Characterizing Alzheimer's disease using quantitative proteomics

Deanna Lisa Plubell

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Michael J. MacCoss, Chair
C. Dirk Keene
James Bruce

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Characterizing Alzheimer's disease using quantitative proteomics

Deanna Lisa Plubell

Chair of the Supervisory Committee:
Michael J. MacCoss
Department of Genome Sciences

Alzheimer's disease is characterized by the accumulation of neuropathologic amyloid-β and tau peptides in the brain. Bottom-up mass spectrometry proteomics methods were used to understand the protein landscape in the brains with different causes of Alzheimer's. Correlating peptide abundances with amyloid-β tryptic peptides reveals additional subgroups of disease in sporadic Alzheimer's cases, with differences across the four brain regions sampled. A cerebrospinal fluid targeted mass spectrometry assay for Alzheimer's disease related proteins was developed as a proof-of-concept of an updated assay development workflow. This work demonstrates the feasibility of using peptide performance on high-resolution instruments to inform assay targets on a unit-resolution instrument. Both projects with Alzheimer's disease demonstrate the importance of proteoforms in human disease, a fact that we argue should be considered more carefully when interpreting or developing bottom-up proteomics experiments.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The author wishes to thank committee members Drs C. Dirk Keene, Jim Bruce, Andy Hoofnagle, Bill Noble, Shao-En Ong, and Devin Schweppe, and of course Mike MacCoss for their valuable feedback and mentorship. The author also wishes to thank Dr Tom Montine for their expertise and time. This work was possible due to the the generous donations by participants of the University of Washington ADRC, Dominantly Inherited Alzheimer's Disease Network, and the Kaiser Adult Changes in Thought study, and this work was supported by National Institutes of Health grants RF1AG053959, U19AG065156, and F31AG069420.

The author also wishes to thank all the members of the MacCoss lab for their support; especially to Genn Merrihew and Jea Park who contributed their expertise in acquiring and processing the human brain data, and Eric Huang and Rich Johnson who aided in mass spectrometry instrument troubleshooting and method advice. Additionally, the numerous colleagues in the Department of Genome Sciences, including the 2018 cohort, the Trainees in Proteomics, Women in Genome Sciences, and Community Organizers of Genome Sciences, for their community and comradery.

Chapter 1. INTRODUCTION TO QUANTITATIVE PROTEOMICS

This chapter is adapted from the following work:

Plubell, D.L., Käll L., Webb-Robertson B., Bramer L.M., Ives A, Kelleher N.L., Smith L. M., Montine T.J., Wu C.C., MacCoss M.J. Putting Humpty Dumpty Back Together Again: What Does Protein Quantification Mean in Bottom-Up Proteomics? J Proteome Res. 2022 Apr 1;21(4):891-898. doi: 10.1021/acs.jproteome.1c00894. Epub 2022 Feb 27. PMID: 35220718

Proteins are the functional units of biology; molecules encoded by genes, responsible for the structure and function of cells. To date 19,778 unique proteins have been annotated from the human genome.[1] Differences in the production and presence of these proteins can alter the molecular processes in cells, tissues, and therefore in organisms. Therefore, it is of great importance to survey the protein composition, or proteome, of these systems since they can provide valuable insights into the biological state.[2] Unlike nucleic acid molecules with four bases for sequencing, proteins are composed of 20 amino acids, making sequencing a complex task.[3] Proteomics as a field has predominantly used mass spectrometry technology for determining the identity and abundance of proteins in a complex mixture.

1.1. Bottom-up mass spectrometry proteomics

The field of mass spectrometry proteomics includes a diverse collection of data acquisition and analysis methods. The most common strategy for measuring the proteome by mass spectrometry is termed “bottom-up” proteomics, which relies on the proteolysis of proteins into peptide fragments by a cleavage enzyme. The use of trypsin to consistently cleave at relatively abundant arginine and lysine residues results in the ability to survey complex protein mixtures robustly and reproducibly. Once the protein mixture is digested to peptides it is usually separated by liquid chromatography. This step is important for complex mixtures to reduce the amount and

complexity of material introduced to the mass spectrometer. This has enabled the detection of more unique species by limiting the number of analytes the instrument will measure at a given time.

The fundamental measurement of mass spectrometers is the mass-to-charge (*m/z*) ratio of ions. All ions detected together at a given time make up a mass spectrum, with the intensities relating to the number of the detected ions at a given *m/z*. The specific mass and charge state of a whole peptide is referred to as a precursor, with peptides having several possible charge states and therefore different possible precursors. These precursors can be fragmented through several different techniques to produce product ions. This fragmentation consistently breaks specific bonds along the amino acid sequence, resulting in product ions corresponding to all possible amino acid subsequences. Mass spectra collected from these peptide fragments are then used to infer what peptides and therefore proteins were present in the original sample. In the early 2000's as large scale peptide identification took off, parsimony was used to assert the set of proteins that could give rise to the peptide data that was observed directly.[4,5] As data increased in scale, controlling for false discovery rate (FDR) at the protein level was determined to be a more conservative way to assert protein presence.[6]

Targeted acquisition methods

Targeted proteomics refers to mass spectrometry proteomics methods designed to measure specific target molecules, based on prior knowledge of that target. In bottom-up proteomics parallel reaction monitoring (PRM) methods rely on the knowledge of the target peptide precursor charge and mass.[7] These methods use the mass spectrometer to select the precursor mass range of the targeted peptide (MS1), performs fragmentation of those precursor ions, and then detects product ions produced from that fragmentation step (MS2 or MS/MS). This method typically relies on the use of instruments with a high-resolution mass analyzer, commonly an orbitrap, and are typically limited to a few select precursors to monitor due to the speed of the instrument. Methods to schedule the targeting of the precursors can improve the number of analytes,[8] but the cost of instruments still makes it a barrier to applications at scale. Selected reaction monitoring (SRM) methods also rely on the knowledge of the target peptide precursor charge and mass, while also requiring the product ion masses. The product ion mass is required because this method is

performed on triple quadrupole instruments, with just the product ions measured by the detector.[9,10] Like PRM methods, SRM is also limited in the number of analytes that can be measured. However, since the target list can be refined to just a couple robust and reproducible product ions per precursor, this means it can potentially measure more peptides, and by extension proteins, of interest. In both acquisition methods fragmentation and detection is performed rapidly over time. The separation by the LC results in peptides commonly eluting in a gaussian distribution of abundance. By sampling these target peptides over time, the intensities of the precursor and product ions will reflect that change in abundance, resulting in gaussian curves for all measured ions, ideally with matching peak apex and width. Multiple ions per precursor can be measured, with their ion intensities across scans used to calculate the area under the curve within a peak boundary. These areas can be summed to result in a single abundance value per precursor.

Discovery acquisition methods

Although the targeting of specific peptides and proteins of interest can be incredibly useful, there are many situations in which the proteins of interest may not be known, or the measurement of the whole proteome to look for novel relationships or molecular signatures is preferred. Data-dependent acquisition (DDA) became the standard for this non-specific, or "discovery" bottom-up proteomics.[11] With this approach a scan is taken of the standard tryptic peptide precursor mass range, and the most intense precursor signal is then selected for fragmentation and detection. A common strategy used in DDA to sample the most possible unique precursors is dynamic exclusion. After the fragmentation and detection of the top n precursors for a given MS1 scan, those precursors are excluded from being targeted for fragmentation for n following scans. This allows for the detection of fragments from lower signal precursors that may be co-eluting but not fragmented if only the most abundant precursors are selected. This approach is extremely powerful for building lists of proteins present in a sample or tissue. For DDA methods quantification of the area of the elution curve can be performed for the precursor masses, but not for the product masses since they are commonly not fragmented for set time after initial detection. An alternative strategy for quantification using DDA methods is the use of isobaric labels, which relies on the comparison of relative intensities across samples with varying isobaric tags. While strategies exist to extend this quantification between individual experiments, it is still limited in the number of samples it can reproducibly measure.[12]

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 scan5. 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. Although the missing values in multiplexing tandem mass tags can be reduced at a protein level, it remains a problem on the peptide level due to the variable peptide sampling between runs.[13]

An alternative to DDA is an acquisition approach known as data independent acquisition (DIA). This method acquires comprehensive fragment ion information by using a repeated cycle of precursor mass ranges for fragmentation (MS/MS). This means that a precursor and its product ions, or transitions, are likely to be sampled multiple times as they elute off the LC column into the MS instrument. The computational analysis of DIA spectra can therefore 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 precursor 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. Like PRM methods, DIA requires reproducible chromatographic separation of peptides for reproducible quantification. Systematic sampling is important when scaling to data sets with samples prepared and run over many batches.

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.[14,15] 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).[16–20] 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.[21]

1.2. Quantifying proteins with proteomics

Today most proteomics experiments are performed with the intent of measuring the differences in abundance of proteins between samples. It is therefore desirable to summarize or aggregate peptide quantities into a single value per protein. Many strategies have been created to accomplish this, with most assuming peptides belonging to the same protein will behave similarly. However, based on historical work in protein biochemistry, 2-dimensional gels, and top-down proteomics, it is estimated that there may be on average up to 100 proteoforms per protein.[22,23] This estimate is based on the possible variations that can occur to a protein's coding sequence or by post-translational modifications (PTMs) on a protein molecule. As most of the amino acid sequence is shared among related proteoforms, a given tryptic peptide can be derived from multiple different proteoforms (Figure 1.1). Once digested in a mixture, the direct connection between a peptide and its originating proteoform(s) is lost, such that the measurements of individual peptides are convolutions of the proteoforms the peptides are present in. This issue of conflation is conceptually similar to the problem of haplotype phasing in genomics.[23]

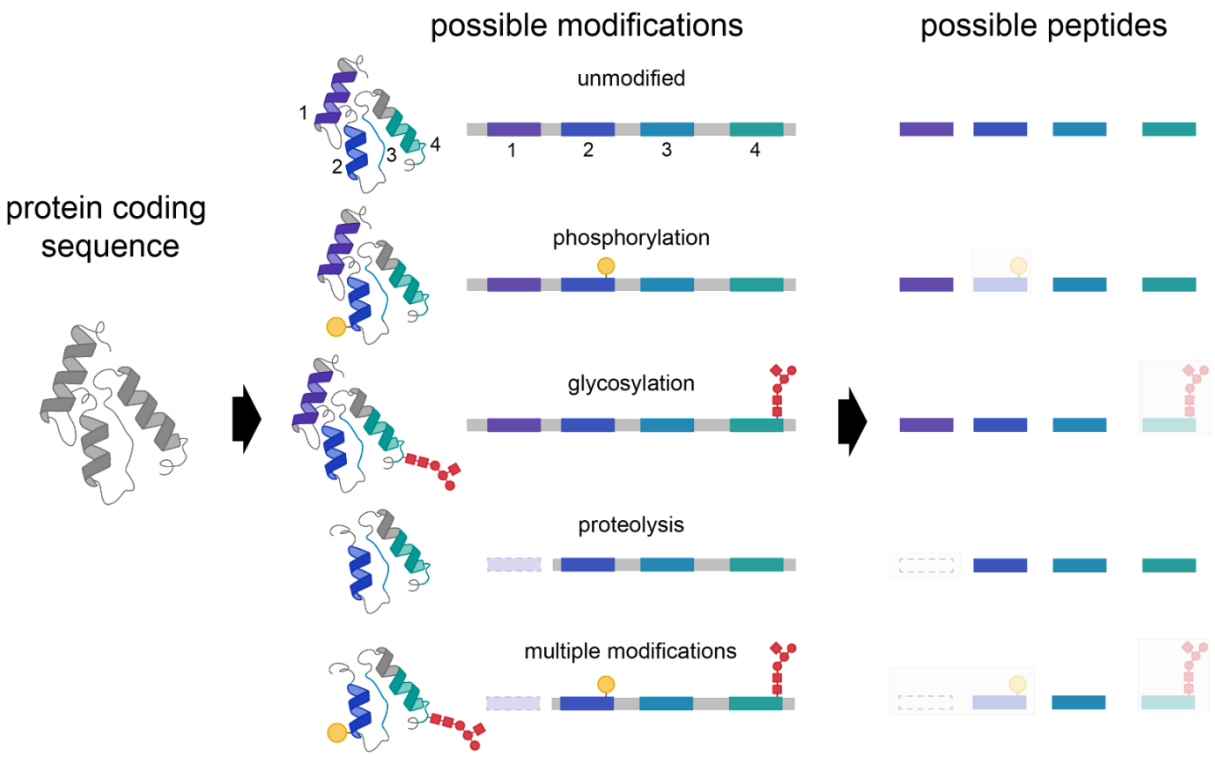


Figure 1.1. Effect of proteoforms on possible peptide detection. A single protein coding gene can be modified to give rise to dozens or many thousands of proteoforms, including those harboring multiple modifications. After proteolysis, proteoforms yield peptides that may be missed in bottom-up proteomics database searching and data processing.

Rationale for combining peptide measurements into a single protein quantity

The idea of aggregating peptide measurements to the protein level is appealing for interpretation and integration of proteomics data with other data types. Since the beginning of quantitative proteomics, scientists have compared the quantification and coverage of proteomics to the latest gene expression data.[24] Intuitively this practice makes sense based on the central-dogma of molecular biology. However, this comparison assumes that for each mRNA transcript there is a single protein quantity for comparison. Despite knowing that there may not be a **single** "protein" derived from the expressed gene, this analysis is standard practice in the field. Such comparisons have demonstrated that the correlation between gene expression and an individual protein measurement is relatively poor.[25] While several explanations have been proposed, it is important to note that most experiments were performed using bottom-up proteomics data

summarized to a single measurement per protein, even though multiple proteoforms likely existed.

Beyond the proposed ease of biological interpretation, there are technical reasons that make aggregating peptides to a protein level measure attractive for quantification. In quantitative proteomics, our ability to find differences is affected by three parameters: 1) the size of the biological effect, 2) the biological and technical variability, and 3) the number of hypothesis tests that are made within the experiment. Thus, it is important to consider how summarizing peptide quantities at the protein level will affect these three parameters. By aggregating peptides mapping to a protein coding sequence into a single measure, especially by common methods that use averaging or summing peptide measurements, outliers or noisy signals are suppressed. Although this effect is diminished with alternative and more sophisticated computational strategies, the overall result remains that there is less variation observed among technical and biological replicates. For example, we observe more variability in the peptide level values compared to the protein level values in technical replicate injections of cerebrospinal fluid (CSF) digests (Figure 1.2). In the case of replicate measures, the reduction in variability is viewed as a positive outcome. Additionally, by aggregating to a single protein measure we reduce the number of hypotheses tested, thereby making the analysis more sensitive to finding changes in protein abundance.

Another reason to aggregate to a protein level has been to reduce the amount of missing data. Sampling is stochastic using data-dependent acquisition and leads to more missing data at the peptide level if the same precursor is not selected in all of the experimental runs. This missing data can have serious implications for the quality of quantitative data. One method to combat this problem is sample multiplexing by isobaric labeling, such as tandem mass tagging peptides. Evaluating large multiplexed experiments in comparison to label-free approaches, the pattern of the missing data appears to be very distinct, but the macrostructure overall is similar in regards to the relationship between abundance and missing data.[26] These multiplexing methods are still limited in the number of samples that can be uniquely tagged, combined, and analyzed at once, and while multiple batches of samples can be acquired, the same peptides are less frequently sampled in different batches compared to proteins.[13]

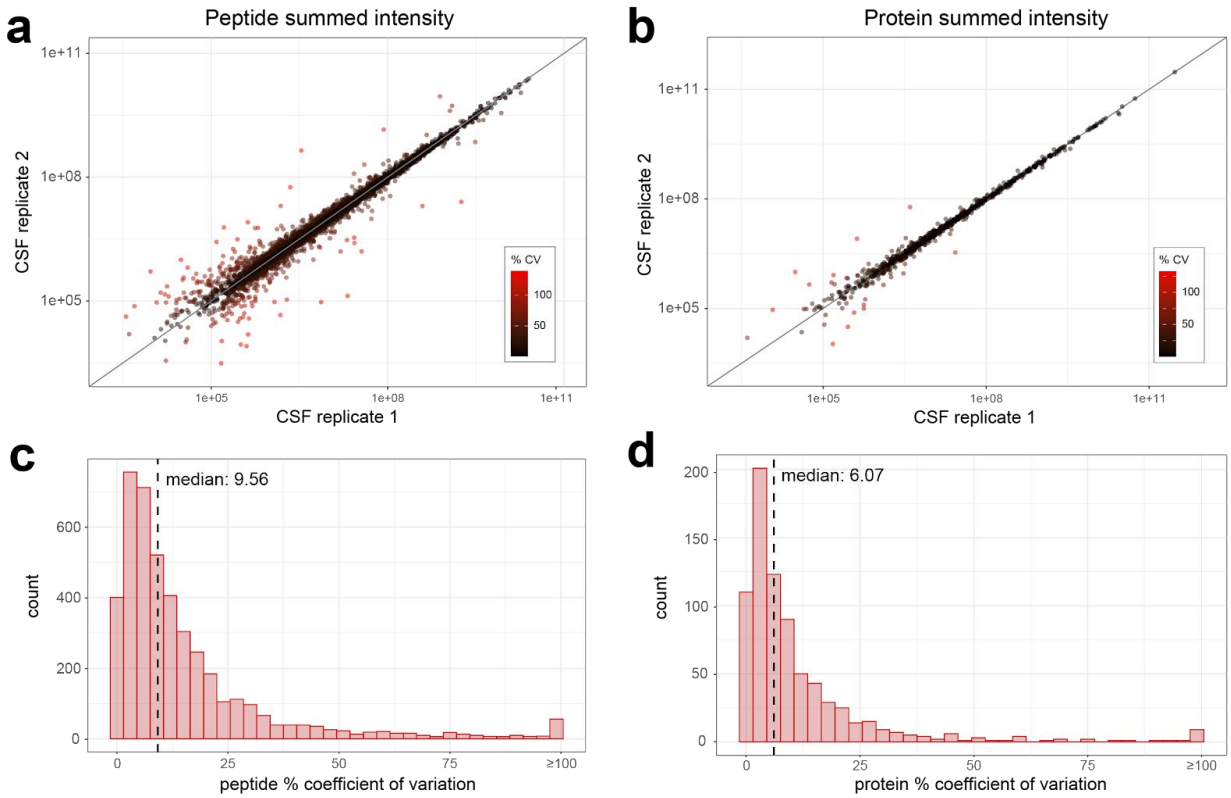


Figure 1.2. Technical variability is reduced when peptide measurements are combined to a protein measurement. A human cerebrospinal fluid sample digest was analyzed by DIA-MS with 8 *m/z* staggered windows (4 *m/z* after demultiplexing). The relationship between a) peptide quantities, or b) summed protein quantities across two replicate instrument runs are plotted, with each peptide colored according to calculated percent coefficient of variation. The distribution of % coefficient of variation for c) peptides and d) summed protein quantities between replicate instrument runs, with the median % coefficient of variation for each indicated by the dashed line.

Interestingly, protein groups with greater numbers of peptides observed tend to be statistically different less often than protein groups with fewer peptides (Figure 1.3). Despite the different types of proteomics data, the difference in scale of the data, and using either a sum-based or reference-based quantification, the fold-change consistently trends towards zero. The loss in quantitative significance in proteins with greater coverage is initially counterintuitive. While the decreased magnitude of change can still be statistically significant, it doesn't mean those differences are representative of every peptide measured from that protein. Greater peptide coverage will likely span more proteoforms, meaning a measured peptide quantity could be derived from multiple proteoforms containing that peptide. Unless all those proteoforms change similarly among conditions, aggregating more peptides to a single protein value can average away

the biological effect. Conversely, if coverage is low, differences in peptides specific to a subset of proteoforms may or may not be captured. If a value is reported at a protein level it makes these differences difficult to compare across studies since the peptides measured may be important for interpreting the results.

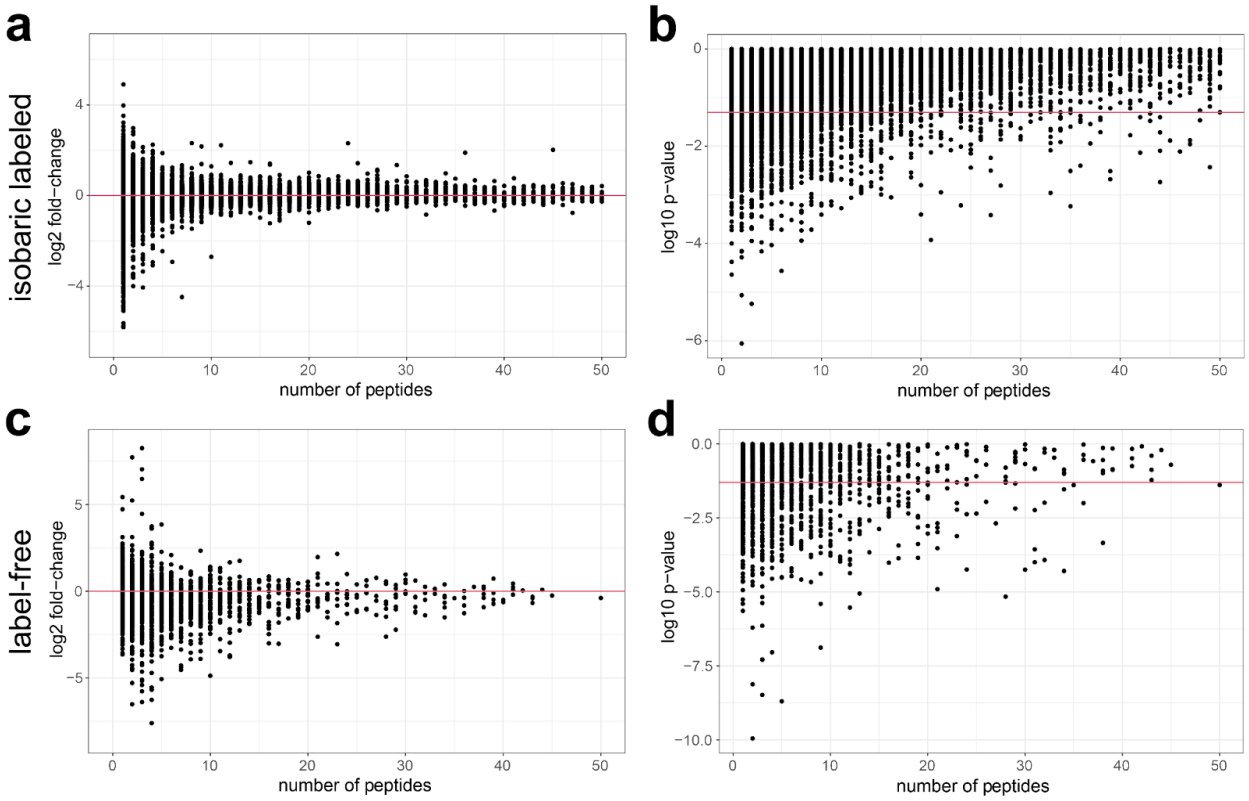


Figure 1.3. The effect size on the protein level is minimized for proteins with greater numbers of peptides. An isobaric-labeled dataset associated with the Clinical Proteomics Tumor Analysis Consortium (CPTAC) consists of 181,389 peptides mapped to 10,495 unique protein identifiers; proteins ranged from having 1 to 563 peptides associated with them. The a) log2 fold-change and b) log10 p-value is based on a comparison of tumor residual disease. The second dataset is label free and smaller, based on a Calu-3 cell culture experiment, also publicly available (MSV000079152). This dataset has 15,953 unique protein identifiers, with proteins represented by 1 to 311 peptides. In this dataset the a) log2 fold-change and b) log10 p-value is based on a Middle East Respiratory Syndrome (MERS) infection to a sham control. Protein sum-based quantification sums all peptide measures per protein coding gene. For b) and d) the red line indicates the significance cutoff corresponding to p=0.05, with significantly different proteins falling below the line. Figures are truncated to 50 for ease of visualization.

The problem of reducing the biological effect when aggregating peptide quantities into a single protein value is analogous to single cell versus bulk tissue analysis. It is well known that tissues are heterogeneous. In a bulk analysis, a large change occurring in a single cell would be indistinguishable from a small change occurring across all cells -- yet these would clearly represent very different biology. By averaging the results from bulk tissue, the ability to assess the degree of heterogeneity on the effect is lost. Furthermore, differences could disappear entirely in the bulk sample because the effect on each cell could be very different. The same is true with reporting protein level quantities from peptides. A change might only be reflected in a proteoform that is best reflected in the quantity of a single peptide. By aggregating the peptides into a single protein level measurement this difference will be 1) misinterpreted as an effect of the entire protein or 2) averaged away and missed entirely.

Limitations of assuming a single quantity per protein coding gene

The fact that many proteins we are detecting are modified cannot be ignored. The estimate of an average of 100 proteoforms per protein coding gene may seem large until one investigates just how many modified peptides have been detected for most proteins.[22] Some notable examples can be seen with clinical biomarkers derived from post-translational modification which are further described below. These examples highlight that just because peptides map to a protein coding sequence, it does not mean that the peptides will be present at the same quantity in a biological system. To the extent this is true, statistical power will be reduced in connecting phenotype to proteomic data.

Alzheimer's disease is fundamentally characterized neuropathologically by the accumulation of amyloid-β plaques and tau neurofibrillary tangles, both of which are composed of post-translationally modified proteins. Specifically, amyloid-β is a peptide derived from the amyloid precursor protein gene. In Alzheimer's disease a series of cleavage events lead to several shorter soluble forms of amyloid precursor protein (sAPPα, sAPPβ), C-terminal fragments (AICD50, CTF 83, CTF 89, CTF 99, p3) and amyloid-β peptides, which contribute to forming the characteristic plaques observed in the brains of diseased individuals.[27] The amyloid-β peptides can be variable lengths depending on specific cleavage site, but commonly occur as a peptide of either 40 or 42 amino acids.[28] In addition to the widely known amyloid-β 40 and amyloid-β 42 peptides, over 20 additional amyloid-β proteoforms have been detected in samples of Alzheimer's brain

samples arising from endogenous cleavage and post-translational modifications.[29,30] Knowing the amyloid precursor protein is heavily processed, it is difficult to determine the origin of many of its tryptic peptides - whether they are derived from an unprocessed amyloid precursor protein, or from one of many processed forms.

If we aggregate all the tryptic peptide measures, we are assuming they are all derived from the unprocessed state, which may not be the most accurate assumption for peptides mapping to amyloid precursor protein. If we look at data from tryptic peptides, we see that some biologically relevant differences would not be accurately represented if our peptide measures are combined to a singular protein level (Figure 1.4). Specifically, in tryptic peptides mapping to the region of the amyloid beta sequence we observe a different abundance profile compared to tryptic peptides mapping to other regions of the protein. In addition to amyloid beta, phosphorylated tau proteoforms in the cerebrospinal fluid of patients have also gained acceptance as diagnostic biomarkers of disease.[31] Additional studies indicate that specific tau phosphosites may be better indicators of disease progression, emphasizing the importance of distinguishing between different pTau isoforms and proteoforms.[32,33]

Ambiguity due to modified protein biomarkers is not a problem unique to Alzheimer's disease, but rather is general to human biology and therefore human disease. The products of processing a precursor protein into polypeptides are important markers in diabetes. C-peptide and insulin are both derived from proinsulin, with C-peptide being a valuable measure of insulin secretion and therefore pancreatic beta cell function.[34] Proglucagon is processed to form up to nine different polypeptide products, including the better known glucagon and GLP1. Both polypeptides have distinct roles in metabolism, and both are drug targets for diabetes and obesity.[35] Additional examples of this type of processing can be found in the kallikrein-kinin system and coagulation pathways.[36] While these examples are well studied, we should not assume that these types of modifications leading to unique biologically relevant proteoforms are uncommon among other less studied proteins.

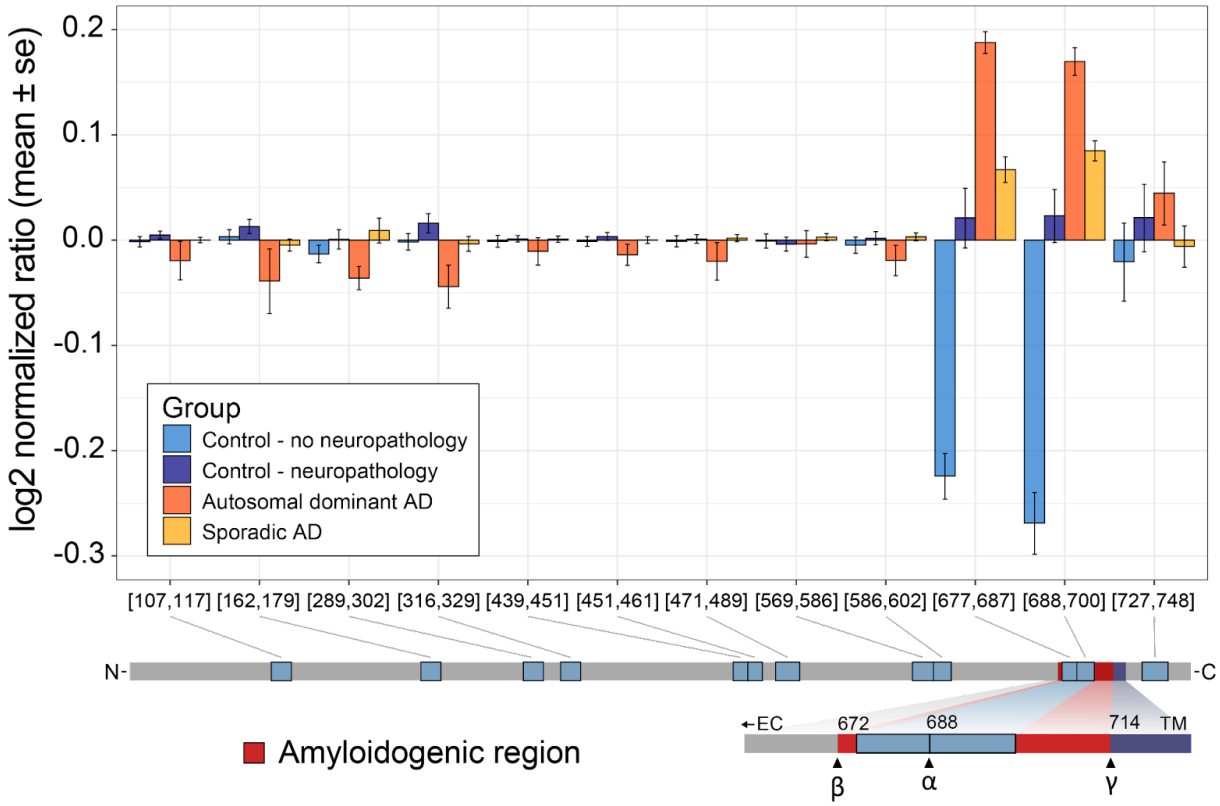


Figure 1.4. Differential abundance profiles of tryptic peptides mapping to amyloid precursor protein. Three experimental groups of patients were analyzed by DIA-MS; Control/No Neuropath with normal cognitive function and no neuropathologic changes of Alzheimer's disease including no amyloid accumulation, Control/Neuropath with normal cognitive function and intermediate or severe level of neuropathologic changes of Alzheimer's disease, Sporadic AD with dementia and intermediate or severe level of neuropathologic changes of Alzheimer's disease, and Autosomal dominant AD with dementia and intermediate or severe level of neuropathologic changes and an autosomal dominant mutation. For all unique peptides mapping to the amyloid precursor protein sequence, peptide measures are normalized to the mean and the mean & standard error are plotted by group. Based on known protein processing we see that the two peptides with large differences map to the amyloidogenic Aβ polypeptide.

When interpreting bottom-up proteomics data we are only able to make conclusions about the peptides we detect, not the proteoforms from which they originate. For example, a study of cerebrospinal fluid in Parkinson's disease found that specific tryptic peptides are differentially abundant in affected individuals compared to healthy, age-matched controls. Specifically, peptides in the C-terminal or N-terminal regions of granin family proteins were found to be decreased in Parkinson's.[37] Importantly, the granin family of proteins is known to play a role in regulating secretion and delivery of peptides and neurotransmitters, and are known to be processed into a

number of derived bioactive peptides (Figure 1.5). As demonstrated in Figure 1.5, if we sum all peptide measures that map to the protein coding sequence of secretogranin 2, then we miss the differences between experimental groups for several of the individual peptides. Instead, aggregating peptides to a single measure per protein coding sequence only accurately reflects the peptide level measurements if all peptides are in agreement (Figure 1.5). In contrast, if we consider peptides detected and quantified from GAPDH protein in the same CSF experiment, we observe the same trend across peptides. Interestingly, GAPDH has been observed not to have many proteoforms by top-down analysis.[38] Although there are known proteoforms, from the peptides we detect we cannot conclude that only one proteoform of GAPDH is present in our samples. Instead, we can only conclude that all the peptides we detect share the same abundance trend.

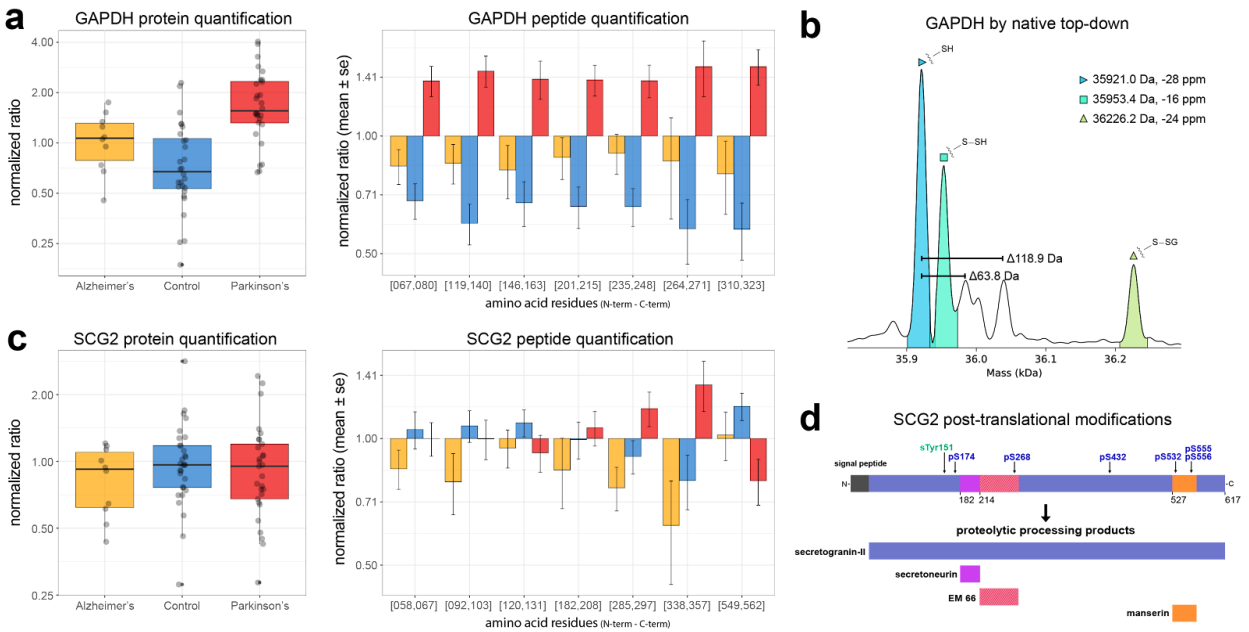


Figure 1.5. Abundance profiles of tryptic peptides mapping to a) GAPDH and b) SCG2 proteins in cerebrospinal fluid. Three groups of human cerebrospinal fluid samples were analyzed by DIA-MS; Alzheimer's disease, Parkinson's disease, and healthy age and sex-matched controls. Unique peptides mapping to the proteins a) GAPDH and c) SCG2 report quantitatively on their relative expression ratios. The protein level display integrates the mean values from all peptide-level results (box-and-whisker plot at left), with the expression ratio for each individual peptide and the group shown in the bar graphs at right. b) GAPDH has been observed as three proteoforms which form homo-tetramers from human cell lines including HEK-tsa. Intact mass spectra of the monomeric form reveal a canonical form, a persulfide-modified form, and a glutathione-modified form. Reported masses represent average masses and ppm mass error from the calculated theoretical average mass. d) SCG2 is proteolytically processed to produce several peptides, has a sulfotyrosine, and can be phosphorylated at several serine residues.

While bottom-up proteomics is arguably the most common method for characterizing protein mixtures, alternative methods have gained interest. These include methods that use antibodies and aptamer affinity to recognize a specific protein or protein domain.[39–41] These methods usually rely on either a single affinity reagent or paired reagents per protein coding gene. It should be noted that any method that constrains complex proteoforms into a single quantitative value per protein coding gene may miss many of the underlying differences. Even assays that use multiple affinity reagents or many tryptic peptides to different domains or modified sites of a protein will likely provide an undersampling of the proteoform species in the sample. For example, the microtubule-associated protein tau is often measured using antibodies that represent phosphorylation at threonine 181 and one that measures so called "total-tau". However, at least 95 post-translational modifications have been discovered on tau,[32] with potentially many 100s of possible proteoforms resulting from the combinations.

The examples given are just a subset of possible causes of differential peptide signals that would generally get aggregated to a singular value. Genomic sequence variation can lead to differing peptide sequences in the population.[42,43] The detection of this variation by bottom-up methods relies on examining individual peptide precursors and fragments. Beyond the mentioned post-translational modifications, other phenomena such as fusion proteins,[44] protein transport and degradation can also produce additional alternate proteoforms. All possible variation and modifications can occur in numerous combinations, further complicating the interpretation of bottom-up data. Thus, even the measurement of every unmodified and modified tryptic peptide along the predicted protein coding gene sequence using either a mass spectrometer or a sequence specific affinity reagent can't put Humpty Dumpty back together again.

Bottom-up proteomics provides peptide measurements and has been invaluable for moving proteomics into large-scale analyses. Commonly, a single quantitative value is reported for each protein coding gene by aggregating peptide quantities into protein groups following protein inference and/or parsimony. However, given the complexity of both RNA splicing and post-translational protein modification, it is overly simplistic to assume that all peptides that map to a singular protein coding gene will demonstrate the same quantitative response. By assuming that all peptides from a protein coding sequence are representative of the same protein, we may miss the discovery of important biological differences. To capture the contributions of existing

proteoforms, we need to the practice of aggregating protein values to a single quantity per protein coding gene should be carefully considered.

