Title

A Peptide-Centric Quantitative Proteomics Dataset for the Phenotypic Assessment of Alzheimer's Disease

Authors

Gennifer E. Merrihew*[1], Jea Park*[1], Deanna Plubell*[1], Brian C. Searle[2], C. Dirk Keene[3], Eric B. Larsen[4], Randall Bateman[5], Richard J. Perrin[6], Jasmeer P. Chhatwal[7], Martin R. Farlow[8], Catriona A. McLean[9], Bernardino Ghetti[10], Kathy L. Newell[10], Matthew P. Frosch[11], Thomas J. Montine[12], and Michael J. MacCoss[1]

*Equal first authorship
1. Department of Genome Sciences, University of Washington, Seattle, Washington 98195, United States
2. Department of Biomedical Informatics, Ohio State University, Columbus, Ohio 43210, United States
3. Department of Laboratory Medicine and Pathology, University of Washington, Seattle, Washington 98195, United States
4. Department of Medicine, University of Washington, Seattle, Washington 98195, United States
5. Department of Neurology, Washington University School of Medicine, 660 South Euclid Avenue, Box 8111, St. Louis, Missouri 63110, United States
6. Department of Pathology and Immunology, Washington University School of Medicine, 660 South Euclid Avenue, Box 8111, St. Louis, Missouri 63110, United States
7. Massachusetts General Hospital, Department of Neurology, Harvard Medical School, 15 Parkman St, Suite 835, Boston Massachusetts 02114, United States
8. Department of Neurology, Indiana University School of Medicine, Indianapolis, Indiana 46202, United States
9. Department of Anatomical Pathology, Alfred Health, Melbourne VIC 3004, Australia
10. Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, Indiana 46202, United States
11. C.S. Kubik Laboratory for Neuropathology, and Massachusetts Alzheimer Disease Research Center, Massachusetts General Hospital, Boston, Massachusetts 02114, United States
12. Department of Pathology, Stanford University, Stanford, CA, 94305, United States
corresponding author(s): Michael J. MacCoss (maccoss@uw.edu) and Thomas J. Montine (tmontine@stanford.edu)

Abstract

Alzheimer's disease (AD) is a looming public health disaster with limited interventions. Progress in therapeutic development requires a more detailed understanding of molecular pathogenesis. Most data for AD pathogenesis in humans is from standard neuropathologic assessments, e.g., neuritic plaques and neurofibrillary tangles or biochemical measurement of a limited number of analytes related to neuropathologic features, e.g., amyloid-β peptides and hyperphosphorylated tau. Standard neuropathologic evaluation is highly valuable as it provides assessment of diseases that afflict an individual brain, but it also is limited in its ability to investigate molecular pathogenesis. Modern mass spectrometry-based proteomics enable quantitative insight into the protein phenotype in brains with AD neuropathology. Careful selection of specimens enabled us to investigate protein signatures specific to autosomal dominant AD dementia (ADD), sporadic ADD with minimal comorbidities, individuals without dementia who had high histopathologic burden of AD, and cognitively normal individuals with no or minimal AD histopathologic burden. All data are deposited in the ProteomeXchange proteomic data repositories (ID: PXD034525).

Background & Summary

Alzheimer's disease (AD) is a major global public health problem. In the US, AD is the seventh most common cause of death for all ages and sexes. In contrast to ischemic heart disease, stroke and several forms of cancer, AD is increasing as a cause of death, of years lived with disability, and of disability-adjusted life years[1]. Success in limiting acute illnesses in the developing world is shifting the burden to non-communicable diseases, with an expected dramatic rise in AD globally by 2025[2]. Existence of forms of AD with known genetic causes or risk, and forms without known genetic underpinnings, highlight the potential for multiple molecular drivers and perhaps multiple pathogenic pathways involved in disease onset and progression. Moreover, longitudinal population-based cohort studies have repeatedly observed that AD is commonly comorbid with pathologic changes of vascular brain injury (VBI), Lewy body disease (LBD), limbic-associated TDP-43 encephalopathy (LATE), and/or hippocampal sclerosis[3,4]. AD is a chronic illness whose ultimate clinical expression as dementia follows years, if not decades, of injury, response to injury, consumption of reserve, and exhaustion of compensation. Determining the molecular profile of its various

forms independent of comorbidities will be fundamental to efforts to develop tailored therapies that specifically target the molecular mechanism(s) of AD.

Thus, a primary focus of this data resource was the selection of tissue specimens based on current guidelines for AD neuropathologic change (ADNC)[3,4] and specific exclusion for the presence of alternative potential causes of dementia, resulting in the carefully annotated examples of AD and controls free of medically significant comorbidities. Additionally, the specimens selected spanned the range of disease severity from the categories designated as high cognitive function (HCF) with no or low ADNC, HCF with intermediate or high ADNC, sporadic AD dementia (ADD, with intermediate or high ADNC), and autosomal dominant ADD (with causal mutations in *PSEN1*, *PSEN2*, or *APP,* and high ADNC). All research participants whose brains were used for this study underwent detailed, research quality, longitudinal cognitive assessments. For individuals without dementia, all had their last evaluation within 2 years of death and had neuropsychological test results in the upper quartile of the cohort to minimize interval conversion. All samples were obtained using a rapid autopsy protocol with postmortem interval less than 8 hours (except autosomal dominant AD due to practical limitations), flash frozen in liquid nitrogen, and kept frozen at -80°C prior to analysis. All samples were matched for age and sex except for those from individuals with autosomal dominant AD who experienced earlier onset.

Due to this careful specimen selection, we now have a unique sample set that can be used to study the molecular underpinnings of autosomal dominant ADD, sporadic ADD, and high burden ADNC with HCF without the confounding comorbidities faced in similar molecular profiling experiments. While previous studies have investigated the molecular profile of AD without excluding comorbidities, the high prevalence of these diseases, each of which can cause dementia on its own when present at high level, likely underlines the specificity of these profiles for AD. Likewise, because autosomal dominant AD is rare, typical molecular profiling studies have focused only on individuals with sporadic ADD, thereby limiting perspective on this heterogeneous disease. These points emphasize the uniqueness of this proteomics dataset for a more comprehensive assessment of the different forms of AD.

To make the most of this unique sample cohort, we used mass spectrometry proteomics methods that give reproducible and highly quantitative data. Most of the large-scale proteomics experiments studying the human brain have been performed using a mass spectrometry approach known as data-dependent acquisition (DDA). This approach is extremely powerful for building lists of proteins present in brain tissue and has been useful when combined with tandem mass tags for modest numbers of samples that can be performed within a “plex”. However, irregular sampling by DDA makes it challenging to provide robust and quantitative measurements across more samples than can fit in a plex. When using DDA, the number of peptides sampled is limited by the MS/MS sampling speed despite the dynamic range and peak capacity of the mass analyzer. A single MS spectrum can contain over one hundred different molecular species, of which only a handful are analyzed by MS/MS prior to the next full scan[5]. This general approach has become extremely powerful for cataloging proteins and modifications, but its irregular sampling results in missing data, requires extensive fractionation to sample low abundance peptides, and results in variable peptide sampling between runs of the same sample. Recently, Lamond, et al., assessed the missing values from performing quantitative proteomics using 10-plex tandem mass tags (TMT) across multiple batches[6]. They showed that by integrating data from different 10-plex TMT batches that >50% of the peptides had missing values – making quantitation on the peptide level between batches challenging to impossible. On the protein level, they showed that after 5 different 10-plex TMT batches, the median number of missing proteins was reduced to ~10% – but required using different peptides to bridge the batches for quantification. While 10% missing protein values is an improvement over 50% missing peptides, the approach constrained the use of TMT to reporting on protein quantities, losing important resolution known to occur in AD that can only be assessed on the peptide level[7].

An alternative to DDA is an acquisition approach known as data independent acquisition (DIA) that acquires comprehensive MS/MS information in a single LC-MS/MS run using a repeated cycle of wide-window MS/MS scans. The computational analysis of DIA spectra can be performed in the same “targeted” manner as fully targeted data; i.e., fragment ion chromatograms for each peptide can be extracted and used for quantification. However, unlike fully targeted data acquisition, DIA analysis can be done for any peptide in the sampled range (e.g., between 400 and 1000 *m/z*), rather than just for a subset of pre-specified peptides. Thus, the reproducible targeting and confident MS/MS-based quantification of parallel reaction monitoring (PRM) can be combined with DDA’s ability to detect and measure thousands of proteins.

Typically, tens to hundreds of biological samples are processed and analyzed using LC-MS/MS in quantitative proteomics experiments. The regularity of DIA enables researchers to make peptide detections in one sample and use that information to inform the detection of the same peptides in other samples. DIA offers four key improvements over DDA. 1) Because peptides are sampled systematically, more peptides are detected in a DIA analysis than a DDA analysis in an equivalent length analysis. 2) The same precursor m/z range is sampled, at the same RT, in all runs – eliminating the issues associated with stochastically sampled DDA data. 3) DIA analysis can make use of previously measured information to improve peptide measurement (e.g. known retention time, known fragmentation patterns, and which peptides provide stable and precise quantitative measurements). 4) Peptide detection can be assessed directly from DIA data, simplifying downstream analysis. These data provide an archive of all detectable molecular species within the measured mass range of the instrument. This methodology benefits from the reproducible and comprehensive sampling of the latest DIA methodology with an innovative approach used to improve peptide precursor selectivity[5]. The combination of the unique specimens with systematically collected mass spectrometry data creates a resource for the scientific community to test new hypotheses about the molecular features of different forms of AD dementia.

Methods

Human Brain Samples.

Brain tissue samples were stratified into 4 groups based on clinical, pathological and genetic data and four brain regions (superior and middle temporal gyri or SMTG, hippocampus at the level of the lateral geniculate nucleus, inferior parietal lobule or IPL and caudate nucleus at the level of the anterior commissure). Cognitive status was determined as dementia or not dementia by DSM-IVR criteria. Individuals diagnosed as not dementia were from the Adult Changes in Thought (ACT) study and were included only if the last research evaluation was within 2 years of death and the last cognitive testing score using the cognitive abilities screening instrument (CASI) was in the upper quartile for the ACT cohort (>90); our definition of HCF. Brains from individuals with HCF who had no or low ADNC were designated “HCF/low ADNC” and those with intermediate or high ADNC were designated “HCF/high ADNC''. All individuals diagnosed with ADD had intermediate or high level ADNC and were further subclassified as sporadic (“Sporadic ADD”) or ADD caused by a mutation in *PSEN1, PSEN2,* or *APP* (“Autosomal Dominant ADD”). Sporadic AD cases were from the ACT study and the University of Washington (UW) AD Research Center (ADRC), and Autosomal Dominant ADD cases were from the UW ADRC and the Dominantly Inherited Alzheimer Network (DIAN). Excluded was any case with LBD or LATE-NC other than involving amygdala, territorial infarcts, more than 2 cerebral microinfarcts, or hippocampal sclerosis. Time from death to cryopreservation of tissue, postmortem interval (PMI), was <8 hr in all cases except for those in the Autosomal Dominant ADD group. Details of sample stratification for the four brain regions (SMTG, Hippocampus, IPL and Caudate) are provided in Tables 1-4.

Ethics oversight

All study cohort participants were collected and provided informed consent under protocols approved by the Institutional Review Board (IRB) at University of Washington, Kaiser Permanente Washington, and Stanford University.

