Alzheimer's disease in the United States: A literature review. RTI Press, RTI Press Publication No. RR-0024–1412.

[4]. (2020) 2020 Alzheimer's disease facts and figures. Alzheimers Dement 16, 391–460.

[5]. Chin AL, Negash S, Hamilton R (2011) Diversity and disparity in dementia: The impact of ethnoracial differences in Alzheimer disease. Alzheimer Dis Assoc Discord. 25, 187–195.

[6]. Rajan KB, Weuve J, Barnes LL, Wilson RS, Evans DA (2019) Prevalence and incidence of clinically diagnosed Alzheimer's disease dementia from 1994 to 2012 in a population study. Alzheimers Dement 15, 1–7. [PubMed: 30195482]

[7]. Gilligan AM, Malone DC, Warholak TL, Armstrong EP (2013) Health disparities in cost of care in patients with Alzheimer's disease: An analysis across 4 state Medicaid populations. Am J Alzheimers Dis Other Demen 28, 84–92. [PubMed: 23196405]

[8]. Husaini BA, Sherkat DE, Moonis M, Levine R, Holzer C, Cain VA (2003) Racial differences in the diagnosis of dementia and in its effects on the use and costs of health care services. Psychiatr Serv 54, 92–96. [PubMed: 12509673]

[9]. Mehta KM, Yeo GW (2017) Systematic review of dementia prevalence and incidence in United States race/ethnic populations. Alzheimers Dement 13, 72–83. [PubMed: 27599209]

[10]. Yaffe K, Falvey C, Harris TB, Newman A, Satterfield S, Koster A, Ayonayon H, Simonsick E (2013) Effect of socioeconomic disparities on incidence of dementia among biracial older adults: Prospective study. BMJ 347, f7051. [PubMed: 24355614]

[11]. Weuve J, Barnes LL, Mendes de Leon CF, Rajan KB, Beck T, Aggarwal NT, Hebert LE, Bennett DA, Wilson RS, Evans DA (2018) Cognitive aging in black and white Americans: Cognition, cognitive decline, and incidence of Alzheimer disease dementia. Epidemiology 29, 151–159. [PubMed: 28863046]

[12]. Barnes LL, Wilson RS, Hebert LE, Scherr PA, Evans DA, Mendes de Leon CF (2011) Racial differences in the association of education with physical and cognitive function in older blacks and whites. J Gerontol B Psychol Sci Soc Sci 66, 354–363. [PubMed: 21402644]

[13]. Manly JJ, Schupf N, Tang M-X, Stern Y (2005) Cognitive decline and literacy among ethnically diverse elders. J Geriatr Psychiatry Neurol 18, 213–217. [PubMed: 16306242]

[14]. Sisco S, Gross AL, Shih RA, Sachs BC, Glymour MM, Bangen KJ, Benitez A, Skinner J, Schneider BC, Manly JJ (2015) The role of early-life educational quality and literacy in explaining racial disparities in cognition in late life. J Gerontol B Psychol Sci Soc Sci 70, 557–567. [PubMed: 24584038]

[15]. Stepler KE, Robinson RAS (2019) The potential of 'omics to link lipid metabolism and genetic and comorbidity risk factors of Alzheimer's disease in African Americans. Adv Exp Med Biol 1118, 1–28. [PubMed: 30747415]

[16]. Turner AD, James BD, Capuano AW, Aggarwal NT, Barnes LL (2017) Perceived stress and cognitive decline in different cognitive domains in a cohort of older African Americans. Am J Geriatr Psychiatry 25, 25–34. [PubMed: 28231871]

[17]. Campdelacreu J (2014) Parkinson's disease and Alzheimer disease: Environmental risk factors. Neurología (English Edition) 29, 541–549.

[18]. Wilkins CH, Schindler SE, Morris JC (2020) Addressing health disparities among minority populations: Why clinical trial recruitment is not enough. JAMA Neurol 77, 1063–1064. [PubMed: 32539100]

[19]. Filshtein TJ, Dugger BN, Jin LW, Olichney JM, Farias ST, Carvajal-Carmona L, Lott P, Mungas D, Reed B, Beckett LA, DeCarli C (2019) Neuropathological diagnoses of demented Hispanic, Black, and Non-Hispanic White decedents seen at an Alzheimer's Disease Center. J Alzheimers Dis 68, 145–158. [PubMed: 30775996]

[20]. Kamara DM, Gangishetti U, Gearing M, Willis-Parker M, Zhao L, Hu WT, Walker LC (2018) Cerebral amyloid angiopathy: Similarity in African-Americans and Caucasians with Alzheimer's disease. J Alzheimers Dis 62, 1815–1826. [PubMed: 29614657]

[21]. Graff-Radford NR, Besser LM, Crook JE, Kukull WA, Dickson DW (2016) Neuropathologic differences by race from the National Alzheimer's Coordinating Center. Alzheimers Dement 12, 669–677. [PubMed: 27094726]

[22]. Barnes LL, Leurgans S, Aggarwal NT, Shah RC, Arvanitakis Z, James BD, Buchman AS, Bennett DA, Schneider JA (2015) Mixed pathology is more likely in black than white decedents with Alzheimer dementia. Neurology 85, 528–534. [PubMed: 26180136]

[23]. Gavett BE, Fletcher E, Harvey D, Farias ST, Olichney J, Beckett L, DeCarli C, Mungas D (2018) Ethnoracial differences in brain structure change and cognitive change. Neuropsychology 32, 529–540. [PubMed: 29648842]

[24]. Robinson RAS, Williams IC, Cameron JL, Ward K, Knox M, Terry M, Tamres L, Mbawuike U, Garrett M, Lingler JH (2020) Framework for creating storytelling materials to promote African American/Black adult enrollment in research on Alzheimer's disease and related disorders. Alzheimers Dement (N Y) 6, e12076. [PubMed: 32995472]

[25]. Besser L, Kukull W, Knopman DS, Chui H, Galasko D, Weintraub S, Jicha G, Carlsson C, Burns J, Quinn J, Sweet RA, Rascovsky K, Teylan M, Beekly D, Thomas G, Bollenbeck M, Monsell S, Mock C, Zhou XH, Thomas N, Robichaud E, Dean M, Hubbard J, Jacka M, SchwabeFry K, Wu J, Phelps C, Morris JC (2018) Version 3 of the National Alzheimer's Coordinating Center's Uniform Data Set. Alzheimer Dis Assoc Disord 32, 351–358. [PubMed: 30376508]