Chapter 2. USING DATA-INDEPENDENT ACQUISITION TO INFORM THE DEVELOPMENT OF TARGETED TRIPLE QUADRUPOLE ASSAYS

Mass spectrometry based targeted proteomics methods provide sensitive and high-throughput analysis of selected proteins. To develop a targeted bottom-up proteomics assay, peptides must be evaluated as proxies for the measurement of a protein or proteoform in a biological matrix. Candidate peptide selection typically relies on predetermined biochemical properties, data from semi-stochastic sampling, or by empirical measurements. These strategies require extensive testing and method refinement due to the difficulties associated with prediction of peptide response in the biological matrix of interest. Gas-phase fractionated (GPF) narrow window data-independent acquisition (DIA) aids in the development of reproducible selected reaction monitoring (SRM) assays by providing matrix-specific information on peptide detectability and quantification by mass spectrometry. To demonstrate the suitability of DIA data for selecting peptide targets, we reimplement a portion of an existing assay to measure 100 Alzheimer's disease proteins in cerebrospinal fluid (CSF).[45] Peptides were selected from GPF-DIA based on signal intensity and reproducibility. The resulting SRM assay exhibits similar quantitative precision to published data, despite the inclusion of different peptides between the assays. This workflow enables development of new assays without additional up-front data acquisition, as demonstrated through a separate assay generated for an unrelated set of proteins in CSF.

2.1. Introduction

Mass spectrometry (MS)-based targeted proteomics is a powerful alternative to the use of immunoassays for the precise and accurate quantification of proteins in complex mixtures.[46] A common targeted proteomics technique is the use of a triple quadrupole MS to collect SRM data. The number of analytes that can be measured in triple quadrupole assays is limited. Thus, following digestion of a protein mixture, precursor and product ion pairs (transitions) are selected from peptides to provide a proxy measurement of the protein of interest. The high selectivity and sensitivity of SRM results in accurate and reproducible peptide quantification.[47] It is possible to

validate the measurement of each analyte through figures of merit, including limit of detection and quantification, ensuring that we can produce robust and reliable results.[48]

Several characteristics of peptides are preferred for inclusion in targeted assays based on their performance in each system. If selecting a limited number of peptides per protein, those peptides should be specific to the intended proteoform or combination of proteoforms to provide a proxy abundance. Additionally, selected peptides should have reproducible signal in the target sample matrix. Part of the reproducibility of the peptide signal will be informed by how well the peptide is separated and eluted by liquid chromatography, and how consistently the peptide is fragmented in the mass spectrometer. However, the primary challenge in establishing a targeted assay is selecting peptides that are 1) good proxies of abundance for specific proteoforms or combinations of proteoforms,[49] 2) produce high intensity signals in the mass spectrometer from the target sample matrix,[50] and 3) chromatograph well, with a gaussian peak shape and no interfering ion signal.[51,52]

Common peptide selection techniques rely on manual or automated analysis of empirical data found in literature review, databases of targeted measures, and discovery experiments.[45,52–54] Confounding factors for peptide and transition selection for SRM include differences in peptide response due to sample matrix, sample preparation, separation, and analysis instrumentation and methods. For example, most public data from discovery experiments was collected by data dependent acquisition (DDA), and this data is often leveraged to manually select proxy peptides or train predictive algorithms. However, the criteria that result in frequent sampling and confident identification of DDA spectra (e.g. a fairly abundant MS1 signal, many characteristic fragments, wide elution peaks) are specifically poor performers for a targeted analysis (e.g. a sharp chromatographic peak, selective precursor > product ion pairs of high abundance).[51,52,55] An alternative strategy to DDA for evaluating peptide suitability is implementing SRM to assess recombinant proteins of interest. Full or near-full length protein is purified and digested, then peptides are measured individually to determine target transitions for a quantitative assay.[51] Although this approach eliminates the issue of translating DDA performance to SRM experiments, it remains a challenge to determine performance of endogenous peptides in the specific matrix of interest. Thus, the selection of peptides for SRM is often a laborious process, typically requiring multiple rounds of data acquisition and peptide target filtering.[45,51,55,56]

Predictive methods trained on data-independent acquisition (DIA) data have been shown to outperform DDA-based predictive approaches.[55] We previously reported a method that used

multiple injections to cover the mass range for typical tryptic peptides using data independent acquisition with narrow isolation windows that is analogous to the PAcIFIC method.[20,57,58] These narrow, effectively 2 *m/z*, isolation windows across the relevant mass range, are similar in size to parallel reaction monitoring (PRM).[59] Data can be collected directly from biological samples of interest, thus taking into account the influence of the sample matrix.

Here we leverage a gas phase fractionated library collected from biological samples to efficiently select peptides for SRM. The indexed retention time (iRT) values and transitions are inherently determined during the generation of the library, making SRM scheduling and transition selection seamless.[60] The workflow minimizes the number of comprehensive measurements to assess which peptides 1) are good responders in the matrix of interest, 2) chromatograph well, and 3) exhibit stability and precision for measurement. To demonstrate the usefulness of this workflow we compare peptides selected from narrow-window DIA to those included in an established, well-characterized SRM assay for Alzheimer's disease in cerebrospinal fluid for a set of proteins.[45] Using the workflow presented herein, an assay for the same proteins of interest was developed, exhibiting equivalent precision and accuracy but bypassing multiple rounds of selection and refinement. Additionally, the technique described is implemented to generate multiple separate quantitative assays for different targets from the same initial data.

2.2. Methods

Cerebrospinal fluid patient samples.

Cerebrospinal fluid (CSF) samples were obtained by lumbar tap from patients diagnosed as either probable Alzheimer's disease or Parkinson's disease, and healthy age-matched controls as a part of the Alzheimer's disease research center at the University of Washington, or the Udall repository at Stanford University. CSF was obtained following NIA AD Center Best Practices Guidelines and samples stored at −80°C. Equal volumes of 30 Parkinson's disease, 11 probable Alzheimer's disease, and 30 non-disease samples were combined to make a pooled reference for method development. All samples were pre-existing and were collected under protocols approved by the Institutional Review Board (IRB) at University of Washington, or Stanford University prior to the start of this project. The UW and Stanford Human Subjects Divisions deems the use of pre-

existing de-identified samples exempt from full IRB review and, thus, treated this project as non-human subjects research.

Cerebrospinal fluid (CSF) sample preparation.

CSF was diluted 1:1 with 0.1% PPS silent surfactant (Expedeon) in 100 mM ammonium bicarbonate with 3.5 ng/µl of [15N]APOA1, and heated to 95°C for 5 minutes to facilitate protein denaturation. All samples were reduced with 5 mM dithiothreitol, alkylated with 15 mM iodoacetamide, and the alkylation reaction quenched with 5 mM dithiothreitol. Proteins were digested with sequencing grade trypsin (Pierce) at a 1:25 enzyme to substrate ratio for 16 h at 37°C. Reactions were quenched and PPS surfactant hydrolyzed with 6N hydrochloric acid to reach pH 2. Samples were desalted by solid phase extraction (SPE) using Waters Oasis 60 µm/30 mg MCX cartridges (Milford, MA). The SPE resin was washed and conditioned with 1 mL of methanol, 1 mL of 2.8% ammonium hydroxide in water, 2 mL of methanol, and 3 mL of 0.1% formic acid (FA) in water. Equilibration of the SPE resin was performed using 1 mL of 0.1% FA in water, and 0.1% FA in methanol. Acidified digested samples were loaded onto SPE resin, washed with 1 ml 0.1% FA in water and 1 ml of 90% acetonitrile/10% water, and peptides eluted with 1 mL of 2.8% ammonium hydroxide in methanol. Peptides were dried by vacuum centrifugation and stored at −80°C. Samples were resuspended to 0.33 µg/µl in 0.1% FA in water with Pierce retention time calibrator (PRTC) peptides spiked into each sample to a final concentration of 15 mM. (15 ng per µl)

Data-independent acquisition and processing.

Tryptic peptides were separated using a Thermo EASY nanoLC 1200 on self-packed 30 cm columns packed with 3 µm ReproSil-Pur C18 silica beads (Dr. Maisch, Ammerbuch, Germany) in a 75 µm inner diameter fused silica capillary (#PF360 Self-Pack PicoFrit, New Objective). Trap columns were created from 150 µm inner diameter fused silica capillary fritted with Kasil on one end and packed with the same C18 beads to 25 mm. Solvent A was 0.1% formic acid in water and solvent B was 0.1% formic acid in 98% acetonitrile. For each injection, 1 µg of peptide was loaded and eluted using a 90 min gradient from 5 to 35% B, followed by a 40 min washing gradient. The 30 cm column was heated to 35°C. Data were acquired by narrow window DIA experiments with a Thermo Q-Exactive HF tandem mass spectrometer, as previously described.[20,61] For each DIA

experiment 6 GPF acquisitions were performed with 4 *m/z* precursor isolation windows at 30,000 resolution, 1e6 target AGC, 55 ms maximum injection times, and NCE of 27. Narrow mass range windows were staggered with optimized window placements as detailed in Pino et al. Two MS1 scan events were collected every 25 MS2 spectra. One of the MS1 scan events spanned the 400-1000 *m/z* range covered by all gas phase fractionated runs. The second MS1 scan event covered just the 100 *m/z* range covered by the specific MS2 isolation windows using the selected ion monitoring (SIM) scan function to improve the MS1 dynamic range. DIA data were demultiplexed with a 10 ppm tolerance following peak picking by ProteoWizard MSConvert.[21] Using Walnut, a re-implementation of PECAN available through EncyclopeDIA (version - 0.6.8),[19] resulting mzMLs were searched using with a Uniprot human proteome (downloaded 6/2016) configured with the default settings: fixed cysteine carbamidomethylation, 10 ppm precursor and product tolerances, using y-ions, and assuming full tryptic digestion with up to 1 missed cleavage. EncyclopeDIA search results were then filtered to a 1% peptide-level FDR using Percolator 3.1 and then filtered again to a 1% protein-level FDR.

Selected reaction monitoring acquisition and processing.

Targeted data were acquired using an SRM method on a Thermo TSQ Altis triple-quadrupole instrument using a Thermo EASY nLC 1200 for separation. Approximately 1 µg of each sample was separated on self-packed 15 cm columns packed with 3 µm ReproSil-Pur C18 silica beads (Dr. Maisch, Ammerbuch, Germany) in a 75 µm inner diameter fused silica capillary (#PF360 Self-Pack PicoFrit, New Objective). Peptides were separated using a 60 min gradient from (2% acetonitrile in 0.1% formic acid to 23% acetonitrile in 0.1% formic acid). The gradient was followed by a wash for 10 min at 80% acetonitrile in 0.1% formic acid and a column re-equilibration at 0.1% formic acid for 10 min. For SRM assays, ions were isolated in both Q1 and Q3 using 0.7 FWHM and peptide fragmentation was performed at 1.5 mTorr in Q2 using calculated peptide specific collision energies. SRM data was acquired for 5 minutes scheduled around the predicted retention time, using both precursor and product ion scan widths of 0.002 *m/z* and a dwell time of 10 ms. Data were extracted and analyzed using the software package Skyline.[14,62] Chromatographic peak intensities from all monitored transitions of a given peptide were integrated and summed to give a final peptide peak area. Peptide areas were normalized to the total-ion-current for each sample. Normalization and figure generation were done with R.

Peptide selection for selected reaction monitoring.

Two peptides targeted in the Spellman AD CSF assay were selected if they also exhibited acceptable behavior in the GPF-DIA experiment (i.e. detected at a 1% FDR and with at least 5 concurrent transitions). These two peptides were selected based solely on the GPF-DIA data, regardless of their inclusion in the Spellman AD assay. From the GPF-DIA experiment peptides were selected if they were detected at a 1% FDR, had at least 5 concurrent transitions without interference, had a %CV of <30% across the 3 GPF replicates, and were ranked in the top 2 peptides per protein based on the summed product ion intensities. Peptides and transitions selected from both the previous assay and from the GPF-DIA data were used to create a single inclusion list. Scheduled windows of (5 min) and limiting the number of concurrent transitions to 200 resulted in 4 separate methods with transitions from the 13 PRTC peptides included in each method. For the pain protein SRM assay peptides were selected with the same criteria as the AD proteins from the GPF-DIA data. Scheduled windows of (5 min) and limiting the number of concurrent transitions to 200 resulted in 2 separate methods, with transitions from the 13 PRTC peptides included in each method.

2.3. Results

Data independent acquisition provides a list of sample specific detectable peptides.

A total of 18 μg of CSF digest was injected across 18 injections for the 3 replicate, 6 gas-phase fractionated (GPF) DIA library. From the three GPF-DIA 6,661 peptides corresponding to 1,176 proteins were detected across all replicates (Figure 2.1). Of the detected peptides 5,595 corresponding to 1,154 proteins were determined to be possible targets since they had at least 3 co-eluting transitions with no interference. Interference was determined by EncyclopeDIA as transitions with a peak apex not co-eluting with the other transitions. For all possible targets, precursor retention times and relative product ion intensities were extracted in Skyline for quantification. A percent coefficient of variation (%CV) was calculated for every peptide to determine signal stability across 3 days in a 4°C autosampler. For all possible target peptides the median %CV was 13.5%, with 4,089 peptides corresponding to 1,018 proteins having %CV ≤ 20 (Figure 2.2).

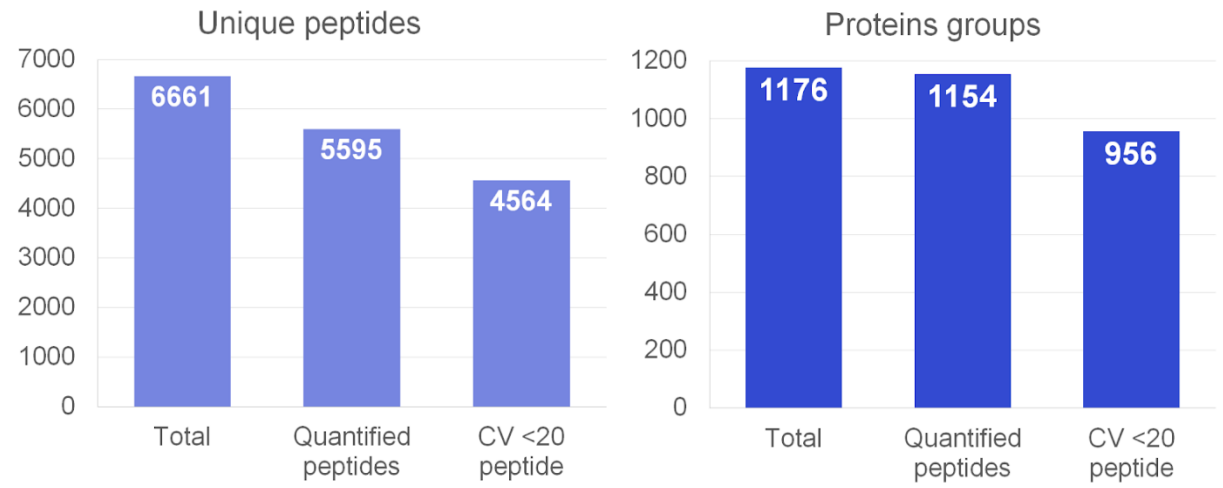


Figure 2.1. Number protein and peptide detections by DIA experiments. The total number of detections from 3 replicate experiments is plotted for both proteins and peptides. The number of proteins with at least one peptide with at least 3 interference-free product ion transitions.

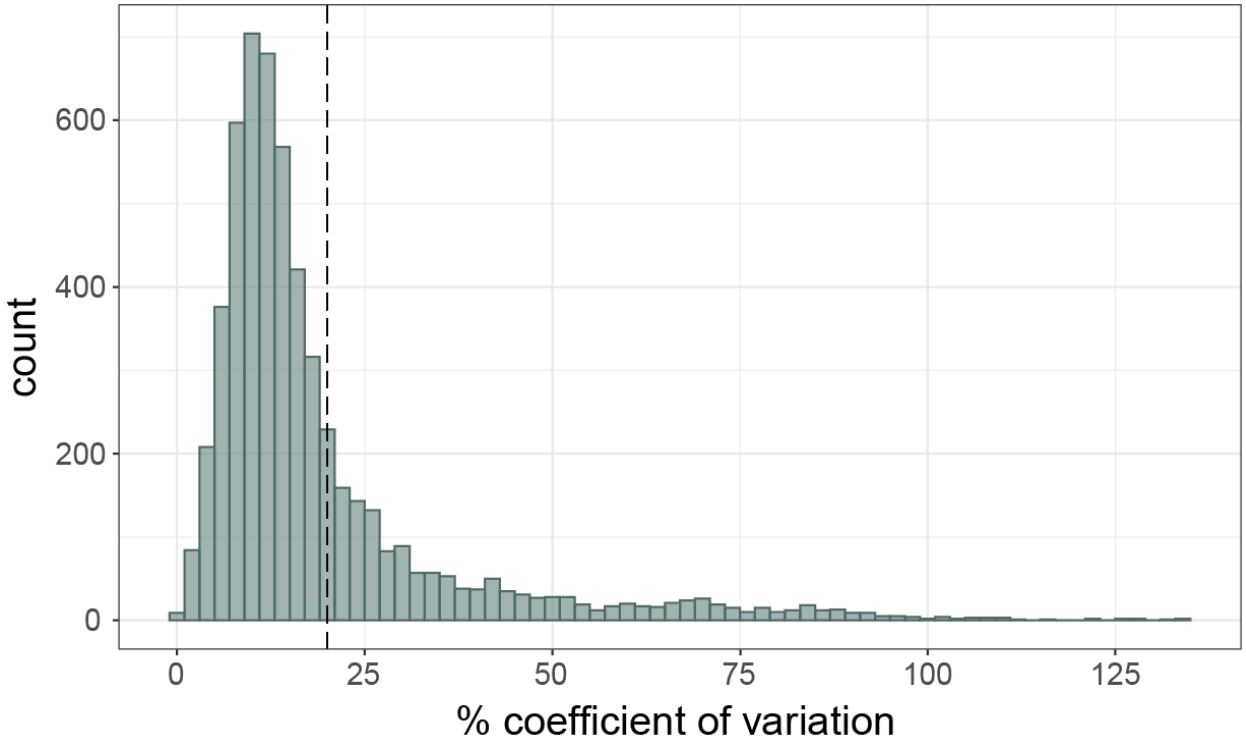


Figure 2.2. Peptide reproducibility across 3 gas-phase fractionated DIA experiments. The percent coefficient of variation for all peptides detected by DIA with at least 3 interference-free transitions. A median CV of 25.7 is observed across all.

Using peptide characteristics from DIA measurements to filter out poor responding peptides.

Using the peptide %CV calculations and information about product ion transition intensities, we can select which peptides we believe will respond well in a triple-quadrupole targeted experiment. Since the %CV is calculated from triplicate DIA experiments over separate days, the %CV captures peptide stability. For the 4564 peptides with calculated %CV less than 20, peptide summed area was used to rank peptides mapping to the same protein coding sequence. Transitions were ranked by measured intensity for each peptide. This information was then used to inform our selection of peptides for targeted analysis (Figure 2.3). For the following assays the selection criteria were the following: for proteins with more than 2 peptides only the top 5 peptides, as ranked by signal intensity, were considered. From the top 5 ranked peptides, the 2 peptides with the lowest %CV were selected to be included in our final assay. For those 2 peptides, the top 5 most intense product ion transitions were included in our targeted assay inclusion list. We chose 5 product ions in part because we observed that the median %CV for peptides with ≥5 interference-free transitions are below 20% (Figure 2.4).

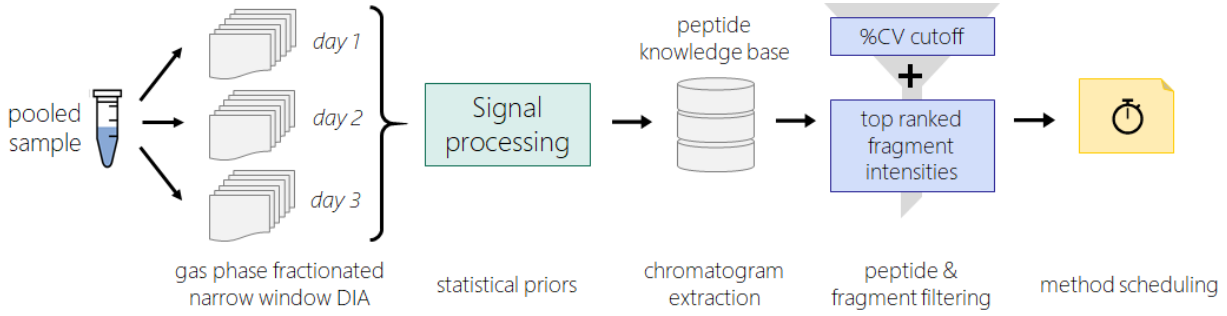


Figure 2.3. Workflow for selecting peptides and transitions for targeted assay. A pooled sample of cerebrospinal fluid from a neurodegenerative disease cohort was subjected to 3 gas-phase fractionated data independent acquisition experiments. Each gas-phase fractionated DIA experiment required 6 injections and spanned the 400-1000 *m/z* range in 100 *m/z* increments. Peptide detection, relative retention time, and precursor product ion pairs are determined through a peptide-centric search method. Search results and chromatographic information is extracted in Skyline to form a knowledge base used for quantitation and visualization. For peptides with 3+ transitions that co-elute and have no interference a %CV is calculated between the summed peptide area across the 3 DIA experiments. For proteins that have more than 2 peptides with %CV less than 30, peptides are ranked by summed area. For the top-ranking peptides, the 5 most intense transitions are selected for inclusion in targeted assay list.

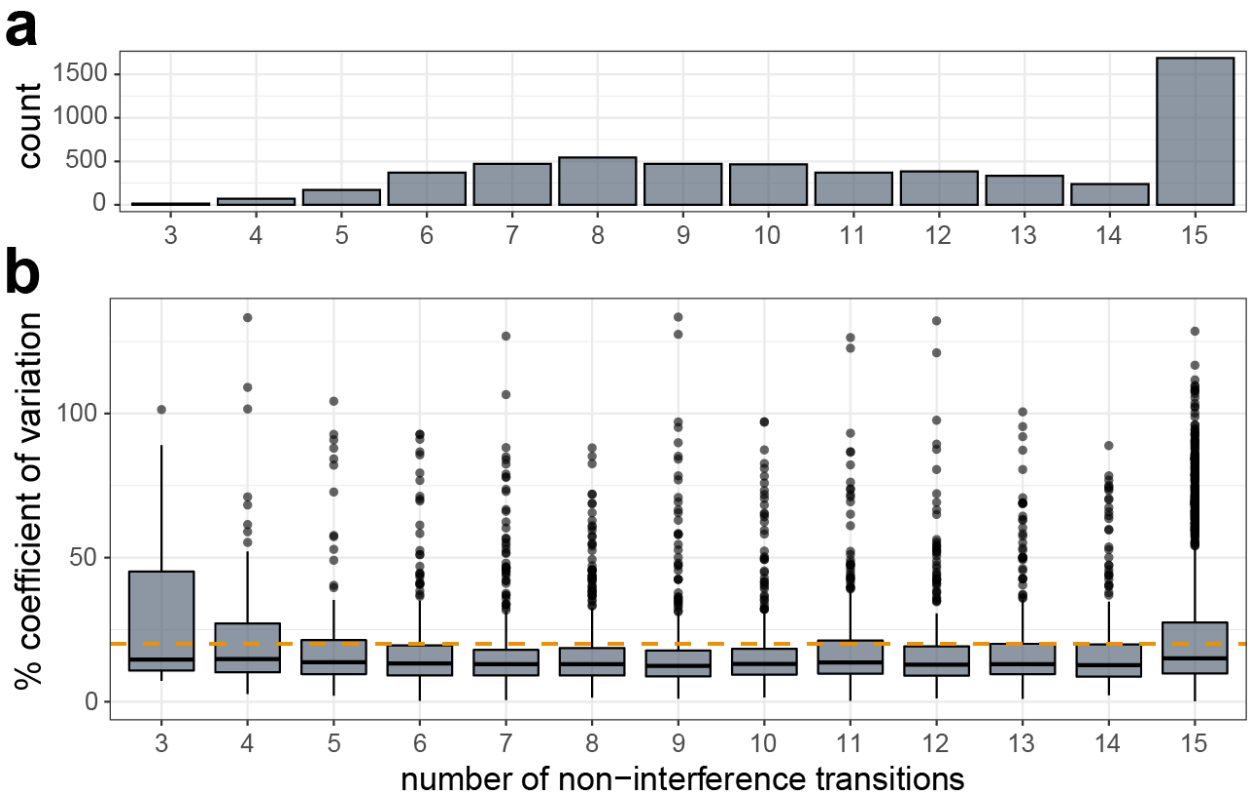


Figure 2.4. Reproducibility is improved with the number of product ion transitions detected per peptide. Using (EncyclopeDIA) software we determine which transitions are interference-free. For peptides that have at least 3 interference-free transitions we extract data in Skyline for quantitation. a) The number of interference-free product ions for each peptide detected by GPF-DIA. The majority (%) of peptides quantified contain more than 5 product transitions. b) The percent coefficient of variation (%CV) from 3 replicate gas-phase fractionated narrow window DIA experiments. The median %CV decreases with an increased number of interference-free transitions per peptide. Gold dashed line represents 20% CV.

Performance of peptides for previously targeted Alzheimer's disease proteins.

To test the application of peptide selection based on DIA data and their subsequent performance by SRM, we generated an assay of proteins previously included in a well characterized assay for use in Alzheimer's dementia research. From the 180 proteins targeted in a published assay by Spellman et al., peptides mapping to 170 were detected by GPF-DIA. Not all proteins included in the previous assay had at least 2 peptides detected by GPF-DIA, so we limited our comparison to 100 proteins that did have at least 2 peptides. For each protein group, 2 peptides were selected based on our criteria; lower than 20% CV and top 2 ranked peptides by summed

product ion intensities. For those peptides the top 5 most intense product ion transitions were added to our targeted assay inclusion list. Approximately 75% of the peptides we selected were unique to our assay, with about 25% overlapping with the previously published assay (Figure 2.5). For the proteins selected for comparison, transitions used in the previous assay were measured by SRM in triplicate. Additionally, transitions from two peptides selected based on performance in our DIA experiments were selected for each protein and measured by SRM in triplicate. For the peptides we selected from DIA the performance is comparable with the previous assay, with a median CV of 3.6%, and 3.2% respectively (Figure 2.5).

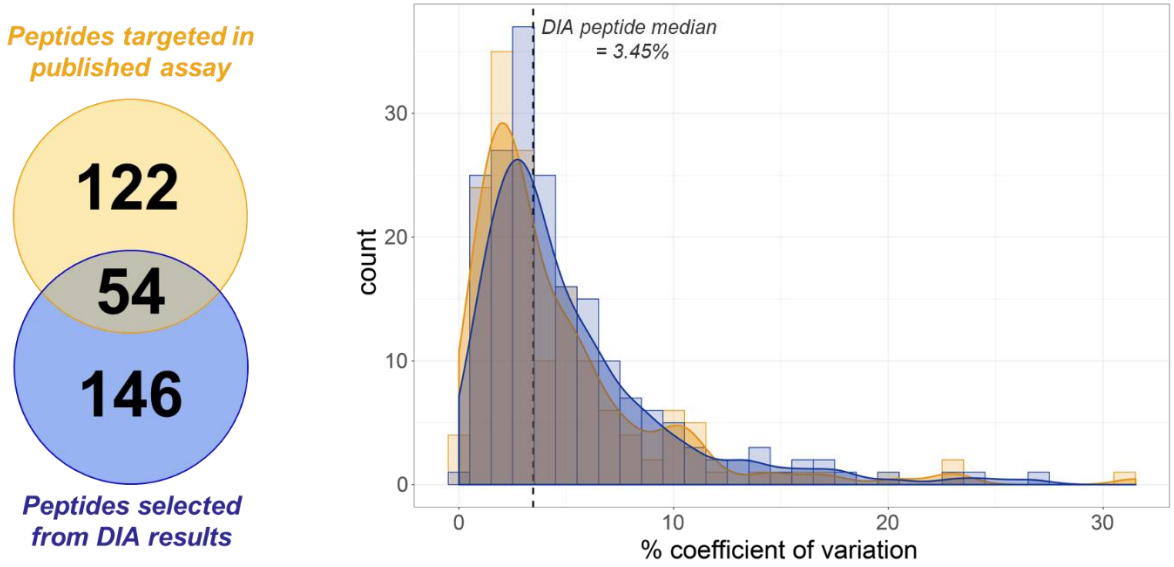


Figure 2.5. Peptides selected from narrow window DIA results perform similarly to previously characterized Alzheimer's disease assay peptide selections. For 100 proteins previously included in a targeted Alzheimer's disease assay, we targeted peptides either selected by the previous study (yellow) or selected using data from our DIA experiments (blue). We measured 5 transitions for peptides previously targeted, and from our selection. The percent coefficient of variation is calculated from 3 replicate injections of pooled cerebrospinal fluid measured by SRM and plotted by selection. Dashed line indicates the median %CV of peptides selected from the DIA results.

DIA data can be re-queried to select peptides for additional protein assays.

One benefit of acquiring DIA data to inform our assay generation is that we obtain valuable peptide information for all peptides detected. This means we can use this data for generating multiple different targeted assays depending on the proteins of interest in any given biological question or application. To demonstrate our ability to generate a separate assay for a different

subset of proteins, we take our same CSF peptide knowledge base data and go through our workflow for proteins previously described as being differential in CSF sampled from sufferers of chronic pain. From the 87 proteins found to be differential in the previous study, we detect peptides for 85. Of those proteins, 67 contain at least 2 peptides with at least 5 co-eluting and interference-free transitions. For targeted assay we selected the best performing 2 peptides per protein based on %CV and relative intensity ranking, leaving us with 134 target peptides and 670 target transitions. For all peptides measured by selected reaction monitoring we find a median %CV of 6.2 across 3 pooled CSF sample acquisitions (Figure 2.6).

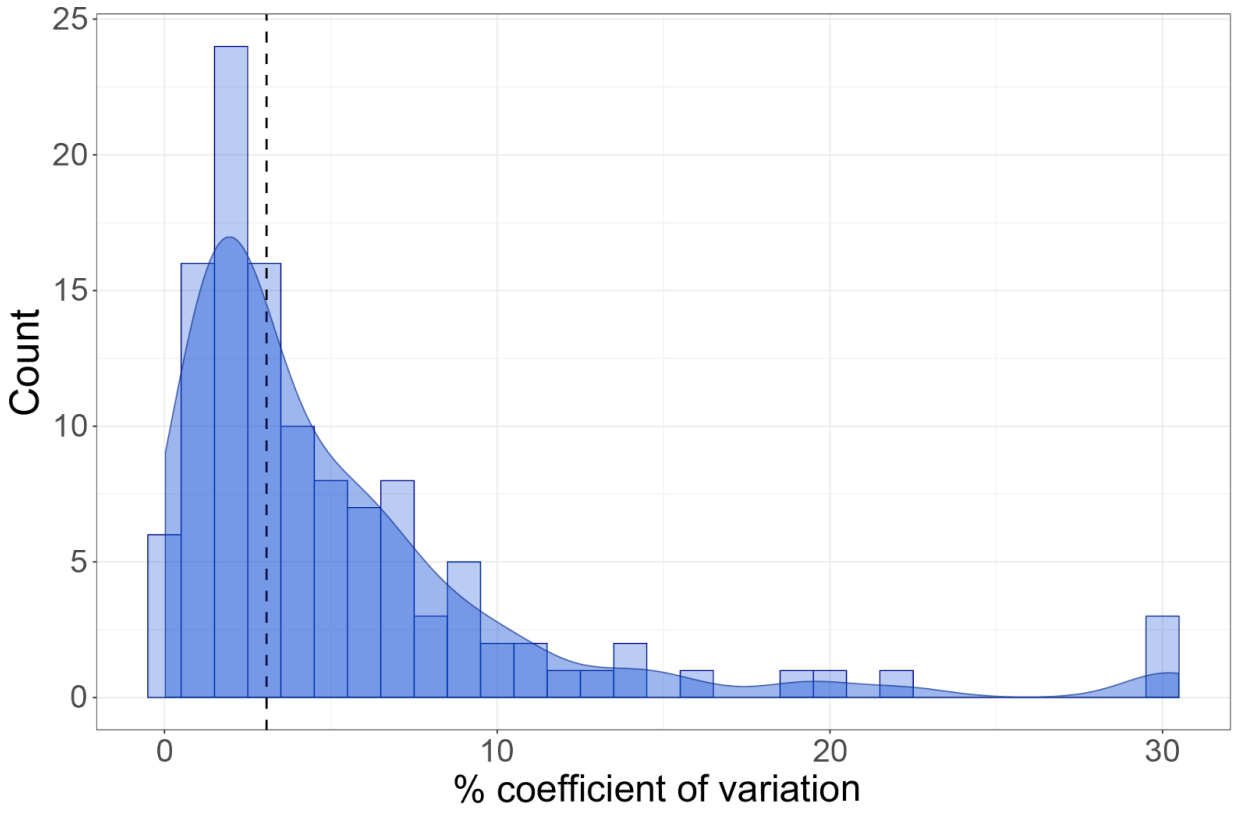


Figure 2.6. Additional assays generated using the same DIA experiments. Peptides for targeted triple quadrupole measurement were selected for 67 previously described pain proteins using information about transition signal and chromatographic performance from DIA. For the 134 peptides measured in triplicate the median %CV of 6.2.

2.4. Discussion

Here we demonstrate the ability to generate a triple quadrupole SRM assay with limited instrument time, minimal technician time, little reagent cost, and relatively small amounts of sample (less than 10 ug of protein). Peptides and transitions with high intensity, good chromatographic performance, and good precision by GPF-DIA were selected for inclusion in an SRM assay. The utility of this GPF-DIA to SRM workflow was demonstrated by reproducing a well characterized, previously published SRM assay for AD based on the DIA data. The re-use of the GPF-DIA was demonstrated through the generation of a separate reproducible assay for the quantification of proteins that are of interest as markers of chronic pain.

This workflow is especially efficient because DIA data is collected using a representative, pooled sample of interest. From this data the chromatographic performance of each detectable peptide in the biological matrix of interest can be rapidly evaluated, which is valuable information for our selection of targets. By sampling the 400-1000 *m/z* range with 2 *m/z* windows we are able to determine which peptide species are detectable in our sample type of interest with the ability to distinguish between different sequences with relatively similar precursor masses.[19,20] The inclusion of indexed retention time standards in the sample analyzed by DIA allows for transfer of relative retention time to different liquid chromatography setups and facilitates scheduling of a SRM assay if necessary.[60] This step also highlights the importance of measuring peptides in the background matrix of interest for the final targeted assay because peptide retention time is dependent on the matrix.[63,64]

The use of DIA methods has increased in recent years, leading to the accumulation of a lot of data that could be used for informing the development of targeted assays. These existing datasets can often be implemented to provide information regarding which peptides are likely to be of interest for a targeted reanalysis of these proteins. Narrow window DIA can be acquired as experimental libraries as a part of larger DIA studies as described by Searle et al. and Pino et al.[20,61] This workflow provides us with quantitative data for comparisons in discovery cohorts, which can further inform our selection of targets. We show that the information gained from one set of DIA experiments can be mined for multiple purposes, making it a valuable resource for future studies. Previous work from Pino et al demonstrate the use of DIA for more extensive characterization of peptide quantification through matrix matched calibration curves.[50] This could be included as an additional step in determining suitable targets for SRM assay development.

Triple quadrupole assays are a valuable method for the rapid and robust, targeted analysis of select analytes. The selection of peptides and target transitions for analysis has remained a bottleneck in assay development. This step often makes use of prior data-dependent acquisition (DDA) information followed by multiple rounds of validation experiments to determine which peptides perform well by selected reaction monitoring (SRM), or it relies on previous SRM assays which will be limited in measured sample types or targets. The workflow presented here greatly reduces the upfront time and effort required to generate a well-performing assay. By acquiring replicate gas-phase fractionated DIA and refining targets in the Skyline environment, we were able to pick target peptides and transitions with good intensity, chromatographic performance, and precision in a triple quad assay with only several days of instrument and data processing time.

Chapter 3. FINDING MOLECULAR SIGNATURES OF ALZHEIMER'S DISEASE WITH DIA

This chapter is adapted from the following work:

Merrihew GE, Park J., Plubell, D.L., Searle B.C., Keene C.D., Larsen E.B., Bateman R., Perrin R.J., Chhatwal J.P., Farlow M.R., McLean C.A., Ghetti B., Newell K.L., Frosch M.P., Montine T.J., MacCoss M.J. A peptide-centric quantitative proteomics dataset for the phenotypic assessment of Alzheimer's disease. Sci Data. 2023 Apr 14;10(1):206. doi: 10.1038/s41597-023-02057-7.

Alzheimer's disease (AD) is a complex, age-related disease that continues to increase in incidence worldwide. Although some cases are caused by known pathogenic variants, most are sporadic, with uncertain etiology, and commonly complicated by comorbid diseases. To better understand the molecular landscape of AD we constructed a unique brain sample set free from neuropathologic comorbidities. These samples consist of autosomal dominant AD dementia, sporadic AD dementia, cognitively normal cases with high AD neuropathologic change (ADNC) - a condition considered resilient to AD, and cognitively normal cases with no or low AD neuropathology. Through quantitative mass spectrometry proteomics, we obtained measures of peptides mapping to Aβ and Tau, capturing differences in the histopathologic burden of disease. Correlations of these peptide abundance differences within sporadic AD cases in the superior and medial temporal gyrus indicate the presence of two separate sporadic AD dementia phenotypes. Additional differences are characterized at a peptide level between sporadic AD dementia and autosomal dominant AD dementia in the inferior parietal lobe and caudate nucleus. These results indicate both shared and distinct molecular features in AD influenced by cause and age of onset of disease.

