Development of amyloidosis typing method and data acquisition strategies using tandem mass spectrometry

Han-Yin Yang

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Michael J. MacCoss, Chair

Andrew N. Hoofnagle

Judit Villén

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Development of Amyloidosis Typing Method and Data Acquisition Strategies using Tandem Mass Spectrometry

Han-Yin Yang

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

A variety of mass spectrometry acquisition methods are designed to maximize multiplexing capacity, quantification accuracy, or detection sensitivity of peptide ions in complex biological matrices. Data-independent acquisition (DIA) is an emerging method that aims to quantify proteins with high accuracy embracing the proteome complexity. With DIA, a novel strategy was proposed to predict amyloidosis types with a majority vote rule. Informative amyloid peptides were selected to cast the vote for specific amyloidosis types based on intensity distributions in disease populations.

In contrast to DIA, the goal of data dependent acquisition (DDA) is to maximize the multiplexing capacity of peptide identification. Although DDA is a powerful data collection strategy, the majority of peptide ions in a complex mixture remain unanalyzed by tandem

mass spectrometry (MS/MS). In this thesis, a novel acquisition approach, data dependent acquisition plus (DDA+), is proposed to improve sampling coverage by combing the characteristics of existing DDA and DIA methods.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would not be here and writing this thesis without the patience, encouragement, and generous help of many people. I would like to thank my parents, sister and Yi-Fen for their unconditional support, love, and understanding.

Graduate school has not always been fun and projects do not always work, I thank my advisor Michael MacCoss for his patience, inspiration, and guidance throughout my graduate training, especially thank him for understanding that pipetting is not my favorite. I would also like to thank my committee members: Andrew Hoofnagle, Judit Villen, William Noble, and Kelly Smith. I appreciate every comment and suggestion they gave to me. They are invaluable sauce of advice that helped me improve and brought different insights to my thesis works. I would like to especially thank Kelly and Andy for collaborations on Amyloidosis project.

I am grateful for working with many of people in MacCoss lab. Without any of them would make the works presented in this dissertation and graduate school incredibly difficult. I would like to thank Lindsay Pino for being the best labmate. She has always been very supportive on everything. I have learned so much from her in both work and outside of work. I thank Brian Searle for the discussions and suggestions on the projects. I also thank him for all the conversations during late work nights. I am lucky to have both of them as my colleagues and friends. I thank Richard Johnson for showing me how to run and fix the nanoflow liquid chromatography mass spectrometry system. It has always been fun to talk with him about his crazy and funny ideas. I thank James Bollinger for his guidance on the beginning of many bench works. I thank Jarrett Egertson for his helps on early stage of my graduate training. I thank Genn Merrihew for keeping lab running smoothly. I thank Nick Shulman for all his helps on using Skyline and C sharp.

It's almost impossible for me to thank everyone. I would like to thank all my collaborators and all the friends in Seattle.

Chapter 1. INTRODUCTION

1.1 MASS SPECTROMETRY-BASED PROTEOMICS

Proteins are highly complex molecules that serve as fundamental functional units in all living organisms. Due to the critical role of proteins in biological functions, in protein interactions, and in regulation etc., one of the most important biological developments comes from studying how proteins contribute to mechanisms leading to disease and finding the potential drug targets for the treatments. Proteomics aims to comprehensively characterize protein structure, modification, localization, interaction, quantification, and activities from complex biological matrices at large scale. Although several techniques such as two-dimensional gel electrophoresis and Edman sequencing are extremely useful to identify and characterize proteins, they have limited throughput[1]. With developments of key techniques over the last several decades (described in detail in the following section), mass spectrometry (MS) has become the most effective and versatile technology for measuring thousands of proteins in complex mixtures.

1.1.1 *Shotgun proteomics*

The term of shotgun proteomics was proposed in 1998 to describe the process of inferring proteins from identified peptides in MS/MS experiments[2]. The experimental workflow was first developed by Hunt and colleagues in 1981 when they sequenced proteolytic apolipoprotein B with secondary ion mass spectra[3]. With the breakthrough of electrospray ionization (ESI)[4], they further developed an online liquid chromatography tandem mass spectrometry (LC-MS/MS) set-up to sequence femtomole amounts of peptides from complex mixture[5], which has now become the most common workflow for current shotgun proteomics.

In a typical shotgun proteomics workflow (Figure 1-1), proteins are digested with proteases into peptides that are separated by reverse-phase liquid chromatography (LC) to further reduce the biochemical complexity in a sample prior to MS analysis. The reverse-phase LC consists of a column filled with hydrophobic resin (also called the stationary matrix) and an elution buffer that generates dynamic hydrophobicity over time (mobile phase). Hydrophilic peptides have less affinity for the resin, therefore, elute from the LC system first. To boost the sensitivity in MS analysis, peptides are pushed through a capillary LC column (often using a column of 75 µm inner diameter) with elution buffer at flow rate of nL/min (nanoLC) [6–8].

Using ESI[4], eluted peptides are de-solvated into the gas phase with minimal or no fragmentation via an ionization process that involves a combination of heat, pressure, and voltage changes between the LC emitter and the inlet of MS. The ionized peptides are then accelerated into the mass analyzer by voltages. The *m/z* values of peptide ions are measured first by the mass analyzer and recorded in an MS "survey scan" or MS1 spectrum. Due to the high complexity of many biological samples, the *m/z* value of an intact peptide ion is not sufficient for peptide identification. To obtain sequence information for peptide identification, the peptide ions are further isolated and fragmented with collision gas. The resulting peptide fragment ions are recorded as MS/MS spectra. Since peptides fragment in a predictable manner, the peptide precursor of a MS/MS spectrum can be identified based on the *m/z* values of its fragment ion pattern in the spectrum[3,9].

To efficiently identify peptide precursors of MS/MS spectra, Eng et al developed the first computational algorithm, SEQUEST, to automatically assign peptide sequences to MS/MS spectra by comparing observed fragment ion patterns with theoretical spectra generated by in silico digestion of protein sequences[10]. With the availability of the human genome sequence

database, this approach can compare empirically obtained spectra with theoretical spectral of all possible peptides derived from the human proteome. To assign the statistical significance of peptide-spectrum matches (PSM), a target-decoy database search strategy[11,12] is often used to generate the null PSM score population by matching spectrum with non-existing peptide sequences (called "decoys", and often generated by reversing the "target" peptide sequence). Comparing the score distribution of real PSMs (targets) with the estimated null population (decoys), the statistical significance (p-value and its associated q-value[13]) of each PSM is computed. Peptide level statistical significance is then further assigned based on PSM[14], and proteins in the sample are inferred from confidently identified peptides. Several database search engines such as X!Tandem[15], MASCOT[16], OMSSA[17], and Andromeda[18] have been developed using the same spectral matching idea but use different scoring functions and statistical methods to assess PSMs.

In addition to qualitative identification of peptides and proteins in complex mixtures, the shotgun proteomics workflow is also a powerful tool for the quantitative analysis of peptides and proteins at a large scale. Several approaches have been developed in quantitative proteomics to measure protein quantities with high sensitivity or high complexity by applying different data acquisition methods and post analysis strategies. In section 1.2, I review the principles of the data acquisition methods that were used or are relevant to this thesis. The quantification approaches associated with these acquisition methods are then discussed in section 1.3.

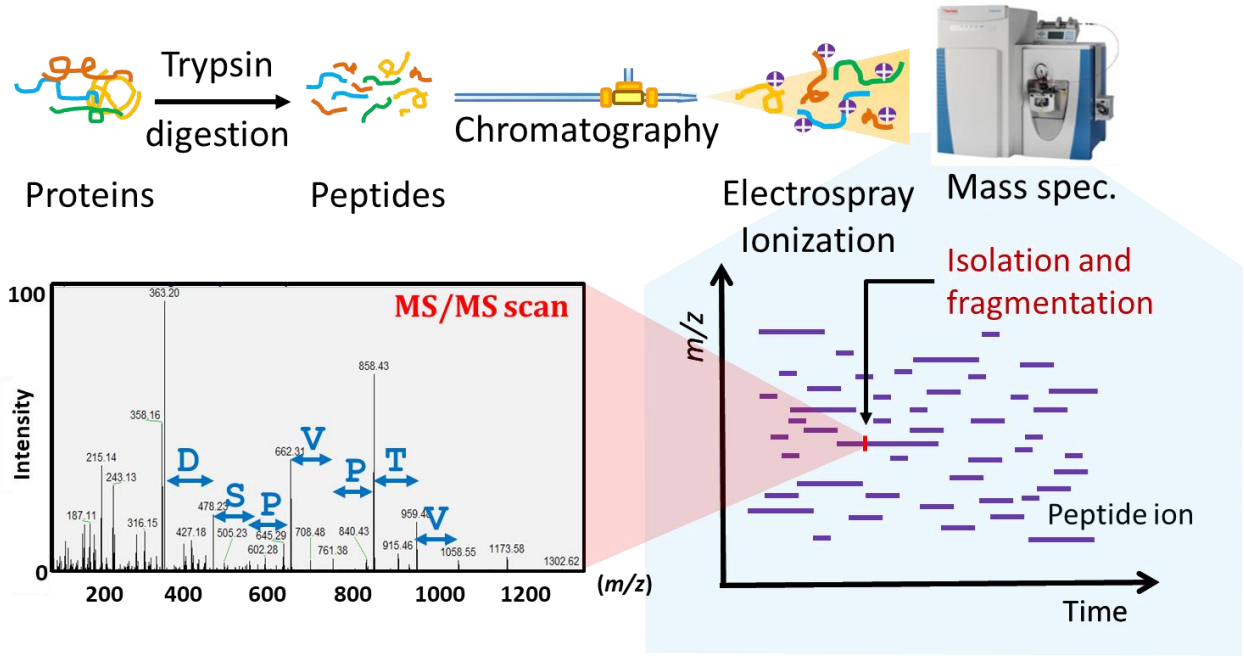

Figure 1-1Shotgun proteomics workflow.