[26]. Besser LM, Kukull WA, Teylan MA, Bigio EH, Cairns NJ, Kofler JK, Montine TJ, Schneider JA, Nelson PT (2018) The Revised National Alzheimer's Coordinating Center's Neuropathology Form-Available data and new analyses. J Neuropathol Exp Neurol 77, 717–726. [PubMed: 29945202]

[27]. Doody RS, Thomas RG, Farlow M, Iwatsubo T, Vellas B, Joffe S, Kieburtz K, Raman R, Sun X, Aisen PS, Siemers E, Liu-Seifert H, Mohs R (2014) Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. N Engl J Med 370, 311–321. [PubMed: 24450890]

[28]. Rosenmann H (2012) CSF biomarkers for amyloid and tau pathology in Alzheimer's disease. J Mol Neurosci 47, 1–14. [PubMed: 22058061]

[29]. Howell JC, Watts KD, Parker MW, Wu J, Kollhoff A, Wingo TS, Dorbin CD, Qiu D, Hu WT (2017) Race modifies the relationship between cognition and Alzheimer's disease cerebrospinal fluid biomarkers. Alzheimers Res Ther 9, 88. [PubMed: 29096697]

[30]. Morris JC, Schindler SE, McCue LM, Moulder KL, Benzinger TLS, Cruchaga C, Fagan AM, Grant E, Gordon BA, Holtzman DM, Xiong C (2019) Assessment of racial disparities in biomarkers for Alzheimer disease. JAMA Neurol 76, 264–273. [PubMed: 30615028]

[31]. Blennow K, Vanmechelen E, Hampel H (2001) CSF total tau, Aβ42 and phosphorylated tau protein as biomarkers for Alzheimer's disease. Mol Neurobiol 24, 87–97. [PubMed: 11831556]

[32]. Wallin ÅK, Blennow K, Andreasen N, Minthon L (2006) CSF biomarkers for Alzheimer's disease: Levels of β-amyloid, tau, phosphorylated tau relate to clinical symptoms and survival. Dement Geriatr Cogn Disord 21, 131–138. [PubMed: 16391474]

[33]. Garrett SL, McDaniel D, Obideen M, Trammell AR, Shaw LM, Goldstein FC, Hajjar I (2019) Racial disparity in cerebrospinal fluid amyloid and tau biomarkers and associated cutoffs for mild cognitive impairment. JAMA Netw Open 2, e1917363. [PubMed: 31834392]

[34]. Wharton W, Kollhoff AL, Gangishetti U, Verble DD, Upadhya S, Zetterberg H, Kumar V, Watts KD, Kippels AJ, Gearing M, Howell JC, Parker MW, Hu WT (2019) Interleukin 9 alterations linked to Alzheimer disease in African Americans. Ann Neurol 86, 407–418. [PubMed: 31271450]

[35]. Simino J, Wang Z, Bressler J, Chouraki V, Yang Q, Younkin SG, Seshadri S, Fornage M, Boerwinkle E, Mosley TH, Jr. (2017) Whole exome sequence-based association analyses

of plasma amyloid-β in African and European Americans; the Atherosclerosis Risk in Communities-Neurocognitive Study. PLoS One 12, e0180046. [PubMed: 28704393]

[36]. Tranah GJ, Yokoyama JS, Katzman SM, Nalls MA, Newman AB, Harris TB, Cesari M, Manini TM, Schork NJ, Cummings SR, Liu Y, Yaffe K, Health, Aging and Body Composition Study (2014) Mitochondrial DNA sequence associations with dementia and amyloid- in elderly African Americans. Neurobiol Aging 35, 442. e441–442.e4428.

[37]. Grewal R, Haghighi M, Huang S, Smith AG, Cao C, Lin X, Lee DC, Teten N, Hill AM, Selenica M-LB (2016) Identifying biomarkers of dementia prevalent among amnestic mild cognitively impaired ethnic female patients. Alzheimers Res Ther 8, 43. [PubMed: 27756387]

[38]. Robinson RA, Amin B, Guest PC (2017) Multiplexing biomarker methods, proteomics and considerations for Alzheimer's disease. Adv Exp Med Biol 974, 21–48. [PubMed: 28353223]

[39]. Kiddle SJ, Sattlecker M, Proitsi P, Simmons A, Westman E, Bazenet C, Nelson SK, Williams S, Hodges A, Johnston C, Soininen H, Kłoszewska I, Mecocci P, Tsolaki M, Vellas B, Newhouse S, Lovestone S, Dobson RJB (2014) Candidate blood proteome markers of Alzheimer's disease onset and progression: A systematic review and replication study. J Alzheimers Dis 38, 515–531. [PubMed: 24121966]

[40]. Rehiman SH, Lim SM, Neoh CF, Majeed ABA, Chin AV, Tan MP, Kamaruzzaman SB, Ramasamy K (2020) Proteomics as a reliable approach for discovery of blood-based Alzheimer's disease biomarkers: A systematic review and meta-analysis. Ageing Res Rev 60, 101066. [PubMed: 32294542]

[41]. Thygesen C, Larsen MR, Finsen B (2019) Proteomic signatures of neuroinflammation in Alzheimer's disease, multiple sclerosis and ischemic stroke. Expert Rev Proteomics 16, 601–611. [PubMed: 31220951]

[42]. Lista S, Faltraco F, Prvulovic D, Hampel H (2013) Blood and plasma-based proteomic biomarker research in Alzheimer's disease. Prog Neurobiol 101–102, 1–17.

[43]. Htike TT, Mishra S, Kumar S, Padmanabhan P, Gulyás B (2019) Peripheral biomarkers for early detection of Alzheimer's and Parkinson's diseases. Mol Neurobiol 56, 2256–2277. [PubMed: 30008073]

[44]. Shi L, Baird AL, Westwood S, Hye A, Dobson R, Thambisetty M, Lovestone S (2018) A decade of blood biomarkers for Alzheimer's disease research: An evolving field, improving study designs, and the challenge of replication. J Alzheimers Dis 62, 1181–1198. [PubMed: 29562526]

[45]. Oeckl P, Otto M (2019) A review on MS-based blood biomarkers for Alzheimer's disease. Neurol Ther 8, 113–127. [PubMed: 31833028]