3.1. Introduction

Alzheimer's disease (AD) is a major global public health problem. 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.[65] AD dementia is clinically complex, with a range of demographics, progression, neurologic features, and neuropathologic features. AD dementia is fundamentally characterized neuropathologically by the accumulation of amyloid-β plaques and tau neurofibrillary tangles, and clinically through cognitive impairment and dementia.

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.

A small proportion of AD patients have a known genetic cause that is autosomal dominant in nature. Almost all known pathogenic variants occur in the genes coding for presenilin-1 (PSEN1), presenilin-2 (PSEN2), or amyloid precursor protein (APP), all of which are involved in the amyloidogenic pathway.[66–68] While these variants are commonly dominantly inherited, de novo pathogenic variants have also been described.[69–71] Pathogenic variants in these genes result in the aberrant processing of APP to amyloid-β, likely due to a change in binding site affinity or specificity, and somehow triggering regionally stereotypical neurofibrillary degeneration.[68,72] However, the accumulation of amyloid-β peptides due to β-secretases and γ-secretases cleavage also occurs in AD with no known genetic cause, called sporadic AD.[66,73] Despite the shared neuropathologic hallmarks between autosomal dominant AD (ADAD) and sporadic AD (SAD), there are often differences in onset and progression that are still not well understood.[73–75] Far and away, the strongest genetic risk for SAD is the ε4 allele of the apolipoprotein E gene (*APOE*).

AD is a highly age-related disease, where the incidence of disease increases with age.[76] The actual age of onset can vary, with a small proportion having symptom onset before the age of 65. An additionally small proportion of these early-onset cases are individuals with autosomal dominant causal variants, but many have some familial component with no clear genetic cause, and others have both no known family history and no known genetic cause.[77] Conversely, the majority of sporadic cases have noticeable symptom onset after the age of 75, and individuals can have some degree of family history without clear causal genetics.[76] 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-predominant, age-associated TDP-43 encephalopathy (LATE), and/or hippocampal sclerosis.[78–80] Since sporadic disease onset occurs at a later age, the incidence of these additional comorbid pathologic changes increases. For individuals with sporadic onset earlier in life, there is less incidence of comorbid pathologic changes.

Existence of forms of AD with known genetic causes or risk, without known genetic underpinnings, and with pathologic comorbidities, highlight the potential for multiple molecular drivers and perhaps multiple pathogenic pathways involved in the most common form(s) of cognitive impairment and dementia. To better understand the molecular signatures of AD we constructed a unique human brain sample set with ADAD dementia, SAD dementia, high cognitive function (HCF) with high AD histopathologic change (resilient to AD or RAD), and HCF with no/low ADNC (healthy control or HC); all cases were selected pathologically to exclude comorbidities so our data reflect only the contributions of AD. Using data-independent acquisition mass spectrometry proteomics we examined differences between AD diagnosis at a peptide and protein level in four brain regions sampled primarily by rapid autopsy.[81] Peptide measurements significantly correlated with known pathological markers of disease separates SAD into two subgroups based on the superior & middle temporal gyrus proteome. These subgroups are distinct primarily in their age of death and their last cognitive function assessment, likely indicating a difference in the severity of disease. Additional differences between ADAD and SAD are described in the inferior parietal lobule and caudate nucleus proteomes, and comparisons across brain regions demonstrate similar signatures of disease in both cerebral cortical regions. This study highlights the use of characterizing the proteome of AD to further understand the differences in disease etiologies.

3.2. 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 most recent 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 as healthy controls (HC) and

those with intermediate or high ADNC were designated resilient to AD (RAD). All individuals diagnosed with AD had intermediate or high level ADNC and were further subclassified as sporadic AD (SAD) or autosomal dominant AD (ADAD) caused by a variant in *PSEN1, PSEN2*, or *APP*. Sporadic AD cases were from the ACT study and the University of Washington (UW) AD Research Center (ADRC), and ADAD cases were from the UW ADRC and the Dominantly Inherited Alzheimer Network (DIAN). Cases were excluded with any neuropathologic changes of Lewy body disease (LBD) or LATE other than involving amygdala, territorial infarcts, more than 2 cerebral microinfarcts, or hippocampal sclerosis. Tissue from hippocampus obtained by coronal section at the level of lateral geniculate nucleus, superior and medial temporal gyri (SMTG), inferior parietal lobule (IPL), and caudate nucleus at the level of the anterior commissure was cryopreserved. Time from death to cryopreservation of tissue, postmortem interval (PMI), was <8 hrs. in all cases except for those with ADAD obtained through DIAN.

Ethics oversight

The brain tissue samples were pre-existing and were collected under protocols approved by the Institutional Review Board (IRB) at University of Washington, Kaiser Permanente Washington, or Stanford University prior to the start of this project. The UW and Stanford Human Subjects Division deems the use of pre-existing de-identified samples exempt from full IRB review and, thus, treated this project as non-human subjects research.

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 AD", 4 "Autosomal Dominant AD", 2 "Resilient to AD", and 3 "Healthy control" samples. Metadata for the samples from the SMTG, Hippocampus, IPL and Caudate brain

regions is provided in Supplementary Tables 1-4 available on Panorama Public in the "Supplementary Data" subfolder.[82] For each region the metadata includes sample batch, age, sex, post-mortem interval, *APOE* genotype, cognitive status, study of origin, and consensus Braak stage and CERAD score.

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 μg 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 minute, 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 an 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 on Panorama Public.

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[17]. 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 for each replicate was exported from Skyline using the custom reporting capabilities of the document grid.[14]

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.[19] 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:

$$F_i = log_2 A_i - [log_2 A_m - log_2 A_j] \quad (3.1)$$

Where A_i = sum of product ion transition area for peptide *I*, A_m = median of areas within LC-MS run *m*, A_j = 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, and by the reproducibility of those peptide abundances across replicate samples. 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) (Supplementary Figure 4 available on Panorama Public) 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 the other biological replicates.[18,19,83] 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.[5] 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. 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.[83]

Statistical analysis

Peptide correlations with Aβ 17-28 for each region were calculated using a Spearman rank based correlation, and significance determined as a Bonferroni adjusted p-value <0.05. Differential abundance testing was performed using linear mixed model analysis (limma) with either the 4 clinical conditions or the 5 conditions generated from separating SAD into two subgroups, and performed on the peptide and protein level. Significance was determined as a Benjamini-Hochberg adjusted p-value <0.05. Hierarchical clustering was performed with Ward linkage on z-score normalized peptide abundances. Principal component analysis was also performed on z-score normalized peptide abundances. Comparison of case characteristics between SAD subgroups were performed using two tailed t-tests. Enrichment analysis of gene ontology terms was performed using the enrichR tool,[84] with biological process terms reported based on their adjusted p-values and low-overlap between terms (less than half of genes co-occurring with a similarly significant term).

Exome sequencing and variant detection

Sporadic AD samples (specifically cases with SMTG tissue available in this cohort) were sequenced by Psomagen using a SureSelect V8-Post library kit and SureSelectXT Target Enrichment on an Illumina NovaSeq, multiplexed paired-end 151 base pair. The sequencing library was version C2 (Dec 2018). Sequence alignment and variant calling was performed for in *APP*, *PSEN1*, *PSEN2*, *SORL1*, *ABCA17*, *TREM2*, and *ABI3* genes by the NW Genomics center using the Seqr platform.

Data Records

The Skyline documents, raw files for quality control, DIA data and supplementary data are available at Panorama Public. ProteomeXchange ID: PXD034525. DOI: https://doi.org/10.6069/wefm-vv52.

DIA data is available in 5 different categories based on the level of post-processing (Figure 3.1e) 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. An interactive dashboard is available for the SMTG and Hippocampus data.

Balanced and controlled experiment design

We 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 3.1). 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 5 available on Panorama Public).

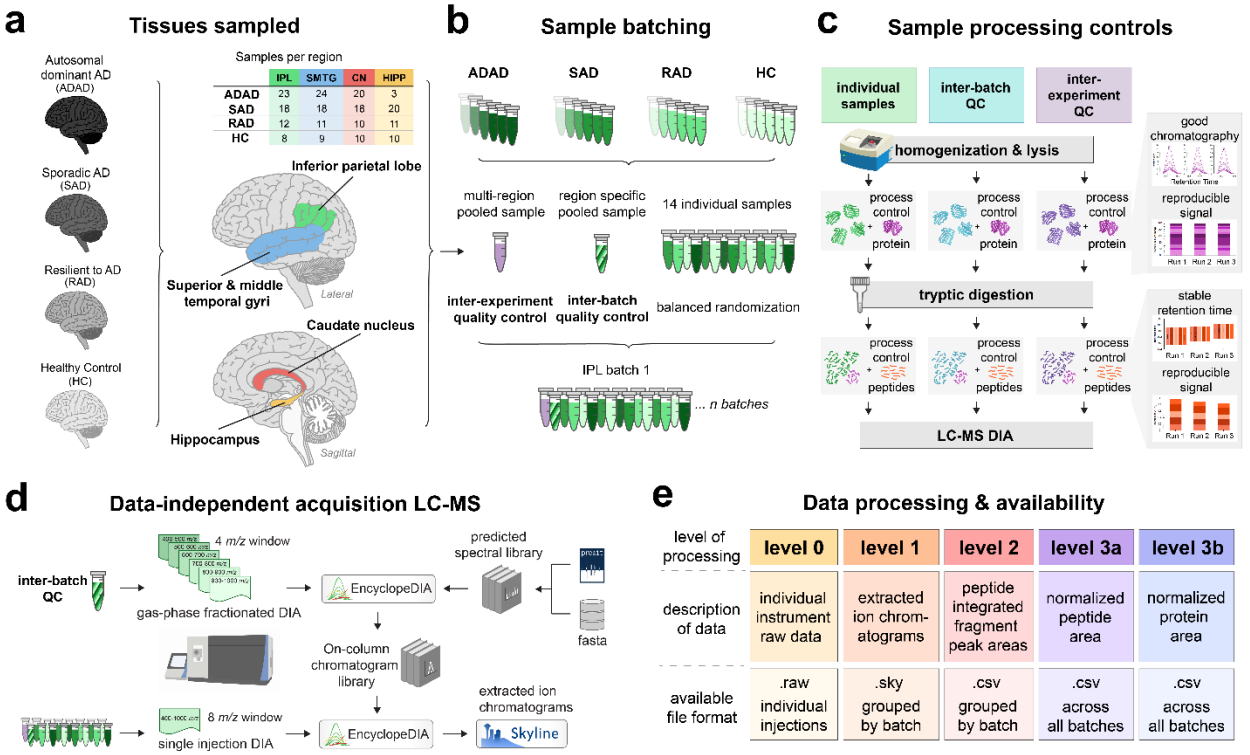


Figure 3.1. Experimental scheme for the collection of the proteomics data using data independent acquisition-mass spectrometry. a) Brain tissue sections from 4 regions were analyzed for all groups. b) Samples were prepared and analyzed in batches of 16, with 14 individual samples per batch selected by balanced randomization. Each batch contained an inter-brain region quality control sample generated from pooling portions of several individual samples from across all 4 brain regions sampled. Each batch also contained a brain region specific 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.

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 external 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 RAD samples, 3 HC samples, 3 ADAD samples and 5 SAD 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 external control was an inter-brain region quality control (“HAD” samples) and composed of a homogenate of a mix of cerebellum and occipital lobe tissue which we had ample material available to use throughout all our brain tissue experiments. The cerebellum and occipital lobe external control is distinct from the rest of the brain tissue regions in the experiment, but this should not hinder the interpretation of the experimental results, as this control is only monitoring the reproducibility of our entire system. The same pool of inter-brain external control was prepared and run in every batch across all brain regions for the entire experiment.

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 3.1b). 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.

3.3. Results

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. 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. The distribution of peptide abundances is aligned with median normalization, as demonstrated with the SMTG data as well as all brain regions (Supplementary Figures 1 and 2 available on Panorama Public).

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%. 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.

Amyloid precursor protein is increased in abundance in ADAD compared to SAD, RAD, and HC.

We quantify 22 unique peptides that map to amyloid precursor protein, with 7 detected in all four brain regions. Peptides span the APP sequence, including peptides mapping to the E1, KPI, E2, JMR, and Aβ regions. ADAD cases overall have higher APP peptide abundance in hippocampus, SMTG, and caudate regions (Figure 3.2). The protein abundance value from all peptides mapping to APP shows a clear trend across these sample groups (Figure 3.2). The increase in APP peptide abundance may be further separated within the ADAD group; with individuals with a PSEN2 variant having a generally higher abundance, while not statistically significant (Figure 3.3). There does not appear to be a consistent trend in either APP or Aβ 17-28 abundance by *APOE* allele within each condition (Figure 3.4).

We observe two peptides with sequence mapping to the amyloid-β region (Figure 3.2). These peptides have very high relative intensity compared to the other APP peptides, and have expected differences across sample groups, with the HC having the lowest abundance, followed by the RAD, the SAD, and ADAD. The amyloid-β tryptic peptide (residues 17-28: sequence LVFFAEDVGSNK) in all brain regions shows more separation between conditions with the same CERAD score (using NIA-AA consensus scoring), compared to between CERAD scores in the same condition (Figure 3.2). The distinction between conditions is most apparent in the C3 scored individuals, with ADAD having higher abundances compared to the SAD or both RAD and HC.

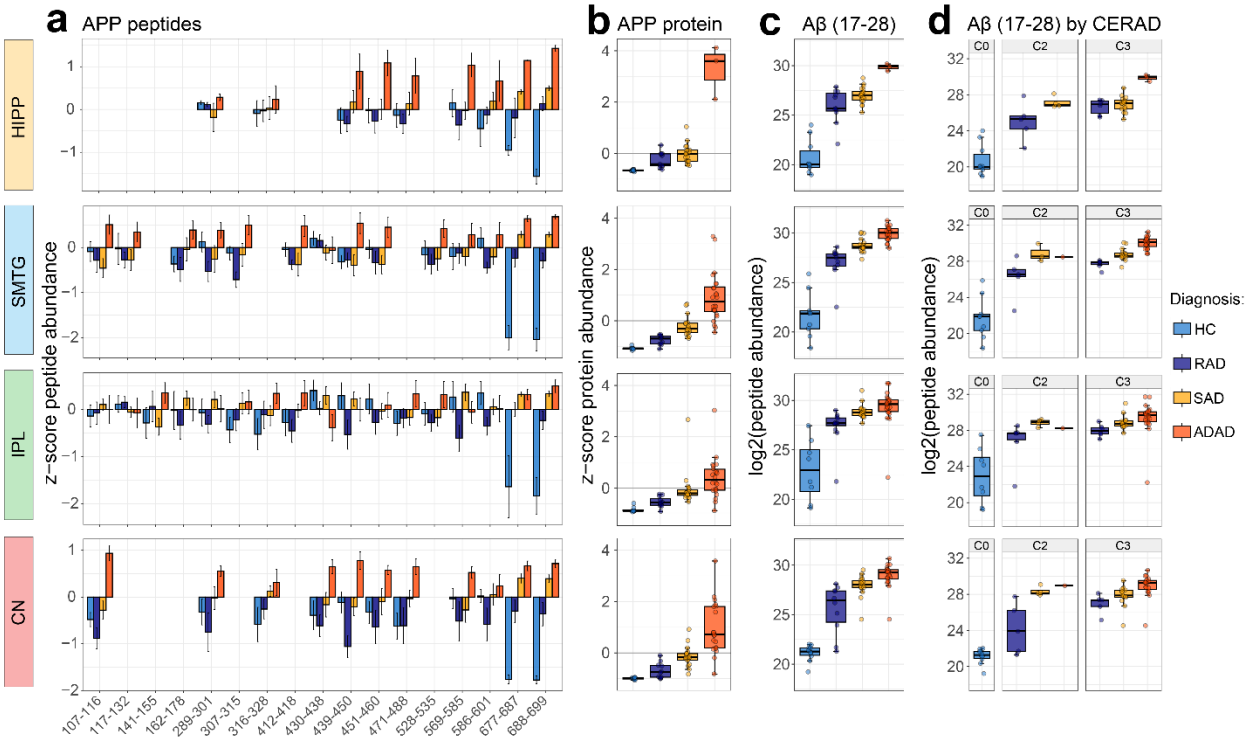


Figure 3.2. Amyloid precursor protein (APP) abundances are highest in autosomal dominant AD (ADAD) cases. A) Tryptic peptides mapping to APP are z-score normalized across samples and plotted by mean +/- se per condition from the c-terminus sequence to the n-terminus sequence. B) Calculated protein abundance z-score normalized for comparison to peptide values. C) The log2 peptide abundance for the peptide LVFFAEDVGSNK mapping to residues 17-28 of Aβ (residues 677-687 of APP). D) The log2 peptide abundance for Aβ 17-28, stratified by case CERAD score. Condition is indicated by bar color; with healthy control (HC) far left, then resilient to AD (RAD), sporadic AD dementia (SAD), and ADAD.

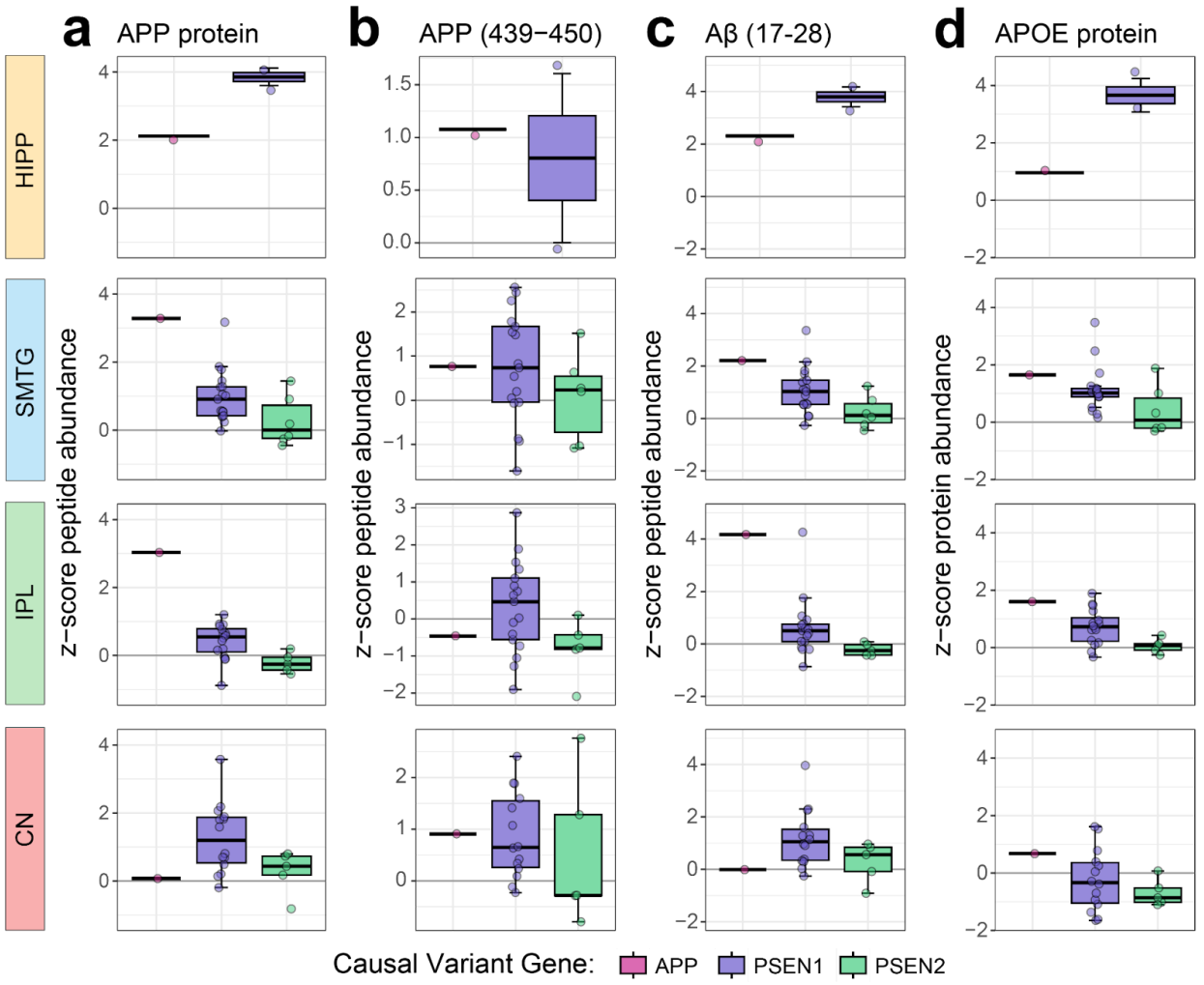


Figure 3.3. Protein and peptide abundance trends in ADAD separated by causal variant gene. A) Calculated APP protein abundance modestly increased in the cases with a causal variant in *PSEN1* compared to *PSEN2*. B) The same trend of an increase in the mean abundance is observed in APP peptide 439-450 in the E3 extracellular APP domain. B) Likewise, the same trend is observed in the Aβ 17-28 peptide, as well as observed in C) APOE protein abundance. (*APP* n=1, all regions; *PSEN1* Hippocampus n=2, SMTG n=17, IPL n=13, caudate n=14; *PSEN2* SMTG n=6, IPL n=6, caudate n=5)

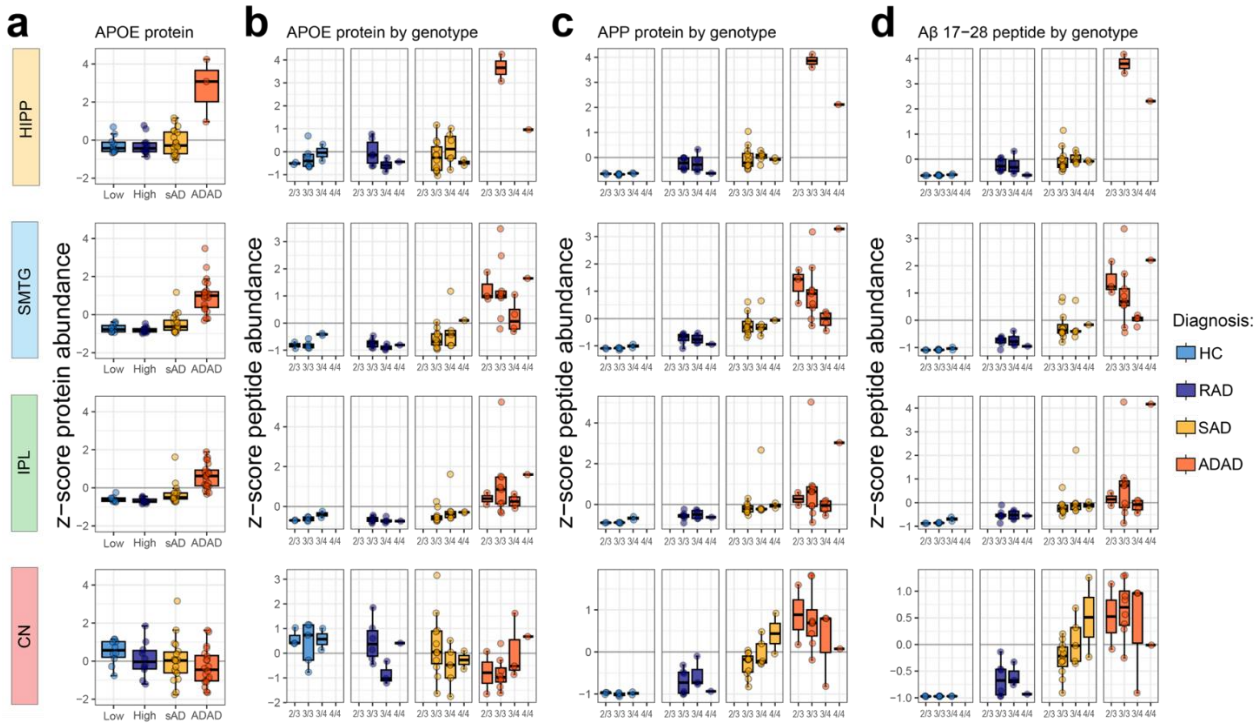


Figure 3.4. Protein and peptide abundance trends in APOE and APP by *APOE* genotype. A) Normalized APOE protein abundance increased in autosomal dominant AD in all but the caudate nucleus region. B) The normalized APOE protein abundances for each condition separated by sample *APOE* genotype. C) Normalized APP protein abundance for each condition separated by sample *APOE* genotype. D) Normalized Aβ 17-28 tryptic peptide abundance for each condition separated by sample *APOE* genotype. Abundances are z-scored normalized across samples. Condition is indicated by bar color with healthy control (HC) far left, then resilient to (RAD), sporadic AD dementia (SAD), and autosomal dominant AD dementia (ADAD).

Tau hyperphosphorylated region and microtubule binding domain show related profiles across samples.

We quantify 34 unique peptides mapping to microtubule associated protein Tau (MAPT), with 22 found across all four brain regions. The quantified peptides span most of the 2N4R Tau sequence isoform MAPT sequence (441 aa, UniprotKB P10636-8); with peptides mapping to all variable domains. Overall, the abundance profiles are most similar between the SMTG and IPL, with some differences in the hippocampus and caudate (Figure 3.5). Most peptides mapping to the microtubule binding regions in R2-R4 (residues 275-368) have similar abundance profiles; HC has the lowest abundance, followed by RAD, SAD, and ADAD. The peptide mapping to R1 (260-267) has the same abundance trend in the Hippocampus and to a lesser extent in the SMTG, but does

not share the same abundance trend across conditions in the IPL or caudate. Peptides mapping to the proline-rich regions of P1 and P2 in the SMTG and IPL have the inverse abundance trends to the R2-R4 peptides, with the lowest abundances in ADAD, followed by SAD, RAD, and HC. Two c-terminus Tau peptides (386-395 and 396-406) also demonstrate abundance profiles more like the P1 and P2 regions, particularly in the SMTG and IPL regions. Three additional peptides are detected and quantified in several of the brain regions that do not map to the 2N4R isoform; instead mapping to the isoform referred to as PNS-Tau (Figure 3.6). These were found due to the inclusion of the PNS-Tau 758 aa sequence in the human proteome FASTA as the canonical MAPT sequence (UniprotKB P10636-1).

A protein-abundance value is calculated from all peptides mapping to Tau, showing a trend across conditions that matches observed abundances for some but not all the peptides (Figure 3.5). Generally, protein abundance follows an expected trend of increased Tau in ADAD, followed by SAD, and then both HCF groups. This trend is similar to that observed in peptides mapping to the MTBR R2-R4, as highlighted by the peptide IGSLDNITHVPGGGNK (354-360) in R4 (Figure 3.5). However, the singular protein abundance does not reflect the peptide abundances observed in the P1-2, and c-terminus regions, as highlighted by the peptide SGYSSPGSPGTPGSR (195-209) (Figure 3.5). The abundance of Tau peptide 354-360 across brain regions compared to the Braak stage for each case shows a slight separation between conditions within Braak stages. These differences are observed primarily between SAD and ADAD in cases with either Braak stage V or VI. The abundance of Tau peptide 195-209, with the inverse abundance profile, does not separate cases within Braak stage any better than Tau peptide 354-360 (Figure 3.6).

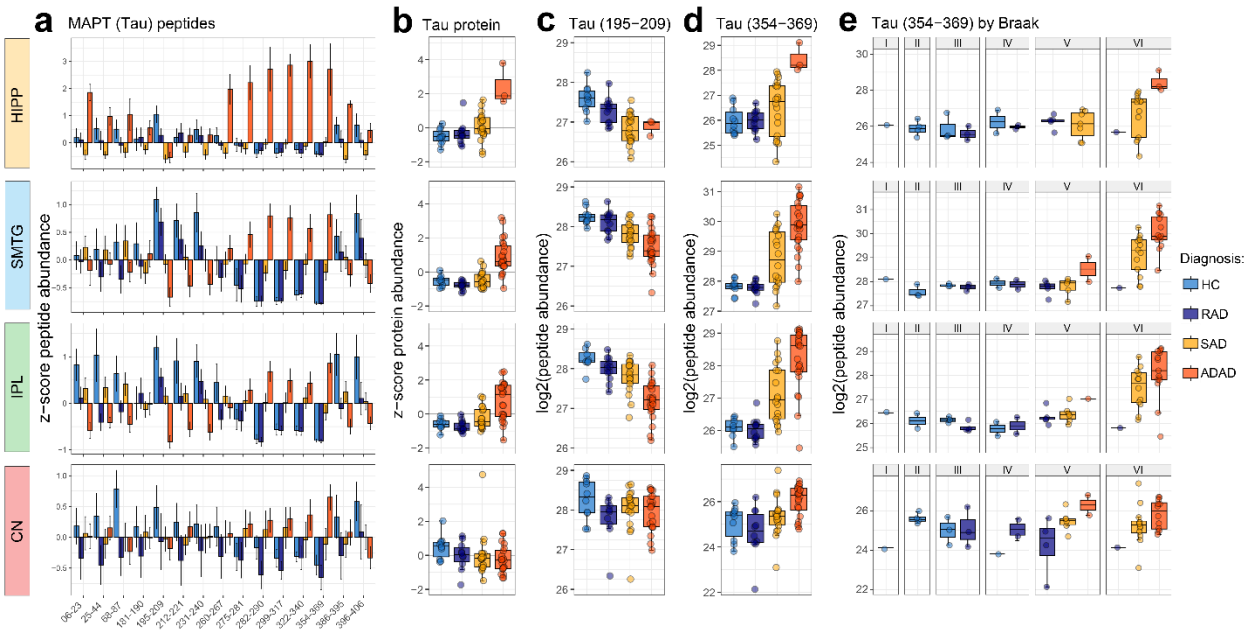


Figure 3.5. Tau peptide abundances vary by region and protein domain. A) Tryptic peptides mapping to Microtubule associated protein tau (TAU) sequence, found across all brain regions. Peptide z-score normalized abundance plotted by mean +/- se per condition from the c-terminus sequence to the n-terminus sequence of isoform 2N4R, UniprotKB P10636-8. B) Calculated protein abundance z-score normalized for comparison to peptide values. C) The log2 peptide abundance for the peptide SGYSSPGSPGTPGSR mapping to residues 195-209 in the proline-rich region of TAU. D) The log2 peptide abundance for the peptide IGSLDNITHVPGGGNK mapping to residues 354-369 in the microtubule binding domain of TAU. E) The log2 peptide abundance for TAU 354-369 but separated by case Braak stage. Condition is indicated by color of the bar; with healthy control (HC) far left, then resilient to (RAD), sporadic AD dementia (SAD), and autosomal dominant AD dementia (ADAD).

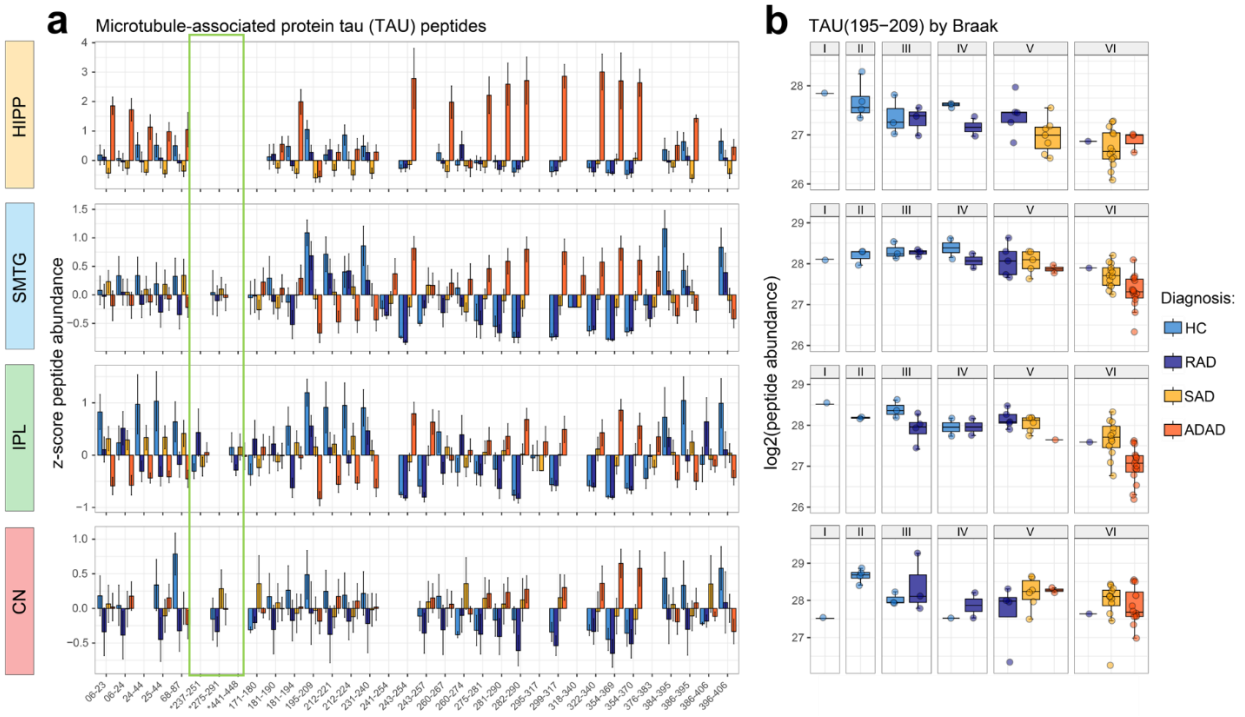


Figure 3.6. Additional TAU tryptic peptide measurements. A) All measured tryptic peptides mapping to microtubule associated protein tau (TAU) sequences, found across all brain regions. Peptide z-score normalized abundance plotted by mean +/- se per condition from the c-terminus sequence to the n-terminus sequence. Peptides with (*) and circled in green are labeled with their residue numbers in the 758 aa TAU sequence. B) The log2 peptide abundance for the peptide SGYSSPGSPGTPGSR (residues 195-209) in the proline-rich region of TAU but separated by case Braak score. Condition is indicated by bar color with healthy control (HC) far left, then resilient to (RAD), sporadic AD dementia (SAD), and autosomal dominant AD dementia (ADAD).

Peptides co-varying with amyloid-β tryptic peptides in SMTG.

The Aβ peptide 17-28 abundance varied by donor neuropathology across all brain regions and was selected as a brain region specific continuous proxy of ADNC. We identify other tryptic peptides with abundances associated with ADNC through correlation analyses with the Aβ peptide (Figure 3.7). In SMTG samples, 1,485 peptides mapping to 558 proteins were significantly correlated with the Aβ 17-28 abundance based on a Spearman's rank correlation with a Bonferroni corrected p-value < 0.05. Of the correlated peptides 50.5% (750) were positively correlated with Aβ 17-28 abundance - generally increased abundances in AD cases, and 49.5% (735) were negatively correlated - generally decreased abundance in AD cases. Of the proteins represented in the positively correlated peptides, the gene ontology enrichment analysis indicated an enrichment in terms related to protein degradation, immune response, and synapse related processes. Proteins

represented in the negatively correlated peptides are enriched in terms primarily related to mitochondrial function and metabolism.

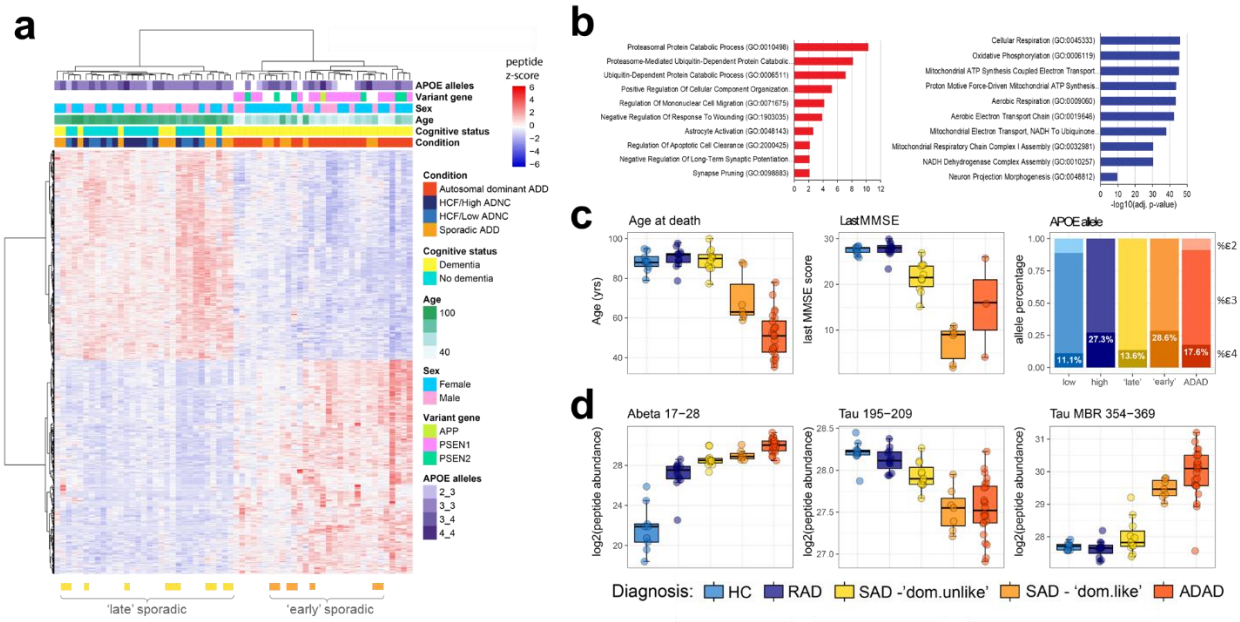


Figure 3.7. Tryptic peptides correlated with Aβ 17-28 in SMTG distinguish between subgroups of sporadic AD. A) Heatmap of significantly correlated peptide abundances with hierarchical clustering of both cases and peptides. B) Gene ontology enrichment analysis of positively and negatively correlated peptides. C) Characteristics for diagnostic groups with SAD separated into (n=7 'dominant-like', n=11 'dominant unlike'), with significant differences between the SAD subgroups for age (p=0.0049) and last MMSE (p=4.2e-05). D) differences in peptide abundance for select Aβ and tau tryptic peptides in diagnostic groups with SAD divided into subgroups.