Sample Metadata, Batch Design and References

Each human brain region was divided into batches of 14 individual samples and 2 pooled references for a total of 16. The first batch of each region was also used to create a region-specific reference pool to be used as a “common reference” and/or single point calibrant, which was homogenized, aliquoted, frozen, and used to compare between batches within a brain region. Human cerebellum and occipital lobe tissue was homogenized, pooled, aliquoted and frozen to be used as a “batch reference” for comparison between batches and other brain regions. Batch design was randomly balanced based on group ratios. For example, batches from the SMTG brain region contained 5 “Sporadic ADD”, 4 “Autosomal Dominant ADD”, 2 “HCF/low ADNC”, and 3 “HCF/high ADNC” samples (Supplementary Table 1). Metadata for the samples from the Hippocampus, IPL and Caudate brain regions is provided in Supplementary Tables 2-4.

Sample Homogenization and Protein Digestion

Two 25 µm frozen sections of brain tissue were resuspended in 120 µl of lysis buffer of 5% SDS, 50mM triethylammonium bicarbonate (TEAB), 2mM MgCl2, 1X HALT phosphatase and protease inhibitors, vortexed and briefly sonicated at setting 3 for 10 s with a Fisher sonic dismembrator model 100. A microtube was loaded with 30 µl of lysate and capped with a micropestle for homogenization with a Barocycler 2320EXT (Pressure Biosciences Inc.) for a total of 20 minutes at 35°C with 30 cycles of 20 seconds at 45,000 psi followed by 10 seconds at atmospheric pressure.

Protein concentration was measured with a BCA assay. Homogenate of 50 µg was added to a process control of 800 ng of yeast enolase protein (Sigma) which was then reduced with 20 mM DTT and alkylated with 40 mM IAA. Lysates were then prepared for S-trap column (Protifi) binding by the addition of 1.2% phosphoric acid and 350 µl of binding buffer (90% Methanol, 100 mM TEAB). The acidified lysate was bound to column incrementally, followed by 3 wash steps with binding buffer to remove SDS and 3 wash steps with 50:50 methanol:chloroform to remove lipids and a final wash step with binding buffer. Trypsin (1:10) in 50mM TEAB was then added to the S-trap column for digestion at 47°C for one hour. Hydrophilic peptides were then eluted with 50 mM TEAB and hydrophobic peptides were eluted with a solution of 50% acetonitrile in 0.2% formic acid. Elutions were pooled, speed vacuumed and resuspended in 0.1% formic acid.

Injection of samples are one ug of total protein (16 ng of enolase process control) and 150 fmol of a heavy labeled Peptide Retention Time Calibrant (PRTC) mixture (Pierce). The PRTC is used as a peptide process control. Library pools are an equivalent amount of every sample (including references) in the batch. For example, a batch library pool consists of the 14 samples from the batch and two references. System suitability (QC) injections are 150 fmol of PRTC and BSA.

Liquid Chromatography and Mass Spectrometry

One µg of each sample with 150 femtomole of PRTC was loaded onto a 30 cm fused silica picofrit (New Objective) 75 µm column and 3.5 cm 150 µm fused silica Kasil1 (PQ Corporation) frit trap loaded with 3 µm Reprosil-Pur C18 (Dr. Maisch) reverse-phase resin analyzed with a Thermo Easy-nLC 1200. The PRTC mixture is used to assess system suitability before and during analysis. Four of these system suitability runs are analyzed prior to any sample analysis and then after every six sample runs another system suitability run is analyzed. Buffer A was 0.1% formic acid in water and buffer B was 0.1% formic acid in 80% acetonitrile. The 40-minute system suitability gradient consists of a 0 to 16% B in 5 minutes, 16 to 35% B in 20 minutes, 35 to 75% B in 1 minute, 75 to 100% B in 5 minutes, followed by a wash of 9 minutes and a 30-minute column equilibration. The 110-minute sample LC gradient consists of a 2 to 7% for 1 minutes, 7 to 14% B in 35 minutes, 14 to 40% B in 55 minutes, 40 to 60% B in 5 minutes, 60 to 98% B in 5 minutes, followed by a 9 minute wash and a 30-minute column equilibration. Peptides were eluted from the column with a 50°C heated source (CorSolutions) and electrosprayed into a Thermo Orbitrap Fusion Lumos Mass Spectrometer with the application of a distal 3 kV spray voltage. For the system suitability analysis, a cycle of one 120,000 resolution full-scan mass spectrum (350-2000 *m/z*) followed by a data-independent MS/MS spectra on the loop count of 76 data-independent MS/MS spectra using an inclusion list at 15,000 resolution, AGC target of 4e5, 20 millisecond (ms) maximum injection time, 33% normalized collision energy with a 8 *m/z* isolation window. For the sample digest, first a chromatogram library of 6 independent injections is analyzed from a pool of all samples within a batch. For each injection a cycle of one 120,000 resolution full-scan mass spectrum with a mass range of 100 *m/z* (400-500 *m/z*, 500-600 *m/z*, 600-700 *m/z*, 700-800 *m/z*, 800-900 *m/z*, 900-1000 *m/z*) followed by a data-independent MS/MS spectra on the loop count of 26 at 30,000 resolution, AGC target of 4e5, 60 ms maximum injection time, 33% normalized collision energy with a 4 *m/z* overlapping isolation window. The chromatogram library data is used to quantify proteins from individual sample runs. These individual runs consist of a cycle of one 120,000 resolution full-scan mass spectrum with a mass range of 350-2000 *m/z*, AGC target of 4e5, 100 ms maximum injection time followed by a data-independent MS/MS spectra on the loop count of 76 at 15,000 resolution, AGC target of 4e5, 20 ms maximum injection time, 33% normalized collision energy with an overlapping 8 *m/z* isolation window. Application of the mass spectrometer and LC solvent gradients are controlled by the ThermoFisher Xcalibur (version 3.1.2412.24) data system. Mass spectrometry run order for all samples is provided in Supplementary Tables 5-8.

Peptide Detection and Quantitative Signal Processing

Thermo RAW files were converted to mzML format using Proteowizard (version 3.0.20064) using vendor peak picking and demultiplexing with the settings of "overlap_only" and Mass Error = 10.0 ppm[5]. On column chromatogram libraries were created using the data from the six gas phase fractionated "narrow window" DIA runs of the pooled reference as described previously[8]. These narrow windows were analyzed using EncyclopeDIA (version 1.4.10) with the default settings (10 ppm tolerances, trypsin digestion, HCD b- and y-ions) of a Prosit predicted spectra library based the Uniprot human canonical FASTA (January 2021). The results from this analysis from each brain region were saved as a "Chromatogram Library'' in EncyclopeDIA's eLib format where the predicted intensities and iRT of the Prosit library were replaced with the empirically measured intensities and RT from the gas phase fractionated LC-MS/MS data. The "wide window" DIA runs were analyzed using EncyclopeDIA (version 1.4.10) requiring a minimum of 3 quantitative ions and filtering peptides with q-value ≤ 0.01 using Percolator 3.01. After analyzing each file individually, EncyclopeDIA was used to generate a "Quant Report'' which stores all the detected peptides, integration boundaries, quantitative

transitions, and statistical metrics from all runs in an eLib format. The Quant Report eLib library is imported into Skyline (daily version 22.2.1.278) with the human uniprot FASTA as the background proteome to map peptides to proteins, perform peak integration, manual evaluation, and report generation. A csv file of peptide level total area fragments (TAFs) for each replicate was exported from Skyline using the custom reporting capabilities of the document grid[9].

Quantitative Data Post-Processing, Normalization, and Batch Correction

Despite precautions taken to ensure equivalent sample preparation, handling and acquisition, additional post-processing was performed to normalize, and batch adjust the quantitative data to remove residual technical noise. Modeling the proportional changes of peptide/protein group intensities, $\log_2$ transformation is applied followed by a Median Deviation (MD) normalization to the peptide total area fragment values (level 2 data) across instrument runs within a brain region (Equation 1) under the assumption that median total area fragment values should be equal sans batch effect from known and unknown sources of variability.

MD Normalization, by calculating the deviation from the median of the sample total area fragment, should neither remove scale information nor de-weigh outlier signals that may be of biological relevance[10]. Here, the MD normalized peptide *F* of each sample is given by the following. The peak areas (A_i) for each peptide *i* are first $\log_2$ transformed and then normalized by equalizing the median peak areas across all samples using the equation:

Equation 1. Fi=log2Ai -[log2Am -log2Aj].

Where: A_i = sum of product ion transition area for peptide *i*

Am = median of areas within LC-MS run *m*

Aj = median of areas between the LC-MS runs.

The effectiveness and validity of the normalization approach is then assessed by evaluating the comparability of the peptide abundance distribution across samples (Figure 2B), and by the reproducibility of those peptide abundances across replicate samples (Figure 3A). Peptide abundances are then adjusted for batch effect by fitting a linear model and "regressing" out the factors with known unwanted sources of variation to return a matrix of residuals. The detection of the presence of batch effect pre- and post-adjustment is assessed by exploring the data variance structure through Principal Variance Component Analysis (PVCA)[11] (Supplementary Figure 4) and Principal component Analysis (PCA) using projections onto the first three principal components. The normalization and batch adjusted peptide abundances are available as the level 3A data file. Using DIA, all observable peptides in one sample will be sampled in all of the other biological replicates[12,13,14]. Due to the comprehensive sampling nature of DIA we can extract information for the same transitions across all samples in an experiment. The resulting zeros in our peptide abundance data therefore represent signals below our limit of detection and are not treated as missing data. After protein group inference, protein abundances are batch corrected using the same method as the peptide data.

Protein Grouping and Inference

The processing and 'roll-up' of DIA data borrows from the established strategies adopted in the DDA field in which the quantification of peptides and their corresponding protein groups is inferred through the modification of IDPicker algorithm[15]. In summary, to quantify the peptide/protein groups, a bipartite graph of peptide-protein interactions is constructed to generate groupings through the parsimony reduction of the graph as it is implemented in MSDaPl[16]. Then, the peptide abundances at the nodes are summed to estimate the abundance of the peptide groups and proteins that match the same set of peptide groups are merged into a single node in the graph, forming an indistinguishable protein group.

Data Records

The Skyline documents, raw files for quality control and DIA data are available at Panorama Public. ProteomeXchange ID: PXD034525. Access URL: https://panoramaweb.org/ADBrainCleanDiagDIA.url.

DIA data is available in 5 different categories based on the level of post-processing (Figure 1e) for each brain region. Level 0 represents the raw data in two different formats - the raw format is directly from the Thermo mass spectrometer and the mzML format is the demultiplexed version of the raw data (Proteowizard version 3.0.20064). Level 1 describes the zipped Skyline document grouped by batch. Level 2 is a csv file grouped by batch of the Skyline output with the integrated peak area for each peptide (row) in each replicate (column). Level 3A is a csv file of the normalized peptide abundance across all batches. Level 3B is a csv file of the normalized protein abundance across all batches.

Quality control Skyline documents, peptide QC plots and instrument raw files for system suitability runs are provided by brain region. The Skyline documents and peptide QC plots for enolase and PRTC process controls are provided by brain region. The instrument raw files for process controls are the same as DIA sample raw files by brain region.