[46]. Shi L, Westwood S, Baird AL, Winchester L, Dobricic V, Kilpert F, Hong S, Franke A, Hye A, Ashton NJ, Morgan AR, Bos I, Vos SJB, Buckley NJ, Kate MT, Scheltens P, Vandenberghe R, Gabel S, Meersmans K, Engelborghs S, De Roeck EE, Sleegers K, Frisoni GB, Blin O, Richardson JC, Bordet R, Molinuevo JL, Rami L, Wallin A, Kettunen P, Tsolaki M, Verhey F, Lleo A, Alcolea D, Popp J, Peyratout G, Martinez-Lage P, Tainta M, Johannsen P, Teunissen CE, Freund-Levi Y, Frolich L, Legido-Quigley C, Barkhof F, Blennow K, Zetterberg H, Baker S, Morgan BP, Streffer J, Visser PJ, Bertram L, Lovestone S, Nevado-Holgado AJ (2019) Discovery and validation of plasma proteomic biomarkers relating to brain amyloid burden by SOMAscan assay. Alzheimers Dement 15, 1478–1488. [PubMed: 31495601]

[47]. Kitamura Y, Usami R, Ichihara S, Kida H, Satoh M, Tomimoto H, Murata M, Oikawa S (2017) Plasma protein profiling for potential biomarkers in the early diagnosis of Alzheimer's disease. Neurol Res 39, 231–238. [PubMed: 28107809]

[48]. Zhao X, Lejnine S, Spond J, Zhang C, Ramaraj TC, Holder DJ, Dai H, Weiner R, Laterza OF (2015) A candidate plasma protein classifier to identify Alzheimer's disease. J Alzheimers Dis 43, 549–563. [PubMed: 25114072]

[49]. Muenchhoff J, Poljak A, Song F, Raftery M, Brodaty H, Duncan M, McEvoy M, Attia J, Schofield PW, Sachdev PS (2015) Plasma protein profiling of mild cognitive impairment and Alzheimer's disease across two independent cohorts. J Alzheimers Dis 43, 1355–1373. [PubMed: 25159666]

[50]. Song F, Poljak A, Kochan NA, Raftery M, Brodaty H, Smythe GA, Sachdev PS (2014) Plasma protein profiling of mild cognitive impairment and Alzheimer's disease using iTRAQ quantitative proteomics. Proteome Sci 12, 5–5. [PubMed: 24433274]

[51]. Zabel M, Schrag M, Mueller C, Zhou W, Crofton A, Petersen F, Dickson A, Kirsch WM (2012) Assessing candidate serum biomarkers for Alzheimer's disease: A longitudinal study. J Alzheimers Dis 30, 311–321. [PubMed: 22426016]

[52]. Hu WT, Holtzman DM, Fagan AM, Shaw LM, Perrin R, Arnold SE, Grossman M, Xiong C, Craig-Schapiro R, Clark CM, Pickering E, Kuhn M, Chen Y, Van Deerlin VM, McCluskey L, Elman L, Karlawish J, Chen-Plotkin A, Hurtig HI, Siderowf A, Swenson F, Lee VMY, Morris JC, Trojanowski JQ, Soares H, Alzheimer's Disease Neuroimaging Initiative (2012) Plasma multianalyte profiling in mild cognitive impairment and Alzheimer disease. Neurology 79, 897–905. [PubMed: 22855860]

[53]. Henkel AW, Muller K, Lewczuk P, Muller T, Marcus K, Kornhuber J, Wiltfang J (2012) Multidimensional plasma protein separation technique for identification of potential Alzheimer's disease plasma biomarkers: A pilot study. J Neural Transm (Vienna) 119, 779–788. [PubMed: 22415063]

[54]. Bennett S, Grant M, Creese AJ, Mangialasche F, Cecchetti R, Cooper HJ, Mecocci P, Aldred S (2012) Plasma levels of complement 4a protein are increased in Alzheimer's disease. Alzheimer Dis Assoc Disord 26, 329–334. [PubMed: 22052466]

[55]. Ijsselstijn L, Dekker LJM, Stingl C, van der Weiden MM, Hofman A, Kros JM, Koudstaal PJ, Sillevis Smitt PAE, Ikram MA, Breteler MMB, Luider TM (2011) Serum levels of pregnancy zone protein are elevated in presymptomatic Alzheimer's disease. J Proteome Res 10, 4902–4910. [PubMed: 21879768]

[56]. O'Bryant SE, Xiao G, Barber R, Reisch J, Doody R, Fairchild T, Adams P, Waring S, Diaz-Arrastia R (2010) A serum protein-based algorithm for the detection of Alzheimer disease. Arch Neurol 67, 1077–1081. [PubMed: 20837851]

[57]. Soares HD, Chen Y, Sabbagh M, Roher A, Schrijvers E, Breteler M (2009) Identifying early markers of Alzheimer's disease using quantitative multiplex proteomic immunoassay panels. Ann N Y Acad Sci 1180, 56–67. [PubMed: 19906261]

[58]. Thambisetty M, Hye A, Foy C, Daly E, Glover A, Cooper A, Simmons A, Murphy D, Lovestone S (2008) Proteome-based identification of plasma proteins associated with hippocampal metabolism in early Alzheimer's disease. J Neurol 255, 1712–1720. [PubMed: 19156487]

[59]. Cutler P, Akuffo EL, Bodnar WM, Briggs DM, Davis JB, Debouck CM, Fox SM, Gibson RA, Gormley DA, Holbrook JD, Hunter AJ, Kinsey EE, Prinjha R, Richardson JC, Roses AD, Smith MA, Tsokanas N, Willé DR, Wu W, Yates JW, Gloger IS (2008) Proteomic identification and early validation of complement 1 inhibitor and pigment epithelium-derived factor: Two novel biomarkers of Alzheimer's disease in human plasma. Proteomics Clin Appl 2, 467–477. [PubMed: 21136851]

[60]. Ray S, Britschgi M, Herbert C, Takeda-Uchimura Y, Boxer A, Blennow K, Friedman LF, Galasko DR, Jutel M, Karydas A, Kaye JA, Leszek J, Miller BL, Minthon L, Quinn JF, Rabinovici GD, Robinson WH, Sabbagh MN, So YT, Sparks DL, Tabaton M, Tinklenberg J, Yesavage JA, Tibshirani R, Wyss-Coray T (2007) Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. Nat Med 13, 1359–1362. [PubMed: 17934472]

[61]. Liao PC, Yu L, Kuo CC, Lin C, Kuo YM (2007) Proteomics analysis of plasma for potential biomarkers in the diagnosis of Alzheimer's disease. Proteomics Clin Appl 1, 506–512. [PubMed: 21136702]