For the peptides significantly correlated with Aβ 17-28 in SMTG, hierarchical clustering separates the cases into two major clusters, with all ADAD clustered and all HCF clustered. SAD cases separate across the two clusters, with 7 clustering with the ADAD ('dominant-like' SAD) and 11 clustering with the HCF ('dominant-unlike' SAD). The same separation of the sporadic samples in the SMTG can also be observed in proteome-wide analyses. Principal component analysis on the peptides correlated with Aβ 17-28 cluster the 'dominant-like' SAD as expected, but principal component analysis on all measured peptides also separates the 'dominant-like' from the 'dominant-unlike' SAD (Figure 3.8). Neither principal component analyses clearly separated the 'dominant-unlike' SAD from the high cognitive function samples. Trajectory analysis was also

performed on all the measured peptides in SMTG, generally separating the samples into 4 groups. This analysis also shows the same 7 ‘dominant-like’ SAD samples more closely associated with the ADAD samples, indicating that the overall proteome signature of those ‘dominant-like’ SAD are more like ADAD than the other 11 ‘dominant-unlike’ SAD samples in the SMTG.

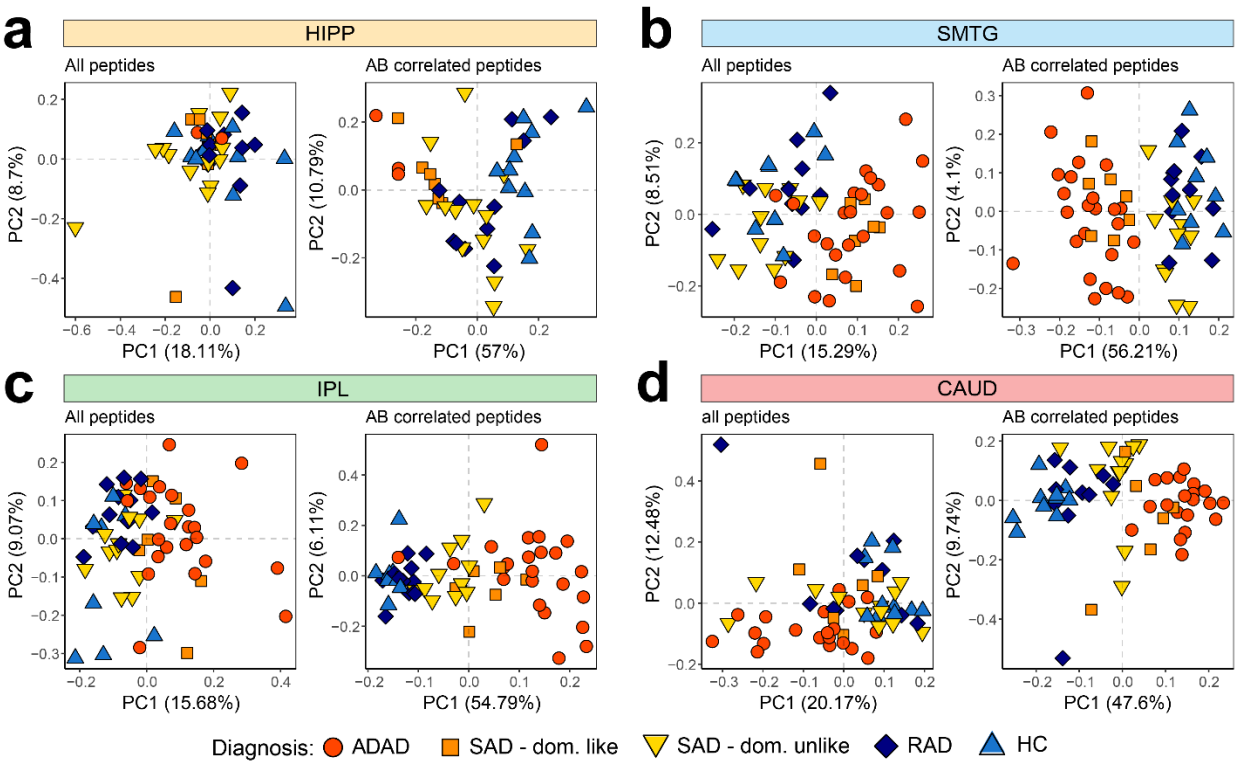


Figure 3.8. PCA of peptide abundances for each brain region. PCA for all samples, calculated either with all measured tryptic peptide abundances (left) and just Aβ 17-28 correlated peptide abundances (right), for A) hippocampus data, B) SMTG data, c) IPL data, and d) caudate nucleus data.

Since the hierarchical clustering, principal component analysis and trajectory analysis of the samples so clearly separated the SAD into two distinct clusters we investigated the features underlying this separation. Any artifacts of the batches used for sample preparation and acquisition were ruled out. From the clinical characteristics we found that the samples clustered with the ADAD samples were significantly younger at the time of death compared to those clustered with the HCF samples (p=0.005, Figure 3.7), but there is still some overlap in the age ranges. However, the last MMSE affiliated with each sample is also significantly different between the two sporadic groups (p=4.2e-05); with complete separation between those that cluster with ADAD and those

that cluster with the HCF. Since the post-mortem interval was longer for most of the ADAD cases, the post-mortem interval between the two SAD clusters was tested but found to be not significantly different (p=0.40) The APOE genotype of the SAD samples show an increased percentage of ε4 alleles in the 'dominant-like' SAD samples, but the cohort size limits meaningful interpretation of allele frequency. When looking at the peptide abundances for Aβ and Tau tryptic peptides the Aβ 17-28 is relatively similar between the 'dominant-like' and 'dominant-unlike' SAD, while both Tau 195-209 and Tau 354-369 are more different between the two SAD subgroups.

Based on the differences in proteome and age in the two SAD sample clusters we investigated whether there was a genetic misclassification of these samples. Exome sequencing was performed on all 18 SAD samples that had their SMTG tissue analyzed by proteomics. Sequence alignment and variant calling was performed for genes known to cause ADAD, and genes with likely causal variants for Alzheimer's disease. For the genes analyzed a total of 8 variants were called in 7 individuals, in *ABCA7*, *SORL1*, and *APP* (Table 3.1). No variants were found in either *PSEN1* or *PSEN2*. Most of the variants found were missense, while one was a synonymous variant and one resulted in a frameshift (p.Leu1403ArgfsTer7) in ABCA7.

Table 3.1 Variants called in genes implicated in AD. Sporadic AD cases were exome sequenced, with variants called for *APP*, *PSEN1*, *PSEN2*, *ABI3*, *ABCA7*, *SORL1*, and *TREM2*.

Group	Gene	Type	Allele count (gnomAD)	Allele frequency (gnomAD)	Clinical significance (ClinVar)	Age (yrs)	Sex	APOE genotype	StudyName
SAD – dom. like	*ABCA7*	frameshift	234	8.80E-04	Conflicting	88	Male	ε3/ε3	UW ADRC
	SORL1	missense	7	2.78E-05	NA	61	Male	ε4/ε4	UW ADRC
SAD – dom. unlike	**ABCA7*	missense	55	1.96E-04	NA	77	Male	ε3/ε3	ACT
	**ABCA7*	missense	55	1.99E-04	Uncertain significance	77	Male	ε3/ε3	ACT
	ABCA7	synonymous	1	3.59E-05	NA	86	Male	ε3/ε4	UW ADRC
	SORL1	missense	715	2.54E-03	Likely benign	91	Female	ε3/ε4	UW ADRC
	SORL1	missense	715	2.54E-03	Likely benign	100	Male	ε3/ε3	ACT
	APP	missense	28	1.11E-04	Uncertain significance	89	Female	ε3/ε3	ACT

To better understand the differences between the seemingly two subtypes of SAD, differential peptide abundance analysis was performed treating the subtypes as two distinct conditions. Comparisons were conducted across the 'dominant-like' SAD, 'dominant-unlike' SAD, ADAD, RAD, and HC. A total of 6,417 peptides mapping to 1,900 proteins have significantly different abundances between the 'dominant-like' and 'dominant-unlike' SAD

(Figure 3.9). Of these peptides about 35% (2184) were increased in the ‘dominant-like’ and mapped to proteins enriched in gene ontology terms related to endosomal processes, aggregation processes, and apoptotic cell processing (Figure 3.9). About 65% (4233) of differential peptides were decreased in the ‘dominant-like’ SAD, mapping to proteins enriched in gene ontology terms related to mitochondrial metabolism (Figure 4). Across all comparisons the largest number of differentially abundant peptides was found between ADAD and the ‘dominant-unlike’ SAD, followed by ADAD and HC, then ADAD and RAD (Figure 3.9). The fewest differentially abundant peptides were found in the comparisons between the HC and RAD, and the ‘dominant-unlike’ SAD compared to both the HC and RAD (Figure 3.9). There are a few differentially abundant peptides between the ADAD and the ‘dominant-like’ SAD. The abundances of the proteins those peptides map to show the peptide is generally moving in the same direction in both; increasing in the ‘dominant-like’ SAD and ADAD compared to the other groups but increasing even more in the ADAD.

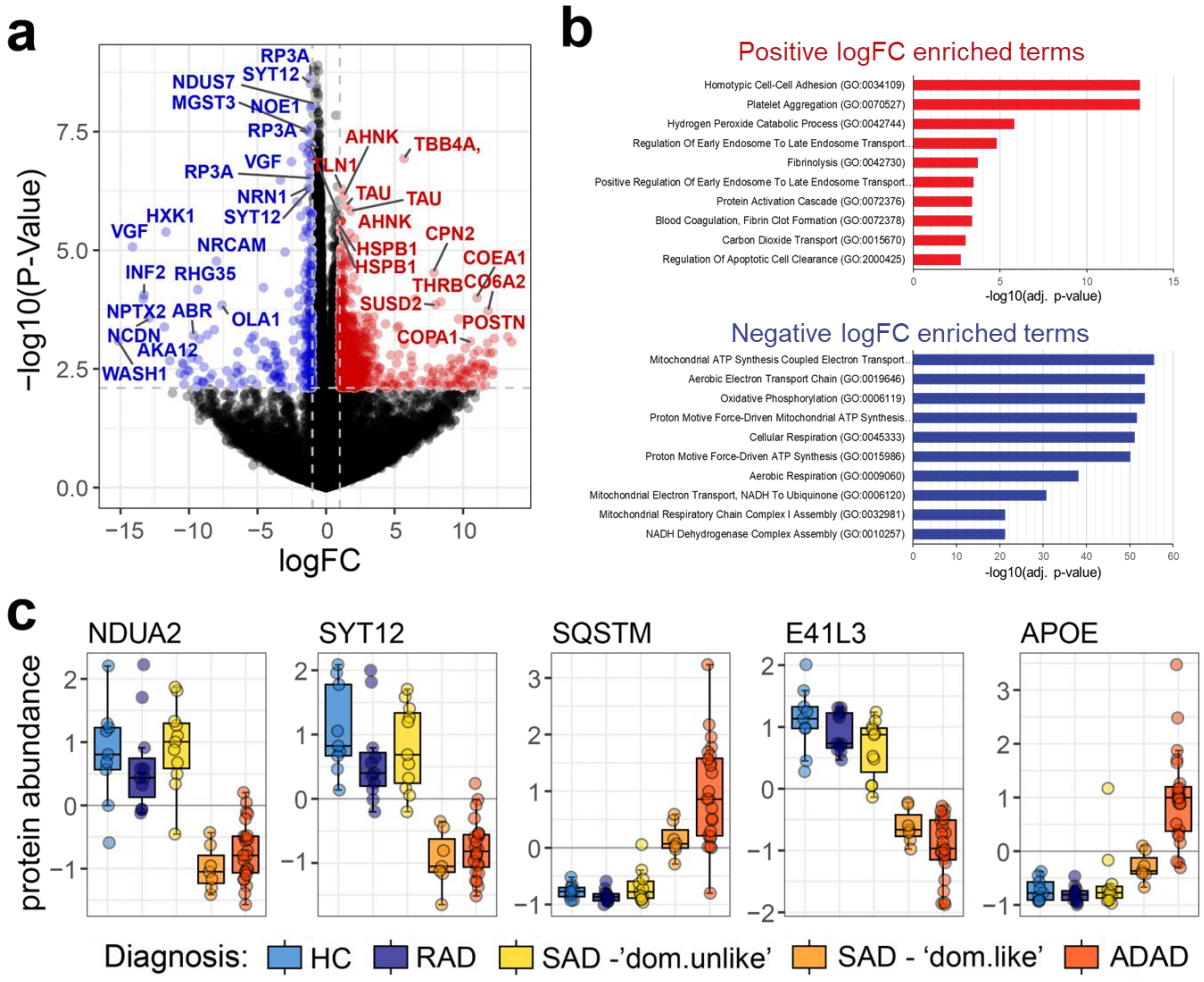


Figure 3.9. Changes in SMTG proteome across sample conditions. A) Tryptic peptides that are differential between 'dominant-like' and 'dominant-unlike' SAD, with red highlighting proteins significantly increased in 'dominant-like' SAD, and blue highlighting proteins significantly decreased in 'dominant-like' SAD. Peptides are labeled with their associated protein group. B) Enriched gene ontology functional analysis terms in the 'dominant-like' and 'dominant-unlike' differential peptide protein groups, with red representing enriched terms from the increased proteins and blue representing enriched terms from the decreased proteins in 'dominant-like' SAD. C) The z-scored protein abundances for proteins significantly altered between 'dominant-like' and 'dominant-unlike' SAD.

Peptides co-varying with amyloid-β tryptic peptides in IPL.

IPL proteomics data somewhat separates the ADAD cases from the HCF cases. Like the SMTG data, peptides significantly correlated with Aβ 17-28 more clearly separate the different conditions (Figure 3.10). A total of 301 peptides, mapping to 128 unique protein groups, have abundances significantly correlated with the Aβ 17-28 peptide abundance. 241 Positively

correlated peptides (100 unique protein groups) are generally increasing in the three AD groups compared to HC and are enriched in protein processing and axon and synapse related processes. 60 negatively correlated peptides (28 unique protein groups) are generally decreasing in the three AD groups compared to HC and enriched for cell structural organization and protein signaling related processes (Figure 3.10).

There are still differences between ‘dominant-like’ and ‘dominant-unlike’ SAD in TAU and ApoE peptides in the IPL, but the Aβ 17-28 correlated peptides in IPL do not stratify SAD cases in the same manner as in SMTG (Figure 3.10). This is also reflected in the low number of significantly different peptides between the ‘dominant-like’ and ‘dominant-unlike’ SAD subgroups in IPL. Instead, the major clusters are composed of most of the ADAD cases and the HCF cases, with a separate cluster of both some SAD and ADAD, and some SAD clustered with HCF cases. 10,575 peptides mapping to 2814 unique protein groups have differential abundance in the IPL based on the four diagnostic group classifications (HC, RAD, SAD, ADAD), with most significant differences found in comparisons with ADAD. 4,369 peptides mapping to 1129 protein groups had significantly altered abundance between SAD and ADAD, with peptides from APOE, TAU, ANK1, and DTNA increased in ADAD compared to SAD, and peptides from ABI2, PSD3, and GRIA2 decreased in ADAD compared to SAD.

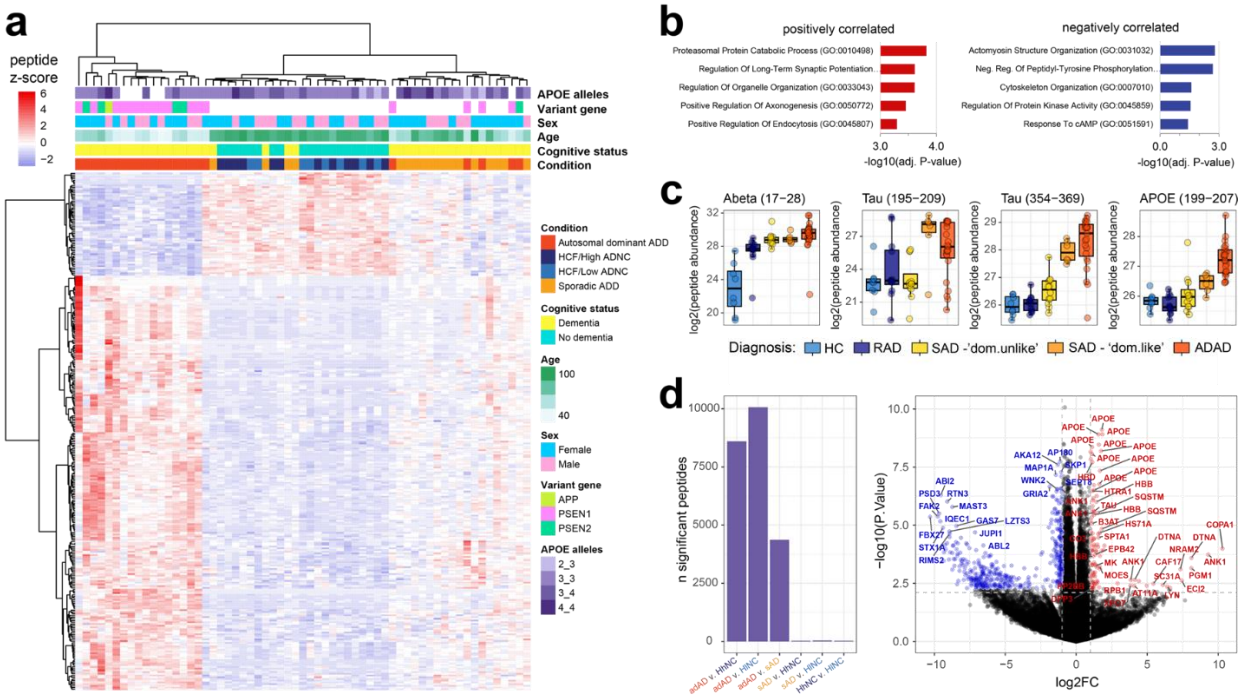


Figure 3.10. Tryptic peptides in the inferior parietal lobe (IPL) do not distinguish between SAD subgroups, but do distinguish between SAD and ADAD. A) Heatmap of significantly correlated peptide abundances with hierarchical clustering of both cases and peptides. B) Differences in peptide abundance for select Aβ, Tau, and APOE tryptic peptides in the split sporadic cases based on the SMTG clustering (n=6 SAD-'dominant-like', n=11 SAD-'dominant-unlike') separated by condition. D) Volcano plot of peptides significantly altered between 'dominant-unlike and 'dominant-like' sporadic AD, with red highlighting peptides increased in in 'dominant-like' compared to 'dominant-unlike, and blue highlighting those decreased in 'dominant-like' compared to 'dominant-unlike, labeled with the protein names the peptides map to.

Peptides co-varying with amyloid-β tryptic peptides in caudate nucleus.

45 peptides have abundances that significantly correlate with Aβ 17-28 in the caudate nucleus samples. Hierarchical clustering of the correlated peptides separates samples mostly by cognitive status, with most HCF (both HC and RAD) clustering together, and most SAD and ADAD clustering together (Figure 3.11). The 45 correlated peptides mapped to 31 unique protein groups, including both MDK and NTN1 with very similar peptide abundances profiles to Aβ 17-28. Signatures of 'dominant-like' and 'dominant-unlike' onset SAD are not detected in the caudate nucleus samples, either in the correlated peptides or in differential abundance analysis (Figure 3.11). Peptides in Aβ, Tau, and APOE also do not have altered abundance across the 'dominant-like' and 'dominant-unlike' SAD. Differential abundance test does find 5,661 peptides with

significantly altered abundance when comparing the original 4 case conditions, with SAD as one group. 2,536 peptides mapping to 1,059 unique protein groups have significantly altered abundance between ADAD and RAD, including peptides that map to LMNA, CXA1, PKHB1, and DFFA, among others.

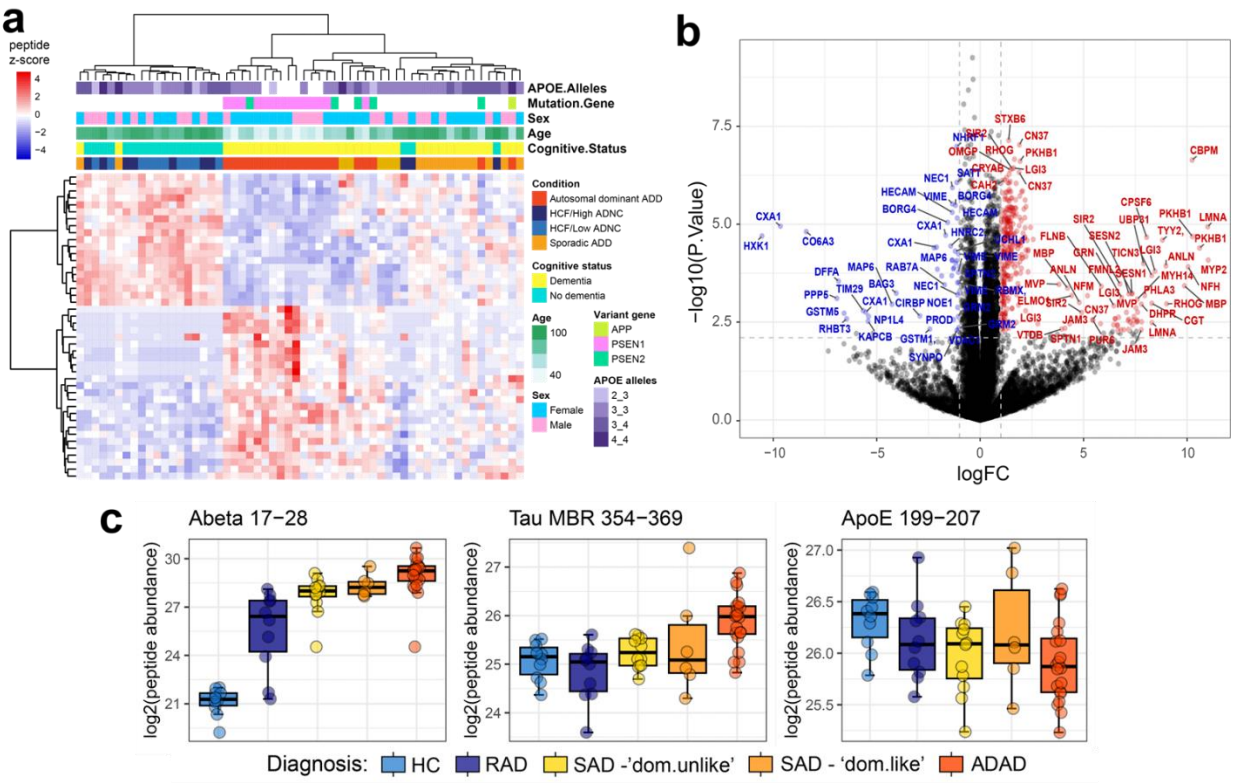


Figure 3.11. Tryptic peptides in the caudate nucleus distinguish between cognitive status. A) Heatmap of significantly correlated peptide abundances with hierarchical clustering of both cases and peptides. B) Differences in peptide abundance for select Aβ, Tau, and ApoE tryptic peptides in the four diagnostic groups (HC, RAD, SAD, and ADAD) with SAD divided into 'dominant-like'(n=6) and 'dominant-unlike (n=11) SAD subgroups. (need to add in the Tau 195 peptide) C) Volcano plot of peptides in the caudate nucleus with significantly altered abundance between 'dominant-like' and 'dominant-unlike' SAD, with red highlighting peptides increased and blue highlighting those decreased in 'dominant-like' compared to 'dominant-unlike', labeled with the protein names the peptides map to.

Unique and shared signatures across brain regions.

Protein groups with significantly altered abundances between the diagnostic groups were compared across the four brain regions. The IPL had the most significantly different proteins (1,794 total), followed by the SMTG (1,547 total), caudate (1,116 total), and hippocampus (109

total). In each analysis the comparison between ADAD and both HCF (HC and RAD) had the largest number of significant differences. The largest overlap in significant proteins is between the two isocortical regions (IPL and SMTG), with 670 shared proteins (Figure 3.12). Beyond those shared proteins, 580 proteins are uniquely altered in IPL, 507 are uniquely altered in SMTG, and 396 proteins are uniquely altered in the caudate. There are an additional 291 proteins significantly altered among IPL, SMTG, and caudate, and only 60 proteins significantly altered among all four brain regions. Proteins uniquely different in each brain region are enriched for different gene ontology terms. Differential proteins unique to the caudate are enriched in terms primarily related to translation, protein transport, and cellular organization (Figure 3.12). Differential proteins unique to the SMTG are enriched in terms primarily related to protein modification, protein localization, and axon and synapse related processes (Figure 3.12). For the IPL, unique differential proteins are enriched in terms primarily related to vesicle transport and protein complex organization (Figure 3.12). Only 8 proteins were differentially abundant in the hippocampus.

The two isocortical regions (SMTG and IPL), hippocampus, and caudate were selected for varying involvement by AD neuropathic changes, but also have varying cellular composition and tissue structure. For this reason, we also examined differences in cell marker protein abundance across conditions in each region. Microtubule-associated protein 2 (MAP2) is commonly used as a general marker protein for neurons.[85–88] As expected, we see strong signal for this protein in each of the four brain regions, with a slight decrease in abundance in SAD and ADAD, but not RAD, compared to HC, specifically in the SMTG and IPL. This is consistent with loss of neurons in symptomatic AD.[89] Glial fibrillary acidic protein (GFAP) is a common marker for astrocytes.[90–92] We observe a slight increase in GFAP in both SAD and ADAD, but not RAD, compared to HC, across all brain regions, but most prominently in hippocampus. Allograft inflammatory factor 1 (AIF1) is a marker protein for microglia,[93] and is not detected in caudate, and does not have a consistent trend across sample groups. Interestingly, monocyte differentiation antigen CD14, which has been used as a marker for microglia,[94] is slightly elevated in SAD and ADAD across brain regions and elevated in the SMTG and IPL brain regions (data not shown). Myelin-oligodendrocyte glycoprotein, which is a marker protein for oligodendrocytes,[95,96] is predominantly observed in the hippocampus, but does not show differences between conditions. Finally, the protein phosphatase 1 regulatory subunit 1B (PPR1B) is a marker of medium spiny neurons, and is predominantly found in the caudate nucleus region, as would be expected.[97]

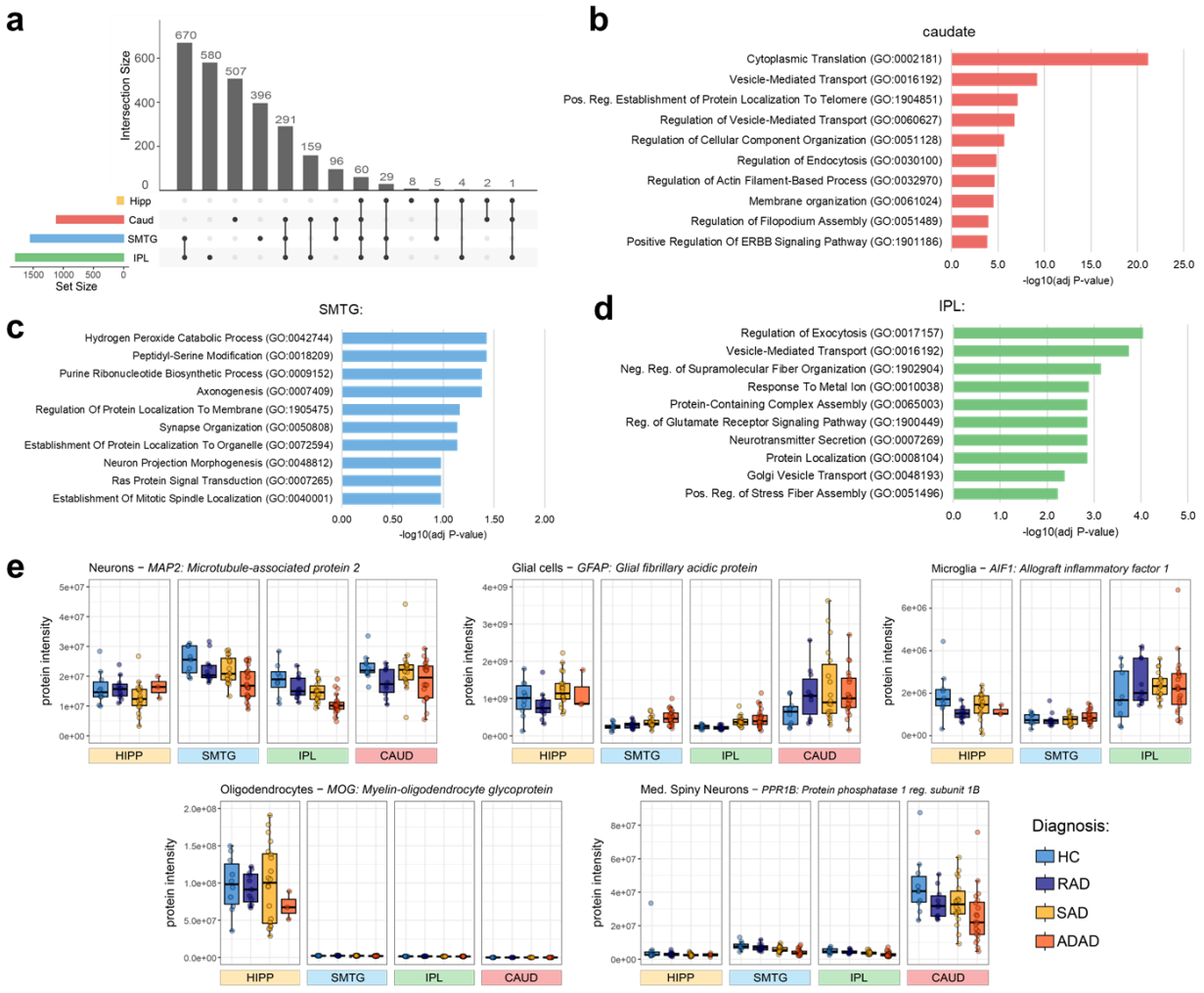


Figure 3.12. Differential protein abundance profiles across brain regions. A) The overlap in proteins with significant differences across sample groups for each brain region. Gene ontology terms enriched in the proteins with significantly altered abundance in B) the caudate nucleus proteome, C) the SMTG proteome, and D) the IPL proteome. There was an insufficient number of proteins differentially expressed in the hippocampus for significant gene ontology enrichment analysis. E) Brain cell type marker protein relative abundances across brain regions stratified by diagnostic groups.

3.4. Discussion

We find distinct differences at a proteome level among four different diagnostic groups, and among four different brain regions. The four diagnostic groups include HC, and three different forms of AD; RAD, SAD, and ADAD - all carefully evaluated to exclude common comorbidities, thereby enabling us to focus on the proteomic signature of AD without confounding. We deliberately selected four regions of brain varyingly impacted by AD neuropathologic change (Aβ plaque; SMTG and IPL > hippocampus > caudate, neurofibrillary degeneration: hippocampus > SMTG and IPL >> caudate), exploiting anatomical variation to gain deeper insights into proteomics signatures. Our focus on quality of case annotation and data generation necessarily limits the number of samples that meet our criteria and rigorous analysis. We recently described this approach using a subset of these samples and protein level data to compare protein signatures of RAD versus HC and SAD.[98] Here we focus on the peptide level changes and critically include the ADAD samples.

Increased regional tissue abundance of Aβ is a hallmark of AD. The Aβ sequence mapping peptides are increased as expected based on neuropathologic selection criteria for this sample set. Importantly, despite the CERAD score being the same across many AD cases, the Aβ peptides still distinguish among the diagnostic groups, confirming at the peptide level our protein level analysis of RAD versus HC and SAD, and importantly extending these data to include ADAD. This is likely due to the more precise measures by mass spectrometry than by immunohistochemistry, allowing us to extend the dynamic range and improve sensitivity especially in cases with highly abundant Aβ, as occurs in ADAD. Moreover, while neuropathologic assessments evaluate aggregates of Aβ peptides,[78,80,99] proteomic measurement of tryptic peptides mapping to Aβ are not limited to Aβ peptides, but also the Aβ sequences from intact APP. However, it is likely that the majority of Aβ tryptic peptide signal is coming from the Aβ peptides due to the difference in relative intensity of these peptides compared to other APP peptides.

Unlike the expected differences in Aβ, the increased abundance of APP in ADAD is not well described. We detect peptides across the whole APP sequence to varying degrees in all brain regions, and find most peptides are increased in ADAD compared to SAD in the hippocampus, SMTG, and caudate nucleus. Previous work has demonstrated the entire APP protein is increased in some causal variant carriers, specifically with APP causal variants,[68,100] however, we observe an increase that is relatively consistent across all causal variants in the sample set. The cause for the increase in APP is not clear, with a preliminary trend of APP being slightly further elevated in

PSEN1 variant cases compared to *PSEN2* carriers. Like APP, APOE is also distinctly increased in ADAD in the hippocampus, SMTG, and IPL, but not in the caudate. There are some preliminary trends in both Aβ and APOE abundance related to *APOE* genotype; however, our sample size is too small to powerfully address this intriguing observation.

An unexpected finding was the distinct differences in SMTG proteomes within the SAD cases. These differences were primarily observed in the peptide abundances that significantly correlated with Aβ 17-28 abundance but were also observed by clustering performed on the whole proteome. Since we had a subgroup of SAD cases that had profiles more like SAD, we wanted to determine if those individuals carried causal variants in *APP*, *PSEN1,* or *PSEN2* that had not been considered previously. Additionally, several genes have been identified as implicated in AD through GWAS; specifically *TREM2, ABCA7, SORL1,* and *ABI3*.[101–107] These 7 genes whose analyzed by whole exome sequencing for the 18 SAD cases. While some variants were found in *ABCA7, SORL1*, and *APP*, none had a known pathogenic status. One potentially functionally significant variant in ATP–binding cassette subfamily A member 7 (ABCA7), mid-exon frameshift (p.Leu1493ArgfsTer7), was identified in one SAD patient. ABCA7 is a membrane protein involved in lipid metabolism and phagocytosis,[108] and has been implicated in APP processing an Aβ clearance.[109,110] The frameshift found in ABCA7 occurs in the large extracellular loop between the first and second transmembrane helix in the second transmembrane domain. Several frameshifts identified in ABCA7 are associated with AD; however, the frameshift found in our sample set has not been implicated as a causal variant.[111] Our proteomic data nominates ABCA7 mid-exon frameshift p.Leu1493ArgfsTer7 as a candidate functional variant relevant to AD.

The two most striking differences in the case characteristics between the subgroups of SAD in SMTG was the age and last Mini-Mental State Examination (MMSE) score. For the SAD subgroup that more closely resembles the SAD SMTG proteome, most cases (5/7) were younger than 70 years of age at death, with the other two at 87 and 88 years. This led us to describe that cluster as an ‘earlier’ SAD subtype, since most in that cluster also likely had an onset of AD dementia at an earlier age. However, since not all were younger than 70 years, and since we don’t have specific information regarding their age of onset, we chose not to specifically refer to them as ‘early onset SAD’, which is often reserved for cases with documented onset prior to age 65 yrs.[112] The fact that almost 28% of our SAD samples are younger than 70 years is unusual, since

early onset SAD is quite rare.[76] We believe that this is an ascertainment bias in which SAD samples were selected for inclusion in this cohort; specifically in requiring a lack of common neuropathological comorbidities. Younger cases of SAD tend to also lack comorbidities that likely increase in incidence with age.[73]

In addition to age, the last documented MMSE score was significantly different between the 'dominant-like' SAD cases and 'dominant-unlike' SAD cases. Most patients with 'dominant-like' SAD had scores below 10, indicating severe dementia at last evaluation. This would also align with the observation of Tau abundance also being differential between SAD groups in SMTG. Clustering analysis of peptides with Tau 354-369 also results in the separation of SAD samples into the same groups (data not shown). Prolific accumulation of both Tau and Aβ is commonly indicative of worse cognitive impairment.[113] While these differences in SAD proteomes were mainly observed in the SMTG, that does not necessarily mean it is distinct to that brain region. Unfortunately, the lack of ADAD cases with available hippocampus material means we are unable to really examine the relationship between the ADAD and SAD proteomes. Based on the PCA clustering of the IPL, the Aβ and tau abundance differences, and peptide abundance differences found by testing with the sporadic cases classified into two categories, there are still differences in the SAD IPL proteome, however they are not as distinct as the SMTG.

There is very little difference in the SAD groups in the caudate nucleus, and fewer differences between SAD and ADAD cases compared to other brain regions. The caudate nucleus is part of the basal ganglia, having a different cellular composition compared to the IPL and SMTG. This is partially highlighted in the overall differences in abundance of certain cell markers in the 4 brain regions. For example, medium spiny neurons (MSNs) are known to be a significant portion of the caudate nucleus,[114] and a common marker protein for MSNs, protein phosphatase regulatory subunit 1B (PPR1B), is most abundant in that proteome. In addition to the differences in cell types and tissue structure in the caudate compared to the other regions, the caudate could be potentially in an earlier 'state' of Aβ and tau accumulation. Both the Aβ and the Tau peptides track with the expected differences; with higher abundance in the AD samples compared to both HCF groups. However, the differences at least in Tau tend to be a smaller magnitude compared to the other brain regions. One could hypothesize that the other brain regions represent a state in which the tissue is already at an end stage of disease, while the caudate would continue to progress towards

that state with additional time. The SAD cases did not have different Aβ and tau in the caudate like they did in the SMTG and partially in the IPL.