Technical Validation

Balanced and Controlled Experiment Design

We have designed our experiment to perform quantitative, peptide-centric proteomics using brain tissue from four different brain regions selected specifically because they represent distinct anatomical regions with varying pathological involvement by AD (Figure 1). The experimental design was intended to compare individual samples from the four different categorical disease groups within each brain region. Samples were prepared in batches of 16 samples which consisted of 14 brain tissue samples and two external control samples. The batch size was determined by the number of samples, 16, that could be prepared within a Barocycler (Pressure Biosciences, Inc.). For each batch, the samples were randomized in a balanced block design (Supplementary Table 1).

Within each batch we included both internal and external controls. Internal controls were added to each sample to provide a QC check of the sample preparation and LC-MS data collection process. These "Process Controls" consisted of the addition of yeast enolase protein after lysis and prior to digestion and the Pierce Retention Time Calibration (PRTC; 15 synthetic stable isotope labeled peptides) peptide mixture following digestion. The "Protein Internal Control" was used to assess the protein digest and peptide recovery and the "Peptide Internal Control" was used to distinguish between sample preparation and measurement issues post-digestion.

The two external controls were different brain lysates that were prepared, measured, and analyzed with the rest of the samples in the batch. One of the controls was a brain region specific pool used to assess between batch quality control. This inter-batch quality control is composed of a randomized balanced pooled sample set for each respective brain region. For example, the inter-batch quality control "TRPR" is composed of 3 HCF/high ADNC samples, 3 HCF/low ADNC samples, 3 AutoDom ADD samples and 5 Sporadic ADD samples from the SMTG. The same inter-batch quality control is run in every batch of the experiment from the SMTG and was used to assess data quality post-normalization. The second control was an inter-brain region quality control ("HAD" samples) and composed of a homogenate of a mix of cerebellum and occipital lobe tissue and the same pool was prepared and run in every batch across all brain regions.

We can determine when our sample preparation and system is not functioning as expected with a combination of system suitability checks, inter-batch quality controls, inter-experiment quality controls and process controls (Figure 1b). Our system suitability consists of a mixture of a BSA tryptic digest and PRTC prior to sample analysis and throughout sample collection at a frequency of once every six to eight samples.

Run Level and Experiment Level Peptide and Protein Detections

For each sample in each brain region several tryptic peptides can be detected at a 1% FDR cut-off. IPL samples ranged from 37840-73168, SMTG ranged from 51582-69590, hippocampus from 32995-59853, and caudate nucleus ranged from 31426-58105 (Table 5). To integrate data across all individual samples within each brain region we control with an experiment level error rate. This leads to the same peptides quantified in all samples within a brain region; 48271 in IPL, 40346 in SMTG, 31863 in hippocampus, and 26135 in caudate nucleus. These peptides map to 6497 quantified proteins in IPL, 5851 in SMTG, 5117 in hippocampus, and 4636 in caudate nucleus (Table 5). The distribution of peptide abundances is aligned with median normalization, as demonstrated with the SMTG data (Figure 2B) as well as all brain regions (Supplementary Figures 1 and 2).

Inter-batch Precision and Reproducibility

The inclusion of inter-batch quality control samples allows us to assess the impact of normalization and batch correction on peptide and protein quantitative reproducibility. For example, the SMTG experiment was split into 5 batches for processing and acquisition. Using the inter-batch control replicate samples, the coefficient of variation can be calculated for all peptides quantified in the SMTG. The distribution of peptide coefficient of variation improves with normalization and batch correction, with the mean decreasing about 8.2%. Likewise, the protein coefficient of variation also improves following batch correction, with the mean decreasing by about 1.25% (Figure 3). Peptide and protein quantities are highly correlated across inter-batch replicates. Inter-batch control replicate samples in SMTG

have peptide Pearson correlation coefficients ranging from 0.867 to 0.950, and protein correlations ranging from 0.894 to 0.960 (Figure 4).

Expected biological differences

Of the peptides quantified, we detect several proteins known to be involved in AD. In the SMTG data we quantify peptides mapping to 250 proteins found in AD-related pathways (Supplemental Figure 5). Preliminary assessment of peptide and protein data show we can distinguish between sample groups (Supplemental Figure 3). Differential abundance analysis between HCF/low ADNC and autosomal dominant ADD in SMTG captures known biology. Protein groups found to be significantly different between the groups include previously documented increases in Amyloid precursor protein (APP/A4), Apolipoprotein E (APOE), SPARC-related modular calcium-binding protein 1 (SMOC1), midkine (MK/MDK) and netrin-1 (NET1/NTN1)[17]. We detect and quantify two peptides mapping to the n- and c-terminal sides of the alpha-secretase cleavage site in amyloid-β sequence. Both peptides have an expected difference in abundance across the sample groups in all four brain regions, with autosomal dominant ADD having the highest distribution, followed by sporadic ADD, HCF/high ADNC, and HCF/low ADNC. Peptides mapping to microtubule-associated protein tau (MAPT) also have some expected differences between the experimental groups. In the SMTG brain region the seven quantified peptides spanning the microtubule binding region of MAPT (residues 243-368) are increased in autosomal dominant ADD, followed by sporadic AD, compared to both HCF groups. This trend has been observed previously and leads to the aggregated protein abundance to be differential in the same manner as the microtubule binding region peptides[18].

Usage Notes

We provide quantitative data both for tryptic peptides directly and for protein groups derived from peptide quantities. Based on extensive existing knowledge of AD, we know that modified forms of the APP or amyloid-β and Tau proteins are important in disease progression. Historically mass spectrometry proteomics data has aggregated multiple tryptic peptides per protein coding sequence to arrive at a singular protein level value. This would result in a loss of important information regarding the status of the individual peptides[7]. By collecting these samples by DIA we can reproducibly quantify individual tryptic peptides across our entire sample set. With this peptide data we observe differentially abundant peptides within a protein coding gene. If aggregated to a singular value this important signal would be lost.

Use Case 1. Sporadic and Autosomal dominant Alzheimer disease dementia differential peptide and protein abundance.

This mass spectrometry data has typically been reported as relative protein group abundance measures, enabling differential abundance testing of protein groups between disease states (Figure 6). In SMTG there are proteins with statistically different abundance between healthy controls and AD. This type of analysis can be extended to all brain regions. A large portion of research investigating the molecular and pathological basis of AD has been generated using model systems informed by genetic causes. Studying individuals with sporadic ADD can be challenging due to the presence of comorbidities. Here we present data generated from both autosomal dominant ADD with minimal comorbidities and sporadic ADD with minimal comorbidities. These data can be used to better understand biological differences or similarities in these two types of ADD.

Use Case 2. Differential peptide and protein abundance with neuropathologic markers and dementia.

The HCF/high ADNC samples have classic histopathologic features of AD - amyloid-β plaques and neurofibrillary tangles - at levels overlapping with individuals who have Sporadic ADD. This enables analyses to measure the differences in protein pathology between individuals who have developed dementia and those who have not developed dementia despite similar levels of high amyloid-β plaques and Tau tangles.

Use Case 3. Mass spectrometry data reuse and reanalysis.

Beyond the analyses presented here, the storage of this data is done in a way that facilitates use of this data for informing subsequent mass spectrometry assay development. All data from this project is available on the Panorama server-based data repository application for targeted mass spectrometry assays[19]. Thus, all extracted ion chromatogram information from every sample across all brain regions is available in an interactive Skyline document format, and readily available for download and reuse. Information about fragment ions and chromatography is important for the development of targeted assays[20,21], making this data valuable to the wider community.

Using DIA mass spectrometry methods allows for the data to be easily reanalyzed. This could be used to search for post-translational modifications or sequence variants not searched for in our current analysis. Additionally, this data can be reanalyzed to look at other unique features, such as isomerized peptides. The feasibility of this was recently shown by Hubbard et al. through reanalysis of a subset of this dataset[22]. Peptide-centric reanalysis is possible due to the comprehensive sampling by data-independent acquisition of a preset range across all samples.

Code Availability

The MSConvert installer and documentation is available from https://proteowizard.sourceforge.io/. EncyclopeDIA is available at https://bitbucket.org/searleb/encyclopedia/. The Skyline-daily installer and documentation are available from https://skyline.ms/skyline.url. The source code for both the MSConvert and Skyline projects are available as part of the Proteowizard project https://github.com/ProteoWizard/pwiz

All code used for the analysis of the data matrix including data QC, normalization, pre-processing, visualization, and figure generation is available online at https://github.com/uw-maccosslab/ADBrainCleanDiagDIA.

Acknowledgements

Support for this research was provided by the National Institute of Aging (RF1 AG053959, U19 AG065156, F31 AG069420, R01 AG075802). The collection of tissue samples were supported by the UW Alzheimer's Disease Research Center (P30 AG066509), the Adult Changes in Thought study U19 AG066567, the Nancy and Buster Alvord Endowment (C.D.K.), the Massachusetts Alzheimer Disease Research Center (P30 AG062421), The Dominantly Inherited Alzheimer Network (U19 AG032438, UF1AG032438), the German Center for Neurodegenerative Diseases (DZNE), Raul Carrea Institute for Neurological Research (FLENI), the Research and Development Grants for Dementia from Japan Agency for Medical Research and Development, AMED, and the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI). This manuscript has been reviewed by DIAN Study investigators for scientific content and consistency of data interpretation with previous DIAN Study publications. We acknowledge the altruism of the participants and their families and contributions of the research and support staff at each of the participating sites for their contributions to this study.

Author contributions

G.E.M., M.J.M, and T.J.M designed the experiment. G.E.M. prepared the brain tissue, ran the mass spectrometer, performed the bioinformatic signal processing, and organized the data repository. J.P. performed data quality control, normalization, batch correction and statistics. J.P and D.P. produced the figures and interpreted the results. B.C.S. designed mass spectrometry methods and software. C.D.K., E.B.L, R.B., J.P.C, M.R.F, C.A.M, B.G., K.L.N. and M.P.F. provided specimens. G.E.M., J.P., D.P., T.J.M. and M.J.M. wrote the paper. All authors reviewed the manuscript.

Competing interests

The authors declare the following competing financial interest(s): The MacCoss Lab at the University of Washington has a sponsored research agreement with Thermo Fisher Scientific, the manufacturer of the instrumentation used in this research. M.J.M. is a paid consultant for Thermo Fisher Scientific.

Figure Legends

Figure 1: Experimental scheme for the collection of the proteomics data using data independent acquisition-mass spectrometry. A) Brain tissue sections from 4 regions were analyzed for all groups. B) Samples were prepared and analyzed in batches of 16, with 14 individual samples per batch selected by balanced randomization. Each batch contained an inter-experiment quality control sample generated from pooling portions of several individual samples from across all 4 brain regions sampled. Each batch also contained an inter-batch quality control sample generated from pooling portions of individual samples within that brain region. C) In addition to quality control samples, both protein and peptide sample processing controls were included in all samples to track system suitability. D) For each batch an on-column data-independent acquisition chromatogram library is generated from overlapping, narrow window gas-phase fractionation of an inter-batch QC. Individual samples are acquired by a single injection wider window data-independent acquisition method. Peptide detection and scoring is performed using EncyclopeDIA and extracted ion chromatograms integrated with Skyline. E) The proteomics data is publicly available on the Panorama web server in 5 different states, each corresponding to the level of post-processing.