[62]. German DC, Gurnani P, Nandi A, Garner HR, Fisher W, Diaz-Arrastia R, O'Suilleabhain P, Rosenblatt KP (2007) Serum biomarkers for Alzheimer's disease: Proteomic discovery. Biomed Pharmacother 61, 383–389. [PubMed: 17614251]

[63]. Liu HC, Hu CJ, Chang JG, Sung SM, Lee LS, Yuan RY, Leu SJ (2006) Proteomic identification of lower apolipoprotein A-I in Alzheimer's disease. Dement Geriatr Cogn Disord 21, 155–161. [PubMed: 16391478]

[64]. Hye A, Lynham S, Thambisetty M, Causevic M, Campbell J, Byers HL, Hooper C, Rijsdijk F, Tabrizi SJ, Banner S, Shaw CE, Foy C, Poppe M, Archer N, Hamilton G, Powell J, Brown RG, Sham P, Ward M, Lovestone S (2006) Proteome-based plasma biomarkers for Alzheimer's disease. Brain 129, 3042–3050. [PubMed: 17071923]

[65]. Generoso JS, Morales R, Barichello T (2020) Biomarkers in Alzheimer disease: Are we there yet? Braz J Psychiatry 42, 337–339. [PubMed: 32667591]

[66]. El Kadmiri N, Said N, Slassi I, El Moutawakil B, Nadifi S (2018) Biomarkers for Alzheimer disease: Classical and novel candidates' review. Neuroscience 370, 181–190. [PubMed: 28729061]

[67]. O'Bryant SE, Mielke MM, Rissman RA, Lista S, Vanderstichele H, Zetterberg H, Lewczuk P, Posner H, Hall J, Johnson L, Fong Y-L, Luthman J, Jeromin A, Batrla-Utermann R, Villarreal A, Britton G, Snyder PJ, Henriksen K, Grammas P, Gupta V, Martins R, Hampel H (2017) Blood-based biomarkers in Alzheimer disease: Current state of the science and a novel collaborative paradigm for advancing from discovery to clinic. Alzheimers Dement 13, 45–58. [PubMed: 27870940]

[68]. Guo LH, Alexopoulos P, Wagenpfeil S, Kurz A, Perneczky R (2013) Plasma proteomics for the identification of Alzheimer disease. Alzheimer Dis Assoc Disord 27, 337–342 [PubMed: 23314060]

[69]. Shah DJ, Rohlfing F, Anand S, Johnson WE, Alvarez MTB, Cobell J, King J, Young SA, Kauwe JSK, Graves SW (2016) Discovery and subsequent confirmation of novel serum biomarkers diagnosing Alzheimer's disease. J Alzheimers Dis 49, 317–327. [PubMed: 26484917]

[70]. Doecke JD, Laws SM, Faux NG, Wilson W, Burnham SC, Lam CP, Mondal A, Bedo J, Bush AI, Brown B, De Ruyck K, Ellis KA, Fowler C, Gupta VB, Head R, Macaulay SL, Pertile K, Rowe CC, Rembach A, Rodrigues M, Rumble R, Szoeke C, Taddei K, Taddei T, Trounson B, Ames D, Masters CL, Martins RN (2012) Blood-based protein biomarkers for diagnosis of Alzheimer disease. Arch Neurol 69, 1318–1325. [PubMed: 22801742]

[71]. Kim S, Nho K, Ramanan VK, Lai D, Foroud TM, Lane K, Murrell JR, Gao S, Hall KS, Unverzagt FW, Baiyewu O, Ogunniyi A, Gureje O, Kling MA, Doraiswamy PM, Kaddurah-Daouk R, Hendrie HC, Saykin AJ (2016) Genetic influences on plasma homocysteine levels in African Americans and Yoruba Nigerians. J Alzheimers Dis 49, 991–1003. [PubMed: 26519441]

[72]. Wisniewski JR, Zougman A, Nagaraj N, Mann M (2009) Universal sample preparation method for proteome analysis. Nat Methods 6, 359–362. [PubMed: 19377485]

[73]. Plubell DL, Wilmarth PA, Zhao Y, Fenton AM, Minnier J, Reddy AP, Klimek J, Yang X, David LL, Pamir N (2017) Extended multiplexing of Tandem Mass Tags (TMT) labeling reveals age and high fat diet specific proteome changes in mouse epididymal adipose tissue. Mol Cell Proteomics 16, 873–890. [PubMed: 28325852]

[74]. Pascovici D, Handler DC, Wu JX, Haynes PA (2016) Multiple testing corrections in quantitative proteomics: A useful but blunt tool. Proteomics 16, 2448–2453. [PubMed: 27461997]

[75]. Wang W, Sue ACH, Goh WWB (2017) Feature selection in clinical proteomics: With great power comes great reproducibility. Drug Discov Today 22, 912–918. [PubMed: 27988358]

[76]. Cao Z, Yende S, Kellum JA, Robinson RAS (2013) Additions to the human plasma proteome via a tandem MARS depletion iTRAQ-based workflow. Int J Proteomics 2013, 654356. [PubMed: 23509626]

[77]. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, Perez E, Uszkoreit J, Pfeuffer J, Sachsenberg T, Yilmaz S, Tiwary S, Cox J, Audain E, Walzer M, Jarnuczak AF, Ternent T, Brazma A, Vizcaíno JA (2019) The PRIDE database and related tools and resources in 2019: Improving support for quantification data. Nucleic Acids Res 47, D442–d450. [PubMed: 30395289]

[78]. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A (2009) E1071: Misc Functions of the Department of Statistics (E1071), TU Wien.

[79]. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M (2011) pROC: An open-source package for R and S+to analyze and compare ROC curves. BMC Bioinformatics 12, 77. [PubMed: 21414208]

[80]. Une K, Takei YA, Tomita N, Asamura T, Ohrui T, Furukawa K, Arai H (2011) Adiponectin in plasma and cerebrospinal fluid in MCI and Alzheimer's disease. Eur J Neurol 18, 1006–1009. [PubMed: 20727007]

[81]. Richens JL, Spencer HL, Butler M, Cantlay F, Vere K-A, Bajaj N, Morgan K, O'Shea P (2016) Rationalising the role of Keratin 9 as a biomarker for Alzheimer's disease. Sci Rep 6, 22962. [PubMed: 26973255]

[82]. Chen M, Xia W (2020) Proteomic profiling of plasma and brain tissue from Alzheimer's disease patients reveals candidate network of plasma biomarkers. J Alzheimers Dis 76, 349–368. [PubMed: 32474469]