Post-translational modification of Tau is thought to affect how it is processed, with some modifications occurring more frequently in Tau tangles in AD.[32] Modifications of Tau have also been implicated in the rate of AD progression.[115] In our Tau tryptic peptide measurements we capture what we believe is signal related to modified peptides in the proline-rich region. Multiple peptides mapping to residues 195-240 have decreased abundance in AD cases compared to the HC cases. This is inverse of the peptides mapping to the microtubule-binding region, such as Tau 354-369, which have the expected increased abundance in AD compared to the HC cases. Since we are detecting and reporting the quantities for the unmodified tryptic peptides, the decrease we are seeing is likely due to a change in stoichiometry of the unmodified:modified peptides. If the majority of Tau in the AD samples are hyperphosphorylated, then the stoichiometry would shift such that the unmodified peptide is relatively decreased. For all brain regions tryptic peptides significantly correlated with Tau 354-369 have very similar trends to the peptides correlated with Aβ 17-28. The relationship of tau abundance in the SMTG and IPL with the subgroups of SAD supports the existing literature on the accumulation of both Aβ and Tau together being more profuse in severe disease.[116]

Another unexpected observation was the detection of several peptides that do not map to the 2N4R Tau sequence of 441 amino acids, but map instead to a longer form of the protein with additional exon inclusion between the N and R regions. This could mean some Big Tau is present in these samples,[117] albeit a low abundance, with no discernible differences between the disease conditions. In the microtubule binding regions, there is a difference in abundance profiles across the four exons. Both 2N3R and 2N4R isoforms are commonly observed in AD, with later stage disease having slightly more 2N3R, which lacks the R2 exon from 274-305.[118,119] However, in our data the peptides that map to that region all strongly distinguish between conditions, indicating that there is R2 sequence present in the samples. Instead, peptides that map to the R1 exon do not have as strong of differences between conditions. These observations are unique to looking at every peptide within a protein coding sequence, since a summed protein abundance from multiple peptides would mask these nuances.[49]

Our findings highlight the importance of looking at AD with different etiologies to better understand how disease is related to aging. The striking difference within the SAD proteomes

indicates the need for further study of the molecular differences between ‘dominant-like’ SAD and ‘dominant-unlike’ SAD with larger well characterized cohorts. One possible comparison could be performed with the currently running Longitudinal Early-Onset Alzheimer’s Disease (LEADS) study.[112] Unfortunately for this current study, once the SAD cases were separated based on their SMTG proteome profiles, there were no longer many significant differences between the ‘dominant-unlike’ SAD and RAD cases. This is unexpected, but could very well be due to this current cohort being underpowered for uncovering molecular differences between these age-matched conditions.[120] However, the careful exclusion of comorbid neuropathology makes this a unique insight. Since many studies with late onset SAD cases likely have additional neuropathologic features, it could be contributing to the distinct molecular profiles seen in large cohort studies. Additional study into the late onset AD with genetic factors could also shed light on dissecting the role of age on disease neuropathology with differing etiologies.[121]

Chapter 4. CLOSING REMARKS

4.1. Outlook and future directions for protein quantification

Over the years, the challenge that not all peptides from the same gene or protein group have the same differential abundance have been an important area of research. The approaches of several proposed methods focus on the exclusion of peptide measurements from inclusion in the aggregate protein quantity if they are outliers from other peptide measurements mapping to the same protein coding gene.[122–127] While these all demonstrate improved protein concentration estimates, they still only report a single protein quantity and ignore peptides that don't agree with that single value. If those outlier peptide measurements are discarded, then true biological signals may be lost. Alternatively, signal could be kept and a weighted distribution applied across all matching isoforms.[128] Another approach taken in previous methods is to try and identify the specific proteoforms present based on peptide quantification across conditions.[129–132] These methods are tolerant to having multiple proteoforms present in a sample, however, once a molecule is digested to peptides it is impossible to track the peptide-protein molecule relationship.

While the challenge of aggregating peptide measurements may not be solved yet, one thing that is apparent is that we should no longer blindly merge all peptides into a single protein level quantity. A solution to the presence of discordant peptides could be to keep all peptides as independent measurements because it is impossible to merge peptides without detailed knowledge of all proteoforms in the sample. While remaining as true to the acquired data as possible, this strategy may prove to be difficult for interpretation of experiments because the role of individual tryptic peptides may be difficult to infer, especially in less studied systems. Additionally, reduced statistical power for differential abundance testing on tens of thousands of peptides compared to thousands of protein groups will also likely result in fewer significant differences. However, there has been recent work towards integrating top-down proteomics with bottom-up proteomic measurements.[133] This strategy could provide higher resolution information about the quantity resulting from specific proteoforms present in a sample, which then can be used to determine how peptides could be combined to more accurately reflect those proteoforms present.

An alternative approach could be to combine peptides that both map to the same gene and co-vary across a diverse set of biological groups or conditions, without designating them as specific proteoforms. We need the ability to generate multiple “peptide groups” for each protein

group -- resulting in 1 to N quantities for each protein where N is the number of peptides. This grouping would require a method that minimized variance and multiple testing while maximizing the biological effect. This approach would not require knowing which proteoforms were present but would still capture quantitative differences observed at the peptide level that would otherwise be eliminated by combining those differences with non-changing peptides within the same gene product. However, this approach could be heavily dependent on having multiple conditions with enough biological replicates and high reproducibility, and may not be suitable for proteins with low peptide coverage.[129] Regardless of how we choose to analyze and report our proteomics data, if peptides are aggregated to a protein quantity, it should be transparent which peptides were used, how they were combined, and the individual peptide quantities should remain accessible. Furthermore, for a specific "protein" it is critical that the same peptides are used to create the protein level quantity for all samples as different peptides will likely reflect different combinations of proteoforms.

While bottom-up proteomics is still the preferred method for characterizing proteomes due to its coverage, robustness across diverse protein physiochemical properties, sensitivity, and quantitative capabilities -- there remain challenges. Moving forward we will need new or repurposed methods, tools, and datasets to better interpret peptide level measurements. Datasets with known differences in peptide measurements will be crucial for validating any new approaches that are proposed to deal with peptide level differences. Additionally, improved data visualization tools are necessary to better distinguish changes inclusive of conserved domains, known PTMs, and structural features within a protein coding gene in the context of a global proteome. Finally, a compiled reference or "atlas" of experimentally observed proteoforms presents a major opportunity for future algorithm development, which the Human Proteoform Atlas recently framed. As technology has advanced, so has our ability to obtain robust measurements across many samples while minimizing missing data. We now need to move towards understanding why these peptide measurements may be different and develop a better suited and more integrative data format.

4.2. Future directions for Alzheimer's Disease proteomics

The future of proteomics in studying Alzheimer's disease (AD) holds immense potential. In chapter 3 we use mass spectrometry proteomics to explore the molecular profile of proteins in AD with varying etiologies, leading to new insights into the complexity of studying sporadic AD without comorbid neuropathology. This demonstrates the usefulness of measuring the proteome at scale in a robust and reproducible manner. It also demonstrates that more molecular mechanisms may still be uncovered in larger and more diverse cohort studies. Beyond general descriptive results, these types of studies can also uncover useful biomarkers and possible therapeutic targets. As discussed in the second chapter, the coupling of these discovery proteomics experiments with targeted assay generation can be easily achieved. This has the potential to hasten the generation of assays for novel biomarkers or for monitoring response to novel therapeutics. With ongoing advancements in technology and interdisciplinary collaborations, proteomics is poised to revolutionize our ability to diagnose, treat, and ultimately prevent Alzheimer's disease.[134]

BIBLIOGRAPHY

1. Omenn, G. S. *et al.* Progress Identifying and Analyzing the Human Proteome: 2021 Metrics from the HUPO Human Proteome Project. *J. Proteome Res.* **20**, 5227–5240 (2021).

2. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).

3. MacCoss, M. J. *et al.* Sampling the proteome by emerging single-molecule and mass spectrometry methods. *Nat. Methods* **20**, 339–346 (2023).

4. Tabb, D. L., McDonald, W. H. & Yates, J. R. DTASelect and Contrast: Tools for Assembling and Comparing Protein Identifications from Shotgun Proteomics. *J. Proteome Res.* **1**, 21–26 (2002).

5. Ma, Z.-Q. *et al.* IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering. *J. Proteome Res.* **8**, 3872–3881 (2009).

6. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003).

7. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. & Coon, J. J. Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Mol. Cell. Proteomics MCP* **11**, 1475–1488 (2012).

8. Remes, P. M., Yip, P. & MacCoss, M. J. Highly Multiplex Targeted Proteomics Enabled by Real-Time Chromatographic Alignment. *Anal. Chem.* **92**, 11809–11817 (2020).

9. Picotti, P. & Aebersold, R. Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **9**, 555–566 (2012).

10. Zhang, H. *et al.* Methods for Peptide and Protein Quantitation by Liquid Chromatography-Multiple Reaction Monitoring Mass Spectrometry. *Mol. Cell. Proteomics MCP* **10**, M110.006593 (2011).

11. Nilsson, T. *et al.* Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods* **7**, 681–685 (2010).

12. Plubell, D. L. *et al.* Extended Multiplexing of Tandem Mass Tags (TMT) Labeling Reveals Age and High Fat Diet Specific Proteome Changes in Mouse Epididymal Adipose Tissue* □ S. *Mol. Cell. Proteomics* (2017) doi:10.1074/mcp.M116.065524.

13. 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).

14. Pino, L. K. *et al.* The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrom. Rev.* **39**, 229–244 (2020).

15. Fernández-Costa, C. *et al.* Impact of the Identification Strategy on the Reproducibility of the DDA and DIA Results. *J. Proteome Res.* **19**, 3153–3161 (2020).

16. Rosenberger, G. *et al.* Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods* **14**, 921–927 (2017).

17. Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **1**, 39–45 (2004).

18. 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-O111.016717 (2012).

19. Ting, Y. S. *et al.* PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat. Methods* **14**, 903–908 (2017).

20. Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat. Commun.* **9**, 5128 (2018).

21. Amodei, D. *et al.* Improving Precursor Selectivity in Data-Independent Acquisition Using Overlapping Windows. *J. Am. Soc. Mass Spectrom.* **30**, 669–684 (2019).

22. Aebersold, R. *et al.* How many human proteoforms are there? *Nat. Chem. Biol.* **14**, 206–214 (2018).

23. Smith, L. M. & Kelleher, N. L. Proteoforms as the next proteomics currency. *Science* **359**, 1106–1107 (2018).

24. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between Protein and mRNA Abundance in Yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).

25. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).

26. Bramer, L. M., Irvahn, J., Piehowski, P. D., Rodland, K. D. & Webb-Robertson, B.-J. M. A Review of Imputation Strategies for Isobaric Labeling-Based Shotgun Proteomics. *J. Proteome Res.* **20**, 1–13 (2021).

27. O'Brien, R. J. & Wong, P. C. Amyloid Precursor Protein Processing and Alzheimer's Disease. *Annu. Rev. Neurosci.* **34**, 185–204 (2011).

28. Ling, Y., Morgan, K. & Kalsheker, N. Amyloid precursor protein (APP) and the biology of proteolytic processing: relevance to Alzheimer's disease. *Int. J. Biochem. Cell Biol.* **35**, 1505–1535 (2003).

29. Portelius, E., Westman-Brinkmalm, A., Zetterberg, H. & Blennow, K. Determination of β-Amyloid Peptide Signatures in Cerebrospinal Fluid Using Immunoprecipitation-Mass Spectrometry. *J. Proteome Res.* **5**, 1010–1016 (2006).

30. Wildburger, N. C. *et al.* Diversity of Amyloid-beta Proteoforms in the Alzheimer's Disease Brain. *Sci. Rep.* **7**, 9520 (2017).

31. Holtzman, D. M. *et al.* Tau: From research to clinical development. *Alzheimers Dement.* **12**, 1033–1039 (2016).

32. Wesseling, H. *et al.* Tau PTM Profiles Identify Patient Heterogeneity and Stages of Alzheimer's Disease. *Cell* **183**, 1699-1713.e13 (2020).

33. Barthélemy, N. R. *et al.* A soluble phosphorylated tau signature links tau, amyloid and the evolution of stages of dominantly inherited Alzheimer's disease. *Nat. Med.* **26**, 398–407 (2020).

34. Leighton, E., Sainsbury, C. A. & Jones, G. C. A Practical Review of C-Peptide Testing in Diabetes. *Diabetes Ther.* **8**, 475–487 (2017).

35. Sandoval, D. A. & D'Alessio, D. A. Physiology of Proglucagon Peptides: Role of Glucagon and GLP-1 in Health and Disease. *Physiol. Rev.* **95**, 513–548 (2015).

36. Weitz, J. I., Fredenburgh, J. C. & Eikelboom, J. W. A Test in Context: D-Dimer. *J. Am. Coll. Cardiol.* **70**, 2411–2420 (2017).

37. Rotunno, M. S. *et al.* Cerebrospinal fluid proteomics implicates the granin family in Parkinson's disease. *Sci. Rep.* **10**, 2479 (2020).

38. Skinner, O. S. *et al.* Top-down characterization of endogenous protein complexes with native proteomics. *Nat. Chem. Biol.* **14**, 36–41 (2018).

39. Gold, L. *et al.* Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLOS ONE* **5**, e15004 (2010).

40. Assarsson, E. *et al.* Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability. *PLOS ONE* **9**, e95192 (2014).

41. Ngo, D. *et al.* Aptamer-Based Proteomic Profiling Reveals Novel Candidate Biomarkers and Pathways in Cardiovascular Disease. *Circulation* **134**, 270–285 (2016).

42. Johansson, A. *et al.* Identification of genetic variants influencing the human plasma proteome. *Proc. Natl. Acad. Sci.* **110**, 4673–4678 (2013).

43. Blanchard, V. *et al.* Kinetics of plasma apolipoprotein E isoforms by LC-MS/MS: a pilot study. *J. Lipid Res.* **59**, 892–900 (2018).

44. Conlon, K. P. *et al.* Fusion Peptides from Oncogenic Chimeric Proteins as Putative Specific Biomarkers of Cancer. *Mol. Cell. Proteomics* **12**, 2714–2723 (2013).

45. Spellman, D. S. *et al.* Development and evaluation of a multiplexed mass spectrometry based assay for measuring candidate peptide biomarkers in Alzheimer's Disease Neuroimaging Initiative (ADNI) CSF. *PROTEOMICS - Clin. Appl.* **9**, 715–731 (2015).

46. Hoofnagle, A. N. *et al.* Multiple-reaction monitoring-mass spectrometric assays can accurately measure the relative protein abundance in complex mixtures. *Clin. Chem.* **58**, 777–781 (2012).

47. Addona, T. A. *et al.* Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat. Biotechnol.* **27**, 633–641 (2009).

48. Carr, S. A. *et al.* Targeted Peptide Measurements in Biology and Medicine: Best Practices for Mass Spectrometry-based Assay Development Using a Fit-for-Purpose Approach. *Mol. Cell. Proteomics MCP* **13**, 907–917 (2014).

49. 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).

50. Pino, L. K. *et al.* Matrix-Matched Calibration Curves for Assessing Analytical Figures of Merit in Quantitative Proteomics. *J. Proteome Res.* **19**, 1147–1153 (2020).

51. Stergachis, A. B., MacLean, B., Lee, K., Stamatoyannopoulos, J. A. & MacCoss, M. J. Rapid empirical discovery of optimal peptides for targeted proteomics. *Nat. Methods* **8**, 1041–1043 (2011).

52. 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).

53. Prakash, A. *et al.* Expediting the Development of Targeted SRM Assays: Using Data from Shotgun Proteomics to Automate Method Development. *J. Proteome Res.* **8**, 2733–2739 (2009).

54. Kusebauch, U. *et al.* Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. *Cell* **166**, 766–778 (2016).

55. 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 MCP* **14**, 2331–2340 (2015).

56. Bollinger, J. G., Stergachis, A. B., Johnson, R. S., Egertson, J. D. & MacCoss, M. J. Selecting Optimal Peptides for Targeted Proteomic Experiments in Human Plasma Using In Vitro Synthesized Proteins as Analytical Standards. in *Quantitative Proteomics by Mass Spectrometry* (ed. Sechi, S.) vol. 1410 207–221 (Springer New York, New York, NY, 2016).

57. Panchaud, A. *et al.* PAcIFIC: how to dive deeper into the proteomics ocean. *Anal. Chem.* **81**, 6481–6488 (2009).

58. Canterbury, J. D., Merrihew, G. E., MacCoss, M. J., Goodlett, D. R. & Shaffer, S. A. Comparison of Data Acquisition Strategies on Quadrupole Ion Trap Instrumentation for Shotgun Proteomics. *J. Am. Soc. Mass Spectrom.* **25**, 2048–2059 (2014).

59. Gallien, S. *et al.* Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. *Mol. Cell. Proteomics MCP* **11**, 1709–1723 (2012).

60. Escher, C. *et al.* Using iRT, a normalized retention time for more targeted measurement of peptides. *PROTEOMICS* **12**, 1111–1121 (2012).

61. 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* mcp.P119.001913 (2020) doi:10.1074/mcp.P119.001913.

62. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinforma. Oxf. Engl.* **26**, 966–8 (2010).

63. Fang, N., Yu, S., Ronis, M. J. & Badger, T. M. Matrix effects break the LC behavior rule for analytes in LC-MS/MS analysis of biological samples. *Exp. Biol. Med.* **240**, 488–497 (2015).

64. Gupta, S., Ahadi, S., Zhou, W. & Röst, H. DIAlignR Provides Precise Retention Time Alignment Across Distant Runs in DIA and Targeted Proteomics. *Mol. Cell. Proteomics MCP* **18**, 806–817 (2019).

65. 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).

66. Haass, C., Kaether, C., Thinakaran, G. & Sisodia, S. Trafficking and Proteolytic Processing of APP. *Cold Spring Harb. Perspect. Med.* **2**, a006270 (2012).

67. Cacace, R., Sleegers, K. & Van Broeckhoven, C. Molecular genetics of early-onset Alzheimer's disease revisited. *Alzheimers Dement.* **12**, 733–748 (2016).

68. Thinakaran, G. & Koo, E. H. Amyloid Precursor Protein Trafficking, Processing, and Function. *J. Biol. Chem.* **283**, 29615–29619 (2008).

69. Aghakhanyan, G. *et al.* PET/MRI Delivers Multimodal Brain Signature in Alzheimer's Disease with De Novo PSEN1 Mutation. *Curr. Alzheimer Res.* **18**, 178–184.

70. Agüero, P. *et al.* De Novo PS1 Mutation (Pro436Gln) in a Very Early-Onset Posterior Variant of Alzheimer's Disease Associated with Spasticity: A Case Report. *J. Alzheimers Dis.* **83**, 1011–1016 (2021).

71. Chen, K.-L. *et al.* Very Early-Onset Alzheimer's Disease in the Third Decade of Life with de novo PSEN1 Mutations. *J. Alzheimers Dis.* **85**, 65–71 (2022).

72. Selkoe, D. J. & Wolfe, M. S. Presenilin: Running with Scissors in the Membrane. *Cell* **131**, 215–221 (2007).

73. Knopman, D. S. *et al.* Alzheimer disease. *Nat. Rev. Dis. Primer* **7**, 1–21 (2021).

74. Nelson, P. T. *et al.* Alzheimer's disease is not "brain aging": neuropathological, genetic, and epidemiological human studies. *Acta Neuropathol. (Berl.)* **121**, 571–587 (2011).

75. Morris, J. C. *et al.* Autosomal dominant and sporadic late onset Alzheimer's disease share a common in vivo pathophysiology. *Brain* **145**, 3594–3607 (2022).

76. 2022 Alzheimer's disease facts and figures. *Alzheimers Dement.* **18**, 700–789 (2022).

77. Thambisetty, M., An, Y. & Tanaka, T. Alzheimer's disease risk genes and the age-at-onset phenotype. *Neurobiol. Aging* **34**, 2696.e1-2696.e5 (2013).

78. 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).

79. Montine, T. J., Sonnen, J. A., Montine, K. S., Crane, P. K. & Larson, E. B. Adult Changes in Thought study: dementia is an individually varying convergent syndrome with prevalent clinically silent diseases that may be modified by some commonly used therapeutics. *Curr. Alzheimer Res.* **9**, 718–23 (2012).

80. Montine, T. J. *et al.* Recommendations of the Alzheimer's Disease–Related Dementias Conference. *Neurology* **83**, 851–860 (2014).

81. Merrihew, G. E. *et al.* A peptide-centric quantitative proteomics dataset for the phenotypic assessment of Alzheimer's disease. *Sci. Data* **10**, 206 (2023).

82. Sharma, V. *et al.* Panorama Public: A Public Repository for Quantitative Data Sets Processed in Skyline. *Mol. Cell. Proteomics* **17**, 1239–1244 (2018).

83. 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).

84. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr. Protoc.* **1**, e90 (2021).

85. Matus, A. Microtubule-Associated Proteins: Their Potential Role in Determining Neuronal Morphology. *Annu. Rev. Neurosci.* **11**, 29–44 (1988).

86. Schoenfeld, T. A. & Obar, R. A. Diverse Distribution and Function of Fibrous Microtubule-Associated Proteins in the Nervous System. in *International Review of Cytology* (eds. Jeon, K. W. & Jarvik, J.) vol. 151 67–137 (Academic Press, 1994).

87. Sánchez, C., Díaz-Nido, J. & Avila, J. Phosphorylation of microtubule-associated protein 2 (MAP2) and its relevance for the regulation of the neuronal cytoskeleton function. *Prog. Neurobiol.* **61**, 133–168 (2000).

88. Dehmelt, L. & Halpain, S. The MAP2/Tau family of microtubule-associated proteins. *Genome Biol.* **6**, 1–10 (2005).

89. Mangalmurti, A. & Lukens, J. R. How neurons die in Alzheimer's disease: Implications for neuroinflammation. *Curr. Opin. Neurobiol.* **75**, 102575 (2022).

90. Brenner, M. & Messing, A. Regulation of GFAP Expression. *ASN Neuro* **13**, 1759091420981206 (2021).

91. Middeldorp, J. & Hol, E. M. GFAP in health and disease. *Prog. Neurobiol.* **93**, 421–443 (2011).

92. Yang, Z. & Wang, K. K. W. Glial fibrillary acidic protein: from intermediate filament assembly and gliosis to neurobiomarker. *Trends Neurosci.* **38**, 364–374 (2015).

93. Hopperton, K. E., Mohammad, D., Trépanier, M. O., Giuliano, V. & Bazinet, R. P. Markers of microglia in post-mortem brain samples from patients with Alzheimer's disease: a systematic review. *Mol. Psychiatry* **23**, 177–198 (2018).

94. Letiembre, M. *et al.* Innate immune receptor expression in normal brain aging. *Neuroscience* **146**, 248–254 (2007).

95. Solly, S. k. *et al.* Myelin/oligodendrocyte glycoprotein (MOG) expression is associated with myelin deposition. *Glia* **18**, 39–48 (1996).

96. Scolding, N. J. *et al.* Myelin-oligodendrocyte glycoprotein (MOG) is a surface marker of oligodendrocyte maturation. *J. Neuroimmunol.* **22**, 169–176 (1989).

97. Willett, J. A. *et al.* Electrophysiological Properties of Medium Spiny Neuron Subtypes in the Caudate-Putamen of Prepubertal Male and Female Drd1a-tdTomato Line 6 BAC Transgenic Mice. *eNeuro* **6**, (2019).

98. Huang, Z. *et al.* Brain proteomic analysis implicates actin filament processes and injury response in resilience to Alzheimer's disease. *Nat. Commun.* **14**, 2747 (2023).

99. Mirra, S. S. *et al.* The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology* **41**, 479–486 (1991).

100. Chhatwal, J. P. *et al.* Variant-dependent heterogeneity in amyloid β burden in autosomal dominant Alzheimer's disease: cross-sectional and longitudinal analyses of an observational study. *Lancet Neurol.* **21**, 140–152 (2022).

101. Karch, C. M., Cruchaga, C. & Goate, A. M. Alzheimer's Disease Genetics: From the Bench to the Clinic. *Neuron* **83**, 11–26 (2014).

102. Jonsson, T. *et al.* Variant of TREM2 Associated with the Risk of Alzheimer's Disease. *N. Engl. J. Med.* **368**, 107–116 (2013).

103. Wightman, D. P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).

104. Bellenguez, C. *et al.* Contribution to Alzheimer's disease risk of rare variants in TREM2, SORL1, and ABCA7 in 1779 cases and 1273 controls. *Neurobiol. Aging* **59**, 220.e1-220.e9 (2017).

105. Carmona, S. *et al.* The role of TREM2 in Alzheimer's disease and other neurodegenerative disorders. *Lancet Neurol.* **17**, 721–730 (2018).

106. Guerreiro, R. *et al.* TREM2 Variants in Alzheimer's Disease. *N. Engl. J. Med.* **368**, 117–127 (2013).

107. De Roeck, A., Van Broeckhoven, C. & Sleegers, K. The role of ABCA7 in Alzheimer's disease: evidence from genomics, transcriptomics and methylomics. *Acta Neuropathol. (Berl.)* **138**, 201–220 (2019).

108. Jehle, A. W. *et al.* ATP-binding cassette transporter A7 enhances phagocytosis of apoptotic cells and associated ERK signaling in macrophages. *J. Cell Biol.* **174**, 547–556 (2006).

109. Fu, Y., Hsiao, J.-H. T., Paxinos, G., Halliday, G. M. & Kim, W. S. ABCA7 Mediates Phagocytic Clearance of Amyloid-β in the Brain. *J. Alzheimers Dis. JAD* **54**, 569–584 (2016).

110. Satoh, K., Abe-Dohmae, S., Yokoyama, S., St George-Hyslop, P. & Fraser, P. E. ATP-binding Cassette Transporter A7 (ABCA7) Loss of Function Alters Alzheimer Amyloid Processing*. *J. Biol. Chem.* **290**, 24152–24165 (2015).

111. Cukier, H. N. *et al.* ABCA7 frameshift deletion associated with Alzheimer disease in African Americans. *Neurol. Genet.* **2**, e79 (2016).

112. Apostolova, L. G. *et al.* The Longitudinal Early-onset Alzheimer's Disease Study (LEADS): Framework and methodology. *Alzheimers Dement.* **17**, 2043–2055 (2021).

113. Gómez-Isla, T. *et al.* Neuronal loss correlates with but exceeds neurofibrillary tangles in Alzheimer's disease. *Ann. Neurol.* **41**, 17–24 (1997).

114. Yager, L. M., Garcia, A. F., Wunsch, A. M. & Ferguson, S. M. The ins and outs of the striatum: Role in drug addiction. *Neuroscience* **301**, 529–541 (2015).

115. Dujardin, S. *et al.* Tau molecular diversity contributes to clinical heterogeneity in Alzheimer's disease. *Nat. Med.* **26**, 1256–1263 (2020).

116. Arriagada, P. V., Growdon, J. H., Hedley-Whyte, E. T. & Hyman, B. T. Neurofibrillary tangles but not senile plaques parallel duration and severity of Alzheimer's disease. *Neurology* **42**, 631–639 (1992).

117. Fischer, I. & Baas, P. W. Resurrecting the Mysteries of Big Tau. *Trends Neurosci.* **43**, 493–504 (2020).

118. Cherry, J. D. *et al.* Tau isoforms are differentially expressed across the hippocampus in chronic traumatic encephalopathy and Alzheimer's disease. *Acta Neuropathol. Commun.* **9**, 86 (2021).

119. Bachmann, S., Bell, M., Klimek, J. & Zempel, H. Differential Effects of the Six Human TAU Isoforms: Somatic Retention of 2N-TAU and Increased Microtubule Number Induced by 4R-TAU. *Front. Neurosci.* **15**, (2021).

120. Johnson, E. C. B. *et al.* Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nat. Med.* **26**, 769–780 (2020).

121. Sung, Y. J. *et al.* Proteomics of brain, CSF, and plasma identifies molecular signatures for distinguishing sporadic and genetic Alzheimer's disease. *Sci. Transl. Med.* **15**, eabq5923 (2023).

122. Forshed, J. *et al.* Enhanced Information Output From Shotgun Proteomics Data by Protein Quantification and Peptide Quality Control (PQPQ). *Mol. Cell. Proteomics* **10**, M111.010264 (2011).

123. Goeminne, LudgerJ. E., Gevaert, K. & Clement, L. Peptide-level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Data-dependent Quantitative Label-free Shotgun Proteomics. *Mol. Cell. Proteomics* **15**, 657–668 (2016).

124. Zhang, B., Pirmoradian, M., Zubarev, R. & Käll, L. Covariation of Peptide Abundances Accurately Reflects Protein Concentration Differences. *Mol. Cell. Proteomics* **16**, 936–948 (2017).

125. The, M. & Käll, L. Integrated Identification and Quantification Error Probabilities for Shotgun Proteomics. *Mol. Cell. Proteomics* **18**, 561–570 (2019).

126. Dermit, M., Peters-Clarke, T. M., Shishkova, E. & Meyer, J. G. Peptide Correlation Analysis (PeCorA) Reveals Differential Proteoform Regulation. *J. Proteome Res.* **20**, 1972–1980 (2021).

127. Tsai, T.-H. *et al.* Selection of Features with Consistent Profiles Improves Relative Protein Quantification in Mass Spectrometry Experiments. *Mol. Cell. Proteomics* **19**, 944–959 (2020).

128. Saltzman, A. B. *et al.* gpGrouper: A Peptide Grouping Algorithm for Gene-Centric Inference and Quantitation of Bottom-Up Proteomics Data. *Mol. Cell. Proteomics* **17**, 2270–2283 (2018).

129. Webb-Robertson, B.-J. M. *et al.* Bayesian Proteoform Modeling Improves Protein Quantification of Global Proteomic Measurements. *Mol. Cell. Proteomics* **13**, 3639–3646 (2014).

130. Bamberger, C. *et al.* Deducing the presence of proteins and proteoforms in quantitative proteomics. *Nat. Commun.* **9**, 2320 (2018).

131. Malioutov, D. *et al.* Quantifying Homologous Proteins and Proteoforms. *Mol. Cell. Proteomics* **18**, 162–168 (2019).

132. Bludau, I. *et al.* Systematic detection of functional proteoform groups from bottom-up proteomic datasets. *Nat. Commun.* **12**, 3810 (2021).

133. Schaffer, L. V., Millikin, R. J., Shortreed, M. R., Scalf, M. & Smith, L. M. Improving Proteoform Identifications in Complex Systems Through Integration of Bottom-Up and Top-Down Data. *J. Proteome Res.* **19**, 3510–3517 (2020).

134. Heil, L. R. *et al.* Evaluating the Performance of the Astral Mass Analyzer for Quantitative Proteomics Using Data-Independent Acquisition. *J. Proteome Res.* (2023) doi:10.1021/acs.jproteome.3c00357.

APPENDIX A. Skyline tutorial for the development of SRM assays from DIA data.

Using data independent acquisition to inform the development of triple quadrupole assays

Selecting peptides and fragments to monitor by targeted triple quad assays can be challenging, proving time and cost intensive. Previous work from the MacCoss lab has shown DIA data to be a better predictor of peptide performance in targeted triple quad assays. DIA methods sample the proteome in a more comprehensive manner, and the type of information we can obtain from these experiments is helpful for target selection.

In this tutorial we will work our way through the workflow we developed to assess and select suitable peptides for SRM analysis based on DIA results. Specifically, we will cover the following concepts:

1. **Getting started with DIA gas-phase fractionated libraries in Skyline**
   To get started we import of our gas-phase fractionated (GPF) DIA library file into Skyline. Here we will go through settings that we can use, import our GPF library, and import our individual run files.
2. **Refining and filtering for high-quality peptide detections in DIA**
   Since we want to select peptides that are reliably detected in our sample we will do some initial filtering to ensure all peptides are viable candidates. This will leave us a subset of peptides that have reproducible measurements.
3. **Peptide filtering for protein targets**
   We then use peptide intensity rankings to aid us in selecting potential SRM candidate peptides for a specific list of proteins.
4. **Scheduling an SRM method based on DIA iRTs**
   Once we have narrowed down our list of candidate peptides we can schedule a SRM method using the iRT standards included in our DIA runs.

Note: the steps for searching Gas-phase fractionated samples with EncyclopeDIA are documented in the Pino et al., 2020 MCP paper. The three gas-phase fractionated samples are searched and saved as separate chromatogram libraries, then combined into a single library using the "Combine multiple libraries" function found in EncyclopeDIA 2.13.30 and later.

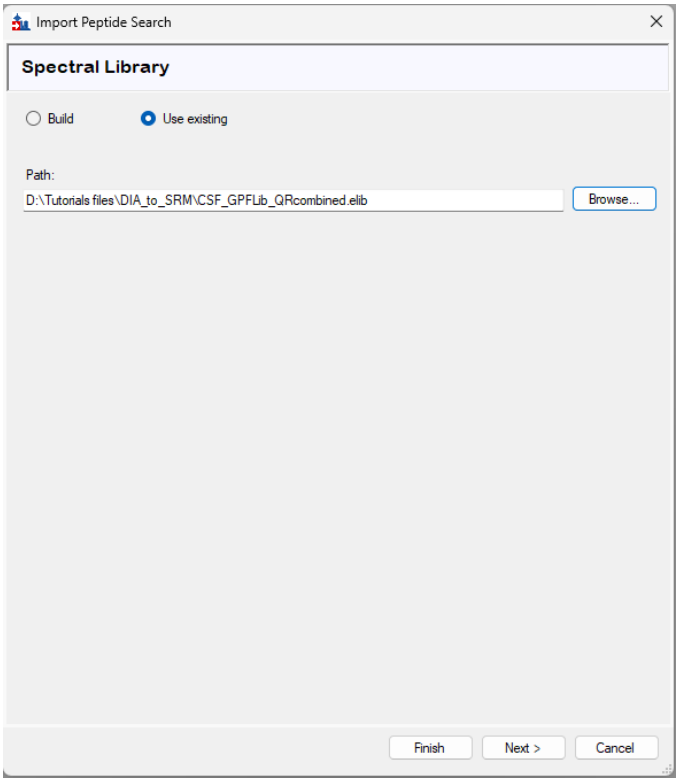
Step 1. Getting started with GPF DIA libraries in Skyline

The goal of this section is to import our GPF DIA search and peak boundary data. We will give Skyline the results and background for our experiment, including parameters for what peptides and transitions to extract. *This could additionally be adjusted to suit other DIA search outputs that provide information about peptide retention time and scoring.*

Setting up the import of data.

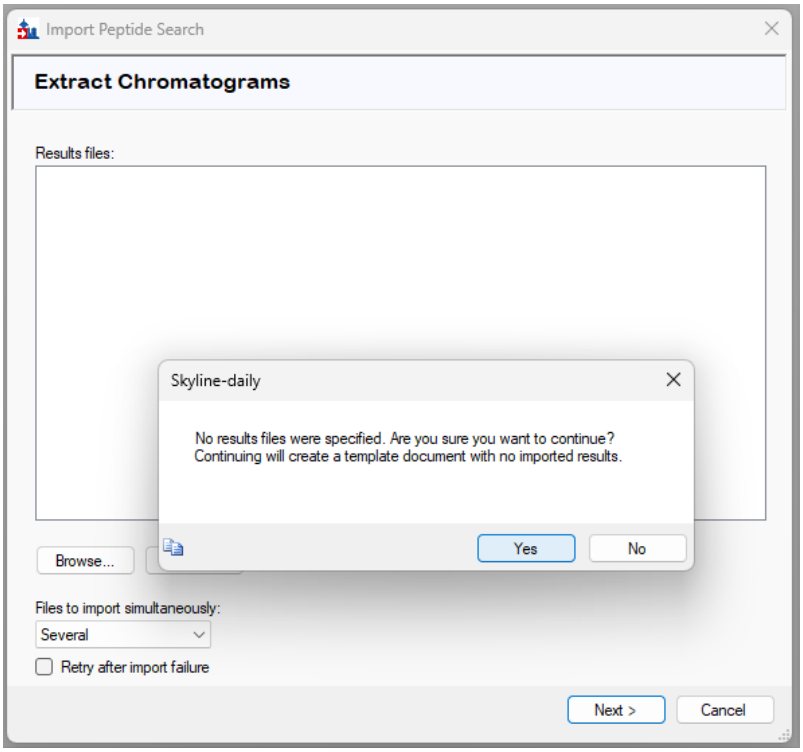
1.1. Start a new instance of **Skyline** and select the **"Import DIA Peptide Search" module**

- You will be prompted to save the file prior to import
- In the new window select **"Use existing"**
- Click **"Browse"** and locate the **"CSF_GPFLib_QRcombined.elib"** in the folder of this tutorial.
- Click **"Next"**



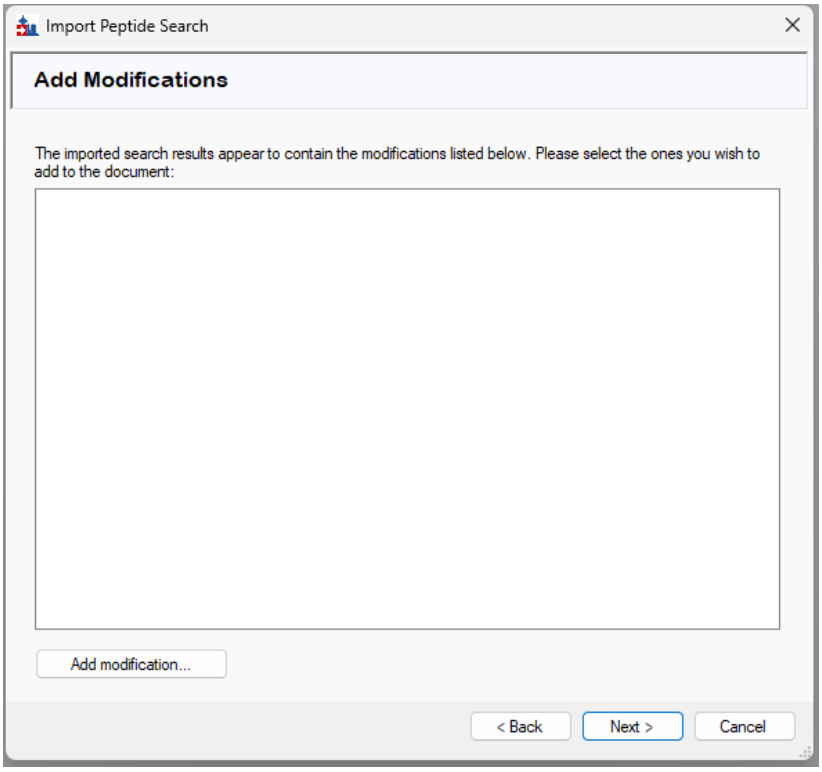
1.2. Extract Chromatograms

- Since we have multiple injections per replicate from our gas-phase fractionation, we will skip adding the results files at this time. Click **"Next"** to continue. A pop-up window will ask if you are sure you want to continue, select **"Yes"**.