Proteins are digested into peptides, separated by liquid chromatography, ionized, and injected into mass spectrometry. The *m/z* values of peptide ions are measured by mass analyzer over time. To obtain sequence information, peptide ions are selected, isolated, and further fragmented with collision gas. The resulting fragment ions are recorded as MS/MS spectra. Since peptides fragment in a predictable manner, the peptide precursor of a MS/MS spectrum can be identified based on the *m/z* values of its fragment ion pattern in the spectrum.

1.2 MASS SPECTROMETRY ACQUISITION METHODS

Acquisition methods describe the particular data collection mode that controls how mass spectrometers analyze peptide ions with finite chromatographic elution time in a LC-MS/MS set up. Several methods have been developed to maximize multiplexing capacity, quantification accuracy, or detection sensitivity of peptide ions, although each has trade-offs with each other. In this section, I will focus on the methods that were used in or are relevant to this thesis.

1.2.1 *Data-dependent acquisition*

Data-dependent acquisition (DDA), developed in 1996[19] is the most widely used method in MS-based proteomics research. Due to the complexity of biological samples and the limited scanning speed of a mass spectrometer, only a small subset of peptide ions can be selected to generate the MS/MS spectra that are required for peptide identification[20]. To increase the chances of identifications, DDA prioritizes peptide ions based on their intensity in the MS survey scan. The most intense N (usually 5 to 20) peptide ions are automatically selected for fragmentation. To avoid redundant sampling on the same peptide ion species, the *m/z* value of selected peptide ion is then excluded from MS/MS analysis for a defined period of time (with a function called "dynamic exclusion"). DDA is a very powerful method and has successfully improved the coverage of peptide species in the analysis. Using DDA with an MS/MS acquisition rate of 20 Hz, a recent study identified more than tens of thousands of yeast peptides in an hour-long MS experiment[21].

Although the peptide selection and exclusion strategy used in DDA boosts the multiplexing capacity of applied proteomics studies, the majority of peptides in complex matrices remains unsampled[20]. Due to under-sampling of peptides and the stochastic random selection method, DDA produces non-negligible variabilities between peptide identifications in

replicated analyses of the exact same sample,[22] where about 30% identification varies between replicates [23]. Moreover, DDA has limited in the dynamic range of the analyzed peptide ions, because the selection scheme is biased towards abundant ions, leaving peptides of lower intensity uncharacterized.

1.2.2 *Data-independent acquisition*

In contrast to DDA, data-independent acquisition[24,25] (DIA) triggers MS/MS analysis on predefined *m/z* isolation windows regardless of peptide ion intensity from survey scans. Several DIA variations that have been developed since 2003, the common goal of DIA methods is to improve the quantitative accuracy, specificity, and dynamic range of peptide ion analysis by maximizing the number of MS/MS events across the chromatographic elution profile of a peptide[26] (maximizing acquisition duty cycle). To analyze the majority of peptides in a sample, DIA targets a wide precursor *m/z* range that covers predicted *m/z* values of most of peptides in protein databases. The desired precursor *m/z* range is divided into several consecutive predefined *m/z* isolation windows. DIA sequentially analyzes isolation windows with MS/MS; all the peptide ions in a given isolation window are fragmented together, and thus the resulting MS/MS spectra contain fragmentation information from multiple precursor ions, which increases the difficulty of interpretation. The complexity of DIA MS/MS spectra can be reduced by using smaller *m/z* isolation windows, however, the time required to analyze all isolation windows across an entire precursor range will increase, and consequently the acquisition duty cycle is decreased which is not preferred. Therefore, it is important to find the balance between isolation window widths and acquisition duty cycle in order to achieve optimization of the mass analyzer's scanning speed.

Finally, the chromatographic elution profile of a particular peptide ion can be extracted from a series of isolation windows at same m/z region over time. The extracted chromatographic peak is then used for peptide quantification (see section 1.3.2).

1.2.3 *Targeted acquisitions*

Similar to DIA, targeted acquisition methods aim to provide better quantitative measurements of peptide ions. Selected reaction monitoring (SRM; also referred to as MRM) and parallel reaction monitoring (PRM) are two major targeted acquisition methods that have emerged as powerful alternatives to traditional immunochemical assays for protein quantification[27–31]. As with all bottom-up proteomics, the peptides are the molecules analyzed by MS. Protein quantities are not directly measured by these experiments. Therefore, selected peptides serve as quantitative surrogates for estimating the concentration of their associated proteins. Targeted acquisition methods are used when proteins (and their peptides) of interest are known before the experiment. With the predetermined list, the mass spectrometer selectively targets *m/z* values of peptides with a narrow isolation window for fragmentation. The targeted peptides are fragmented in repeating cycles, and the resulting fragment ions are used to build chromatographic elution profiles from their MS/MS spectra over time. The chromatographic elution profiles are used for quantification (see section 1.3.2).

The SRM and PRM modes are designed for different instrumentations[32]. In the SRM mode, data is acquired on a triple-quadrupole (QqQ) mass spectrometer. After collision in the second quadrupole, a handful of fragment ions are selected by the third quadrupole one by one, and transferred to the ion detector. The pair of selected *m/z* values of precursor (selected in the first quadrupole) and associated fragment ion (produced in the second quadrupole, and selected in the third quadrupole) is referred as a “transition”. Since each transition is measured by the ion

detector individually, increasing the number targeted transitions of a given peptide compromises analysis time. Alternatively, PRM is performed on a quadrupole-Orbitrap[33] or quadrupole-TOF[34] hybrid mass spectrometer, where the mass analyzer (Orbitrap or TOF) can analyze all fragment ions simultaneously, therefore, once a selected peptide is isolated and fragmented, all the fragment ions are transferred to the mass analyzer and measured all together. Using PRM the scanning time required for a given peptide is fixed regardless of the number of transitions.

1.3 QUANTITATIVE PROTEOMICS USING MASS SPECTROMETRY

One of the major goals in proteomics studies is to compare the relative quantities of proteins in samples representing different biological states. With significant advances in analytical power and instrumentation, mass spectrometry has emerged as a powerful peptide/protein quantification tool in biological research and clinical applications[35]. However, the mass spectrometric signals measured between different peptides are not comparable with each other mainly due to the different ionization efficiencies intrinsic to peptides with diverged physicochemical properties. For comparable quantification, the comparison of peptides between different MS experiments[36] must be based on same peptide species.

Mass spectrometry-based quantification methods can be grouped into two categories: those that use isotopic labeling and those that are label-free[36,37]. This section will focus on the label-free quantitative techniques used in this thesis. There are two major different label-free approaches. The first is spectral counting, which uses the number of identified MS/MS spectra (spectral counts) of a given peptide to represent peptide quantity. The other approach is to extract peptide chromatographic peaks from adjacent/continuous mass spectrometric measurements over elution time, where peptide abundance is inferred from the area under the curve or from the

intensity at the apex. The principles, advantages and limitations of both label-free approaches are reviewed in following sections.

1.3.1 *Spectral counting*

Spectral counting[23] (SC) estimates protein quantification by counting the number of identified MS/MS spectra of associated peptides, which can be easily derived from the database search results of a DDA experiment. Therefore, SC requires very minimal data analysis effort and resources compared with all other MS-based quantification methods, which makes it an attractive option for quantification. However, SC has suffered from limitations inherent from DDA. For example, peptides/proteins with lower intensity are invisible to quantification due to DDA biases toward abundant peptide ions. Furthermore, SC saturates at highly abundant peptides[38,39] because of the dynamic exclusion used in DDA (see section 1.2.1). As a result, the dynamic range of quantification is 2 to 3 orders of magnitude[23], and limited at both ends of abundance. In addition, the randomness of DDA makes spectral counting a less accurate quantification method compared with others[36].

Several techniques have been proposed to address the limitations of SC, such as applying normalization methods on the raw spectral counts[40,41], converting absolute counts to an index[41,42], and/or only counting a subset of peptides for each protein[43]. Although the proposed corrections have shown significant improvements regarding dynamic range and accuracy, SC is still a semi-quantitative method. In conclusion, SC is more suitable for proteins that produce multiple proteolytic peptides and have significant changes (2.5 to 5 fold) between conditions[38].