Figure 2: Peptides detected in individual runs and across the entire dataset for SMTG. A) To make comparisons across individual runs we enforce an experiment level false discovery rate (FDR) cutoff. Although zeros exist in this remaining data, these represent a lack of signal above background. B) Kernel density plots and box plots for each individual show the distribution of peptide abundances before and after log2 transformation and median normalization. Condition groups are highlighted with different colors in the log2 transformation and median normalization box plots.

Figure 3: Effect of normalization and batch correction on the inter-batch variance for SMTG. A) The effect of median normalization and batch correction on the inter-batch peptide coefficient of variance. B) Effect of batch correction on the inter-batch protein coefficient of variance. The relationship between coefficient of variation and the log2 median abundance is visualized with a loess fit of a contoured density plot (red line).

Figure 4: Correlation of the quantitative results from five different sample preparation and analysis batches following normalization and batch correction for SMTG. A) Pearson correlation of the data between the batches on the peptide level. B) Pearson correlation of the same data on the protein group level between all batches. Log2 intensity for peptides and proteins.

Figure 5: Summary data of quantitative changes observed for SMTG. A) Volcano plot showing proteins that are statistically different between the autosomal dominant Alzheimer's disease (AD) dementia (ADD) and the high cognitive function (HCF)/low AD neuropathologic change (ADNC) SMTG. B) Protein groups with the greatest differences between the extreme autosomal dominant ADD and the HCF/low ADNC SMTG. C) Heatmap showing the clustering of 33 SMTG samples using the 115 proteins that displayed a significant difference between ADD and the HCF/low ADNC samples. Significance cutoffs are $p<0.05$, $logFC>0.5$.

Figure 6: Amyloid-β and tau peptide abundances recapitulate known molecular changes for the SMTG. A) Two tryptic peptides mapping to amyloid-β (6-16: HDSGYEVHHQK, 17-28: LVFFAEDVGSNK) are quantified across all four brain regions. B) Tryptic peptides mapping to tau demonstrate some common signatures within protein domains. These signatures are lost if aggregated to a protein level. Tryptic peptides are labeled based on their first and last amino acid residues. Mean + se of z-score.

Supplementary Figure 1. Raw peptide quantities from all brain regions and batch density. The top panel of box plots shows the distribution of log2 peptide abundances for each individual in each brain region. The bottom panel of density plots shows the distribution of log2 peptide abundances by batch in each brain region.

Supplementary Figure 2. Peptide quantities and batch density following MD Normalization. The top panel of box plots shows the distribution of log2 peptide abundances after MD normalization for each individual in each brain region. The bottom panel of density plots shows the distribution of log2 peptide abundances after MD normalization by batch in each brain region.

Supplementary Figure 3. Plot of first two principal component scores of the SMTG data set after controlling for batch factor. The scores are colored by biological replicate, donor age, experimental batch and condition group factors. Confidence ellipses are shown for cofactors of interest.

Supplementary Figure 4. Principal variance component analysis (PVCA) of protein group abundance from the SMTG samples. Estimated contribution of proportional variance by age, batch condition, and sex effects are shown. The confounding batch cofactor present in unadjusted data set (A) is "regressed out" using a simple linear regression model (B) prior to hypothesis testing.

Supplementary Figure 5: Known Alzheimer's disease proteins shown in a Kegg pathway. Proteins quantified in the SMTG experiments map to protein groups highlighted in red on pathway hsa05010.

Tables

Table 1. Brain donor characteristics for the superior and middle temporal gyri (SMTG).

	HCF/low ADNC	HCF/high ADNC	Sporadic ADD	Autosomal Dominant ADD

N	9	11	18	24
Age *(mean ± sd)*	88 ± 5	90 ± 5	82 ± 13	51 ± 11
Sex *(M:F)*	4:5	5:6	11:7	15:9
Post mortem interval hr *(mean ± sd)*	3.9 ± 0.9	4.9 ± 1.5	4.5 ± 1.2	15.7 ± 9.3
APOE ε4 alleles *(n (%))*	2 (11.1%)	6 (27.3%)	7 (19.4%)	6 (17.6%)*
ε3:ε4 *(n)*	2	4	5	4
ε4:ε4 *(n)*	0	1	1	1
missing	0	0	0	7
Mutations (n x *gene)*	NA	NA	NA	17 x *PSEN1*, 6 x *PSEN2*, 1 X *APP*
Braak Stage				
B1 *(n)*	4	0	0	0
B2 *(n)*	5	5	0	0
B3 *(n)*	0	6	18	24
CERAD Score				
C0 *(n)*	9	0	0	0
C2 *(n)*	0	5	3	1
C3 *(n)*	0	6	15	23

*For these; if there were missing genotypes the percentage is of the reported genotypes

Abbreviations: Alzheimer's disease (AD), AD dementia (ADD), AD neuropathologic change (ADNC), high cognitive function (HCF)

Table 2. Brain donor characteristics for the hippocampus.

	HCF/low ADNC	**HCF/high ADNC**	**Sporadic ADD**	**Autosomal Dominant ADD**
N	10	11	20	3
Age *(mean ± sd)*	87 ± 6	90 ± 5	81 ± 12	53 ± 7
Sex *(M:F)*	5:5	5:6	12:8	2:1
Post mortem interval hr *(mean ± sd)*	3.9 ± 0.9	5.1 ± 1.3	4.2 ± 1.4	8.5 ± 5.4
APOE ε4 alleles *(n (%))*	2 (10.1%)	6 (27.3%)	10 (25.0%)	2 (33.3%)

ε3:ε4 *(n)*	2	4	6	0
ε4:ε4 *(n)*	0	1	2	1
missing	0	0	0	0
Mutations (n x *gene)*	NA	NA	NA	2 x *PSEN1*, 1 X *APP*
Braak Stage				
B1 *(n)*	5	0	0	0
B2 *(n)*	5	5	0	0
B3 *(n)*	0	6	20	3
CERAD Score				
C0 *(n)*	10	0	0	0
C2 *(n)*	0	5	4	0
C3 *(n)*	0	6	16	3

Table 3. Brain donor characteristics for the inferior parietal lobule (IPL).

	HCF/low ADNC	**HCF/high ADNC**	**Sporadic ADD**	**Autosomal Dominant ADD**
N	8	12	18	23
Age *(mean ± sd)*	89 ± 4	89 ± 6	80 ± 13	50 ± 10
Sex *(M:F)*	4:4	6:6	11:7	15:8
Post mortem interval hr *(mean ± sd)*	3.9 ± 0.9	4.9 ± 1.5	4.5 ± 1.2	15.7 ± 9.3
APOE ε4 alleles *(n (%))*	2 (12.5%)	6 (20.8%)	10 (27.8%)	6 (18.8%)*
ε3:ε4 *(n)*	2	4	6	4
ε4:ε4 *(n)*	0	1	2	1
missing	0	0	0	7
Mutations (n x *gene)*	NA	NA	NA	17 x *PSEN1*, 5 x *PSEN2*, 1 X *APP*
Braak Stage				
B1 *(n)*	3	0	0	0
B2 *(n)*	5	6	0	0

B3 *(n)*	0	6	18	23
CERAD Score				
C0 *(n)*	8	0	0	0
C2 *(n)*	0	6	4	1
C3 *(n)*	0	6	14	22

*For these; if there were missing genotypes the percentage is of the reported genotypes

Table 4. Brain donor characteristics for the caudate nucleus.

	HCF/low ADNC	**HCF/high ADNC**	**Sporadic ADD**	**Autosomal Dominant ADD**
N	10	10	18	20
Age *(mean ± sd)*	85 ± 6	90 ± 5	81 ± 11	51 ± 11
Sex *(M:F)*	6:4	5:5	10:8	13:7
Post mortem interval hr *(mean ± sd)*	3.9 ± 0.8	5.2 ± 1.4	4.4 ± 1.3	15.2 ± 9.8
APOE ε4 alleles *(n (%))*	2 (10.0%)	5 (25.0%)	9 (25.0%)	5 (17.9%)*
ε3:ε4 *(n)*	2	3	5	3
ε4:ε4 *(n)*	0	1	2	1
missing	0	0	0	6
Mutations (n x *gene)*	NA	NA	NA	17 x *PSEN1*, 6 x *PSEN2*, 1 X *APP*
Braak Stage				
B1 *(n)*	6	0	0	0
B2 *(n)*	4	5	0	0
B3 *(n)*	0	5	18	20
CERAD Score				
C0 *(n)*	10	0	0	0
C2 *(n)*	0	5	4	1
C3 *(n)*	0	5	14	19

*For these; if there were missing genotypes the percentage is of the reported genotypes

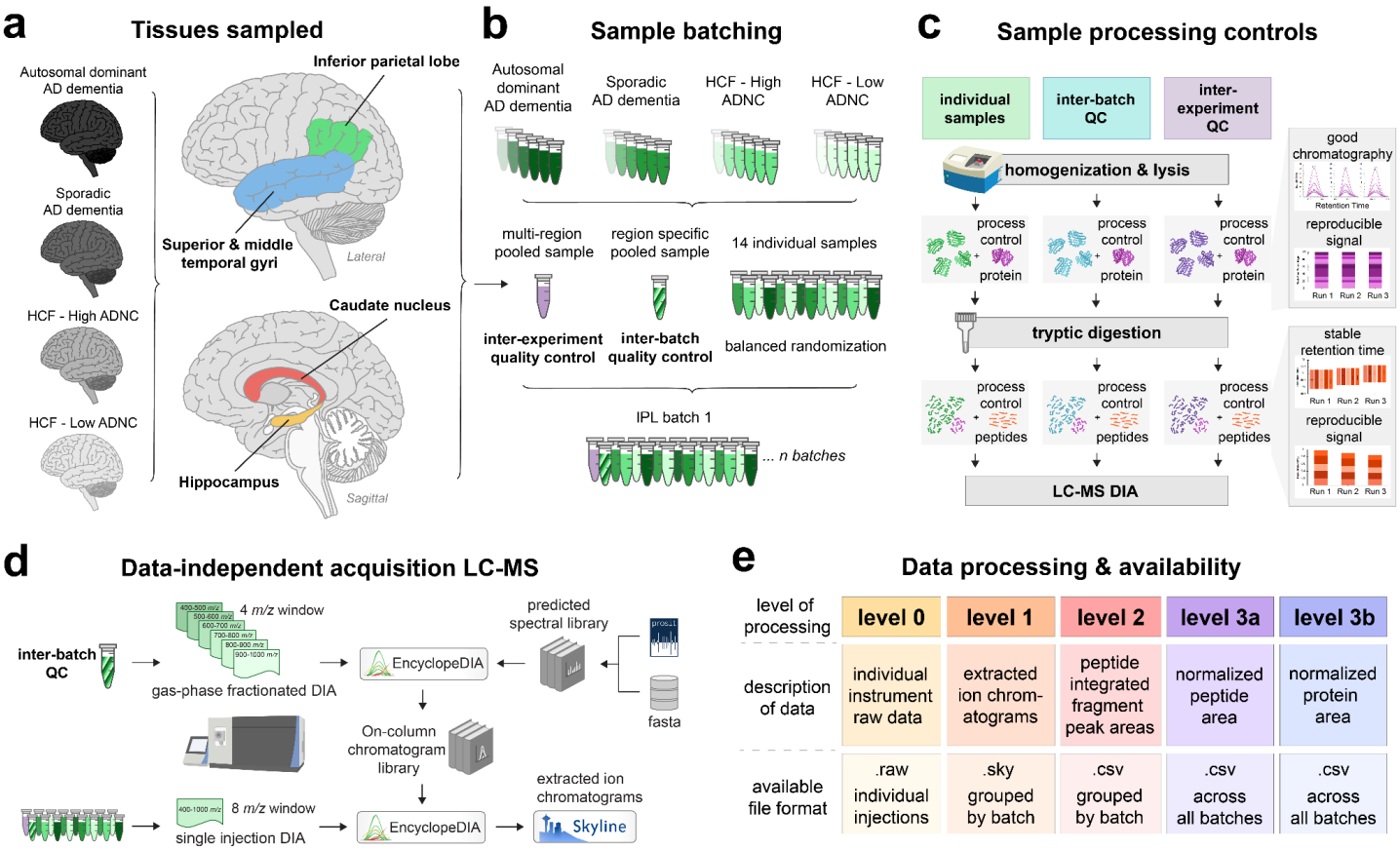
Table 5. Run Level and Experiment Level Peptide and Protein Detections