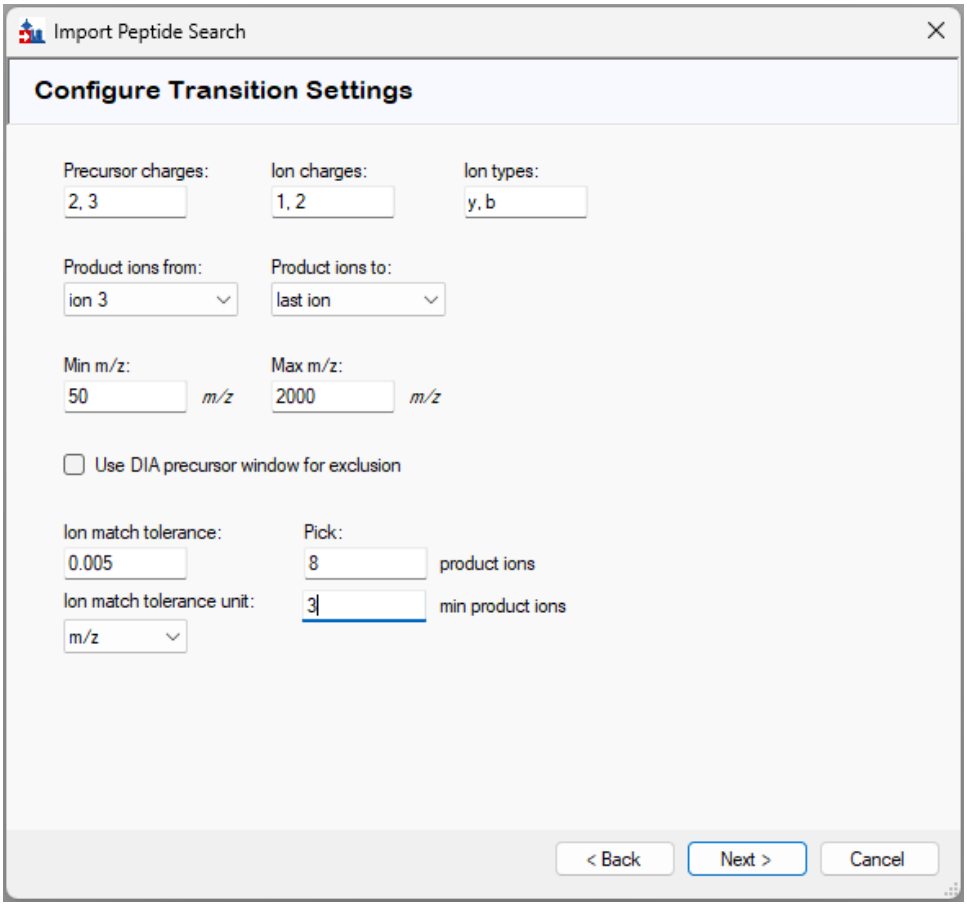
1.3. Add Modifications

- For this tutorial we did not include any modifications in our peptide search step; so we will just click **"Next"** to continue.
-



1.4. Configure Transition Settings

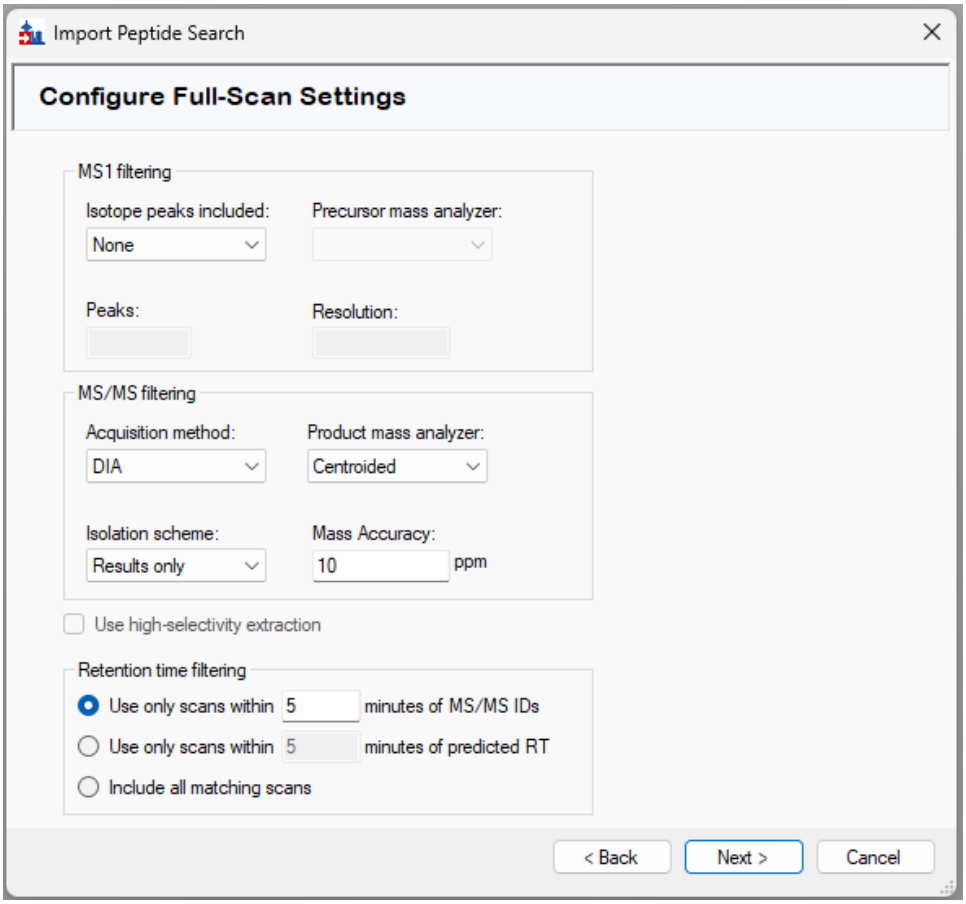
- For this dataset we will just extract the product ions using the following options:
    - Precursor charges: **“2, 3”**
    - Ion charges: **“1, 2”**
    - Ion types: **“y, b”**
    - Product ions from: **“ion 3”**
    - Product ions to: **“last ion”**
    - Min m/z: **“50”**
    - Max m/z: **“2000”**
    - Ion match tolerance: **“0.005 m/z”**
    - Pick: **“8 product ions; 3 min product ions”**
    - Click **“Next”**



Some of these options may be slightly different than what we would use for a regular DIA analysis; for example limiting us to the 3rd ion – this is a common heuristic used in targeted triple quad assays – our end goal!

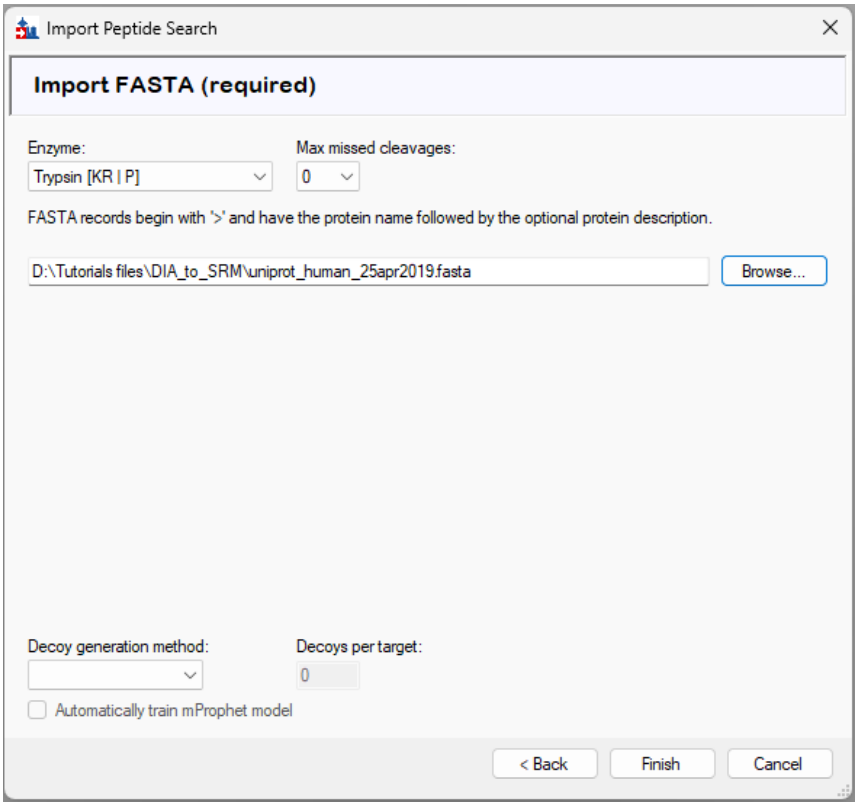
1.5. Configure Full-Scan Settings

- We will just be extracting the MS/MS from our results. Select the following options:
  - MS1 filtering: **"None"**
  - Acquisition method: **"DIA"**
  - Product mass analyzer: **"Centroided"**
  - Isolation scheme: **"Results Only"**
  - Mass Accuracy: **"10" ppm**
  - Retention time filtering: **"Use only scans within [5] minutes of MS/MS IDs"**
  - Click **"Next"**

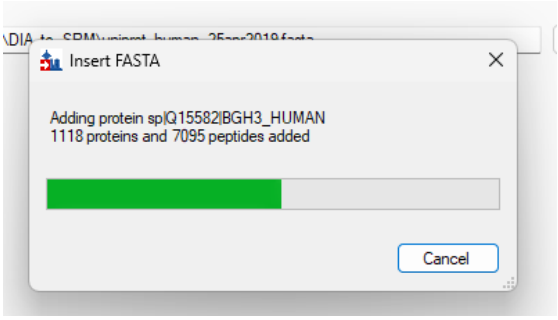


1.6. Import FASTA (required)

- For this dataset select **enzyme "Trypsin [KR|P]"** and set **Max missed cleavages to "0"**.
- Click **"Browse"** and navigate to the **"uniprot_human_25apr2019.fasta"** file.



- Click **"Finish"**
    - A pop-up will show the progress of the fasta processing

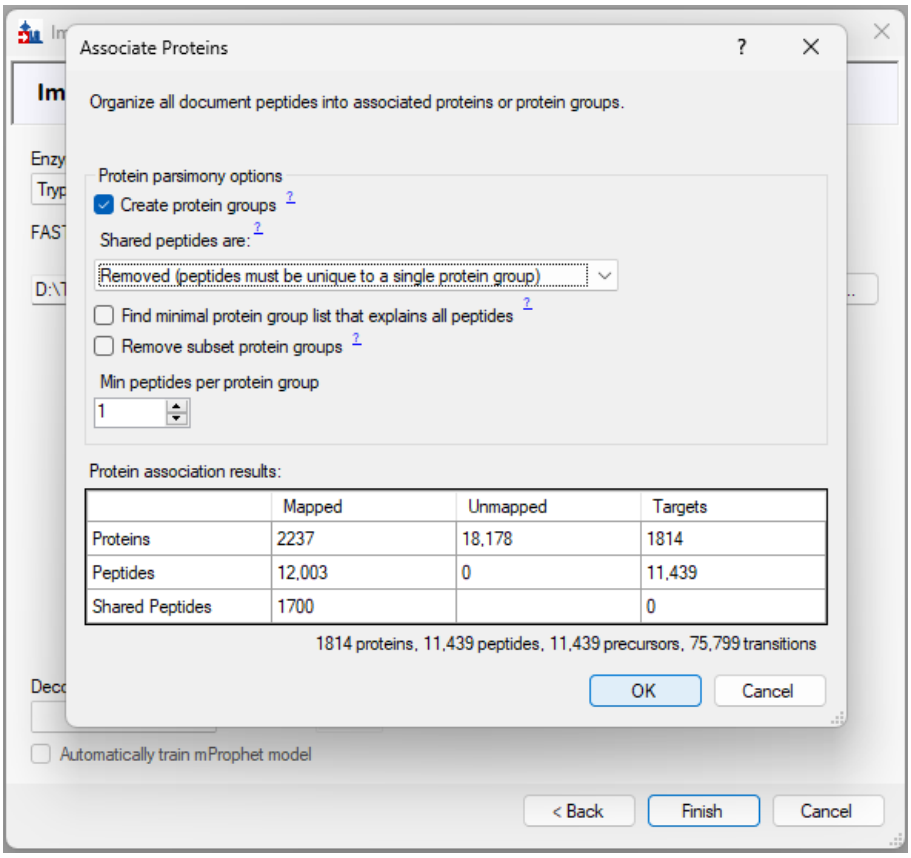


1.7. Associate Proteins

- A new pop-up will appear for associating proteins, us the following options:
    - Select **"Create protein groups"**
    - Shared peptides are: "**Removed (peptides must be unique to a single protein group)"**
    - Min peptides per protein group: **1**

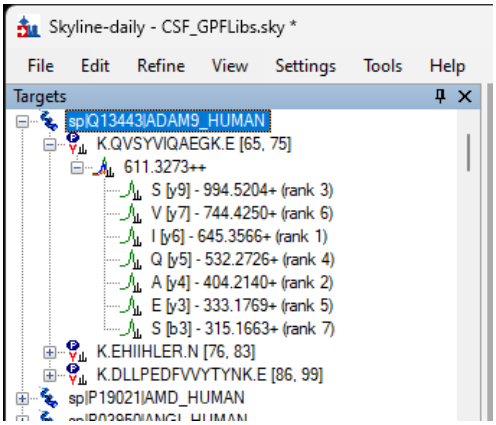
You will notice that the numbers in the bottom Protein association results table will change with the selection of different options for how to treat proteins and peptides.

"Mapped" = have peptides detected in the library, "Unmapped" = in the fasta but no peptides mapping to them, "Targets" = after protein grouping and pass above criteria, "Shared peptides" = in more than one protein entry in the fasta.



- Click "OK"
-

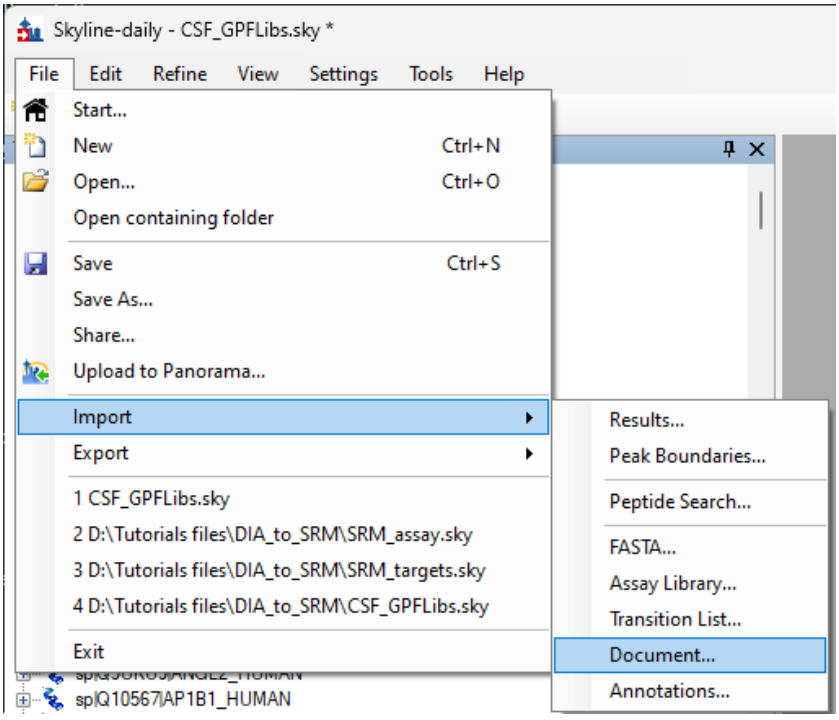
The targets section of your Skyline document should now be populated with the proteins, peptides, precursors, and fragments detected in your library, in this case from EncyclopeDIA results. There is some additional filtered out based on our selected parameters – e.g. 0 missed cleavages.

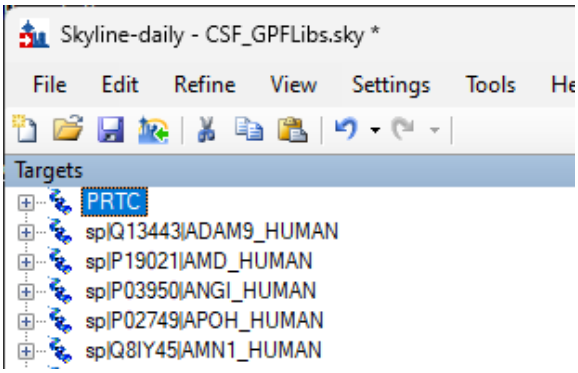


1.8. Adding PRTC peptides.

For scheduling our SRM assays we will need an indexed retention time standard, such as PRTC. We included PRTC in the samples in this DIA experiment, so now we need to add them to our document since they were not included in the peptide search/library. We want to do this prior to importing results and extracting chromatograms to ensure they are extracted. There are a couple ways to do this, but to save time/energy I prefer to copy and paste them from an existing document. If they are added based on sequence or a fasta it requires manual annotation for the heavy labeled residues.

- Go to **File > Import > Document**
    - Find the file "**PRTC.sky**" and click "**Open**"



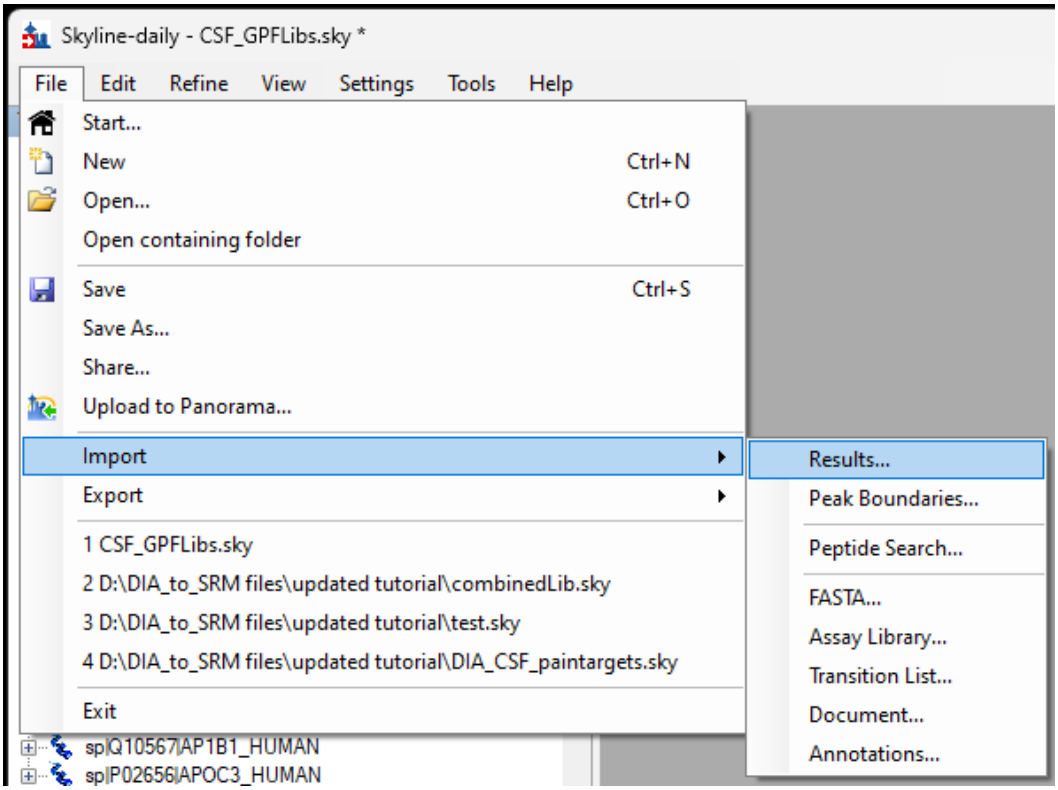


- **Save** the document.

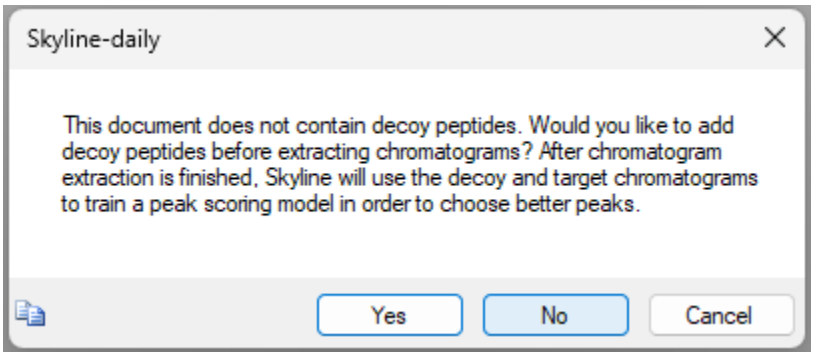
1.9. Importing gas-phase fractionated DIA results into Skyline

Now that we have our environment all set up we are ready to import and extract the chromatograms for all of the peptides we have populated from our library. Since we gas-phase fractionate with 6 separate injections, we will import the results as multi-injection replicates.

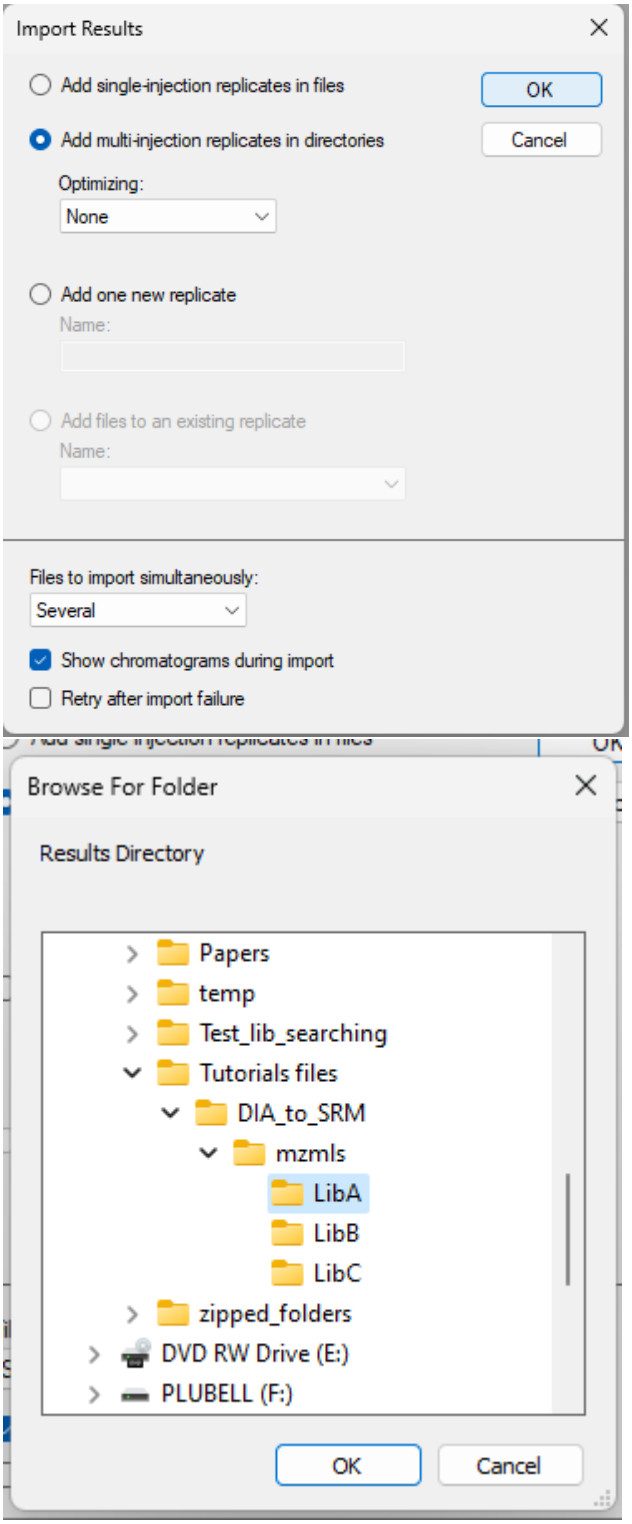
- Navigate to **File>Import>Results**



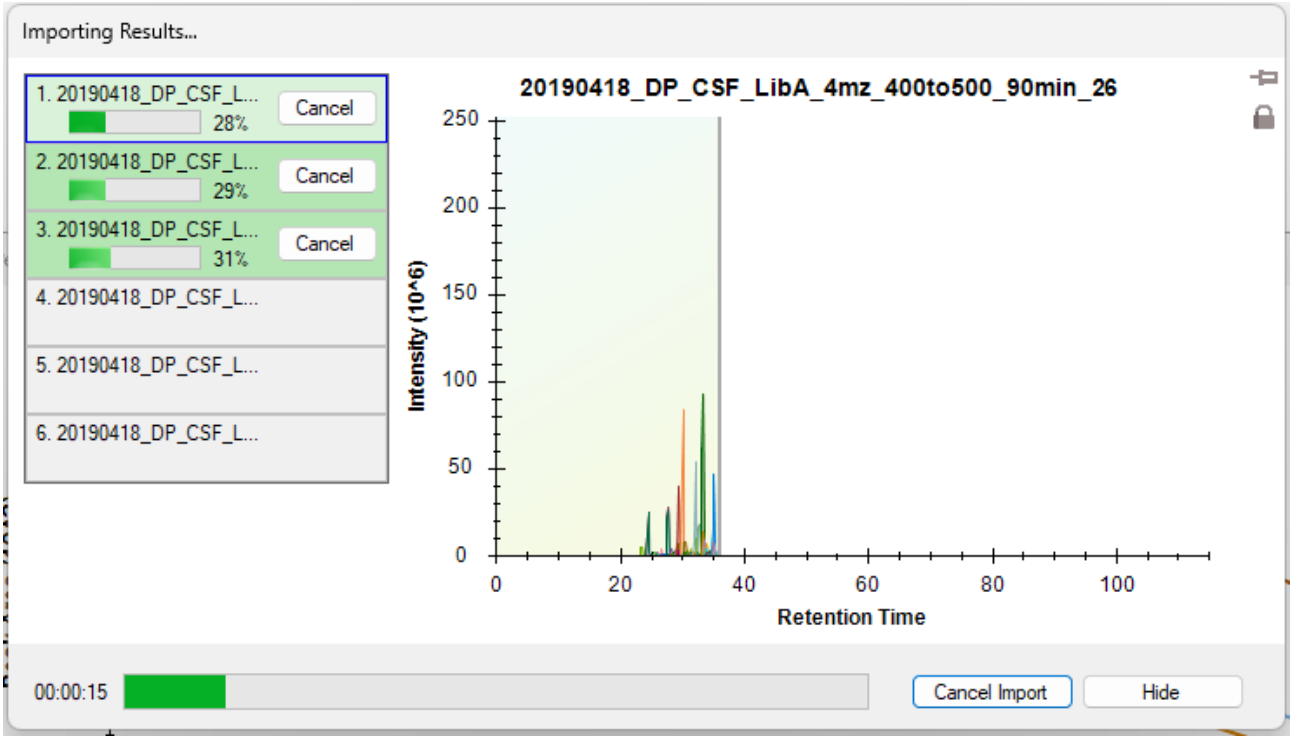
- In the pop-up asking if you want to add decoy peptides select **"No"**.



- Select "**Add multi-injection replicates in directories**"
- Navigate to the "**mzmls**" subfolder and select the three subfolders; **LibA**, **LibB**, and **LibC**



- Import may take a couple minutes.



- **Save the document!**

We now have all the DIA results matching our document settings. Next, we will go through those results to filter for just reproducible peptides.

Step 2. Refining and filtering for high-quality peptide detections in DIA

The power of this method is that we can be confident the peptides for these target proteins will be measured in our own samples because we have already observed them in our own DIA experiment. Additionally, we can inform our selection of peptides based on some of the prior knowledge we have gained through our DIA measurements. Specifically, we have fragment ion intensity and reproducibility information - as measured by percent coefficient of variation (%CV) calculated for 3 replicate injections. This information then helps us select peptides for each protein that will most likely have good performance in an SRM assay.

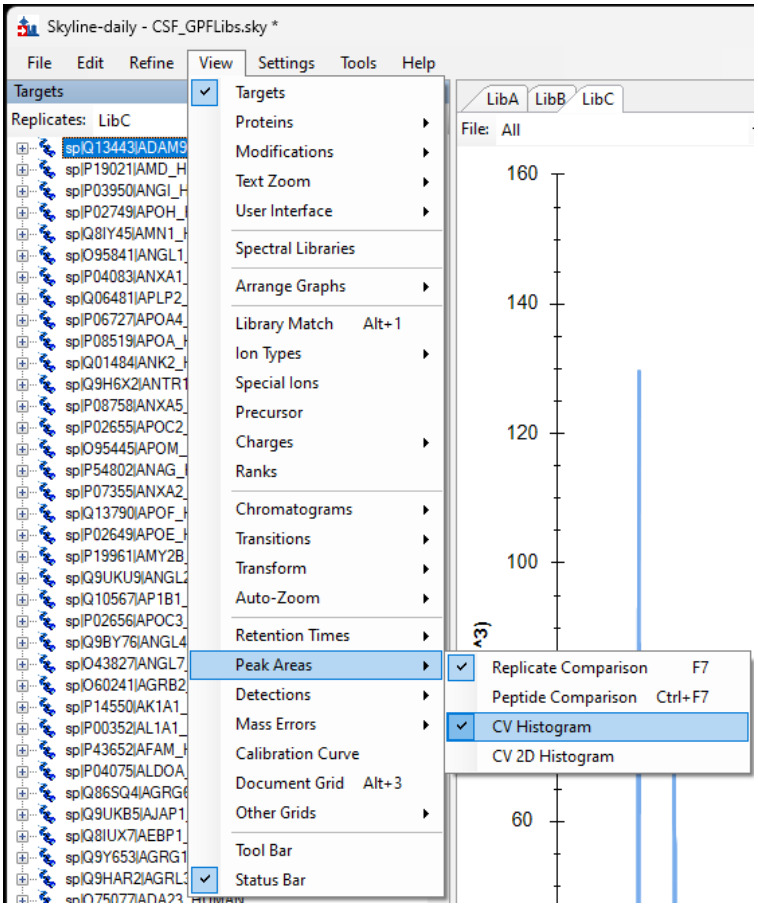
For the peptide refinement we will start by just doing some global refinements. This will leave us with a list of peptides that can serve as a set of known reproducible peptides that we can revisit for many different proteins as needed in the future. However, this filtering could also just be done for specific proteins as we demonstrate in Step 3. I prefer to first filter by peptide %CV, then by peptide

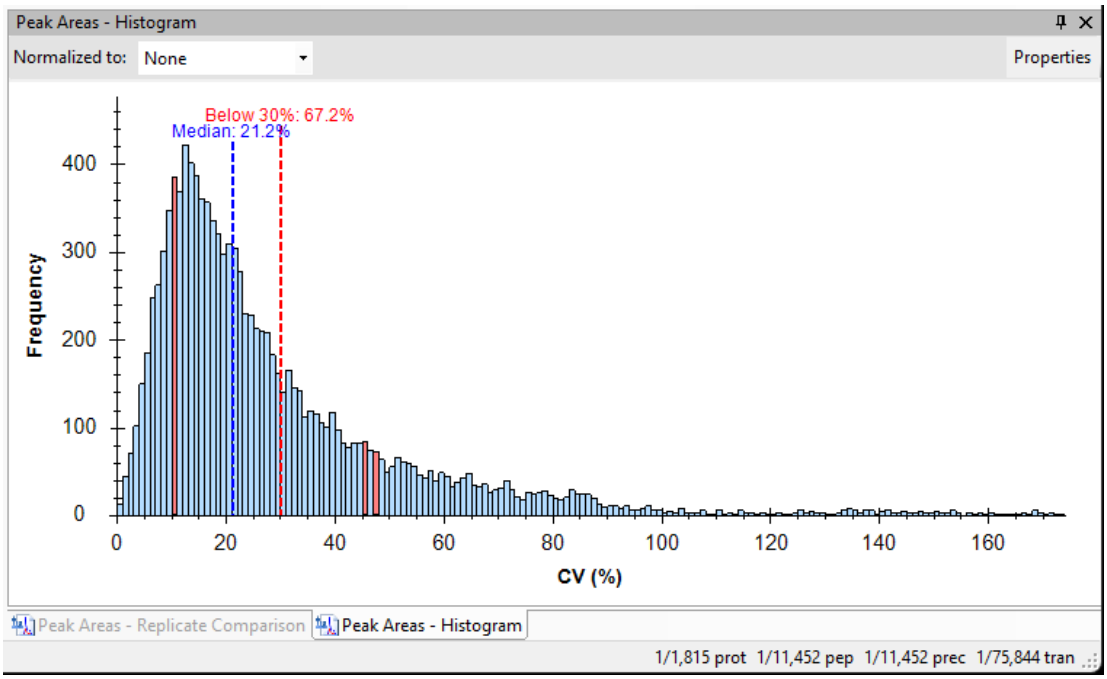
intensity rank, however the filtering can be done in the opposite order as well. Generally, I do %CV first to get rid of bad detections or to catch any poor chromatography or reproducibility, then the peptide intensity to pick the best of the best for each protein and then fragment intensity rank for each peptide. Additionally, the filtering parameters can be relaxed, or specific peptides of biological interest can be kept/added back. However, an important caveat is that we can only use this method for proteins we detect in our DIA experiment.

2.1. Peptide coefficient of variation (%CV)

Skyline automatically calculates the %CV for every peptide if there are replicates. This information can be found in optional columns in the document grid. First let's get a feel for the general %CV distribution in this document.

- Navigate to **View>Peak Areas>CV Histogram.**



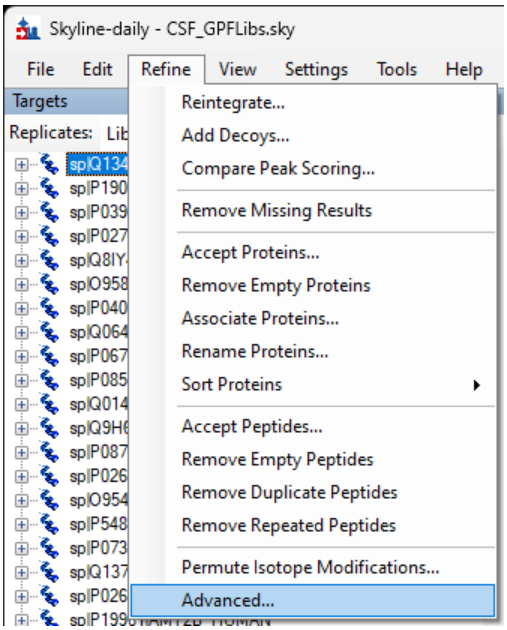


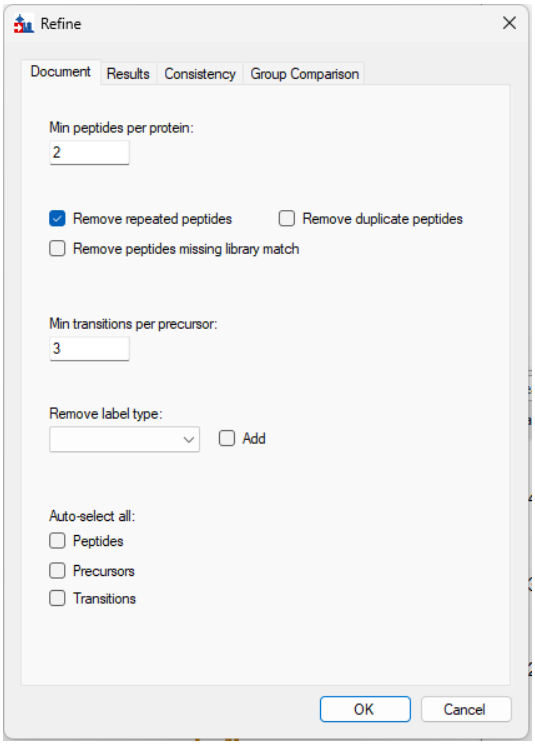
The default view will have the red dashed line at 20% CV, you can adjust it to show 30% by clicking "Properties" and adjusting the "CV cutoff" to 30.

3.2. Target refinement

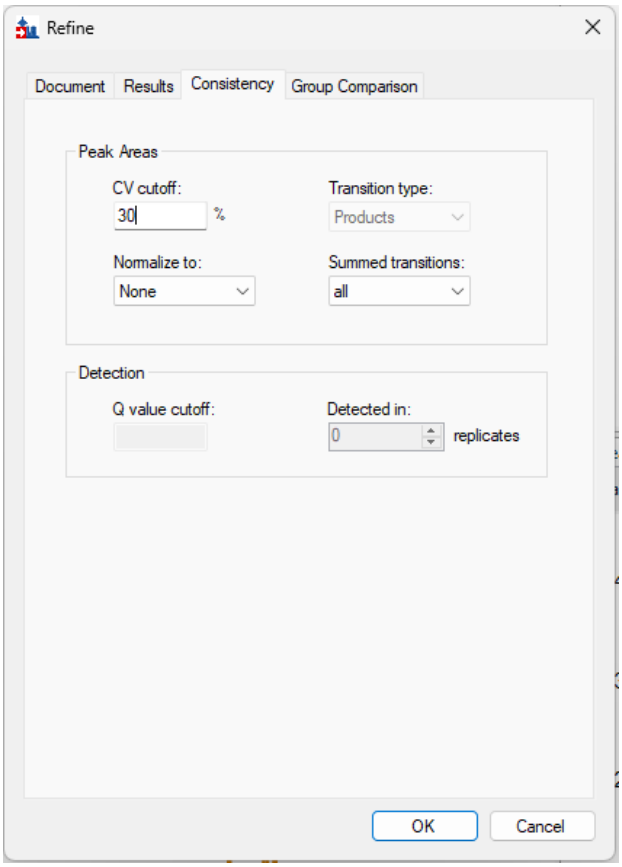
As a general cutoff let's impose the following rules for the peptides we would want to include in our database of possible peptides suitable for targeted assays: there must be 2 peptides per protein, and peptides must be below 30% CV across the three replicates.