1.3.2 *Chromatographic peak intensity*

By chromatographic peak we refer to the signal intensity of at a specific *m/z* over time, in which the signal derives from a peptide ion intensity in MS survey scans, or the fragment ion intensity in MS/MS scans, or ion detector. The integrated area under the chromatographic peak curve represents the peptide quantity. Chromatographic peaks can be extracted from data acquired with DDA, DIA, or targeted acquisitions.

With DDA, chromatographic peaks at the MS level are extracted from adjacent MS survey scans, and peptide sequence information from identified MS/MS spectra that are mapped to this MS-level peak. The peptide identity mapping relies on matching *m/z* values and elution times between chromatographic peaks and precursor ions of MS/MS spectra. When comparing peptides between different LC runs, peptide chromatographic peaks are paired with their counterparts in other runs using retention time alignment algorithms and identification results. This approach estimates peptide quantity based on direct measurements (intensity; ions per second) in the MS, and reports better quantification accuracy than the spectral counting method. Considering the complexity of biological samples, interference from co-eluting peptides and misalignment between MS runs are major potential issues; therefore, high mass accuracy, resolving power, and robust LC systems are essential for quantification by MS level chromatographic peaks.

In targeted acquisitions and DIA, the chromatographic peaks are built from fragment ion signals, which are either selected by the third quadrupole in SRM acquisition mode or extracted from MS/MS scans in PRM and DIA. The MS/MS chromatographic peak has better specificity compared to chromatographic peaks at the MS level, because the additional level of *m/z* selection (specific fragment ion *m/z*) is applied when building the chromatogram. Among the SRM, PRM

and DIA methods, SRM uses a relative narrow isolation window (0.2~1 *m/z* window) to isolate targeted precursor ions and fragment ion, and only one targeted fragment ion is analyzed at a time; therefore, SRM has higher selectivity and sensitivity for targeted peptides compared to PRM and DIA[32].

Similar to SRM, PRM isolates targeted precursor ions, but uses a relatively wide isolation window (typically 2 *m/z*) to generate MS/MS scans that contain signal from all the fragment ions of the targeted precursor. Since multiple fragment ions are measured in PRM, the chromatographic peak can be built from many transitions (whereas usually 2~4 transitions are used in SRM). The high correlation between multiple transition chromatograms increases the specificity of peptides selection. Finally, similar to PRM, DIA generates MS/MS scans and extracts chromatogram from MS/MS scans over time. However, DIA has much wider isolation window (typically between 10 and 25 *m/z*) compared to PRM, thus the resulting MS/MS scan are more likely contain fragment ions from multiple precursors. For this reason, DIA may require a more sophisticated transition selection method to find appropriate transitions for peptide quantity when extracting chromatographic peaks from DIA MS/MS spectra.

In summary, MS/MS chromatograms are more specific than MS-level chromatograms due to additional level of *m/z* selection on fragment ions. Furthermore PRM and SRM seem to be better acquisition approaches than DIA for quantitative proteomics due to the relatively wide isolation windows used in DIA. To summarize, SRM and PRM provide the best sensitivity among all the methods, DIA is next most sensitive, and then the MS chromatogram from DDA is the least sensitive for quantitative proteomics. It may be noted that in terms of multiplexing capacity of quantification methods, DIA has a much higher capacity than both SRM and PRM methods, and might be similar to DDA.

1.4 AMYLOIDOSIS

Amyloidosis is a group of diseases characterized by protein misfolding and aggregation in the extracellular region, which causes progressive tissue damage and organ dysfunction. Amyloidosis cases have been described in the literatures since 1639. The term of "amyloid" was first introduced by Rudolf Virchow in 1854 when he observed a blue color after staining corpora amylacea with iodine, a reaction that usually resembles starch[44]. So far, more than thirty biochemically distinct soluble proteins have been reported as amyloidogenic proteins, which under unknown conditions transform into insoluble amyloid fibrils with characteristic beta-sheet structures[45]. The mechanism causing this protein misfolding and disease is still unclear, but serum amyloid P has been recognized as one of factors that stabilizes amyloid fibrils, and thus is considered a universal marker for amyloidosis[46]. Amyloidosis is often difficult to recognize due to its broad range of clinical syndromes including heart failure, nephrotic-range proteinuria with or without renal dysfunction, postural hypotension, and hepatomegaly. Therefore, a histologic demonstration of amyloid fibrils is required to diagnose amyloidosis, which is usually accomplished by observing apple-green birefringence in Congo-red stained tissues under polarized light. In some medical centers, the preferred screening procedure is the abdominal fat aspiration which is a relatively noninvasive approach and the test result can be obtained within in a day. However, if the fat aspiration screening is negative, a targeted biopsy of the affected organ is required for further confirmation[45].

Amyloidosis is classified into different types based on the causative amyloidogenic protein in the affected organs. Since treatment and prognosis are different between types, accurate typing is essential for designing an effective therapeutic strategy[47]. Historically, immunohistochemistry (IHC) has been the most frequently used method to identify causative

proteins. However, sensitivity and specificity of IHC methods are problematic due to the complexity of human tissue and the conformation changes of amyloidogenic proteins during tissue fixation or disease. In addition, antibodies are not available for less common types[45]. To address these issues, Vrana et al [48] developed a mass spectrometry-based typing assay that has shown superior sensitivity and specificity compared to IHC. In the following sections, I will review the mass spectrometry-based typing assay that is described in previous studies[48,49]. In addition, I will also review some important amyloidosis types, especially immunoglobulin light chain (AL) and serum amyloid A (AA) amyloidosis, which I used in the study described in Chapter 3.

1.4.1 *Shogun proteomics based amyloidosis typing assay*

Laser microdissection (LMD) and mass spectrometry (MS)-based amyloidosis typing was developed in 2009 by Ahmet Dogan's group [48] at the Mayo Clinic. Once amyloid fibrils are confirmed with Congo-red staining, the Congo-red positive regions are isolated and collected with LMD. To identify the protein components in the affected tissue, resulting samples are subject to MS analysis using a shotgun proteomics workflow with data-dependent acquisition. Proteins in Congo-red positive regions are inferred from the subsequently identified peptide species. The protein abundance is then estimated by spectral counts from all associated peptides. Since amyloid fibrils are enriched during the LMD process, the amyloidosis typing is determined based on the amyloidogenic protein with the highest spectral counts within an individual sample. For example, if the immunoglobulin light chain kappa constant region protein (IGKC) has the most spectral counts compared to other detected amyloidogenic proteins in the sample, the patient will be typed as AL-kappa. Since 2008, there are now thousands of amyloidosis cases that have been successfully analyzed with LMD-MS workflow at the Mayo Clinic. This LMD-

MS approach, which combines the sampling specificity of LMD and the analytical power of MS, has shown promising results and enhanced the ability of typing amyloidosis accurately[45]. Although the LMD-MS based assay has become the new standard of care for amyloidosis typing in clinics, it may also suffer from several fundamental limitations discussed in Chapter 3.

1.4.2 *Amyloidosis types*

In contrast to localized amyloidosis, systemic amyloidosis deposits amyloid fibrils in multiple organ across entire body. Amyloidosis classification was historically based on the organ distribution of amyloid fibrils and clinical observations, in which amyloidosis was considered a localized subtype if only one specific organ is involved in the disease. Based on associated medical conditions, amyloidosis can also be grouped into primary, secondary, age-related (senile amyloidosis), hereditary (familial amyloidosis), and neurologic types. Primary amyloidosis is not associated with other medical conditions, whereas secondary amyloidosis occurs under other conditions such as chronic inflammation or long-term hemodialysis[45]. Modern classification systems are biochemically based, which depend on the identification of the fibril precursor proteins in the affected organs. The current amyloidosis naming system starts with a capital 'A', referring to amyloidosis, and followed by the abbreviation of the amyloidogenic protein[50]. For example, transthyretin (TTR) amyloidosis is typed as ATTR.

1.4.2.1 Immunoglobulin light chain amyloidosis (AL)

Immunoglobulin light chain amyloidosis (AL) is the most frequently observed type of primary systemic amyloidosis. AL can be further classified into AL-kappa and AL-lambda where the precursors are immunoglobulin light chain kappa and lambda, respectively. The aggregated kappa or lambda light chains are usually produced by clonal plasma cell in bone marrow (although rarely can be produced by B lymphoid cells). Typically, about 5 to 10 % of plasma

cells are affected in disease. The heart and the kidney are the two organs most frequently affected by light chain amyloid fibrils [45,51], and the aggressiveness of light chain amyloid causes rapid organ failure in just a few weeks. In a healthy individual, the ratio of expressed kappa to light chain is about 2:1, but lambda becomes the dominant one in AL patients (specifically, the kappa to lambda ratio is about 0.33 in AL patients), which could be a potential maker for diagnosis [51,52].

Since the majority of AL cases are caused by somatic mutation in the immunoglobulin light chain variable region but not the constant domain, it has been thought that the amyloid fibril consist of light chain variable sequences. However, recent mass spectrometry based studies show strong evidence for sequences beside variable regions (constant regions and joint regions) in amyloid plaques from many patients [45,48,53], which indicates that sequences in constant regions may also be involved in fibrillogenesis. In the Chapter 3, I also show that a few constant region sequences are enriched in amyloid fibrils from AL patients, which provides extra evidence for disease involvement of constant regions.