[83]. Wang Z, Liu F, Ye S, Jiang P, Yu X, Xu J, Du X, Ma L, Cao H, Yuan C, Shen Y, Lin F, Zhang R, Li C (2019) Plasma proteome profiling of high-altitude polycythemia using TMT-based quantitative proteomics approach. J Proteomics 194, 60–69. [PubMed: 30605725]

[84]. Wang T, Shen H, Deng H, Pan H, He Q, Ni H, Tao J, Liu S, Xu L, Yao M (2020) Quantitative proteomic analysis of human plasma using tandem mass tags to identify novel biomarkers for herpes zoster. J Proteomics 225, 103879. [PubMed: 32585426]

[85]. Chen C, Geng L, Xu X, Kong W, Hou Y, Yao G, Feng X, Zhang H, Liang J (2020) Comparative proteomics analysis of plasma protein in patients with neuropsychiatric systemic lupus erythematosus. Ann Transl Med 8, 579–579. [PubMed: 32566606]

[86]. Ye S, Ma L, Zhang R, Liu F, Jiang P, Xu J, Cao H, Du X, Lin F, Cheng L, Zhou X, Shi Z, Liu Y, Huang Y, Wang Z, Li C (2019) Plasma proteomic and autoantibody profiles reveal the proteomic characteristics involved in longevity families in Bama, China. Clin Proteomics 16, 22. [PubMed: 31139026]

[87]. Shen N, Chen N, Zhou X, Zhao B, Huang R, Liang J, Yang X, Chen M, Song Y, Du Q (2019) Alterations of the gut microbiome and plasma proteome in Chinese patients with adolescent idiopathic scoliosis. Bone 120, 364–370. [PubMed: 30481617]

[88]. Lu J, Zhang W, Zhang L, Lou Y, Gu P, Nei W, Qian J, Xu J, Zhao X, Zhong H, Wang H, Han B (2020) Plasma proteomics analysis for screening anlotinib responders in non-small cell lung cancer. Research Square, doi: 10.21203/rs.3.rs-26596/v1

[89]. Wang Z, Zhang R, Liu F, Jiang P, Xu J, Cao H, Du X, Ma L, Lin F, Cheng L, Zhou X, Shi Z, Liu Y, Huang Y, Ye S, Li C (2019) TMT-based quantitative proteomic analysis reveals proteomic changes involved in longevity. Proteomics Clin Appl 13, 1800024.

[90]. Chen M, Xia W (2020) Proteomic profiling of plasma and brain tissue from Alzheimer's disease patients reveals candidate network of plasma biomarkers. J Alzheimers Dis 76, 349–368. [PubMed: 32474469]

[91]. Zhou B, Zhou Z, Chen Y, Deng H, Cai Y, Rao X, Yin Y, Rong L (2020) Plasma proteomics-based identification of novel biomarkers in early gastric cancer. Clin Biochem 76, 5–10. [PubMed: 31765635]

[92]. Nazeri A, Ganjgahi H, Roostaei T, Nichols T, Zarei M, Alzheimer's Disease Neuroimaging Initiative (2014) Imaging proteomics for diagnosis, monitoring and prediction of Alzheimer's disease. Neuroimage 102 Pt 2, 657–665. [PubMed: 25173418]

[93]. Moya-Alvarado G, Gershoni-Emek N, Perlson E, Bronfman FC (2016) Neurodegeneration and Alzheimer's disease (AD). What can proteomics tell us about the Alzheimer's brain? Mol Cell Proteomics 15, 409–425. [PubMed: 26657538]

[94]. Oeckl P, Halbgebauer S, Anderl-Straub S, Steinacker P, Huss AM, Neugebauer H, von Arnim CAF, Diehl-Schmid J, Grimmer T, Kornhuber J, Lewczuk P, Danek A, Ludolph AC, Otto M (2019) Glial fibrillary acidic protein in serum is increased in Alzheimer's disease and correlates with cognitive impairment. J Alzheimers Dis 67, 481–488. [PubMed: 30594925]

[95]. Zhang R, Barker L, Pinchev D, Marshall J, Rasamoelisolo M, Smith C, Kupchak P, Kireeva I, Ingratta L, Jackowski G (2004) Mining biomarkers in human sera using proteomic tools. Proteomics 4, 244–256. [PubMed: 14730686]

[96]. McAlister GC, Nusinow DP, Jedrychowski MP, Wühr M, Huttlin EL, Erickson BK, Rad R, Haas W, Gygi SP (2014) MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. Anal Chem 86, 7150–7158. [PubMed: 24927332]

[97]. Brenes A, Hukelmann J, Bensaddek D, Lamond AI (2019) Multibatch TMT reveals false positives, batch effects and missing values. Mol Cell Proteomics 18, 1967–1980. [PubMed: 31332098]

[98]. Shen L, Liao L, Chen C, Guo Y, Song D, Wang Y, Chen Y, Zhang K, Ying M, Li S, Liu Q, Ni J (2017) Proteomics analysis of blood serums from Alzheimer's disease patients using iTRAQ labeling technology. J Alzheimers Dis 56, 361–378. [PubMed: 27911324]

[99]. Barnes LL, Shah RC, Aggarwal NT, Bennett DA, Schneider JA (2012) The Minority Aging Research Study: Ongoing efforts to obtain brain donation in African Americans without dementia. Curr Alzheimer Res 9, 734–745. [PubMed: 22471868]

[100]. Albert SM, Costa R, Merchant C, Small S, Jenders RA, Stern Y (1999) Hospitalization and Alzheimer's disease: Results from a community-based study. J Gerontol A Biol Sci Med Sci 54, M267–M271. [PubMed: 10362011]

[101]. Jones RN(2003)Racialbiasintheassessmentofcognitive functioning of older adults. Aging Ment Health 7, 83–102. [PubMed: 12745387]

[102]. Jones RN (2003) Racial bias in the assessment of cognitive functioning of older adults. Aging Ment Health 7, 83–102. [PubMed: 12745387]

[103]. Manly JJ, Jacobs DM, Touradji P, Small SA, Stern Y (2002) Reading level attenuates differences in neuropsychological test performance between African American and White elders. J Int Neuropsychol Soc 8, 341–348. [PubMed: 11939693]