Brain Tissue Region	Run Level FDR Peptide Detection Range	Experiment Level Peptide Detections	Experiment Level Protein Detections
SMTG	51582-69590	40346	5851
Hippocampus	32995-59853	31863	5117
IPL	37840-73168	48271	6497
Caudate	31426-58105	26135	4636

References

1. Vos, T. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet* **386**, 743–800 (2015).
2. Scheltens, P. *et al.* Alzheimer's disease. *The Lancet* 397, 1577–1590 (2021).
3. Hyman, B. T. *et al.* National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimers Dement* **8**, 1–13 (2012).
4. Montine, T. J. *et al.* Recommendations of the Alzheimer's disease-related dementias conference. *Neurology* **83**, 851–860 (2014).
5. Amodei, D. *et al.* Improving Precursor Selectivity in Data-Independent Acquisition Using Overlapping Windows. *J Am Soc Mass Spectrom* **30**, 669–684 (2019).
6. Brenes, A., Hukelmann, J., Bensaddek, D. & Lamond, A. I. Multibatch TMT Reveals False Positives, Batch Effects and Missing Values. *Mol Cell Proteomics* **18**, 1967–1980 (2019).
7. Plubell, D. L. *et al.* Putting Humpty Dumpty Back Together Again: What Does Protein Quantification Mean in Bottom-Up Proteomics? *J Proteome Res* **21**, 891–898 (2022).
8. Pino, L. K., Just, S. C., MacCoss, M. J. & Searle, B. C. Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. *Mol Cell Proteomics* **19**, 1088–1103 (2020).
9. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
10. Webb-Robertson, B.-J. M., Matzke, M. M., Jacobs, J. M., Pounds, J. G. & Waters, K. M. A statistical selection strategy for normalization procedures in LC-MS proteomics experiments through dataset-dependent ranking of normalization scaling factors. *Proteomics* **11**, 4736–4741 (2011).
11. Bushel P (2013) Pvca: principal variance component analysis
12. Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* **11**, O111.016717 (2012).
13. Egertson, J. D. *et al.* Multiplexed MS/MS for improved data-independent acquisition. *Nat Methods* **10**, 744–746 (2013).
14. Ting, Y. S. *et al.* PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat Methods* **14**, 903–908 (2017).
15. Ma, Z.-Q. *et al.* IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res* **8**, 3872–3881 (2009).
16. Sharma, V., Eng, J. K., Maccoss, M. J. & Riffle, M. A mass spectrometry proteomics data management platform. *Mol Cell Proteomics* **11**, 824–831 (2012).
17. Johnson, E. C. B. *et al.* Large-scale deep multi-layer analysis of Alzheimer's disease brain reveals strong proteomic disease-related changes not observed at the RNA level. *Nature Neuroscience* **25**, 213–225 (2022).
18. Wesseling, H. *et al.* Tau PTM Profiles Identify Patient Heterogeneity and Stages of Alzheimer's Disease. *Cell* **183**, 1699-1713.e13 (2020).
19. Sharma, V. *et al.* Panorama: a targeted proteomics knowledge base. *J Proteome Res* **13**, 4205–4210 (2014).
20. Bereman, M. S., MacLean, B., Tomazela, D. M., Liebler, D. C. & MacCoss, M. J. The development of selected reaction monitoring methods for targeted proteomics via empirical refinement. *Proteomics* **12**, 1134–1141 (2012).
21. Searle, B. C., Egertson, J. D., Bollinger, J. G., Stergachis, A. B. & MacCoss, M. J. Using Data Independent Acquisition (DIA) to Model High-responding Peptides for Targeted Proteomics Experiments. *Mol Cell Proteomics* **14**, 2331–2340 (2015).
22. Hubbard, E. E. *et al.* Does Data-Independent Acquisition Data Contain Hidden Gems? A Case Study Related to Alzheimer's Disease. *J Proteome Res* **21**, 118–131 (2022).

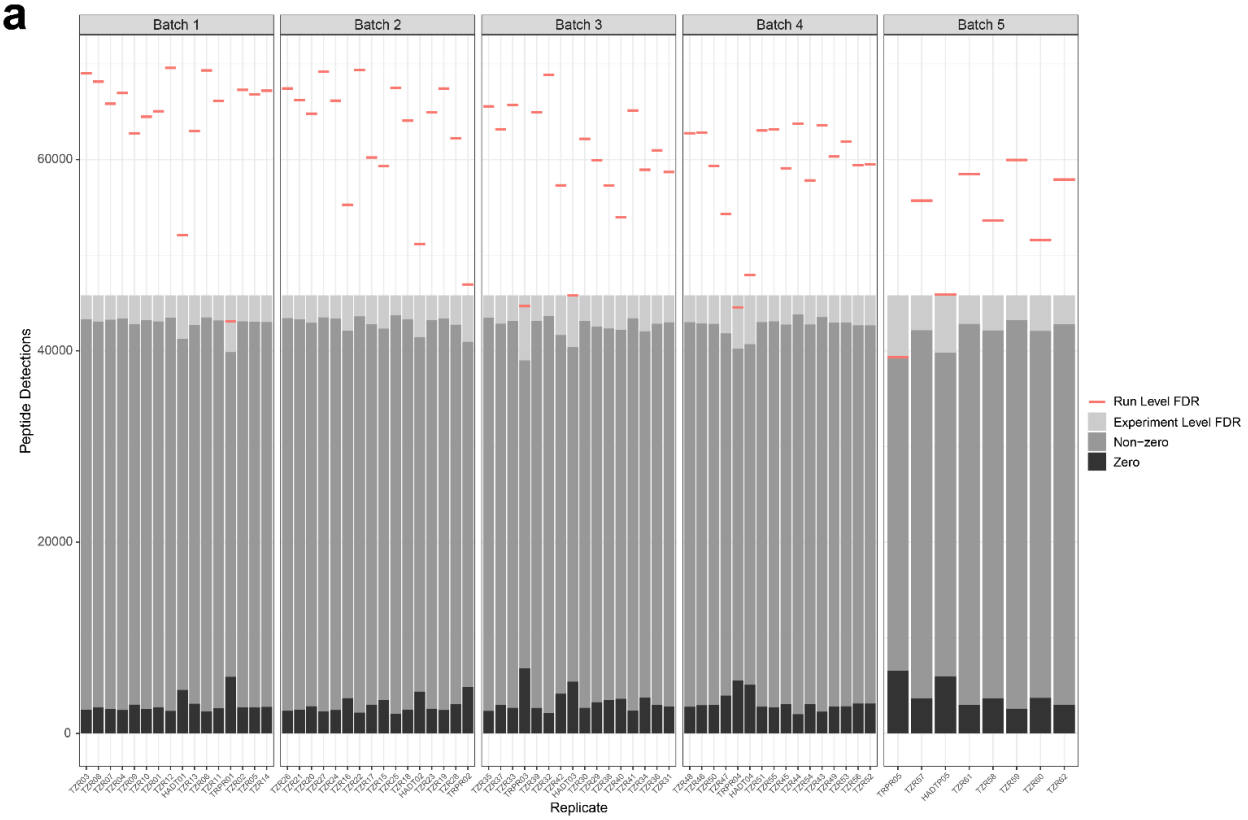
Figure 1



level of processing	level 0	level 1	level 2	level 3a	level 3b
description of data	individual instrument raw data	extracted ion chrom-atograms	peptide integrated fragment peak areas	normalized peptide area	normalized protein area
available file format	.raw individual injections	.sky grouped by batch	.csv grouped by batch	.csv across all batches	.csv across all batches

Figure 2

a


Batch 1
Batch 2
Batch 3
Batch 4
Batch 5
Peptide Detections
60000
40000
20000
0
Replicate
Run Level FDR
Experiment Level FDR
Non−zero
Zero

b

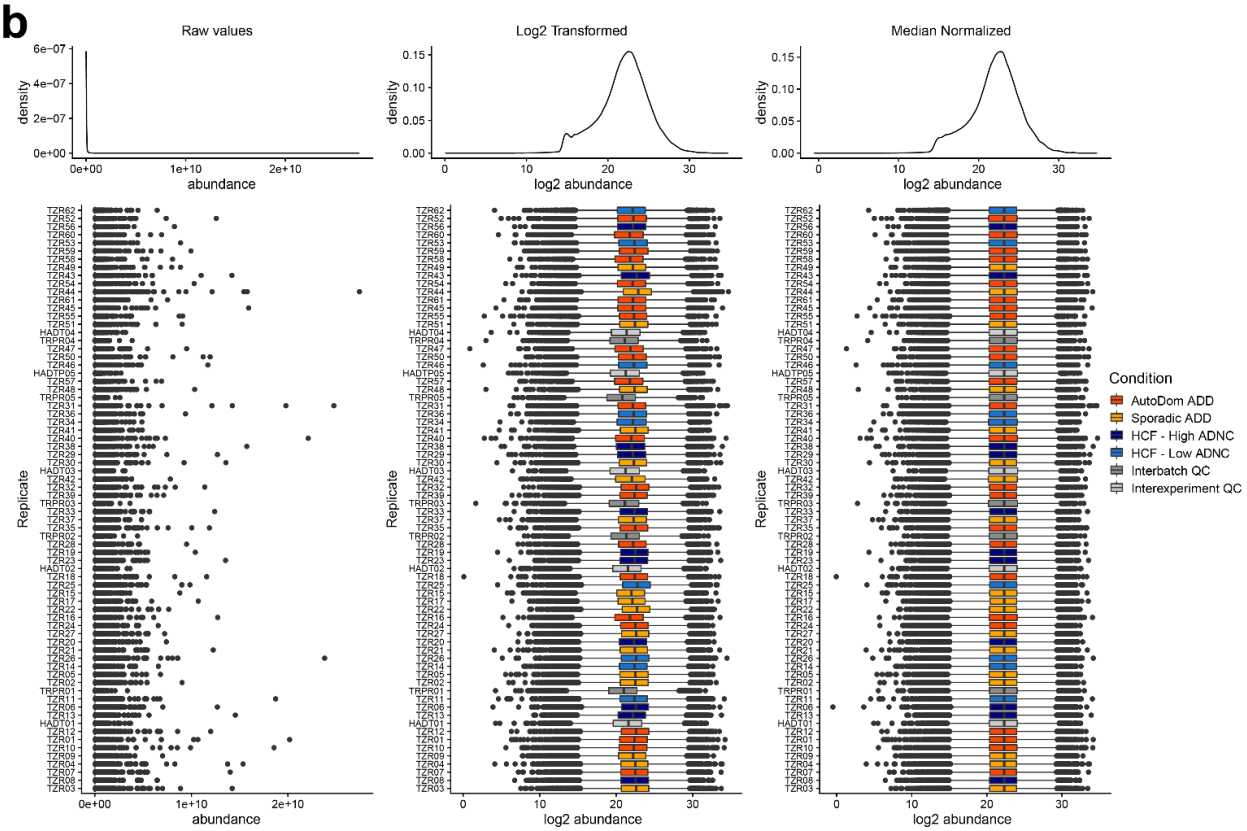

Raw values
Log2 Transformed
Median Normalized
density
abundance
log2 abundance
Replicate
Condition
AutoDom ADD
Sporadic ADD
HCF - High ADNC
HCF - Low ADNC
Interbatch QC
Interexperiment QC