1.4.2.2 Serum amyloid A amyloidosis (AA)

Serum amyloid A amyloidosis (AA) is the major subtype of secondary amyloidosis, and is one of the most severe complications of chronic inflammatory diseases. AA fibrils mostly deposit in the kidney, liver and spleen[54]. There are three serum amyloid A genes and one pseudogene in humans, where serum amyloid A1 (SAA1) is the most common form associated with AA. For the SAA1 gene, there are three major alleles (1.1, 1.3, and 1.5) that differ by one single amino acid substitution at two different locations[55]. Previous studies show that the three alleles have different predispositions for AA in different populations, specifically that Caucasian rheumatoid arthritis patients show higher risk if they have SAA allele 1.1 compared to having other alleles[56];

whereas the highest risk for AA in Japanese is having SAA allele 1.3[55]. The N-terminal 76 amino acids of SAA has been thought to be the major component of amyloid fibrils[57], but a recent study using a bacterial expressed system to demonstrate multiple SAA fragments and full-length SAA can form amyloid-like fibrils, suggesting that the truncated fibril version may be produced after fibril formation[58].

Chapter 2. AN APPROACH FOR PEPTIDE IDENTIFICATION COMBING DATA-DEPENDENT AND DATA-INDEPENDENT ACQUISITION ON Q-EXACTIVE HF

2.1 INTRODUCTION –

Mass spectrometry-based shotgun proteomics workflows are predominant in discovery proteomic research. In this workflow proteins are digested first with proteolytic enzymes, and the resulting peptides are separated by liquid chromatography, ionized and subjected to MS analysis. To identify peptide species from complex mixtures, peptide ions in the MS survey scan are selected for fragmentation, producing MS/MS spectra that contain peptide sequence information. Due to the high complexity of biological samples, only a subset of peptide ions is selected for MS/MS analysis and most peptide ions remain unanalyzed.

Data-dependent acquisition (DDA) is a peptide selection acquisition mode where peptide ions are prioritized based on their measured intensity in the MS survey scan. The intense peptide ions are isolated with a narrow (typically 2 m/z units) isolation window, fragmented, and analyzed with MS/MS. To avoid repeatedly analyzing the same peptide species, the *m/z* values of previously analyzed peptide ions are excluded from consideration for a short period of time. This dynamic exclusion and intensity based selection scheme used by DDA has successfully identified thousands of peptides from complex mixtures in one MS run. Although DDA is a powerful approach for shotgun proteomics, previous studies have shown that the majority (~90%) of peptide ions in complex mixture remain unanalyzed with DDA, and as much as 30% of the analyzed peptides vary between replicated analyses of same sample due to the semi-random selection of DDA. This difference in peptides sampling is the major hurdle for

comparative proteomic studies, because it is unclear if unseen peptides are not sampled because they are missing in the biological sample or because they were not selected for MS/MS analysis during DDA.

As an alternative to DDA, data-independent acquisition (DIA) is designed to improve the quantitative capability of shotgun proteomic studies. Instead of optimizing peptide selection for MS/MS analysis, DIA systematically analyzes predefined *m/z* regions independent of any information obtained in an MS survey scan. In DIA, the desired precursor *m/z* range is divided into a list of relatively wide (10 *m/z* or lager) *m/z* regions which are sequentially analyzed one after one throughout the entire acquisition. All the peptide ions in a predefined *m/z* regions are co-isolated and fragmented for MS/MS analysis. Compared with DDA, DIA could potentially generate MS/MS spectra for all peptide ions within a desired precursor *m/z* range, but co-fragmentation of multiple peptide ions in such a wide isolation window generates complex MS/MS spectra, which is difficult to use for peptide sequencing by most conventional database search tools.

In this chapter, I explored a new acquisition approach, called data-dependent acquisition plus (DDA+), which aims to improve peptide sampling coverage and reproducibility between replicate samples by combining the systematic sampling of DIA with the data-dependent characteristics of DDA (Figure 2-1). This acquisition algorithm is implemented using the mass spectrometer's application programming interface (API), which allows us to communicate with the Q-Exactive HF mass spectrometer in real-time. Subsequent identification of acquired MS/MS spectra from DDA+ and DDA was conducted with COMET. I compared the performance of DDA+ to DDA regarding sampling coverage of peptide features, the reproducibility of peptide identification, and the identification rate of acquired MS/MS spectra. With DDA+, the number of

sampled peptide features in a Hela digest is comparable to DDA, but the identification rate of sampled peptide features were lower in DDA+ compared to DDA. I investigated the differences of identification rates between these two acquisition methods and suggested further directions for improvement in this area.

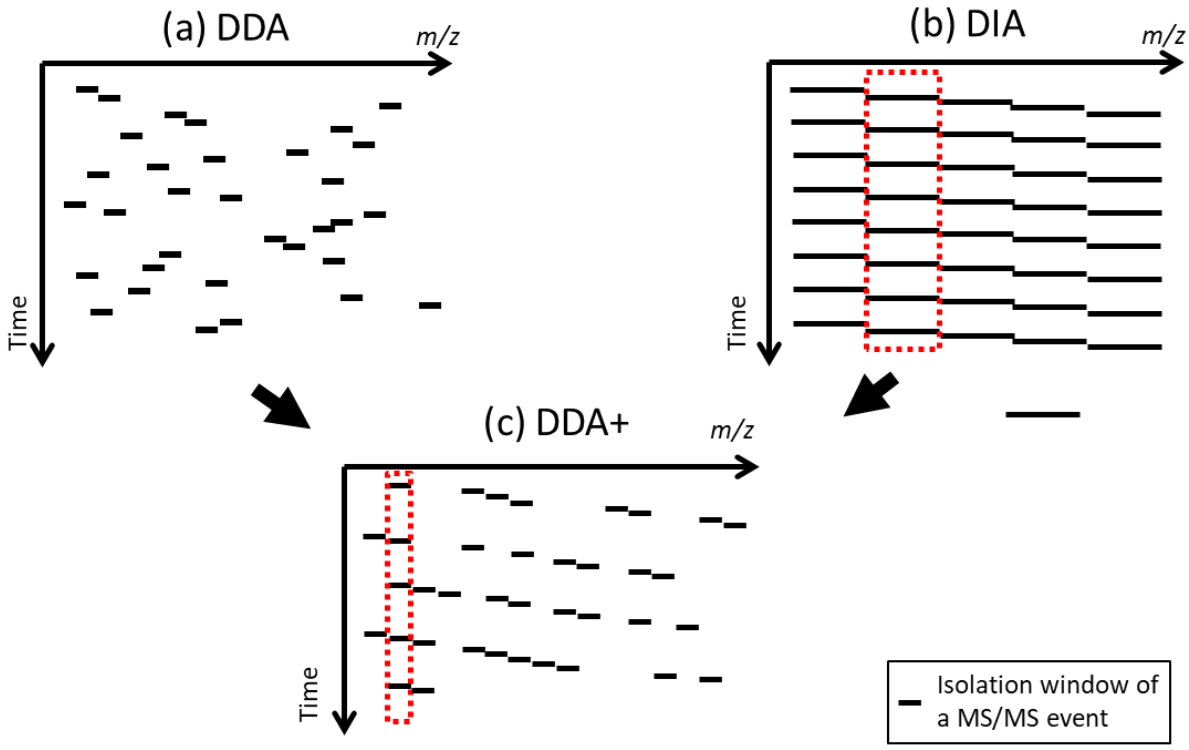

Figure 2-1 The precursor m/z space is analyzed by MS/MS events with three different acquisition methods

(a) DDA centers isolation windows on selected peptide ions, (b) DIA systematically collects MS/MS scans on consecutive isolation windows independent of peptide information from an MS1 scan, (c) DDA+ uses consecutive, non-overlapping, narrow windows to cover the precursor m/z range, but only triggers MS/MS events on windows having peptide ions detected in an MS1 scan. The red rectangle indicates that the isolation windows in DIA and DDA+ are aligned with each other

2.2 MATERIAL AND METHODS

2.2.1 *Implementation of DDA+ for Q-Exactive HF mass spectrometer*

The proposed data-dependent acquisition plus method (DDA+) triggers MS/MS events based on *m/z*, intensity and retention time information using a series of recent MS survey scans. In DDA+ mode, the precursor m/z range from 400 to 1,400 Th is covered by 500x2 m/z-wide successive windows. The isolation window boundary placement is optimized based on a previous study[59] to reduce the likelihood that peptide ions could fall on the window edges. DDA+ sequentially acquires MS/MS spectra on a subset of predefined 2 *m/z*-wide windows where one or many peptide features are detected within that defined 2-*m/z* region in recent MS survey scans (Figure 2-2). A peptide feature is determined when a peptide-like isotopic envelope is consistently observed in more than three consecutive MS survey scans. Therefore, to aid in peptide feature detection, MS survey scans are acquired every 20th scan, which translates to about a MS scan per second at a scanning speed of 20 Hz.