[104]. Hohman TJ, Cooke-Bailey JN, Reitz C, Jun G, Naj A, Beecham GW, Liu Z, Carney RM, Vance JM, Cuccaro ML, Rajbhandary R, Vardarajan BN, Wang L-S, Valladares O, Lin C-F, Larson EB, Graff-Radford NR, Evans D, De Jager PL, Crane PK, Buxbaum JD, Murrell JR, Raj T, Ertekin-Taner N, Logue MW, Baldwin CT, Green RC, Barnes LL, Cantwell LB, Fallin MD, Go RCP, Griffith P, Obisesan TO, Manly JJ, Lunetta KL, Kamboh MI, Lopez OL, Bennett DA, Hardy J, Hendrie HC, Hall KS, Goate AM, Lang R, Byrd GS, Kukull WA, Foroud TM, Farrer LA, Martin ER, Pericak-Vance MA, Schellenberg GD, Mayeux R, Haines JL, Thornton-Wells TA, Alzheimer Disease Genetics Consortium (2016) Global and local ancestry in African-Americans: Implications for Alzheimer's disease risk. Alzheimers Dement 12, 233–243. [PubMed: 26092349]

[105]. Juster RP, McEwen BS, Lupien SJ (2010) Allostatic load biomarkers of chronic stress and impact on health and cognition. Neurosci Biobehav Rev 35, 2–16. [PubMed: 19822172]

[106]. Danese A, McEwen BS (2012) Adverse childhood experiences, allostasis, allostatic load, and age-related disease. Physiol Behav 106, 29–39. [PubMed: 21888923]
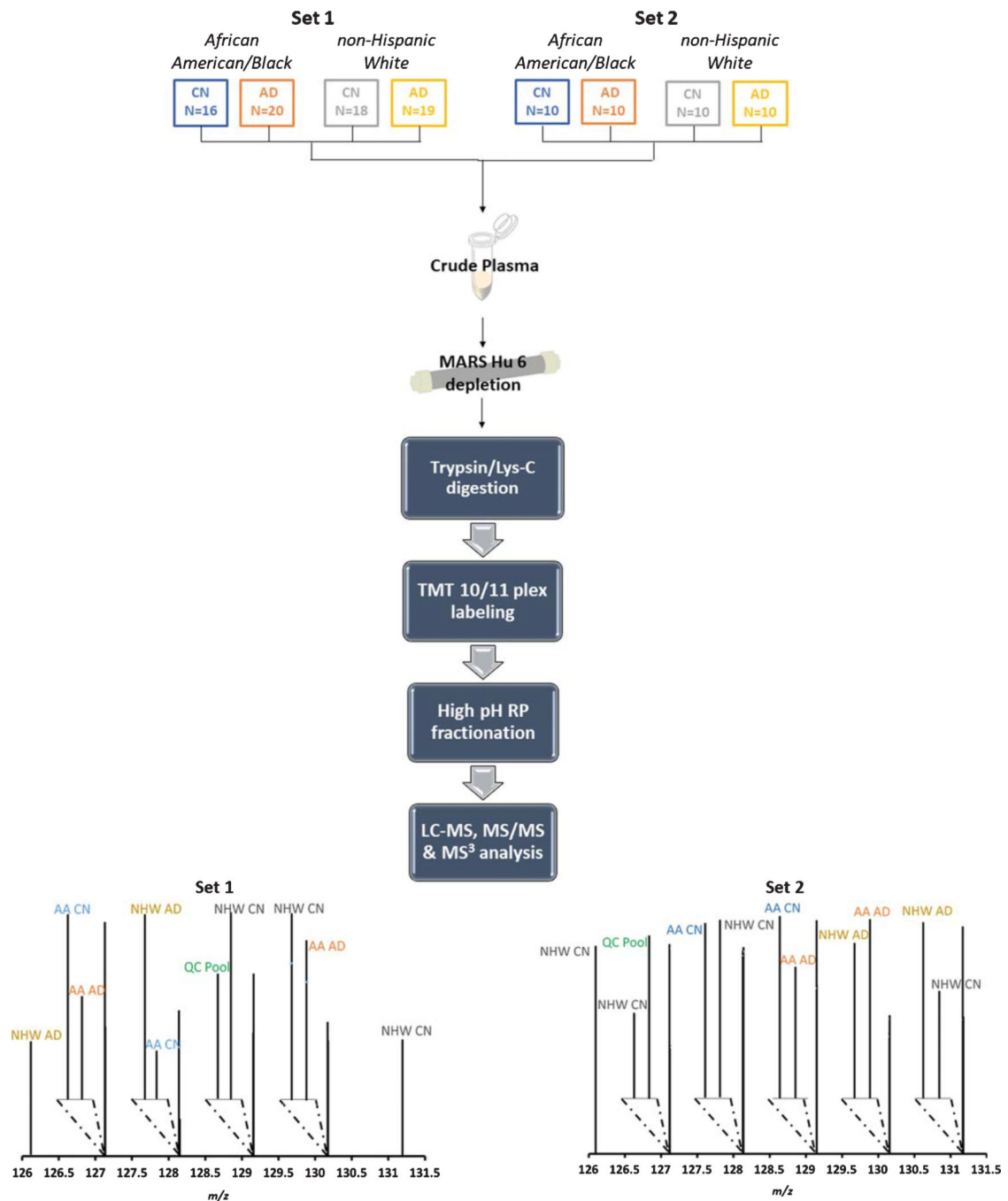
Fig. 1.

Overview of the plasma proteomics workflow. Samples from four study groups-African American/Black Alzheimer's disease (AD) and cognitively normal (CN), non-Hispanic White AD and CN-were obtained from the University of Pittsburgh ADRC. Samples were divided into Set 1 ($N = 73$) and Set 2 ($N = 40$) for this study. Samples were randomized into eight batches for Set 1 and four batches for Set 2. There was one QC pool sample in each batch and representation of one sample from each study group in each batch. The samples were randomly assigned TMT channels for both experiments. The experimental