Figure 3

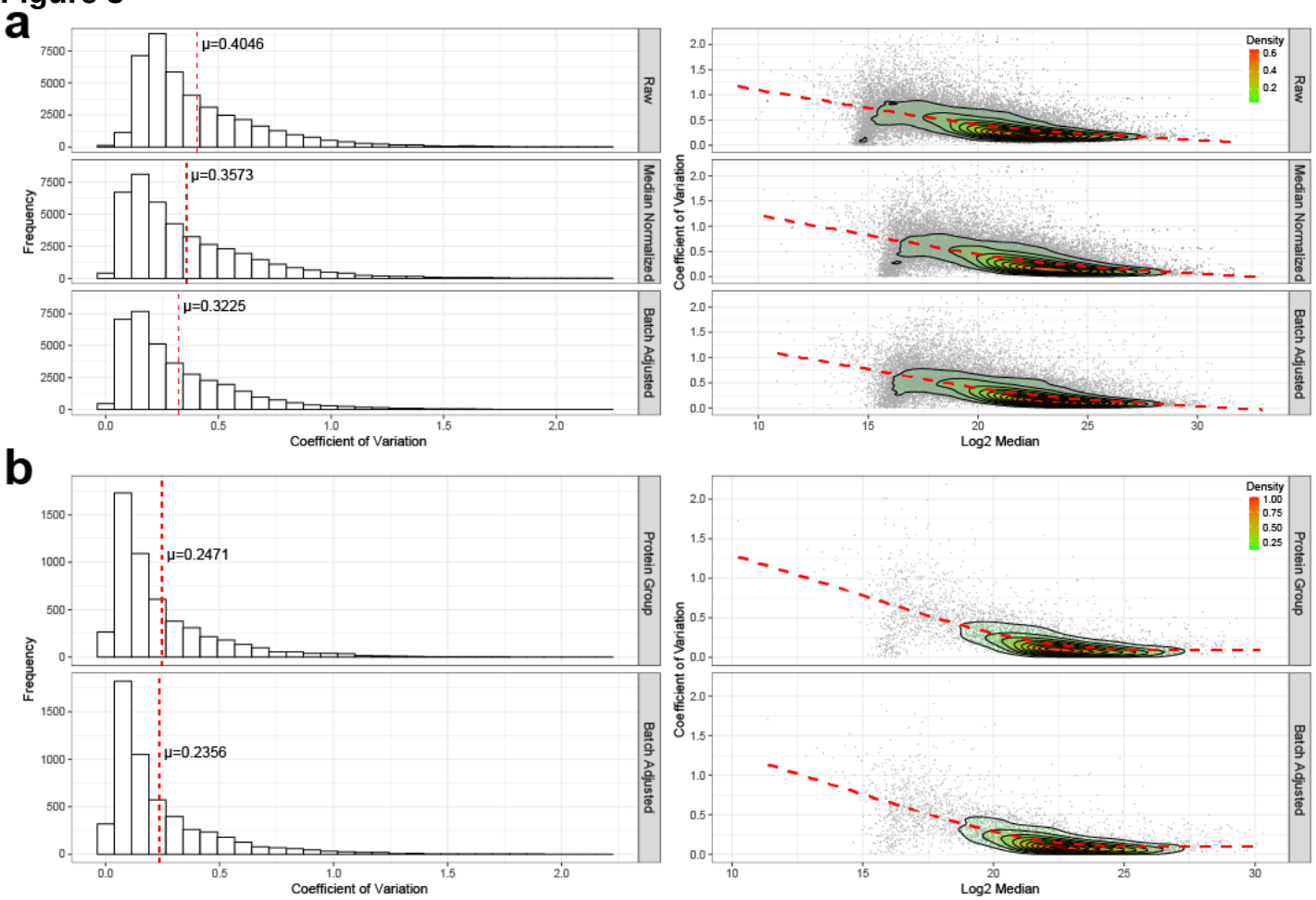

a
μ=0.4046
μ=0.3573
μ=0.3225
Raw
Median Normalized
Batch Adjusted
Frequency
Coefficient of Variation
Density
Log2 Median
b
μ=0.2471
μ=0.2356
Protein Group
Batch Adjusted