DDA+ was implemented with ESAPI (Exactive Series Application Programing Interface, Thermo) to analyze incoming MS survey scans from the Q-Exactive HF and then request MS/MS analysis on desired 2-m/z regions on the fly. Further details about the acquisition procedures are described in following sections.

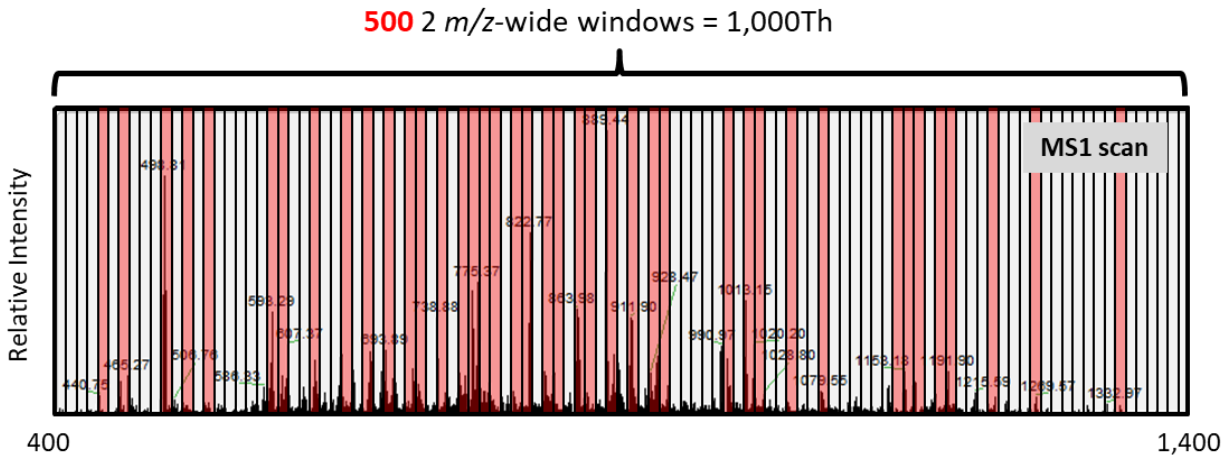

Figure 2-2 Subset of Isolation windows is selected for MS/MS analysis.
The mass range of a survey scan (400-1,400 Th) is divided into 500 2 m/z-wide windows. The isolation window colored with red color has qualified peptide feature and is selected for MS/MS analysis

2.2.1.1 Peptide feature detection

To distinguish peptide ions from other undesired ions such as ionized metabolites and chemicals in the experiment, the Hardklor[60] algorithm was adapted for real-time peptide detection in DDA+. In brief, the observed isotopic envelopes in the incoming MS survey scans are compared to predicted peptide isotope distribution, which is approximated using the averaged atomic composition of amino acids. The isotopic envelops with similarity scores equal to or higher than 0.9 are considered to be peptide ions. To reduce the likelihood of false positive peptide ion detections from this isotope distribution matching, DDA+ incorporates time domain information to filter out inconsistent peptide ions over time. The result is that only peptide ions detected in more than three consecutive MS survey scans are considered to be a peptide feature (Figure 2-3). Multiply charged peptide features (+2, +3, and +4) are considered for subsequent analysis in this study.

2.2.1.2 Acquiring and analyzing spectra in parallel

The Q-Exactive HF mass spectrometer is equipped with an Orbitrap mass analyzer for both MS and MS/MS analysis. Since the Orbitrap is the only mass analyzer, in theory, there will be a short period of idle time for the Orbitrap after an MS survey scan is acquired before the DDA+ algorithm must determine the isolation window for the dependent-MS/MS scan. This idle time for the Orbitrap might be negligible in the conventional DDA method, but it may not be trivial in DDA+ due to the complexity of computational procedures required to perform the MS/MS isolation window selection. To reduce the possible overhead on acquisition scan times, a multithreading system is implemented in DDA+, where one thread is responsible for peptide detection and isolation window selection using information in MS survey scans, which it then uses to add spectrum acquisition requests of all the selected 2-m/z windows that will be queued in a candidate list for MS/MS analysis.

Another thread will send spectrum acquisition requests to the instrument one by one based on queued order, therefore, the Orbitrap continuously acquires scans as long as the candidate list is not empty, and does not wait for analysis results from the most recent MS survey scans. In the case where no isolation windows are queued in the candidate queue, the Orbitrap continuously acquires MS survey scans until more spectra are requested in candidate list.

2.2.1.3 Updating Acquisition queue of isolation windows every second

In a complex biological mixture, more than a hundred peptide features can be detected at once, therefore, many spectrum requests could be queued to the candidate list. With a scanning rate of 20 Hz, a spectrum request could stay in the candidate list for many seconds after a peptide feature is detected. One concern is that, during the time between detecting the peptide in a survey

scan and that peptide's spectrum request reaching the top of the queue, the peptide may have eluted off of the column. To avoid the long waiting time between peptide detection and spectra collection, DDA+ acquires one MS scan per second, and updates the candidate list right after. The updated spectrum requests will start from the unanalyzed *m/z* regions in previous candidate list (Figure 2-3).

2.2.2 *Peptide standard and liquid chromatography*

A Hela digest standard (Pierce, #88328) was reconstituted with 40 µL 0.1 % formic acid (buffer A in liquid chromatography system) to make a final peptide concentration of 0.5 µg/µl. A 3µL Hela digest injection was made with an 8 µL trapping column wash and a low rate of 2 µL/min. The peptides were separated using a reversed-phased liquid chromatography system (Waters nanoACQUITY) running a 60 min linear gradient of 2-40% buffer B and a 5-min gradient of 40-60% buffer B. Mobile phase A consisted of water with 0.1% formic acid. Mobile phase B consisted of acetonitrile with 0.1% formic acid. The homemade trapping and analytical columns were made using 3cm (150um inner diameter) and 15 cm (75um inner diameter) fused silica capillary column (Polymicro Technologies, Phoenix, AZ), packed with 90Å 4µm C12 beads (Jupiter Proteo; Phenomenex, Ventura, CA) and 120Å 3µm C18-AQ resin (Dr. Maisch; GmbH, Germany), respectively. The eluted samples were analyzed with the Q-Exactive HF mass spectrometer (Thermo Scientific). Two replicate analyses of Hela digest were performed using both DDA and DDA+.

The m/z range of MS survey scan (MS1) is divided into 500 2-m/z windows. The colored isolation windows indicate one or more peptide ions are detected within that region. The two

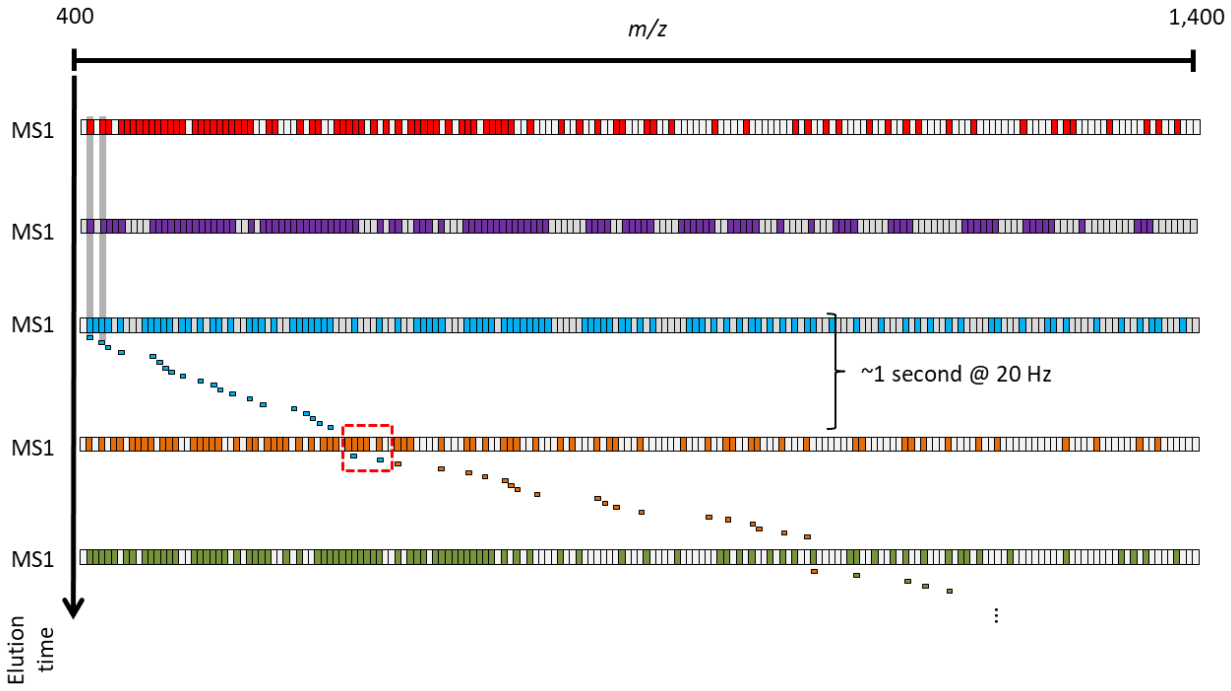

Figure 2-3 Acquisition scheme of DDA+.

grey strips are examples illustrate peptide ions are detected in three consecutive MS1 scans, therefore, a MS/MS scan is triggered to analyze that isolation window. Small colored blocks in between MS1 event are MS/MS events. The color on MS/MS events associate to the MS1 scan when MS/MS was queued into candidate list. The red rectangle box indicates an example that the triggered MS/MS events are not based on the most recent MS1 scan. The candidate list is updated once the analysis on MS1 scan is done, therefore, majority of MS/MS events afterward are still selected based on most recent MS1.