workflow was maintained the same except for the digestion step, where in solution digestion was used for Set 1, while FASP digestion was employed in Set 2. The plasma samples were immunodepleted of the six most abundant proteins, followed by proteolytic digestion. This was followed by isobaric tagging using either TMT 10/11 plex labels, followed by high pH reversed-phase fractionation. The resulting peptides were loaded into an Ultimate 3000 RPLC system coupled to an Orbitrap Fusion Lumos mass spectrometer for LC-MS, MS/MS and MS^3 analysis. Example representative MS^3 reporter ion spectra for TMT-10 plex sample (Set 1) and TMT-11 plex sample (Set 2) is also provided, demonstrating analysis of multiple samples using a single injection.
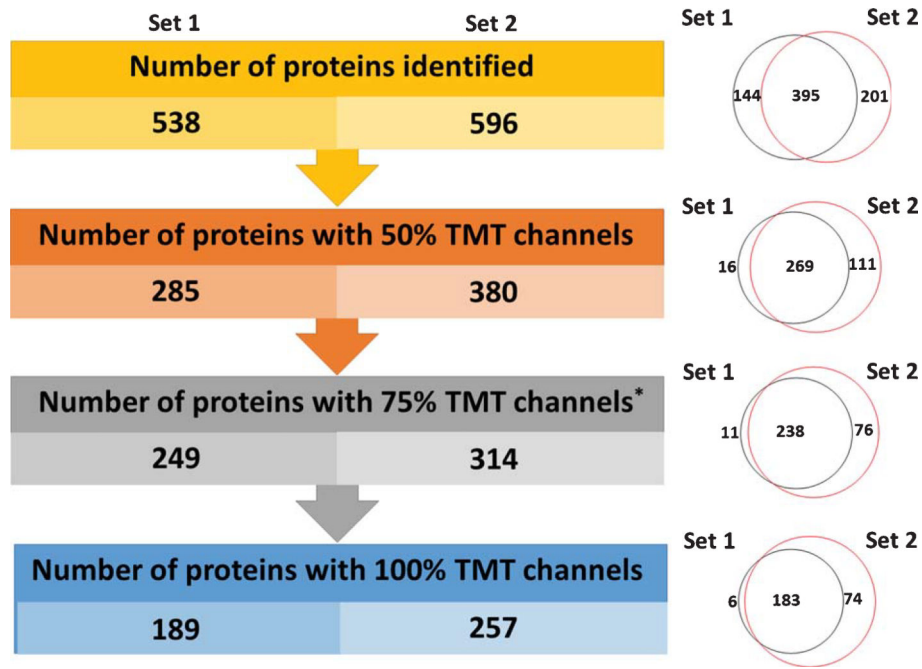
Fig. 2.
Summary of the number of identified proteins in both sample sets. On the left, are the number of high confidence identified proteins as a function of missing channels for TMT reporter ions. Values are provided for Set 1 and Set 2. On the right are Venn diagrams, displaying the overlap in common proteins at each level from Set 1 and Set 2.
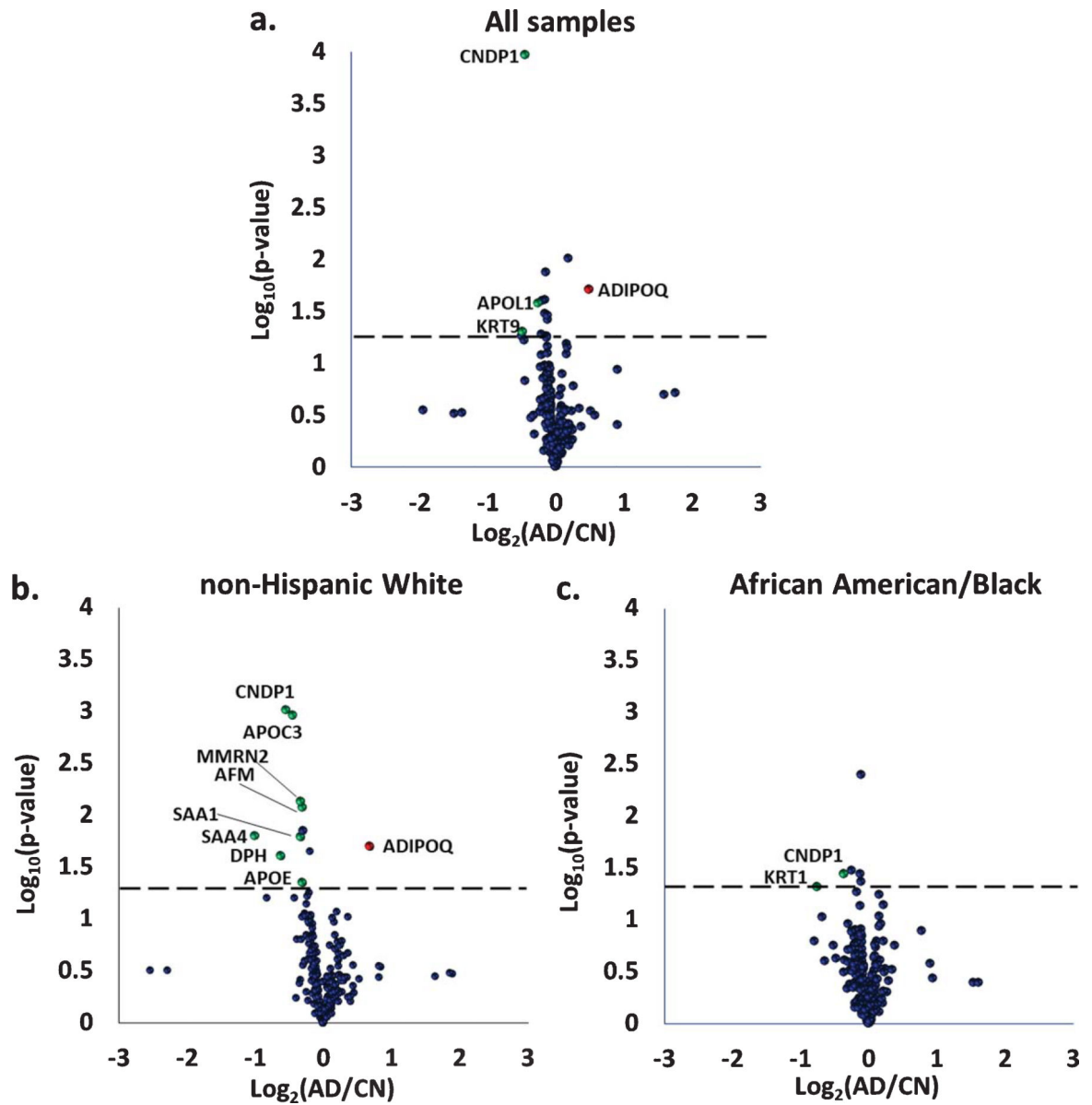
Fig. 3.
Volcano plots of differentially-expressed proteins between Alzheimer's disease (AD) and cognitively normal individuals (CN) for the entire set of samples in a) Set 1 ($N = 39$ AD, $N = 34$ CN); b) data from the non-Hispanic White group only, Set 1 ($N = 19$ AD, $N = 18$ CN); and c) data from the African American/Black group only, Set 1 ($N = 20$ AD, $N = 16$ CN). Red circles coincide with proteins higher in AD compared to CN, while green circles coincide with proteins lower in AD. CNDP1, Beta-Ala-His dipeptidase; KRT9, Keratin type I cytoskeletal 9; APOL1, Apolipoprotein L1; ADIPOQ, Adiponectin; KRT1, Keratin type II cytoskeletal 1; APOC3, Apolipoprotein C3; MMRN2, Multimerin-2; AFM, Afamin; SAA1, Serum amyloid A-1 protein; SAA4, Serum amyloid A-4 protein; DPH, Dopamine beta-hydroxylase; APOE, Apolipoprotein E.
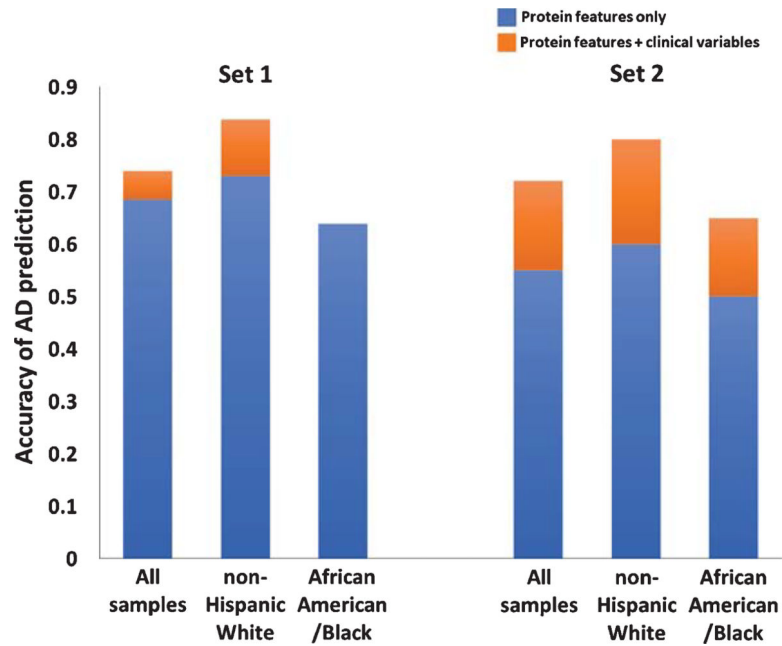
Fig. 4.
Histogram displaying classification accuracy for predicting AD in Set 1: $N = 73$ samples and Set 2: $N = 40$ samples. Blue bars: Accuracy determined when only the four differentially expressed proteins (beta-ala-his dipeptidase, keratin type I cytoskeletal 9, apolipoprotein L1, and adiponectin) are included in the model. Orange bars: Additional improvement in accuracy when clinical variables (age, sex, education, and *APOE*) are also included in the model.