- Navigate to **Refine>Advanced,** a new pop-up window will appear with multiple refinement options
- In the **Document** tab select the following options:
    - Min peptides per protein: **2**
    - Select **Remove repeated peptides**
    - Min transitions per precursor: **3**

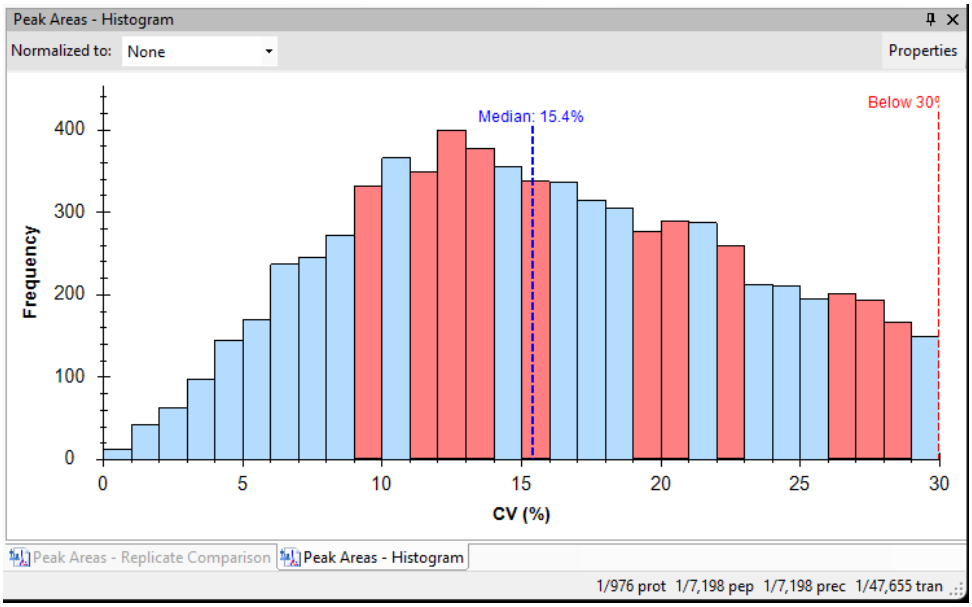




- In the **Consistency** tab select the following options:
  - CV cutoff: **30%**
  - Transition type: **Products**
  - Normalize to: **None**
  - Summed transitions: **all**



- Click **OK**
- **Save as…** to save the filtered document, e.g. : **"CSF_GPFLibs_filtered.sky"**
-



We have now filtered all the out peptides and proteins that are less reproducible. Specifically, you can see in the bottom right hand corner of the Skyline document that we have **976 proteins** that have 2+ peptides with at least 3 transitions and a %CV <30% (**7,198 peptides**).

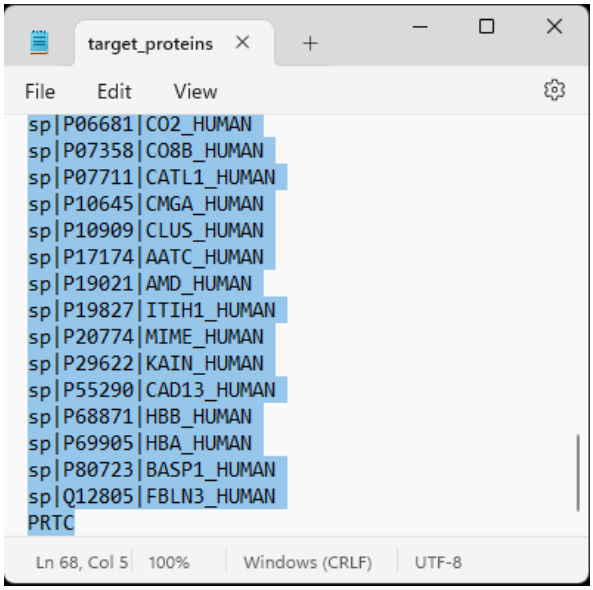
Step 3. Peptide filtering for protein targets

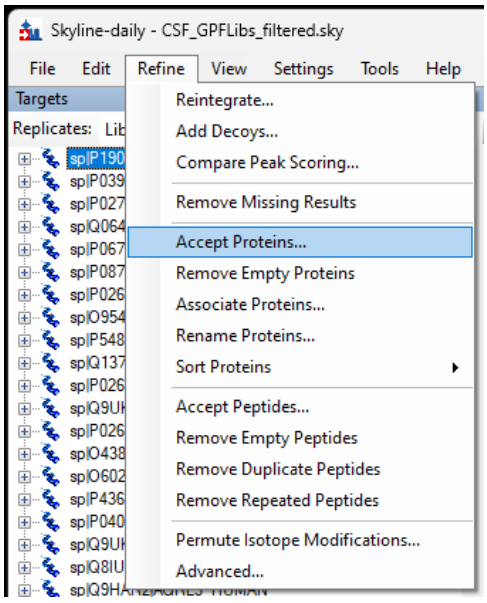
For this example, we are proposing that we are interested in a subset of 67 proteins in CSF that have previously been described as differentially abundant in a study of interest. Now we can focus just on selecting peptides from those proteins

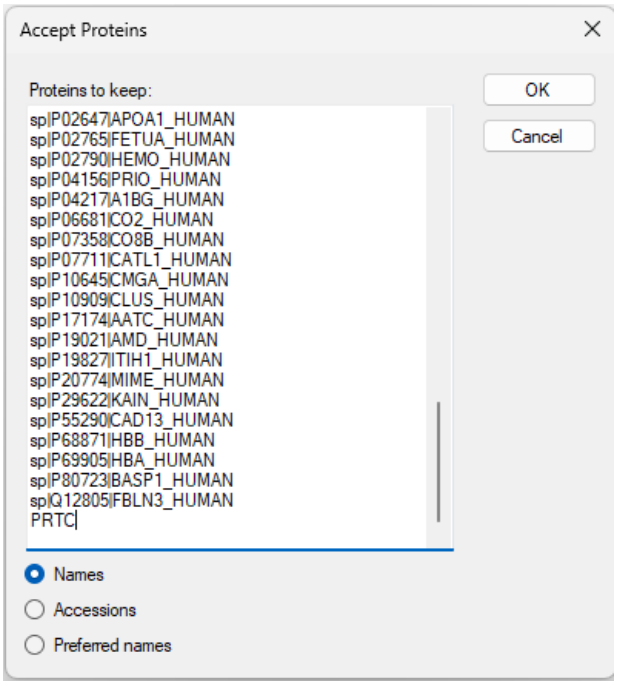
3.1. Protein filtering

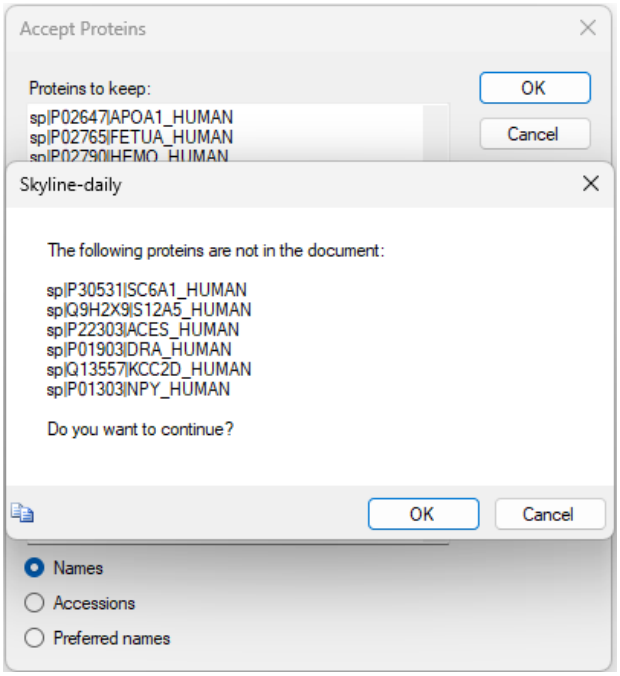
For this step we will only keeping proteins in our document if they are present in our list.

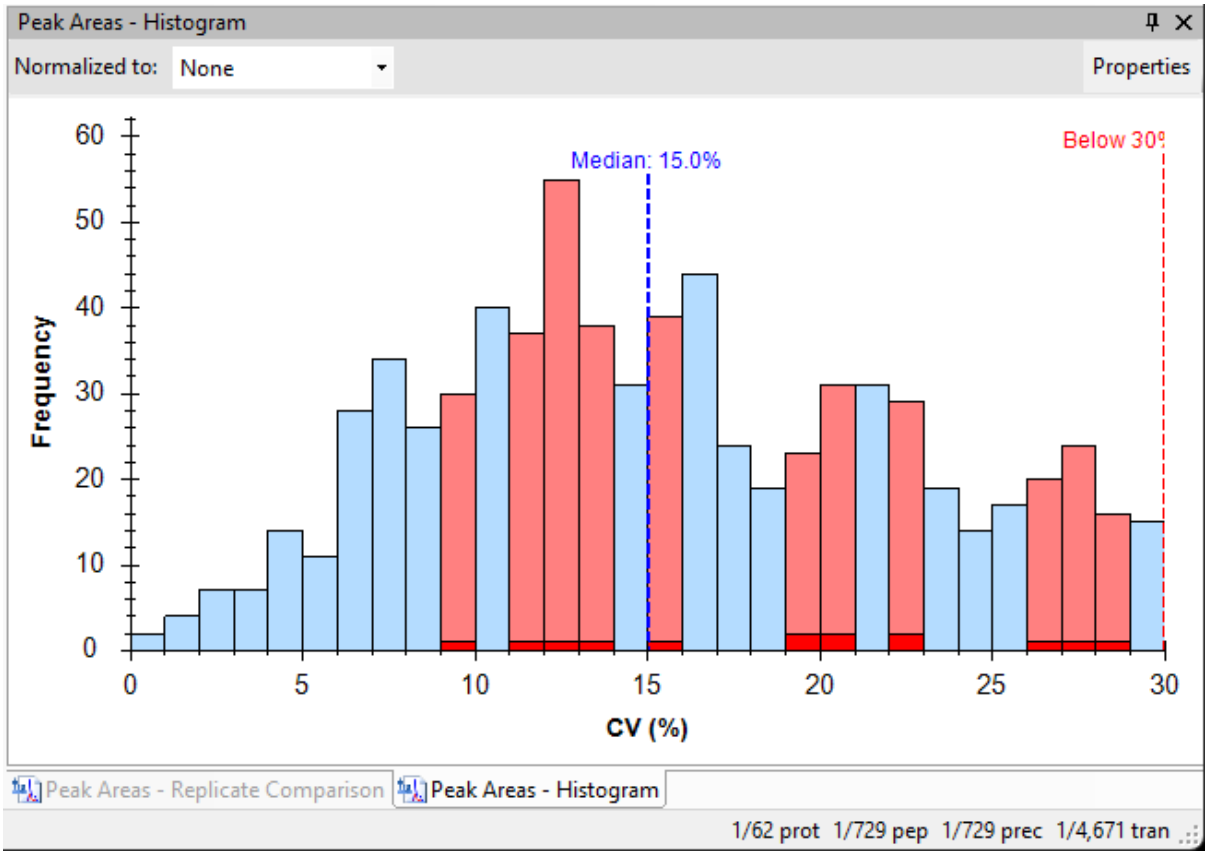
- Open the "**target_proteins.txt"** file that contains a list of protein names.
- Select all and copy: **"ctrl+A"** then **"ctrl+C"**
- Switch back to the Skyline document, then navigate to **Refine>Accept proteins…**
- In the new pop-up window paste (**"ctrl+V"**) the copied list, and select **"Names"**
- Click **OK;** a new pop-up will appear listing the proteins in the list that are not in the document, click **OK** to continue.






Accept Proteins
Proteins to keep:
sp|P02647|APOA1_HUMAN
sp|P02765|FETUA_HUMAN
sp|P02790|HEMO_HUMAN
sp|P04156|PRIO_HUMAN
sp|P04217|A1BG_HUMAN
sp|P06681|CO2_HUMAN
sp|P07358|CO8B_HUMAN
sp|P07711|CATL1_HUMAN
sp|P10645|CMGA_HUMAN
sp|P10909|CLUS_HUMAN
sp|P17174|AATC_HUMAN
sp|P19021|AMD_HUMAN
sp|P19827|ITIH1_HUMAN
sp|P20774|MIME_HUMAN
sp|P29622|KAIN_HUMAN
sp|P55290|CAD13_HUMAN
sp|P68871|HBB_HUMAN
sp|P69905|HBA_HUMAN
sp|P80723|BASP1_HUMAN
sp|Q12805|FBLN3_HUMAN
PRTC
OK
Cancel
Names
Accessions
Preferred names

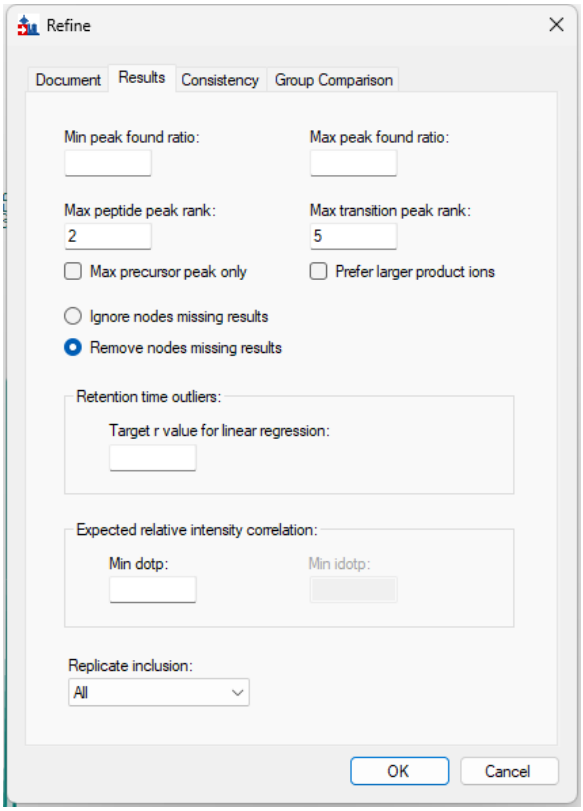

Accept Proteins
Proteins to keep:
sp|P02647|APOA1_HUMAN
sp|P02765|FETUA_HUMAN
OK
Cancel
Skyline-daily
The following proteins are not in the document:
sp|P30531|SC6A1_HUMAN
sp|Q9H2X9|S12A5_HUMAN
sp|P22303|ACES_HUMAN
sp|P01903|DRA_HUMAN
sp|Q13557|KCC2D_HUMAN
sp|P01303|NPY_HUMAN
Do you want to continue?
OK
Cancel
Names
Accessions
Preferred names

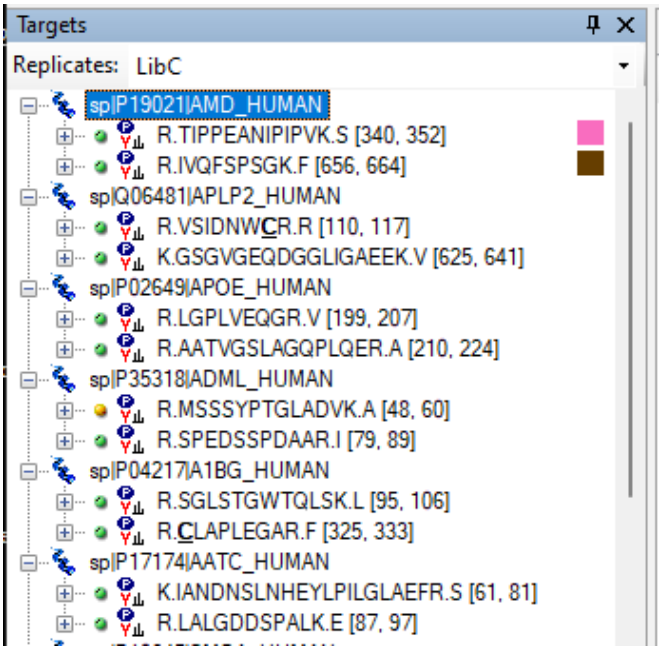

Peak Areas - Histogram
Normalized to:
None
Properties
Median: 15.0%
Below 30%
Frequency
CV (%)
0
10
20
30
40
50
60
0
5
10
15
20
25
30
Peak Areas - Replicate Comparison
Peak Areas - Histogram
1/62 prot 1/729 pep 1/729 prec 1/4,671 tran

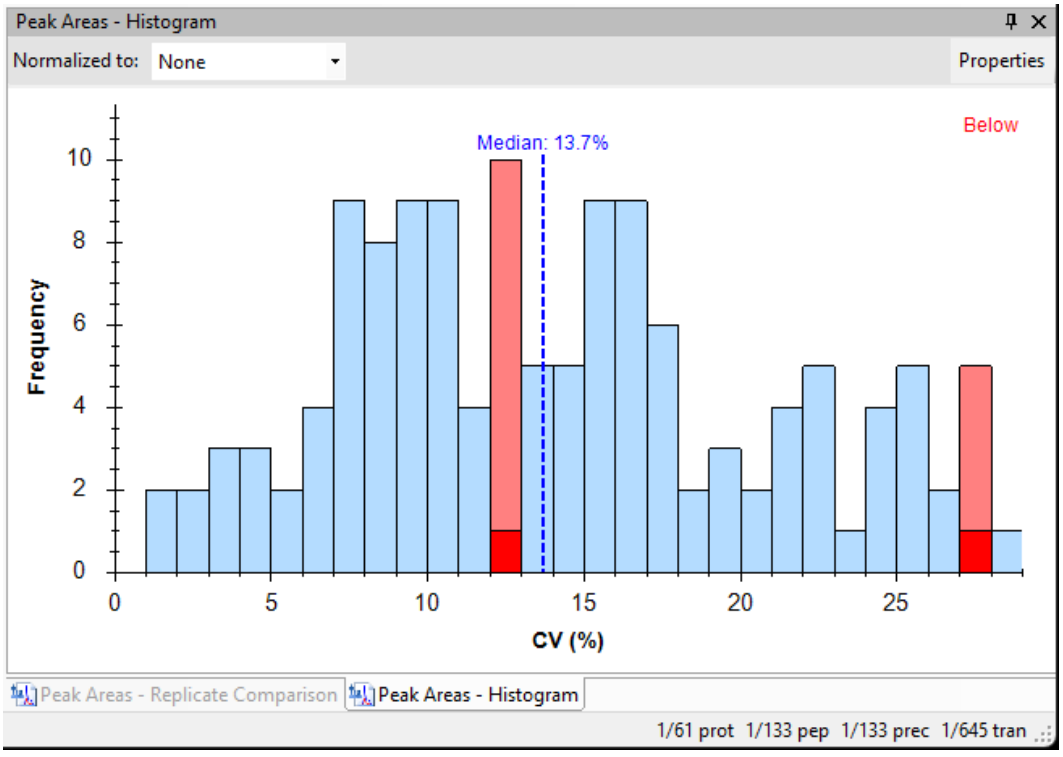
We now have a document that only contains peptides matching the proteins from our list. Additionally, these are only peptides with good reproducibility as enforced by our %CV filtering. Depending on the number of target proteins or the aim of the project additional filtering for peptides and transitions can be done at this point. In this case we are simply going to limit our peptide targets to two per protein, and limit the transitions to 5 per peptide. This is cutting out some manual assessment for the sake of this tutorial, but additional manual assessment and literature searching is highly encouraged for a smaller assay or for accounting for potential confounding biology of the proteins – like understanding which isoforms or which portion of the protein will be detectable with the peptides selected.

3.2. Peptide ranked intensity filtering

- Navigate to **Refine>Advanced…**
- In the **Results** tab set the following options:
    - Max peptide peak rank: **2**
    - Max transition peak rank: **5**
    - **Remove nodes missing results**.







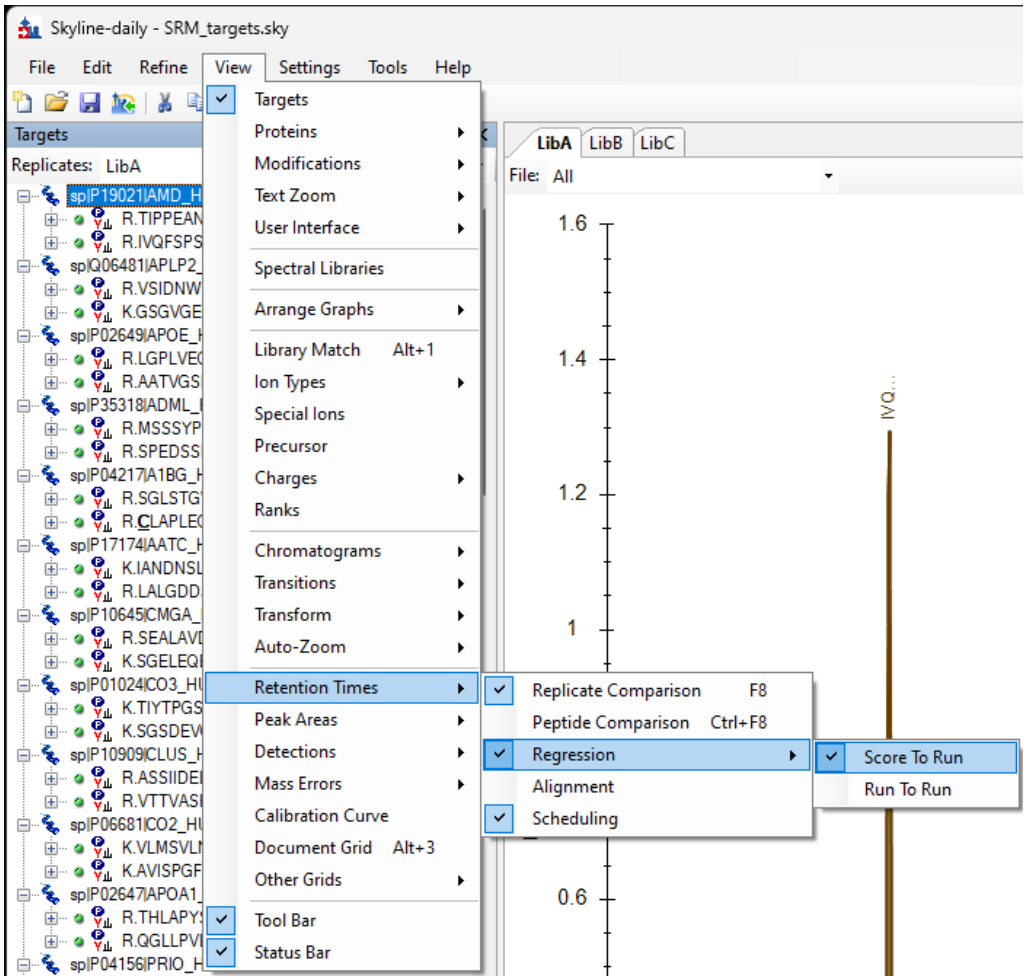
- Save file as “SRM_targets.sky”
-

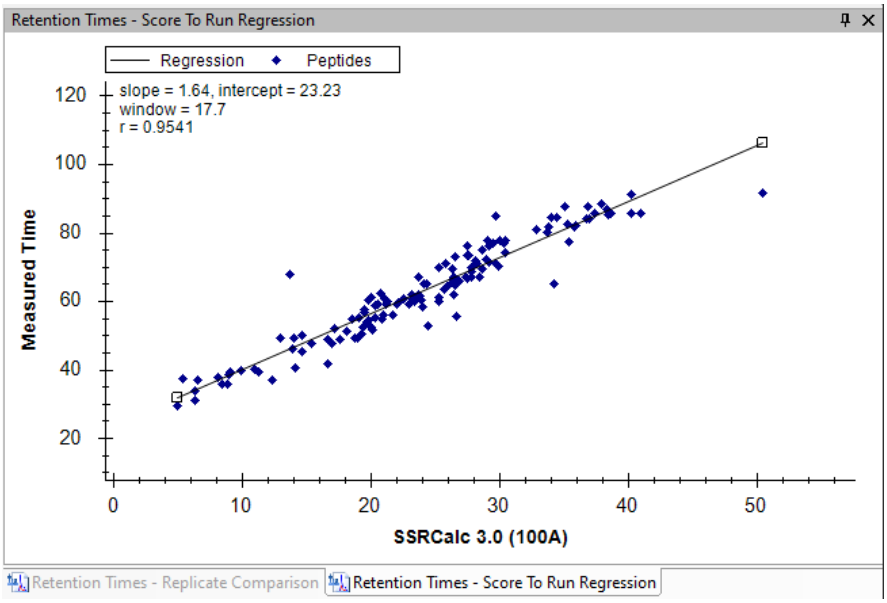
We should now have a document with 61 proteins with 2 peptides, and 13 PRTC peptides. We will use this as our list for our SRM assay. The reason I keep 5 transitions for each peptide is to ensure I have some room to adjust our list after an initial round of SRM assay testing.

Step 4. Scheduling an SRM method based on DIA iRTs

There are two strategies to make use of the retention time measured in our DIA experiments. One easy way may be using it as a library to confirm peaks are correctly picked in an unscheduled SRM assay. Another way we can use them is by directly scheduling a SRM run using an iRT standard included in the DIA runs. This requires a couple runs of just the iRT sample on the HPLC and instrument we desire to run our SRM assay on. In this case it is an Easy nLC and a Thermo Altis triple quad. For more information about iRT and how to use iRT in your own experiments you can follow both the "iRT retention time prediction" tutorial or webinar #7 of the same name found on the Skyline website. In this example we will build a calculator from our DIA results, then use it for scheduling with a shorter gradient SRM method.

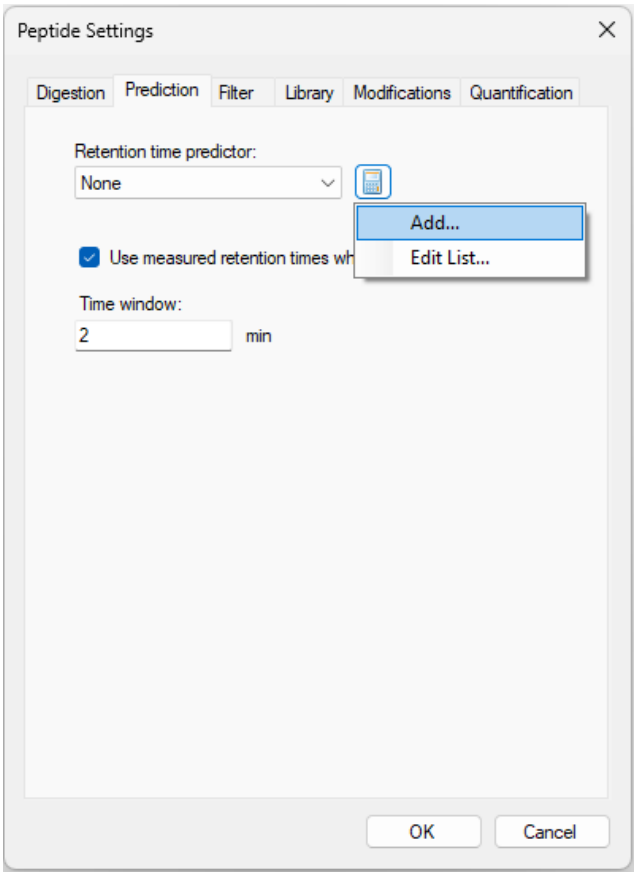
- In the **Retention Times - Replicate Comparison** window, right click and select **Graph** > **Regression** > **Score To Run**.





4.1. Build an iRT Calculator.

- Go to **Settings** > **Peptide Settings** > **Prediction**
- Click on the icon of a **calculator** and select **Add...**

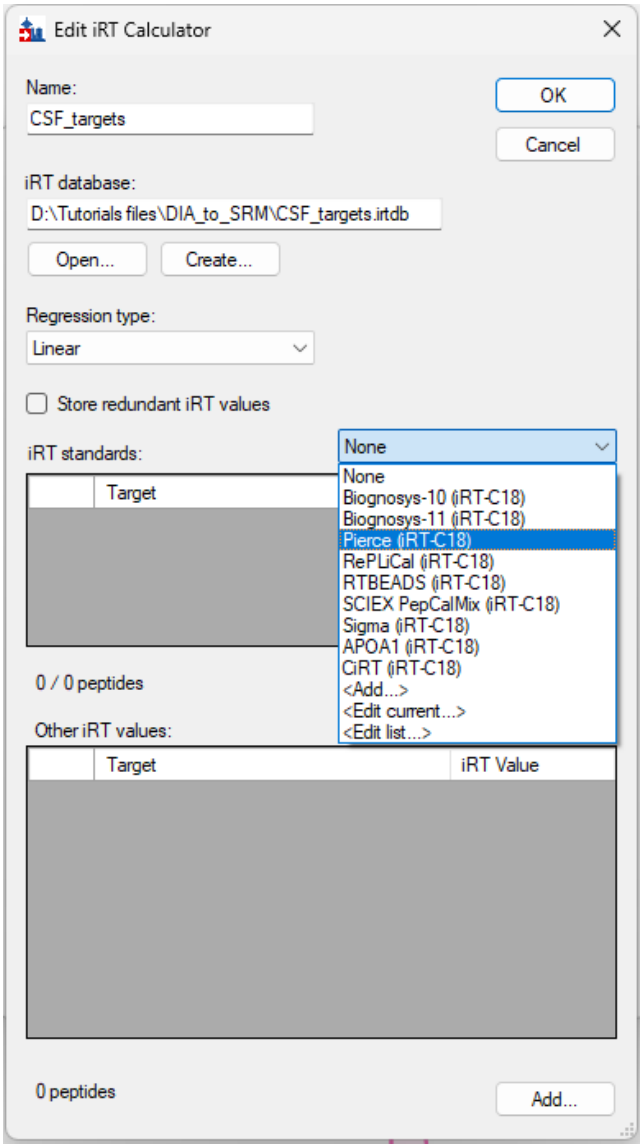


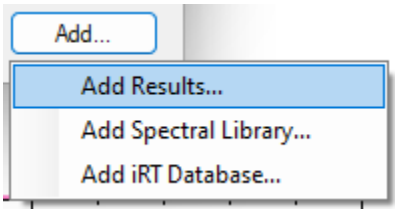
- Give your iRT calculator a **name** (I'll call it **CSF_targets**) and then click "**Create…**"

- In the pop-up window, "**Create iRT Database**", we'll also give the iRT calculator **FILE** a name.

  In the step above, we just gave it a nickname that we'll see in Skyline. Here, we'll name the actual computer file (it's an *.irtdb file) so make sure not to use any strange characters. I'll save my calculator to this tutorial folder and name it the same as before: **CSF_targets.**

- Click "**Save**" when you're done.
- In the **Edit iRT Calculator** window, select **Pierce (iRT-C18)** from the dropdown menu across from **iRT standards:**

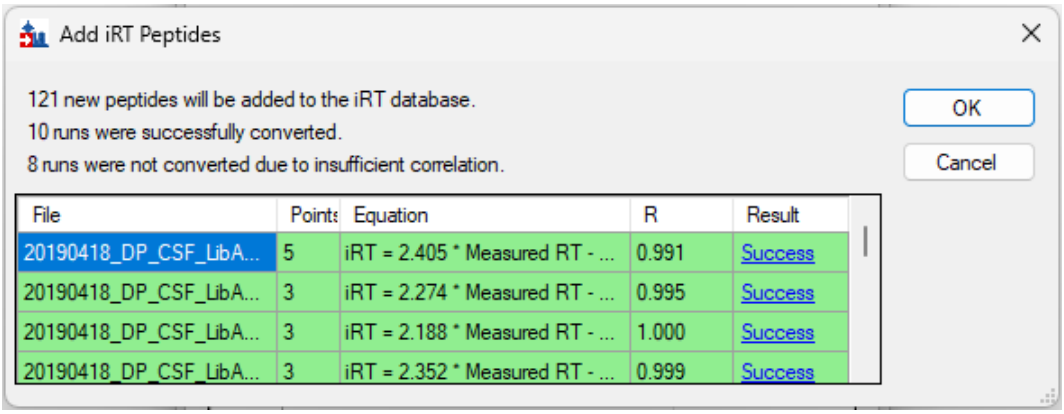




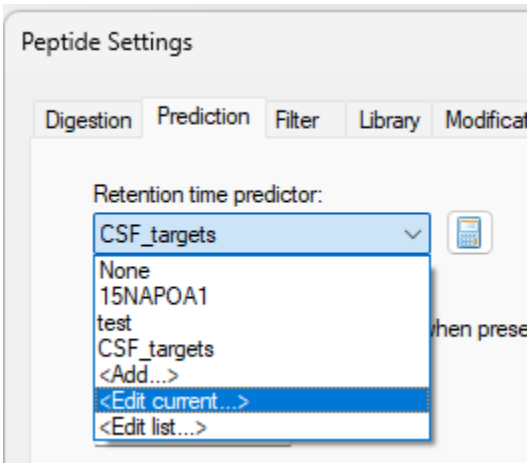
We have just set our iRT values! Recall that iRT values are unitless, they don't really mean anything by themselves besides being a relative value relating to the order of elution; the values range from -27.60 to 100.

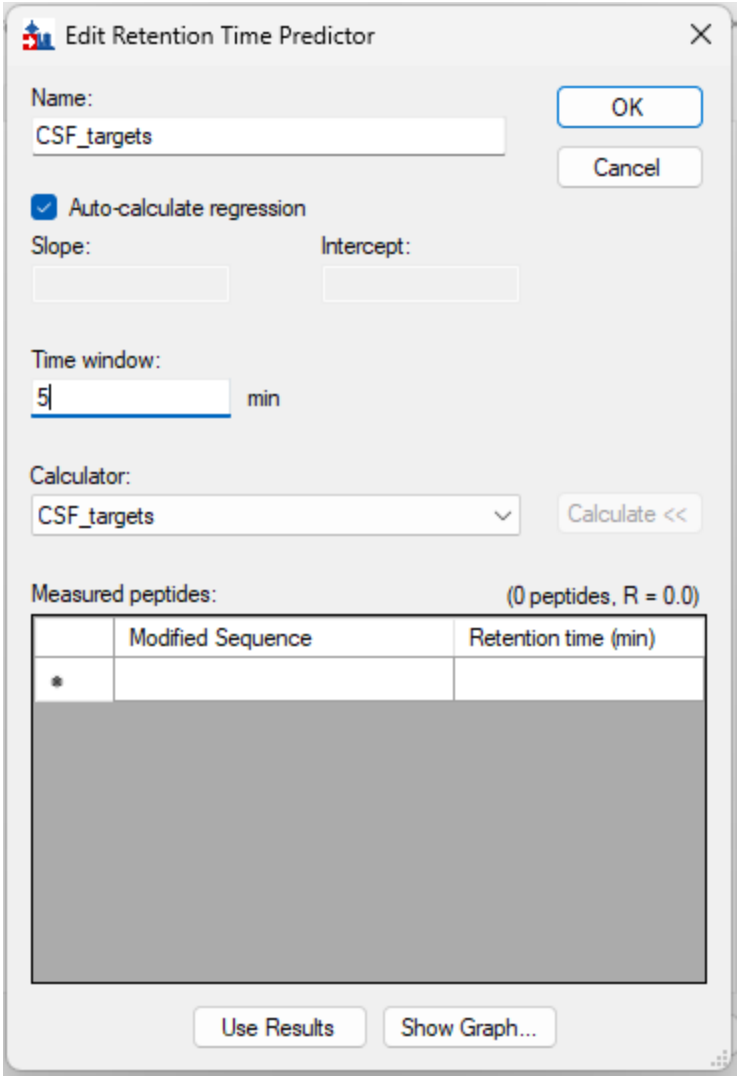
4.2. Add the peptides to the iRT calculator.