2.2.3 *Data acquisition parameters*

Mass spectral data was acquired with DDA and DDA+ methods on the same Q-Exactive HF. The DDA parameters were set up with the vender-provided method editor tool that comes with the instrument computer. The DDA+ parameters were set up through the self-implemented tool that sends spectra request, including all instrument parameters via Exactive API installed on the instrument computer. To compare the performance between these two methods, I strove to set all parameters comparably in the two methods; however, there are a few parameters were not

accessible or slightly different when using Exactive API. Therefore, I computed the equivalent parameters when using API. In both DDA and DDA+, spectral data was acquired with a cycle of one high resolution MS survey scan (400 – 1,400 *m/z*, 120,000 resolution at 200 *m/z*) followed by twenty MS/MS scans (15,000 resolution at 200 *m/z*), a method commonly referred to as "Top 20 DDA" The automatic gain control, and maximum injection time for MS and MS/MS scans were set to 3e6 and 1e5, and 50ms and 25ms, respectively. The precursor ion (in DDA mode) or peptide feature (in DDA+ mode) with charge states of 2, 3, and 4 were considered for MS/MS analysis. The normalized collision energy was set to 27.

The differences between the two methods are detailed as follows. The isolation window width was 2 *m/z* in both methods, but DDA center the isolation window to precursor ion *m/z* while DDA+ uses a predefined window. Targeted precursor *m/z* was put into a dynamic exclusion list for 15 seconds in DDA. The under fill rate of ions for a MS/MS scan was 10 % in the DDA method, and the equivalent value for DDA+ was 1,000 ions per second. The real-time charge state determination was set to "preferred" in DDA, but was required in DDA+, in which DDA used an algorithm built into the instrument for charge state detection while DDA+ used the Hardklor algorithm for charge state detection. The mass range of MS/MS scans was computed in real-time for both DDA and DDA+, but DDA uses an instrument built-in algorithm, and DDA+ used derived equations from previously collected DDA spectral datasets. The mass range equations for 2+, 3+ and 4+ precursor ions are shown in Figure 2.4. DDA+ used the predominant charge state of peptide features within a selected isolation window to determine the mass range for the MS/MS scan.

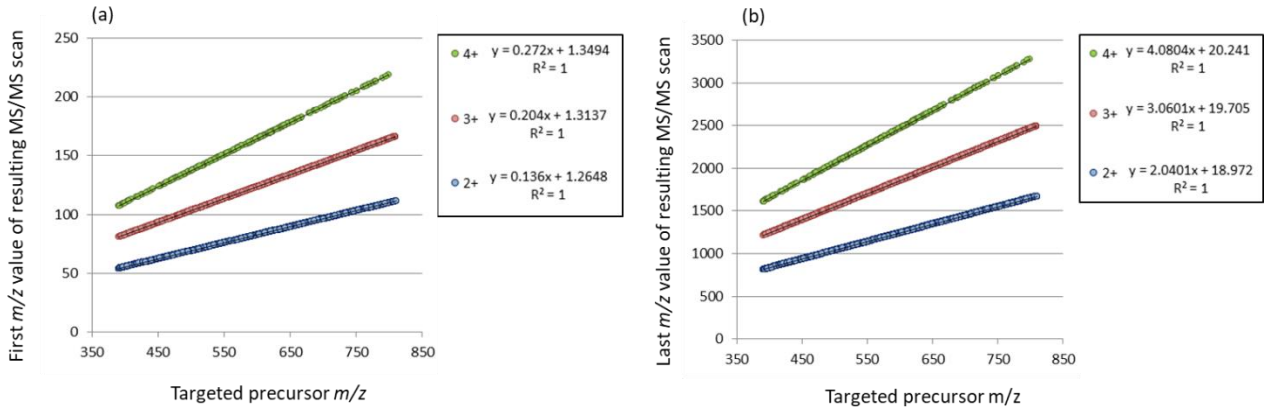

Figure 2-4 The relationship between precursor m/z value and mass range of the resulting MS/MS scan.

Each dot on the plot represents a MS/MS scan in acquired data. The charge state of MS/MS scan was the detected charged state by instrument built-in algorithm, which is recorded in the raw file MS/MS scan header. Linear relationship between precursor m/z and mass range, (a) first m/z and (b) last m/z, of resulting MS/MS were observed in different charge states. The derived linear equations were used in DDA+ to define the mass range of request MS/MS scan on the fly. In all the equations, y is predicted first or last m/z value, and x is the targeted m/z value which is the center m/z value of isolation window.

2.2.4 *Peptide identification on MS/MS spectra*

All acquired MS/MS spectra in both DDA and DDA+ experiments were subjected to the same analysis pipeline, in which raw spectral data is first converted to a compressed peak list format (cms1 and cms2) using MSConvert[61]. The monoisotopic mass of MS2 precursors are corrected by Bullseye[62] (v1.30) prior to database searching. MS2 spectra were searched against a human proteome sequence database downloaded from the UniPort database (http://www.uniprot.org/) using COMET[63] (2014.01 rev. 0). The search parameters are configured to allow fully and semi-tryptic peptides with a maximum of one missed cleavage and a static carbamidomethyl modification (+57.02 Da) of cysteine. The mass tolerance for precursor and MS/MS matching

are 10 ppm and 0.2 Da respectively. The search result from target and decoy (reversed sequence) databases were analyzed using Percolator[14] (v2.04) to assign a q-value to each identified spectrum. Confidently identified peptides (q-value equal or less than 0.01) were used in DDA and DDA+ performance evaluation.

2.2.5 *Post-acquisition peptide feature detection and association with MS/MS scans.*

To analyze how peptide sampling impacts MS/MS spectra identification, post-acquisition peptide feature detection was performed on the acquired DDA and DDA+ datasets using Bullesys[62]. Peptide features are defined by Bullseye using the consistently observed peptide ion envelopes (in MS survey scans) over elution time, which is the same concept I implemented in DDA+ for real-time detection of peptide features. The detected peptide features can be grouped into unanalyzed and analyzed peptide features depending on whether it is isolated for MS/MS scans (Table 2-2). The MS/MS scans can also be classed into different types based on the number of peptide features with in isolation regions (Figure 2-5). The precursor *m/z* of MS/MS scans are re-assigned based on detected peptide features. Database searching was performed on MS/MS scans with these re-assigned precursor *m/z* values. The precursor intensity is the intensity of the peptide ion envelope in the most recent MS survey scan. In this analysis, we only considered 2+, 3+, and 4+ peptide features. Although peptide detection in real-time and post-acquisition are based on the same idea, post-acquisition detection can utilize the complete spectral data. Therefore, more criteria for peptide feature matching were applied to remove potential false positive detections. As a result, the post- acquisition should detect less but more accurate peptide features than real-time detection.

2.3 RESULT

2.3.1 *Analysis on acquired MS/MS spectra*

Two replicates of Hela digest were acquired using DDA and DDA+, in which DDA+ acquired about 30% more MS/MS and 38% less MS scan than DDA (Table 2-1). To know whether isolation window placement affected the complexity of the resulting MS/MS scans, I counted the number of co-fragmented precursor ions for each MS/MS scan. Since the instrument vendor does not report the number of ion species isolated for MS/MS analysis, Bullseye[62] was used to map peptide features to MS/MS scans by matching *m/z* and elution time. The results (Figure 2-5) show that an average of 46% and 44% of MS/MS scans were from a single precursor, and 47% and 34% of MS/MS scans were from multiple precursor ions in DDA and DDA+, respectively. Although DDA+ generated a smaller proportion of chimeric spectra in the data, a large proportion (average of 22%) of MS/MS scans in DDA+ did not match to any peptide feature in the spectral data, therefore, no precursor *m/z* was assigned to those scans which were discarded before database searching. This results in about 42,000 and 46,000 MS/MS scans subjected to database search in DDA and DDA+, respectively (Table 2-2). The number of three MS/MS scan types (MS/MS scans with no detected precursor, MS/MS scans with a single precursor, and finally MS/MS scans with multiple precursors) were plotted against elution time in Figure 2-6 showing that a majority of no precursor-matched MS/MS scans were observed an early elution times in DDA+. In addition, Figure 2-6 also shows that DDA+ acquired significantly more MS/MS scans than DDA at early elution times, and slightly more MS/MS scans at the very end of elution.