Table 1

Demographics of plasma sample cohorts

	African American/Black		non-Hispanic White		p*
	Cognitively normal	Alzheimer's disease	Cognitively normal	Alzheimer's disease	
Sample size	26	30	28	29	
Sex (Male/Female)	8 / 18	8 / 22	9 /19	12 / 17	0.69
Mean age at draw (SD)	72.5 (7.4)	74.5 (7.8)	71.1 (9.1)	76.2 (8.7)	0.58
Mean year of most current blood draw (SD)	2006.4 (7.4)	2007.4 (6.6)	2008.8 (6.6)	2007.2 (6.0)	0.55
Mean years of education (SD)	13.3 (2.0)	13.4 (3.9)	16.1 (2.4)	13.2 (2.6)	0.0004
MMSE mean (SD)	27.1 (2.4)	14.5 (7.1)	27.8 (4.5)	13.6 (5.9)	$1.94E^{-22}$
Diabetes[a]	N = 15	N = 18	N = 19	N = 13	0.22
Hypercholesterolemia[a]	N = 15	N = 18	N = 19	N = 13	0.65
Hypertension[a]	N = 15	N = 18	N = 19	N = 13	0.046
APOE genotype					
2/3	4	4	5	1	
2/4	2	0	0	0	
3/3	14	10	18	11	
3/4	5	9	4	15	
4/4	1	7	1	2	

[a]Presence of comorbidity prior to or during blood draw.

MMSE, Mini-Mental State Exam; APOE, Apolipoprotein E.

Table 2

List of differentially-expressed proteins in race-stratified groups (Set 1)

Accession	Protein name	African American		Non-Hispanic White	
		Fold change[#]	p[*]	Fold change[#]	p[*]
Q96KN2	Beta-Ala-His dipeptidase	**0.78**	**0.04**	**0.69**	**9.78E-4**
P04264	Keratin type II cytoskeletal 1	**0.59**	**0.05**	0.85	0.5
Q15848	Adiponectin	1.19	0.3	**1.61**	**0.02**
P43652	Afamin	0.98	0.77	**0.81**	**0.01**
P02656	Apolipoprotein C3	1.04	0.72	**0.74**	**1.1E-3**
P02649	Apolipoprotein E	1.02	0.93	**0.81**	**0.05**
P09172	Dopamine beta-hydroxylase	1.17	0.32	**0.69**	**0.03**
Q9H8L6	Multimerin-2	1.02	0.86	**0.8**	**0.01**
P0DJI8	Serum amyloid A-1 protein	1.06	0.78	**0.5**	**0.02**
P35542	Serum amyloid A-4 protein	1.01	0.86	**0.8**	**0.02**

[#]Fold change cut off 1.23

[*]p < 0.05.

Bold denotes proteins that meet both fold-change cut-off and p-value<0.05.

Table 3

Summary of results from machine learning

| | Non-Hispanic White (NHW)* | | | | African American/Black (AA)* | | | |
| | No extra variables | | Age/sex/edu/APOE | | No extra variables | | Age/sex/edu/APOE | |
Set 1	**NHW**	**AA**	**NHW**	**AA**	**NHW**	**AA**	**NHW**	**AA**
AUC	0.91	0.49	0.97	0.61	0.83	0.59	0.91	0.47
Accuracy	86%	47%	86%	56%	73%	67%	84%	47%
Set 2	**NHW**	**AA**	**NHW**	**AA**	**NHW**	**AA**	**NHW**	**AA**
AUC	0.73	0.31	0.94	0.43	0.63	0.84	0.40	0.56
Accuracy	65%	40%	85%	45%	55%	70%	45%	50%

*
Differentially-expressed proteins from Set 1 were used for the SVM classification.

The full list of differentially-expressed proteins is provided in Table 2. Set 1: $N = 73$ samples, Set 2: $N = 40$ samples. Edu, education; *APOE*, Apolipoprotein E4 genotype; AUC; area under the curve.