- Click the **Add** in the bottom right hand corner of Edit iRT Calculator window
- In the dropdown menu select **Add Results…**
    - In the **Add iRT Peptides** pop-up window, select "**OK**"

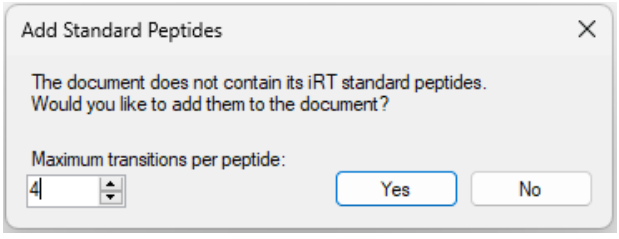


- Click "**OK**" at the top of the **Edit iRT Calculator** window.
- In the Peptide Settings window the **"Pain_PRTC"** should now be selected in the retention time predictor.
- Click to dropdown and select **<Edit current>**
    - Adjust **"Time window"** to **5** min, and click "**OK**"

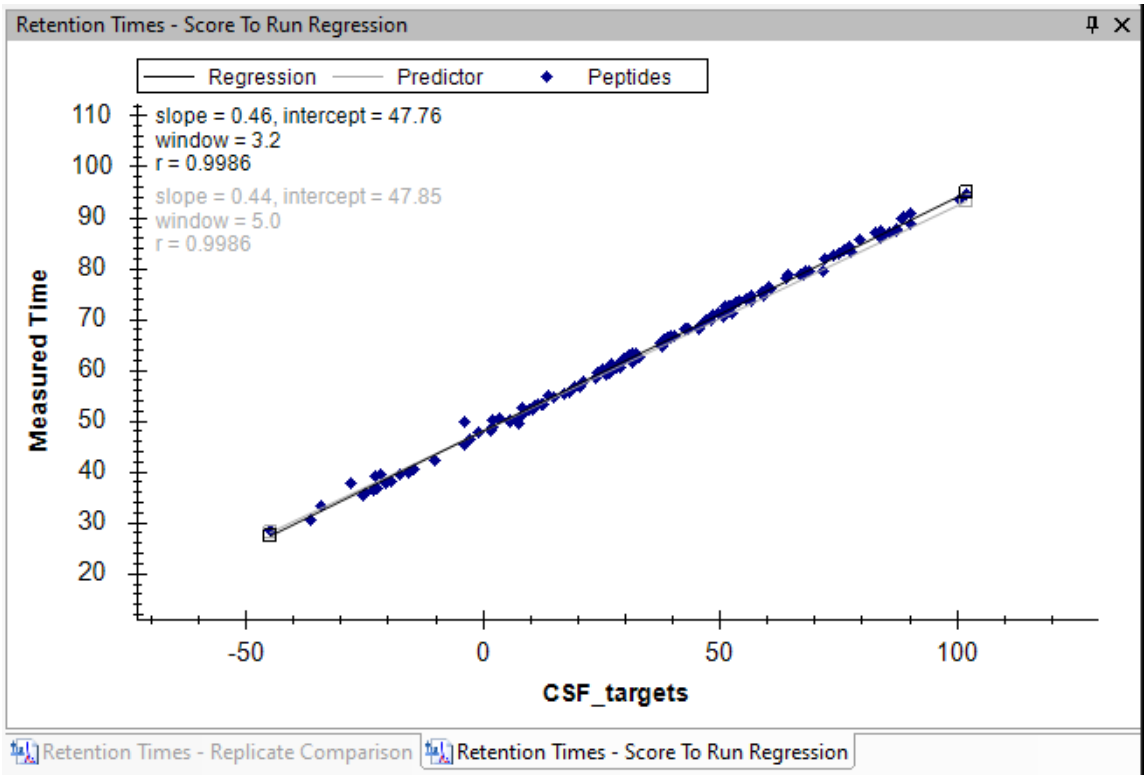




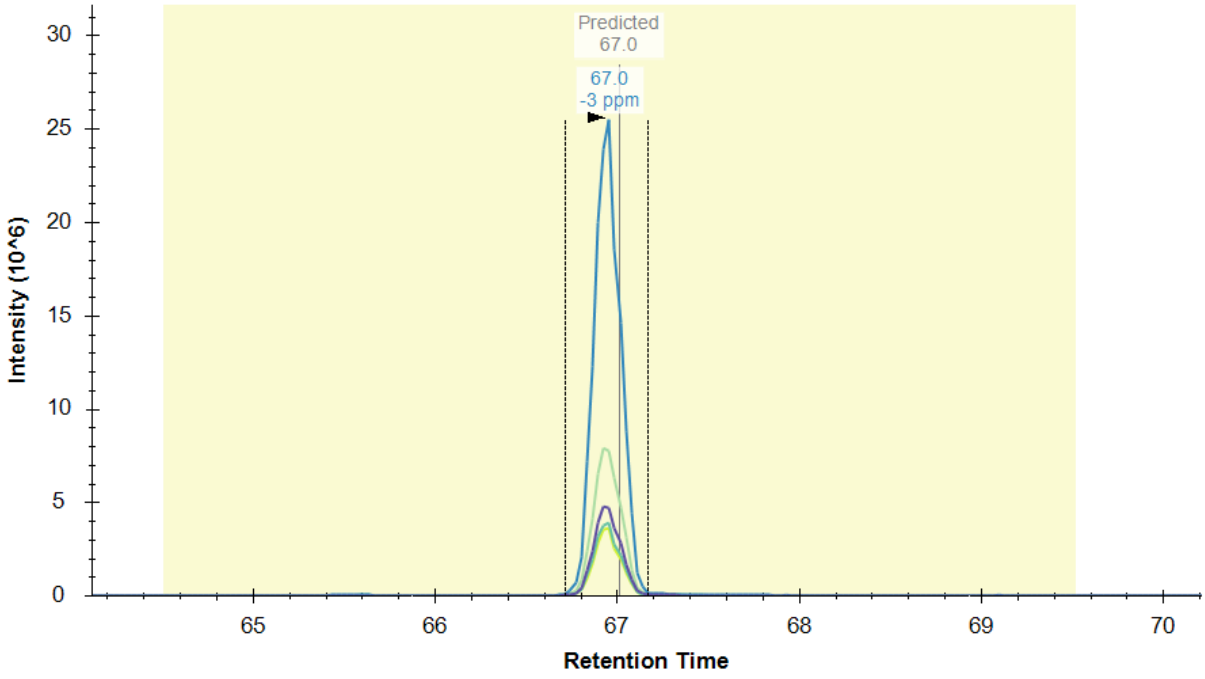
- In the prediction tab set the **"Time window"** to **5** min
- Click "**OK**"
    - In the Add Standard Peptides window, adjust the Max transitions to **4** and click "**OK**"



- **Right click** on the **Regression** plot, select **Calculator** > **CSF_targets**
- Your regression should now be updated



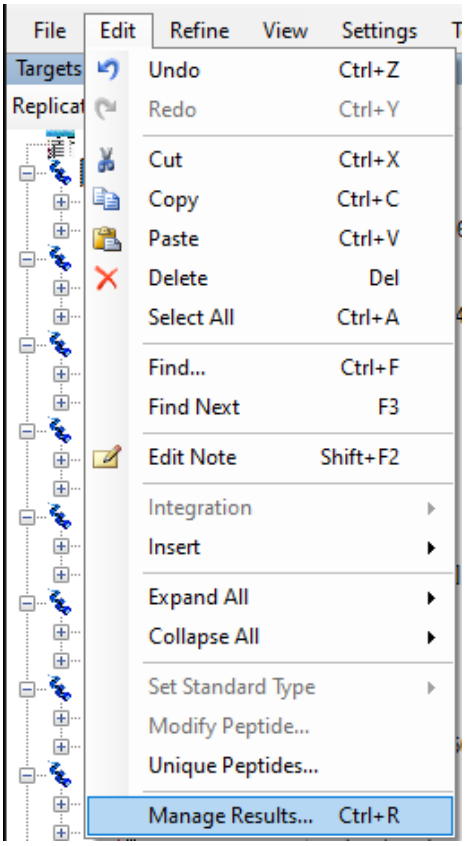
- If you look at the chromatograms you can see that there is a beige/light yellow section on either side of the now predicted retention time for each peptide.

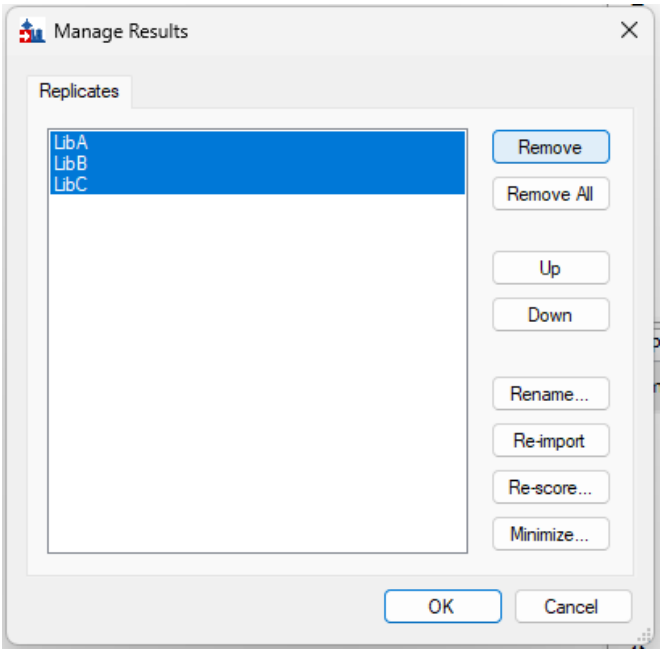


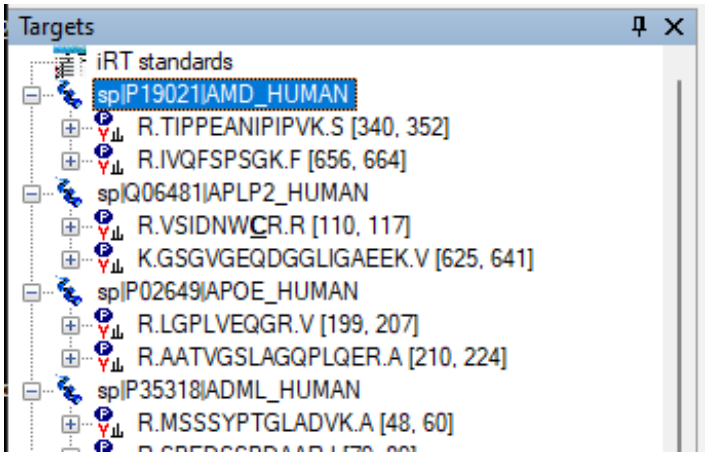
4.3. Scheduling for the LC and mass spec used for the SRM assay.

For this step we acquire and import unscheduled, iRT standards-only data files on your new gradient, HPLC, and column set up.

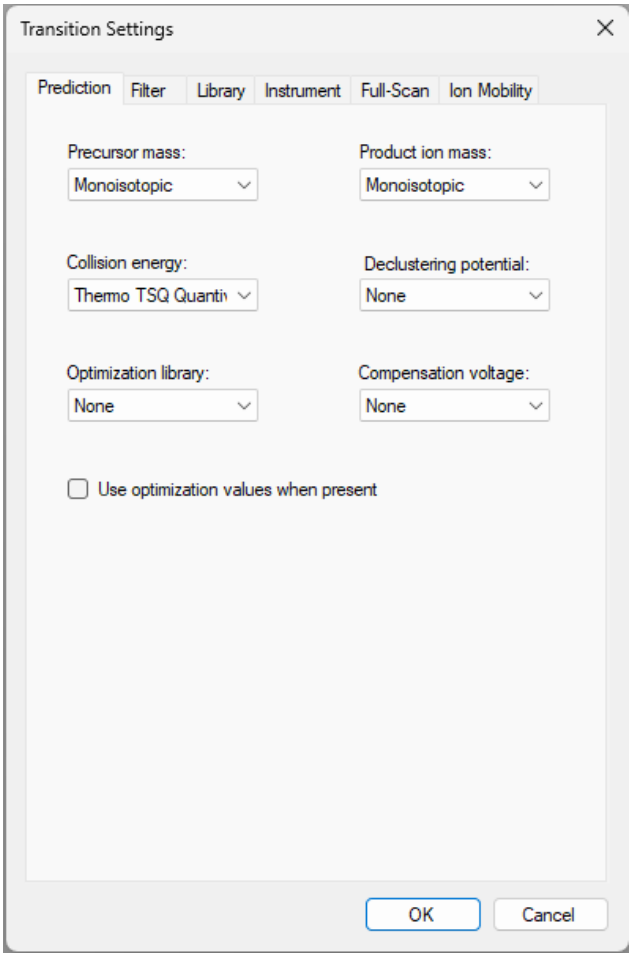
- Remove the DIA results by navigating to **Edit>Manage Results…**
- In the dialogue box click "**Remove All**" and click "**OK**"

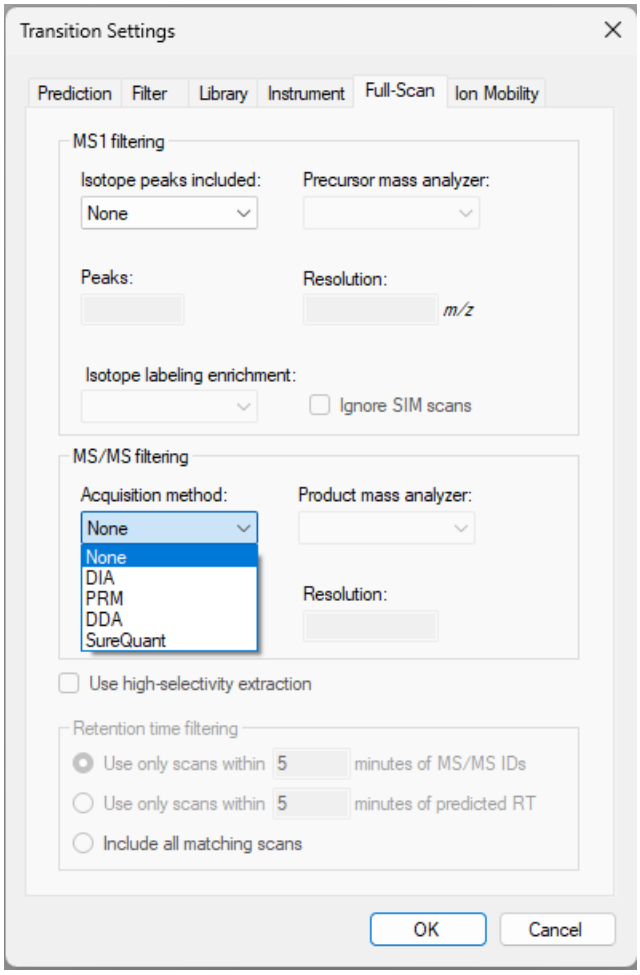






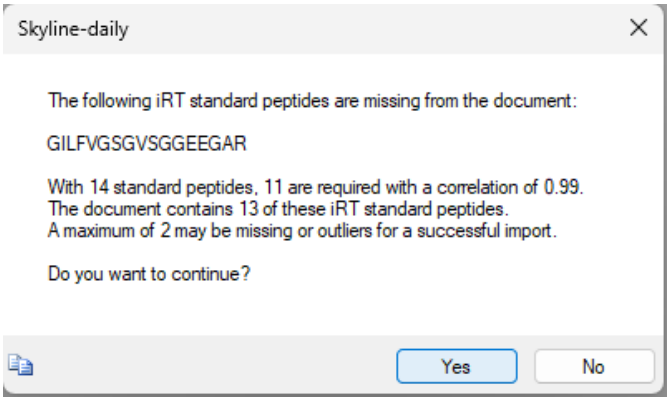
- Go to **Settings** > **Transition settings**
- In the **Prediction** tab use the following settings:
    - Precursor mass: **Monoisotopic**
    - Product ion mass: **Monoisotopic**
    - Collision energy: **Thermo TSQ Quantiva**



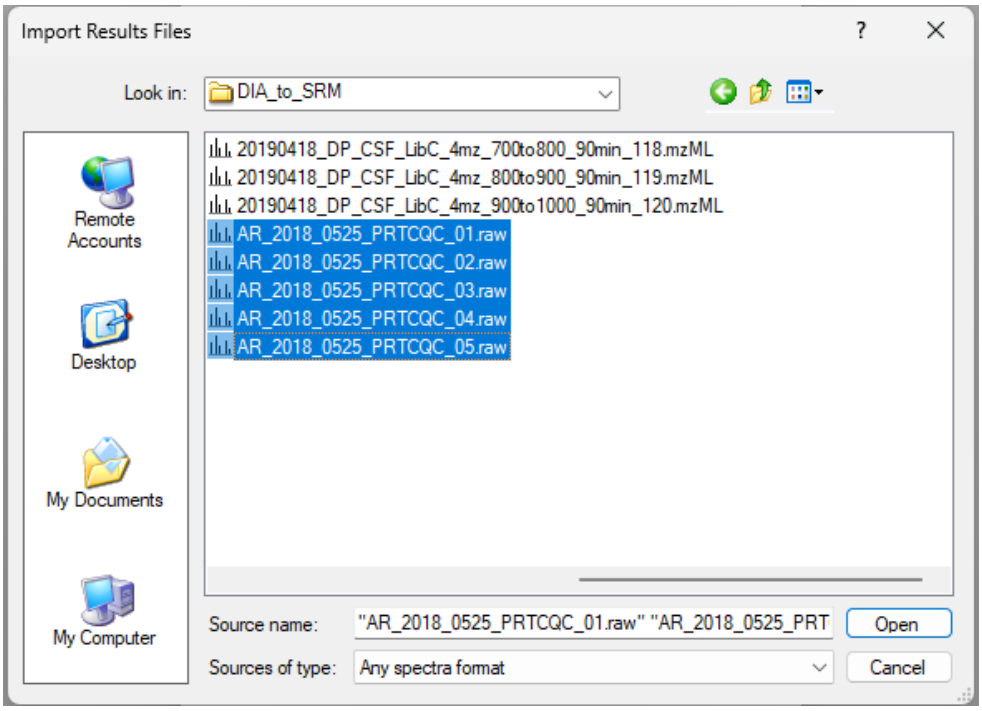


- In the **Full-Scan** tab use the following settings:
    - Acquisition method: **None**

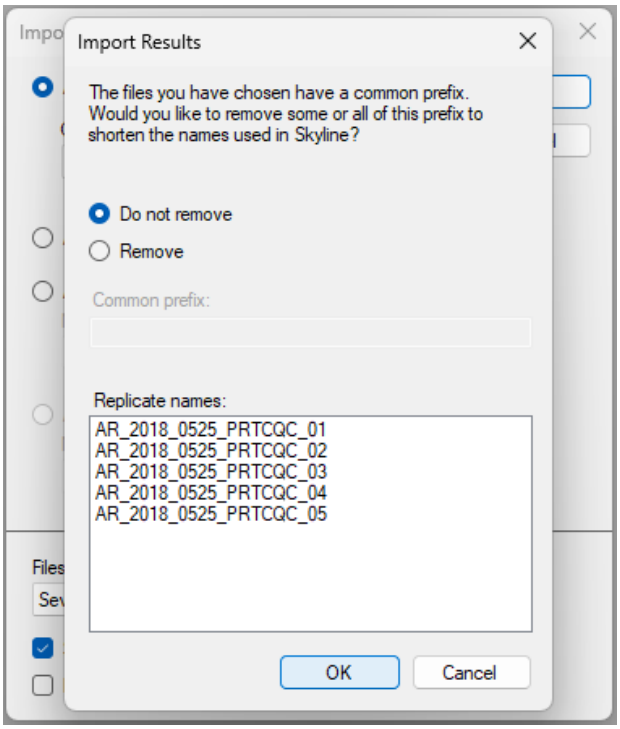
- Click "**OK**"
- Go to **File** > **Import** > **Results**, select "**Add single-injection replicates in files**", select "**OK**"
- If this dialogue box pops up select "**Yes**".

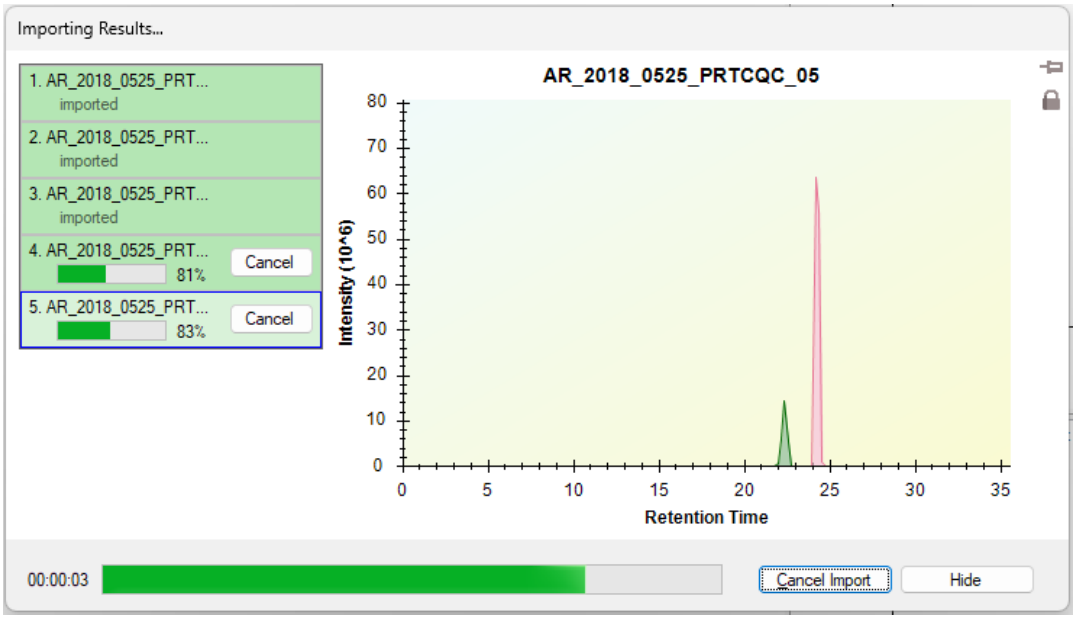


- In the pop-up asking about adding decoy peptides select "**No**"
- Navigate to the folder for this tutorial to import the 5 .raw files like "AR_2018_0525_PRTCQC_01.raw".
- Select the 5 PRTCQC files and click "**Open**"

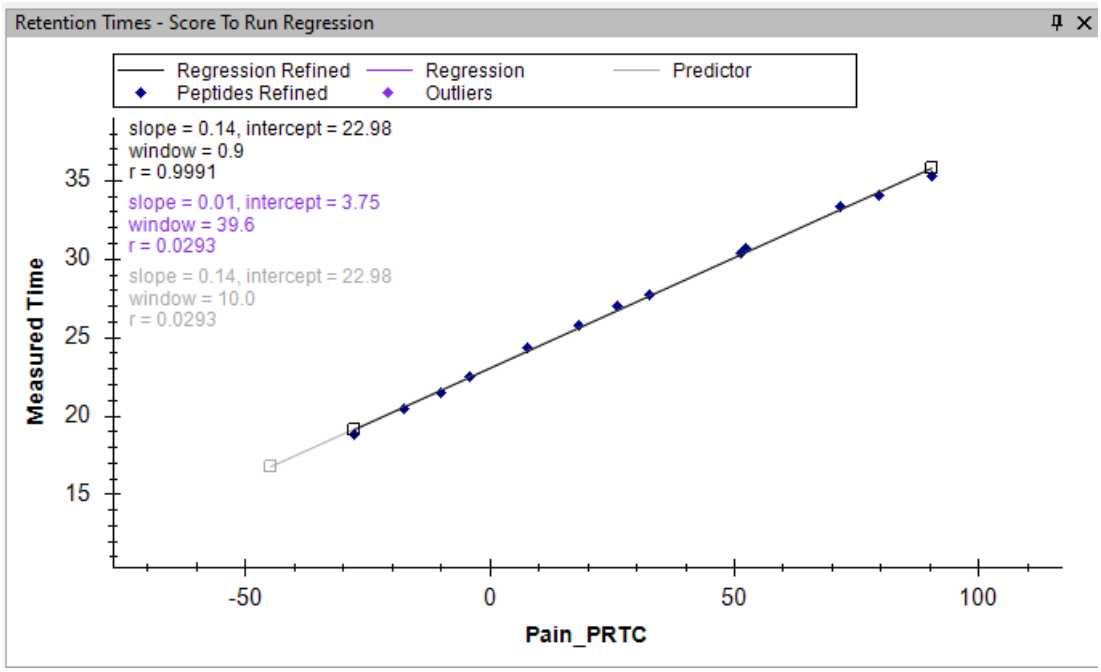


- In the pop-up asking about a common prefix select "**Do not remove**"
- Click "**OK**"



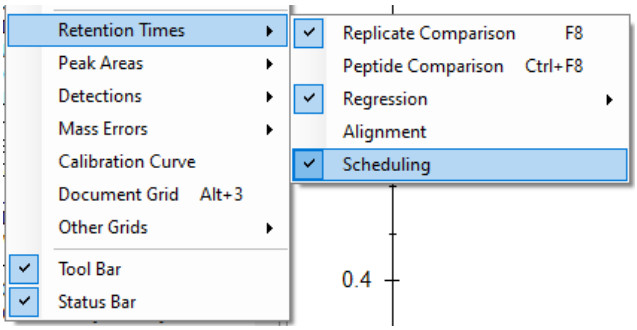


- Once the files are imported the regression graph should adjust to the new data.



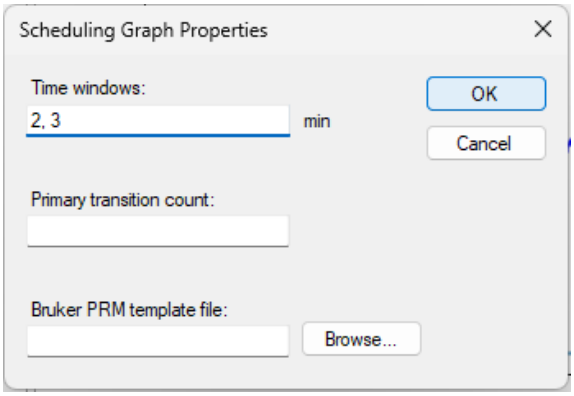
4.4. Exporting scheduled transition lists for scheduled SRM method.

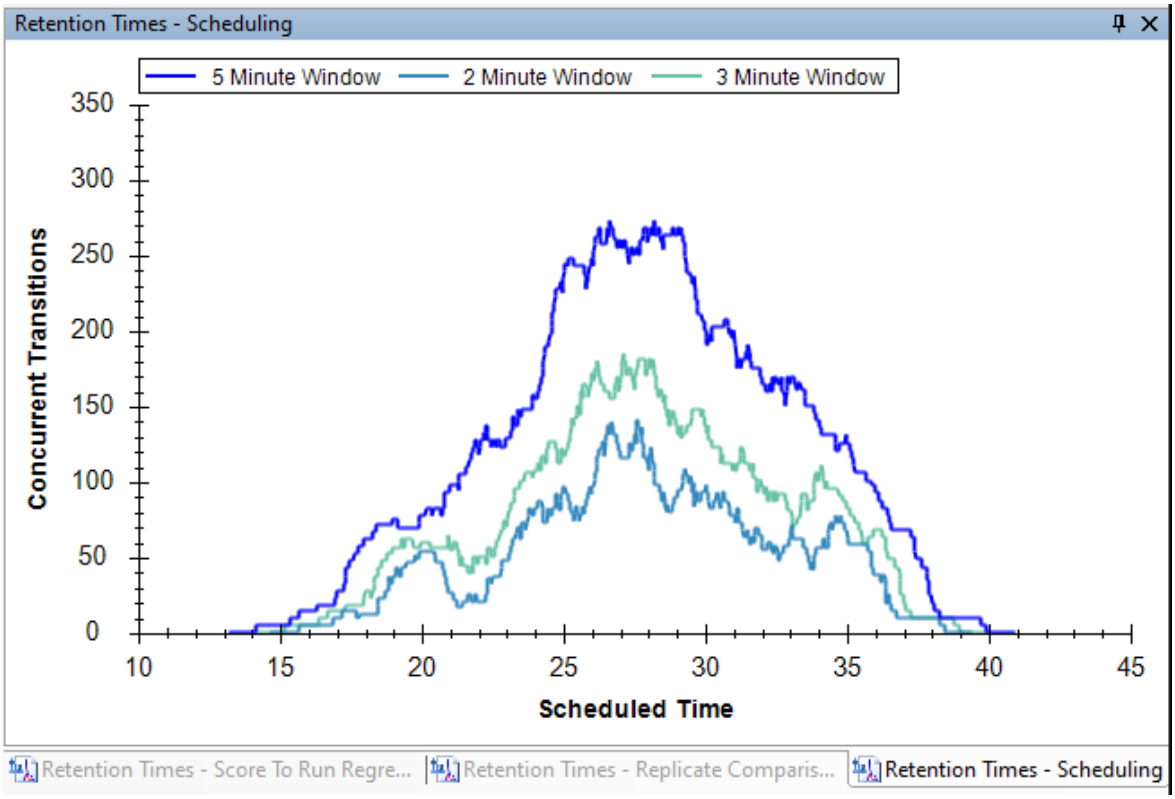
- In the **View** > **Retention times** > **Scheduling.**



We always want to make sure to choose an appropriate number of concurrent transitions for your instrument and your chromatography. Remember, the goal is to balance the points-across-the-peak considering the peak width and your instrument's dwell time. We can examine the effect of window width on the number of concurrent transitions in Skyline.

- Adjust the scheduling window widths by right clicking in the **Retention Times - Scheduling** window and
    - Select **Properties**
    - entering the desired retention time window widths (min) and clicking **OK**.

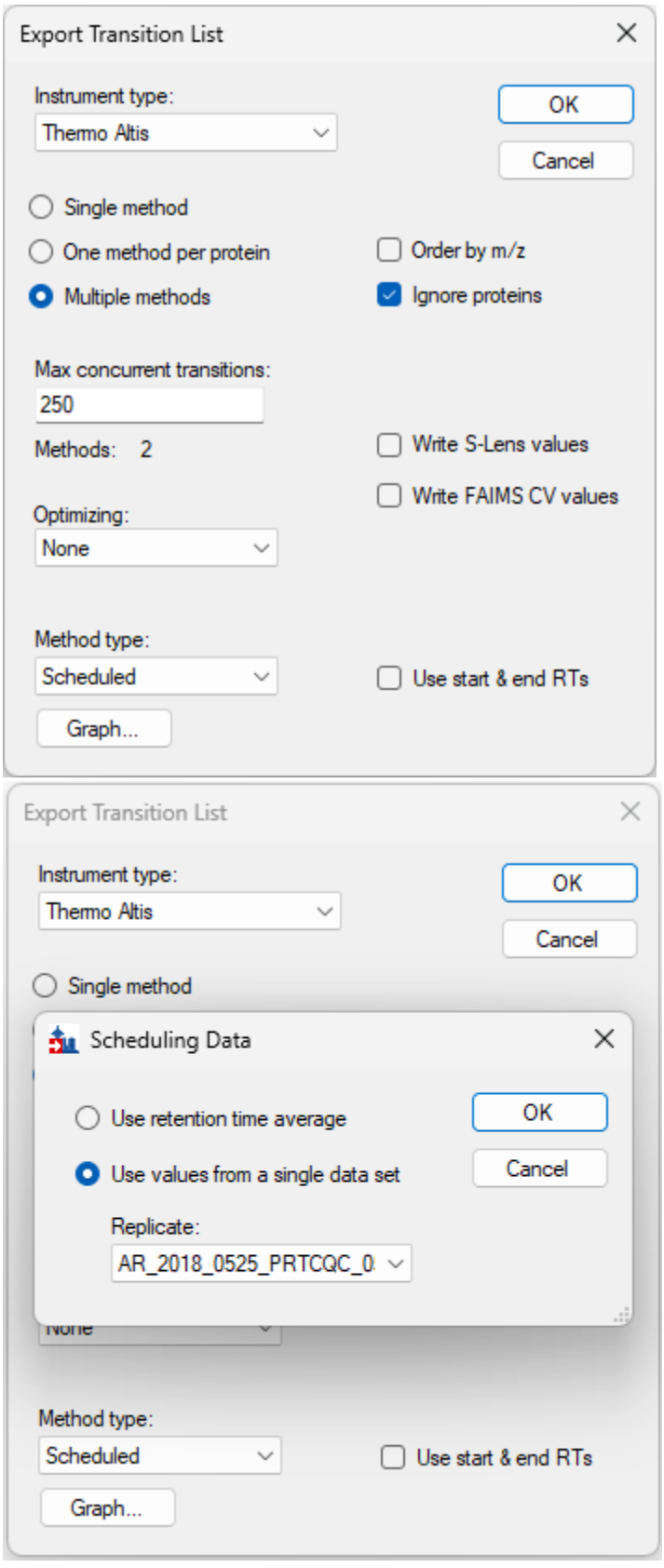




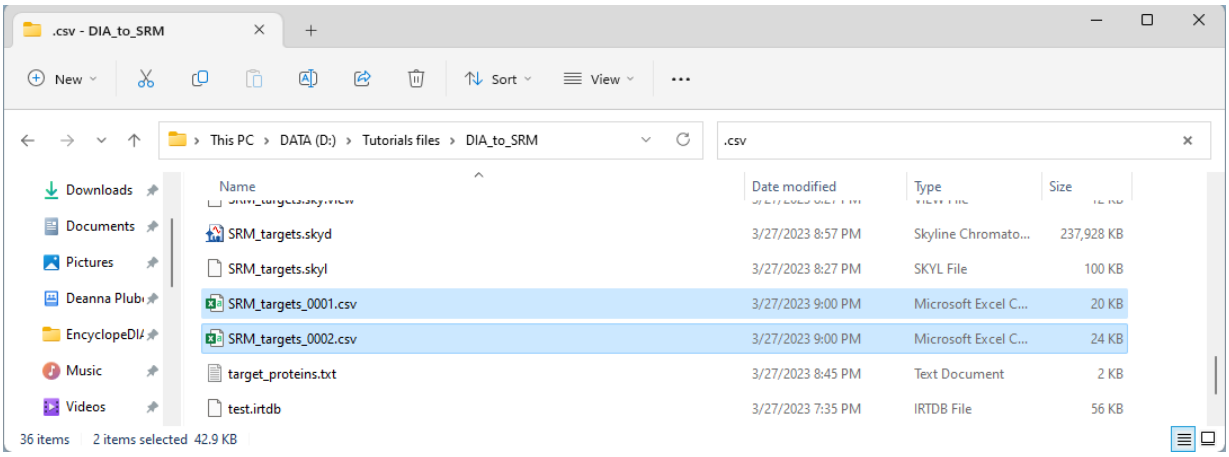
- Go to **File** > **Export** > **Transition List**
    - Instrument type: **Thermo Altis**
    - Select **"Multiple methods**", then play around with the "**Max concurrent transitions**". The number of methods generated will change depending on how many transitions you allow.
    - Once you have selected your number of concurrent transitions click "**OK**".

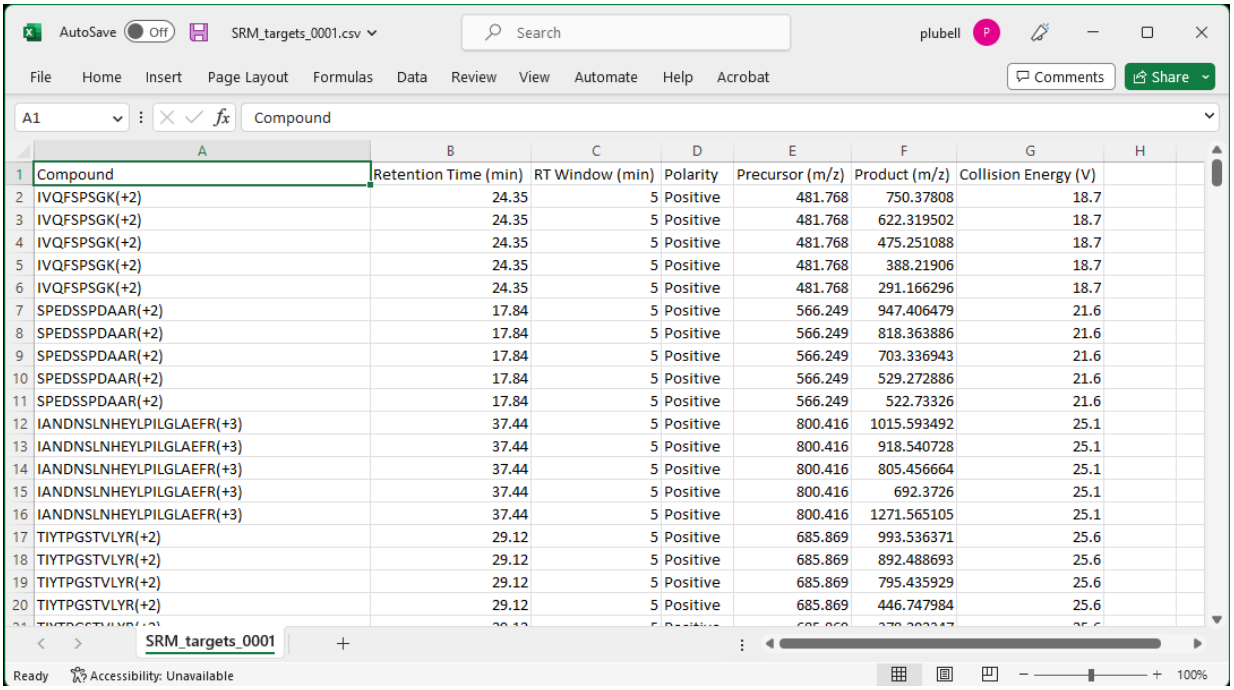
- In the **Scheduling Data** dialogue box that pops-up select **Use retention time average** and click "**OK**".

- Name your isolation list and click save *Note: if there are multiple methods being written (see above) then Skyline will automatically append numbers to the end of the file name, starting with 0001.*


Export Transition List
Instrument type:
Thermo Altis
OK
Cancel
Single method
One method per protein
Multiple methods
Order by m/z
Ignore proteins
Max concurrent transitions:
250
Methods: 2
Write S-Lens values
Write FAIMS CV values
Optimizing:
None
Method type:
Scheduled
Use start & end RTs
Graph...
Export Transition List
Instrument type:
Thermo Altis
OK
Cancel
Single method
Scheduling Data
Use retention time average
Use values from a single data set
OK
Cancel
Replicate:
AR_2018_0525_PRTCQC_0
Method type:
Scheduled
Use start & end RTs
Graph...

- In your computer's file explorer, navigate to where you just saved the transition lists. There are multiple transition lists -- one for each sample injection. Notice that even though there's multiple methods, each method contains the iRT standard fragments.





Compound	Retention Time (min)	RT Window (min)	Polarity	Precursor (m/z)	Product (m/z)	Collision Energy (V)
IVQFSPSGK(+2)	24.35	5	Positive	481.768	750.37808	18.7
IVQFSPSGK(+2)	24.35	5	Positive	481.768	622.319502	18.7
IVQFSPSGK(+2)	24.35	5	Positive	481.768	475.251088	18.7
IVQFSPSGK(+2)	24.35	5	Positive	481.768	388.21906	18.7
IVQFSPSGK(+2)	24.35	5	Positive	481.768	291.166296	18.7
SPEDSSPDAAR(+2)	17.84	5	Positive	566.249	947.406479	21.6
SPEDSSPDAAR(+2)	17.84	5	Positive	566.249	818.363886	21.6
SPEDSSPDAAR(+2)	17.84	5	Positive	566.249	703.336943	21.6
SPEDSSPDAAR(+2)	17.84	5	Positive	566.249	529.272886	21.6
SPEDSSPDAAR(+2)	17.84	5	Positive	566.249	522.73326	21.6
IANDNSLNHEYLPILGLAEFR(+3)	37.44	5	Positive	800.416	1015.593492	25.1
IANDNSLNHEYLPILGLAEFR(+3)	37.44	5	Positive	800.416	918.540728	25.1
IANDNSLNHEYLPILGLAEFR(+3)	37.44	5	Positive	800.416	805.456664	25.1
IANDNSLNHEYLPILGLAEFR(+3)	37.44	5	Positive	800.416	692.3726	25.1
IANDNSLNHEYLPILGLAEFR(+3)	37.44	5	Positive	800.416	1271.565105	25.1
TIYTPGSTVLYR(+2)	29.12	5	Positive	685.869	993.536371	25.6
TIYTPGSTVLYR(+2)	29.12	5	Positive	685.869	892.488693	25.6
TIYTPGSTVLYR(+2)	29.12	5	Positive	685.869	795.435929	25.6
TIYTPGSTVLYR(+2)	29.12	5	Positive	685.869	446.747984	25.6

- Your new transition lists are now ready for use in creating a new instrument method in Xcalibur

Congrats! You have a scheduled inclusion list for a new assay!