Table 2-1 Number of acquired MS/MS scans by DDA and DDA+

	DDA replicate1	DDA replicate2	DDA+ replicate1	DDA+ replicate2
Number of MS scan	6,140	6,256	3,808	3,718
Number of MS/MS scan	45,390	44,839	59,407	59,569

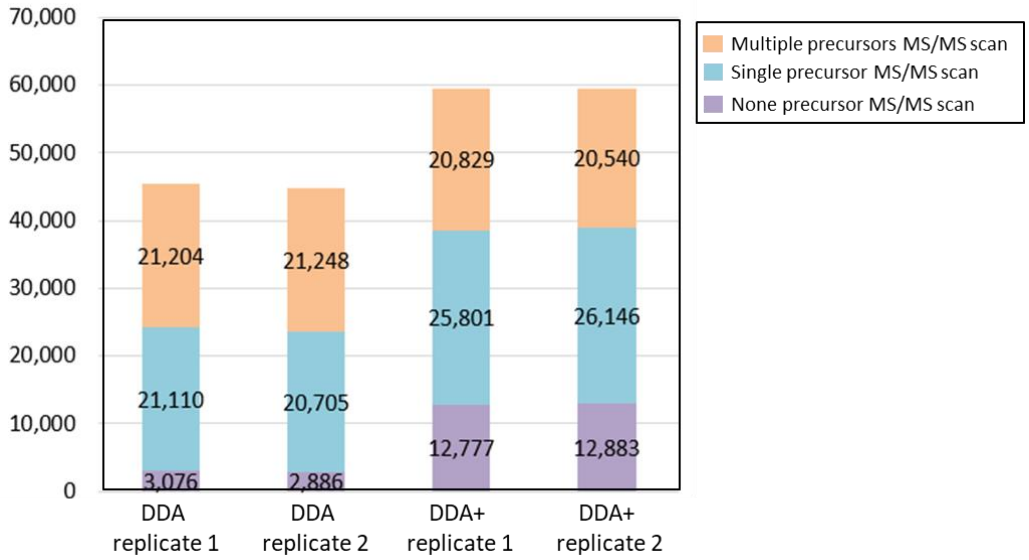

Figure 2-5 Number of acquired MS/MS scan in different types.

The MS/MS scans are classed into three types based on number of precursor ion reassigned to MS/MS scan.

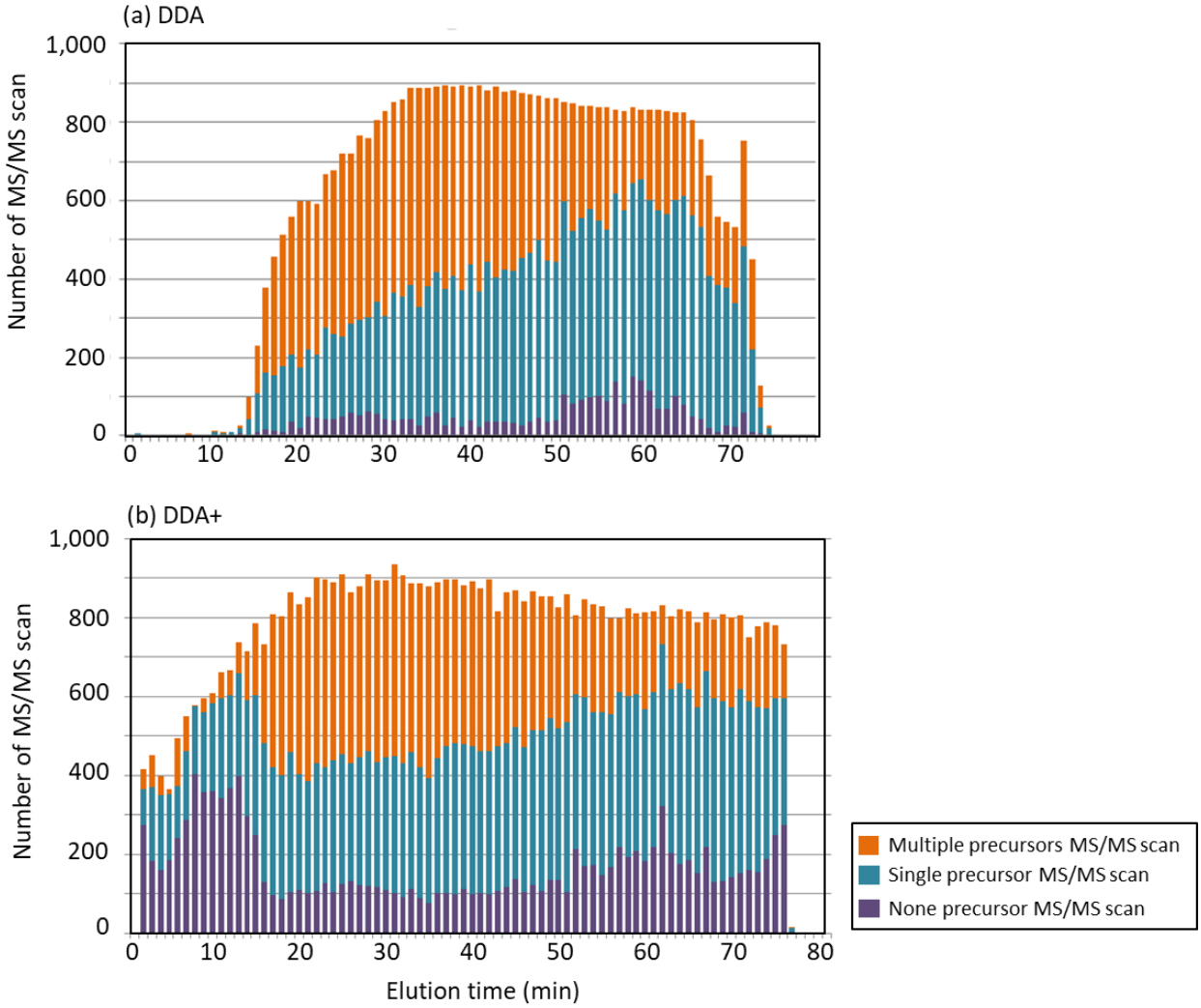

Figure 2-6 Number of MS/MS scans acquired over elution time.

The histogram plots shown here were from the first replicated Hela data acquired with (a) DDA and (b) DDA+. The number of MS/MS scans were plotted over elution time and colored based on the number of precursor ion reassigned by Bullseye.

2.3.2 *Identification of MS/MS spectra*

Database searches were performed on the MS/MS scans with one or more reassigned precursor ions. The number of identified scan, peptide-spectrum match (PSM), peptide sequences, and peptide identification reproducibility are listed in Table 2-2. The identification results are counted in three different levels that are scan, PMS, and peptide. One scan may be associated with multiple PSMs if more than a peptide was confidently matched to the same scan by COMET and Percolator. The reproducibility of identified peptide sequences between two replicates is the percentage of commonly identified sequence in union of peptides in both replicates

Although DDA+ generated more MS/MS scans, there are fewer of them identified by database search engine. To understand the underling factors, we investigated the relationship between identification rates with several parameters, the results are shown in section 2.3.4.

Table 2-2MS/MS scan identification results in DDA and DIA+ replicates.

	DDA replicate1	DDA replicate2	DDA+ replicate1	DDA+ replicate2
Number of queried MS/MS scan	42,314	41,953	46,630	46,686
Number of identified MS/MS scan	27,285 (64.48%)	27,397 (65.30%)	18,014 (38.63%)	18,278 (39.15%)
Number of identified PSM	32,248	32,287	19,942	20,249
Number of identified peptide	16,316	16,981	12,481	11,869
Number of peptide identified in both replicates	13,615 (65.4%)		8,711 (55.7%)	

2.3.3 *Analysis on detected peptide features*

Peptide features were classed into analyzed or non-analyzed peptide features depending on whether there were one or more MS/MS scans acquired on those peptide features. The number of analyzed and non-analyzed peptide features eluting over time are plotted in Figure 2-7. Only peptide features with 2+, 3+, or 4+ charges were considered in analysis. The results show that the number of analyzed features comparable in DDA and DDA+, but that DDA+ gives higher sampling coverage than DDA (Table 2-3). Peptides were considered identified if one or more associated MS/MS scans were confidently matched during database searching. The results show that more features are identified in DDA data. In addition, the sampling rate of peptide features across *m/z* dimension is plotted in Figure 2-8 that showing that the DDA+ sampling rate is biased toward peptide features with smaller *m/z* values compared to DDA.