Figure 4

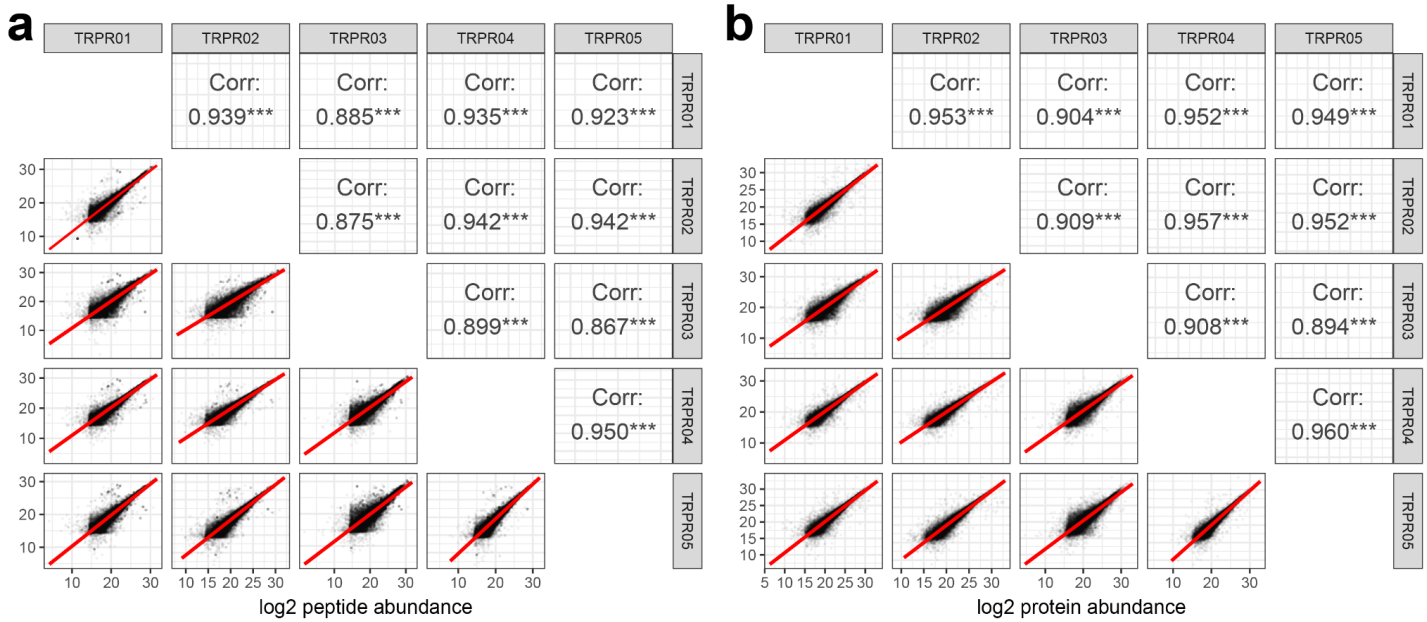


Figure 5

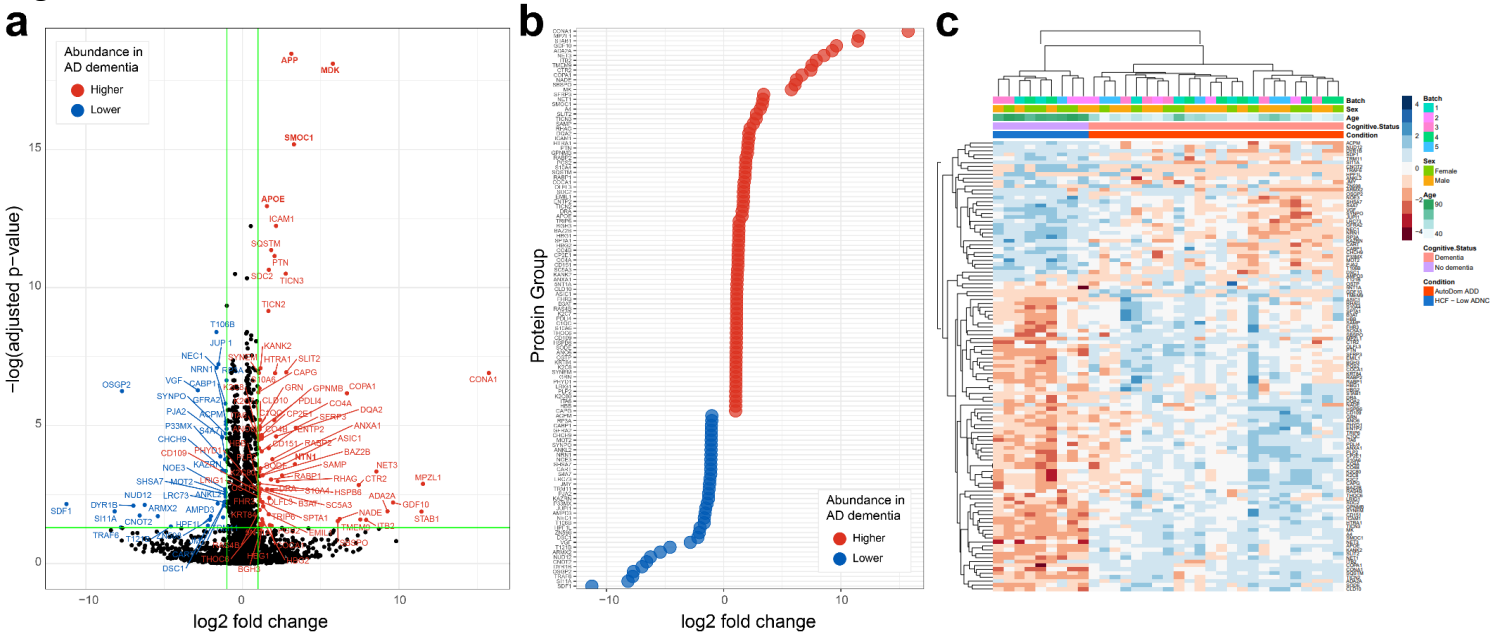
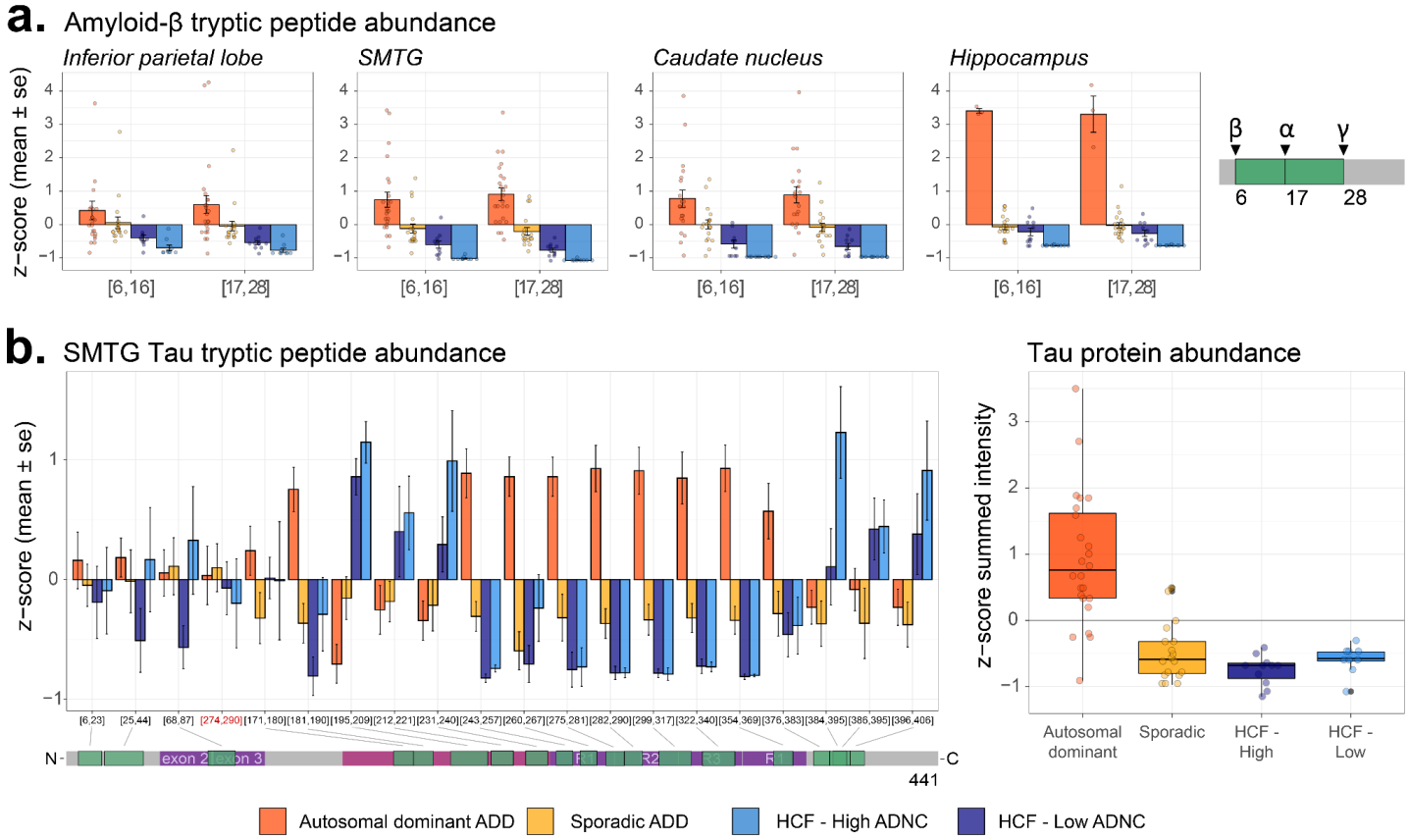
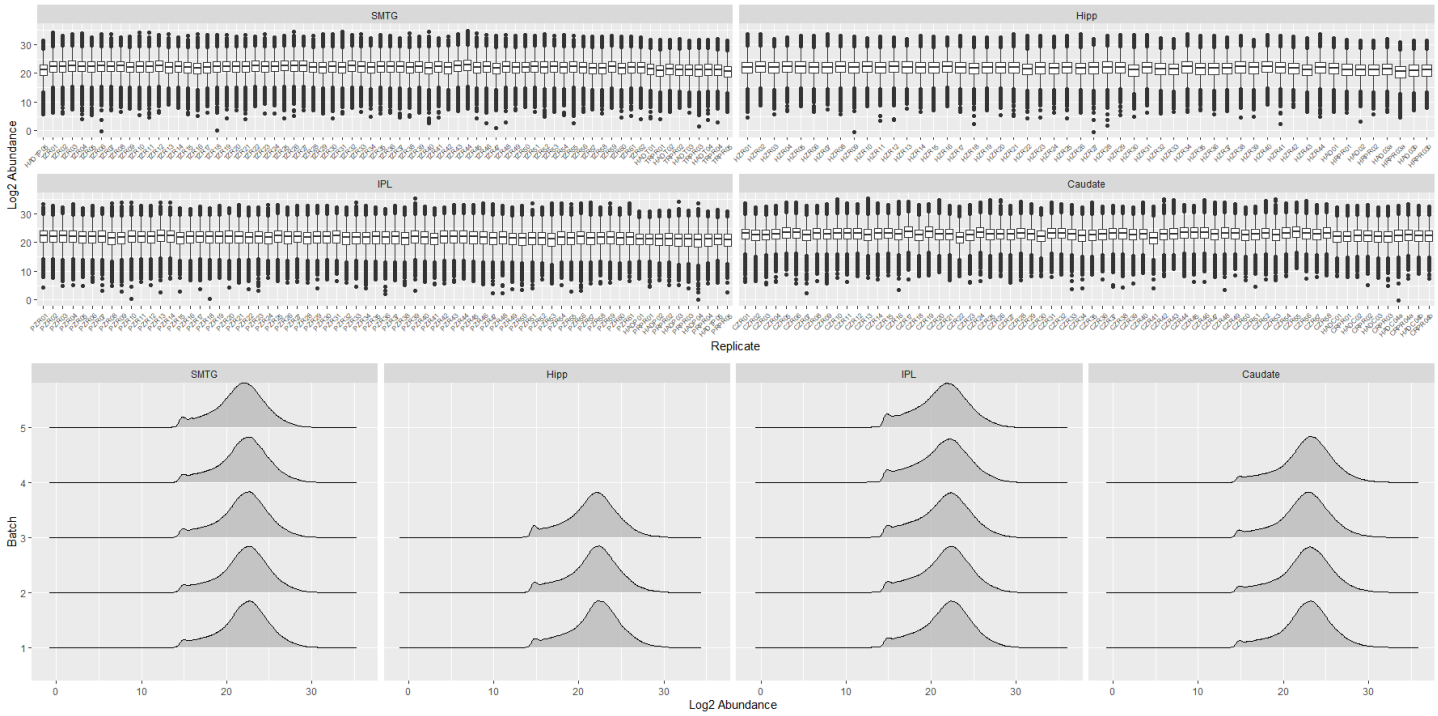


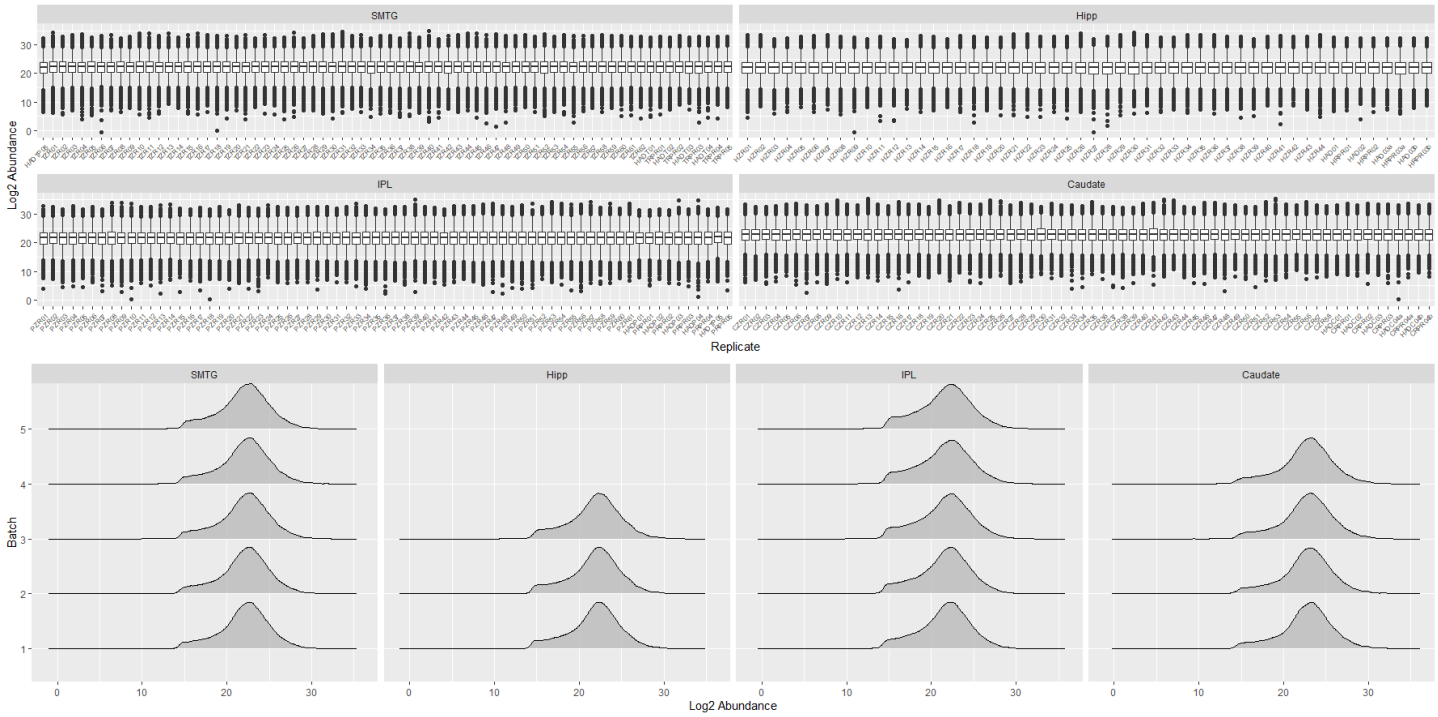
Figure 6


a. Amyloid-β tryptic peptide abundance
Inferior parietal lobe
SMTG
Caudate nucleus
Hippocampus
z−score (mean ± se)
[6,16]
[17,28]
β
α
γ
6
17
28
b. SMTG Tau tryptic peptide abundance
z−score (mean ± se)
N-
-C
441
exon 2
exon 3
R1
R2
R3
R4
Tau protein abundance
z−score summed intensity
Autosomal dominant
Sporadic
HCF - High
HCF - Low
Autosomal dominant ADD
Sporadic ADD
HCF - High ADNC
HCF - Low ADNC

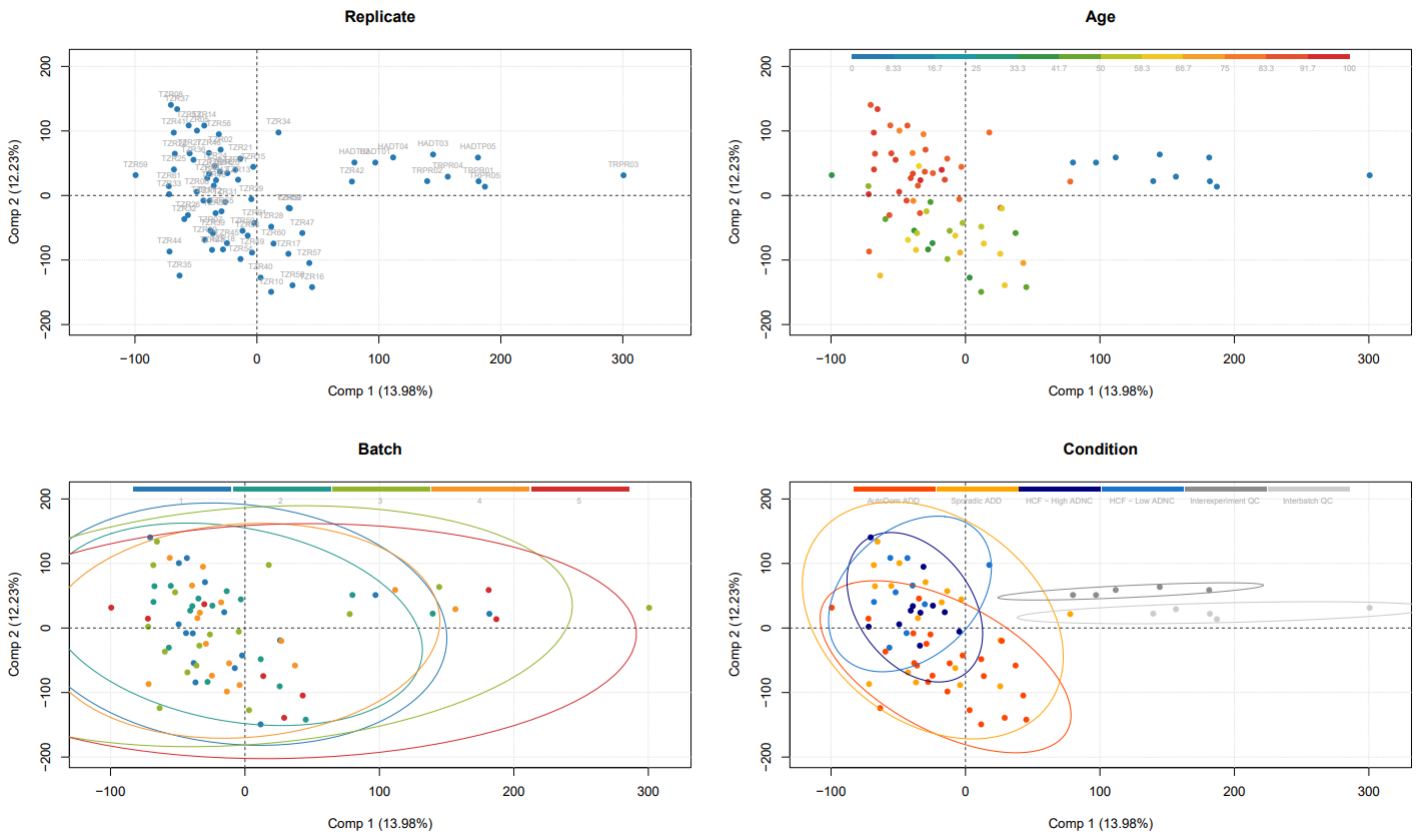
Supplementary Figure 1



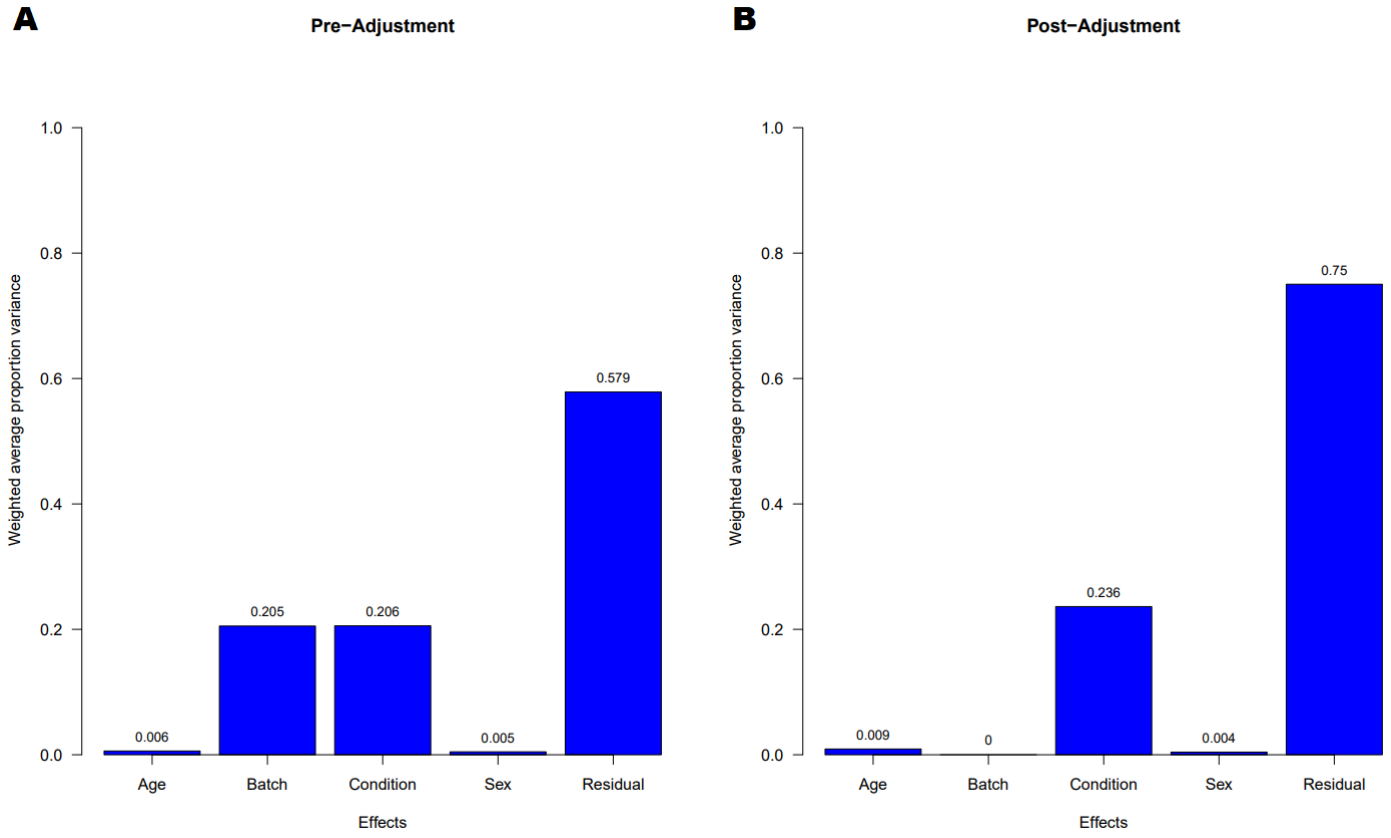
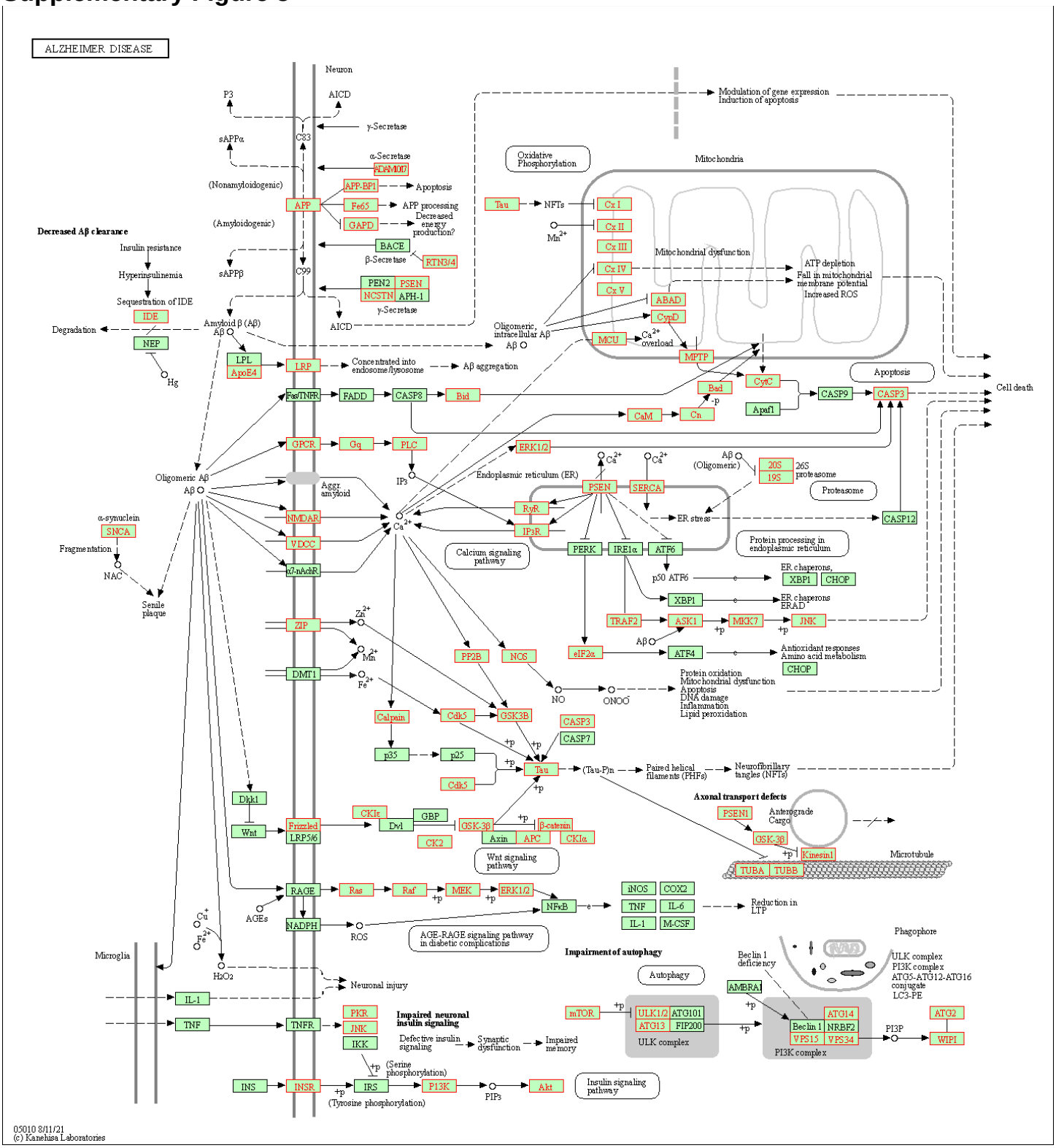
Supplementary Figure 2



Supplementary Figure 3



Supplementary Figure 4



Supplementary Figure 5