Table 2-3 Peptide features in DDA and DDA+ replicates

	DDA replicate1	DDA replicate2	DDA+ replicate1	DDA+ replicate2
Number of detected peptide feature	49,657	50,369	44,898	42,911
Number of feature analyzed by MS/MS	37,895 (76.31%)	38,603 (76.64%)	38,974 (86.81%)	37,366 (87.08%)
Number of Identified peptide feature	20,337 (49.95%)	21,088 (41.87%)	14,939 (33.27%)	14,504 (33.8%)

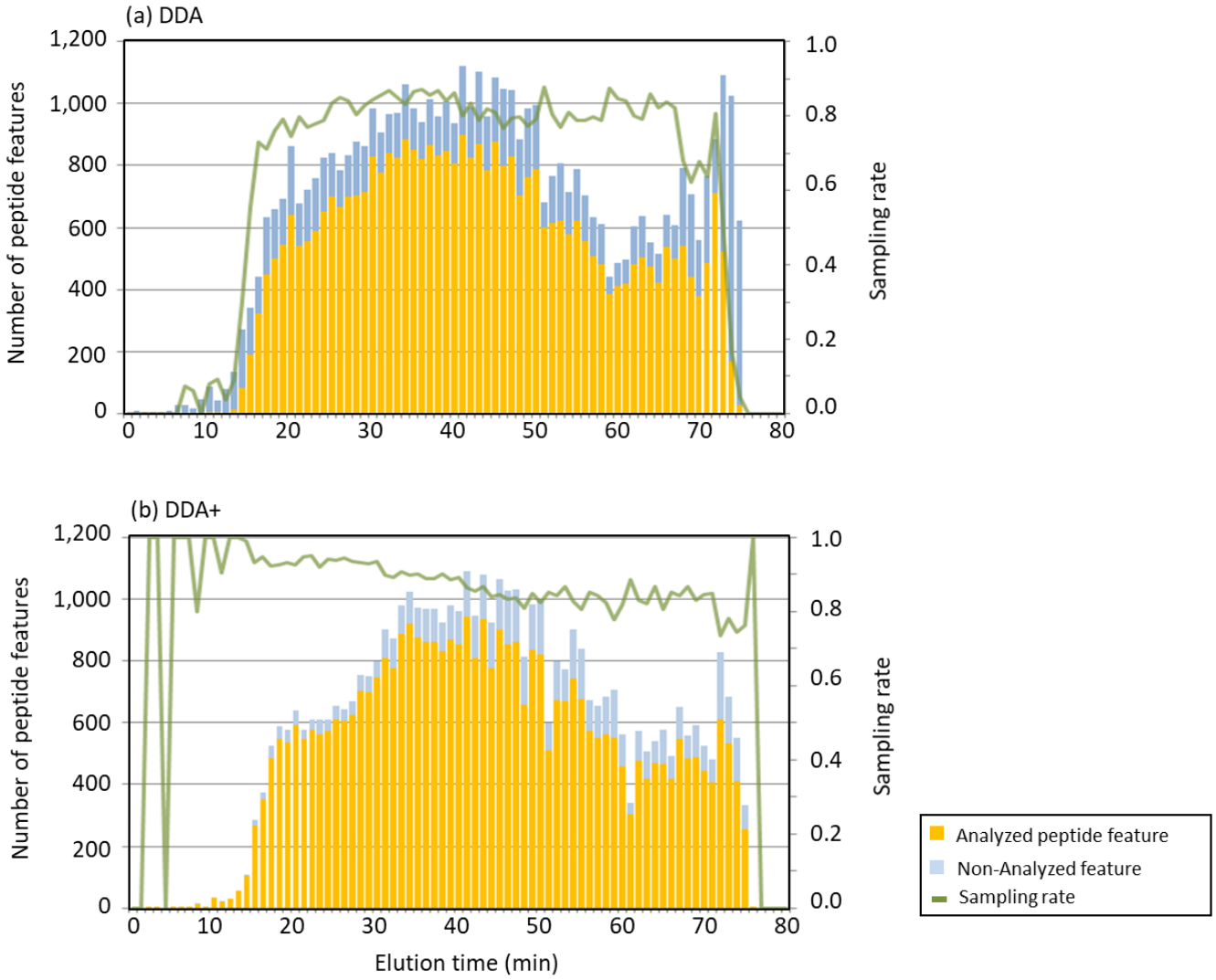

Figure 2-7 Number of analyzed and non-analyzed peptide feature over elution time.

The histograms shown here are the number of detected peptide features over elution time in first replicated HeLa data acquired with (a) DDA and (b) DDA+. The analyzed and non-analyzed peptide features are colored with yellow and light blue, respectively. The sampling rate is the green line plot along the secondary y-axis. Sampling rate is the number of analyzed peptide feature divided by the total number of peptide features.

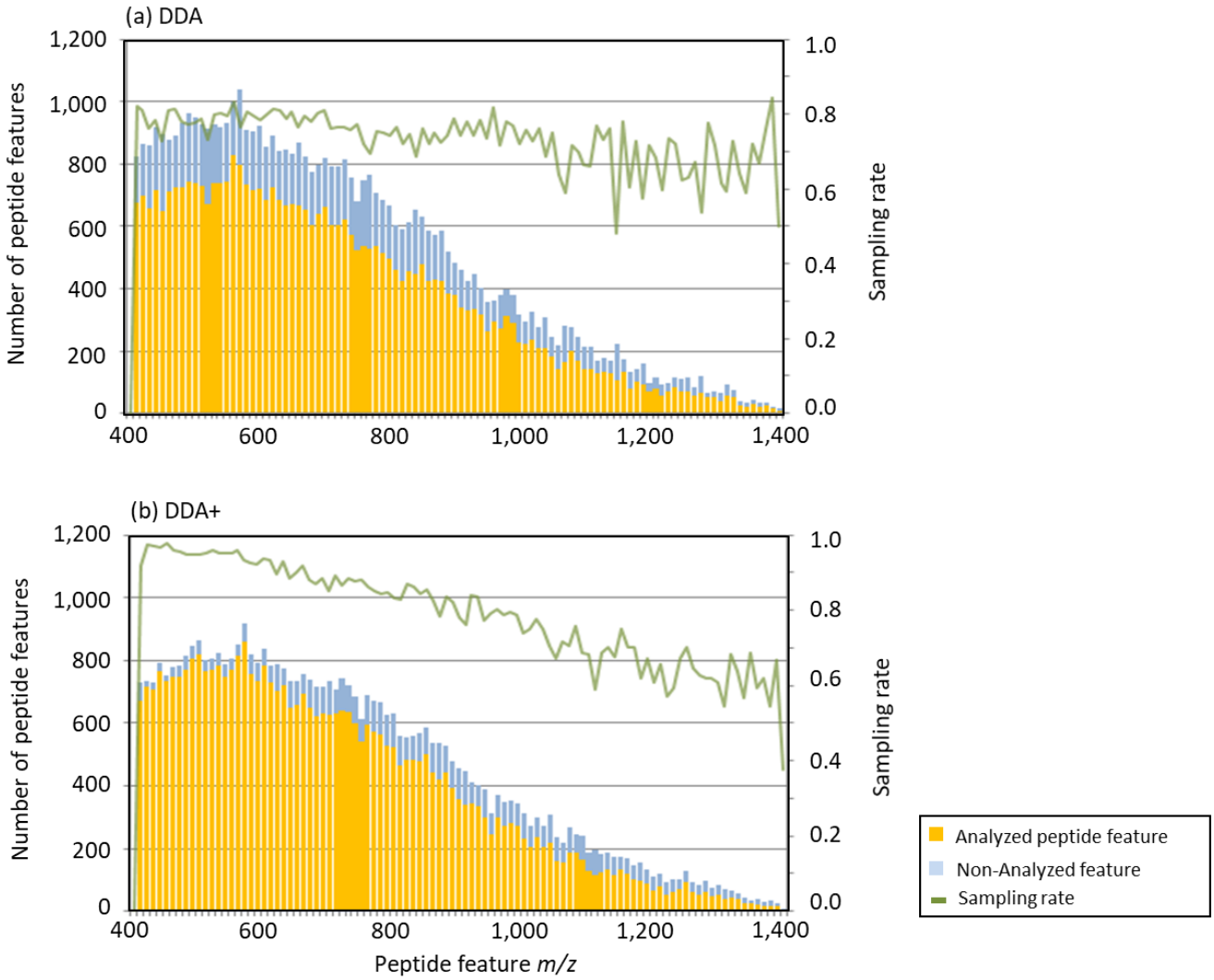

Figure 2-8 Number of analyzed and non-analyzed peptide feature over peptide feature m/z. The histograms shown here are the number of detected peptide features over *m/z* dimension in first replicated HeLa data acquired with (a) DDA and (b) DDA+. The analyzed and non-analyzed peptide features are colored with yellow and light blue, respectively. The sampling rate (the green line plot along secondary y-axis) is the number of analyzed peptide feature divided by the total number of peptide features.

2.3.4 *The identification rate with different parameters*

Although DDA+ samples more peptide features and generates more MS/MS spectra within a specific sample, the resulting identification rate of DDA+ spectral data is lower than that from DDA. To understand the underlying factors impacting identification in these two datasets, I investigated the relationships between identification rates with different parameters, include the MS/MS elution time, relative location of precursor ions in isolation window, relative sampling location on the chromatographic peak, and the intensity of precursor ion envelop. The results in both Hela replicates are similar, therefore, I focus below on the results from first Hela replicate acquired by DDA and DDA+ for clarity.

2.3.4.1 MS/MS scan elution time

The number of MS/MS scans confidently identified over elution time is shown in Figure 2-9. In both DDA and DDA+, higher identification rates are observed in the middle of the gradient, where the majority peptides elute off the LC column. The identification rate in DDA+ data is lower than DDA, especially in the middle of LC gradient. Although many MS/MS scans acquired in the early and late parts of the LC gradient in DDA+, the number of identified scans is very low on both sides.

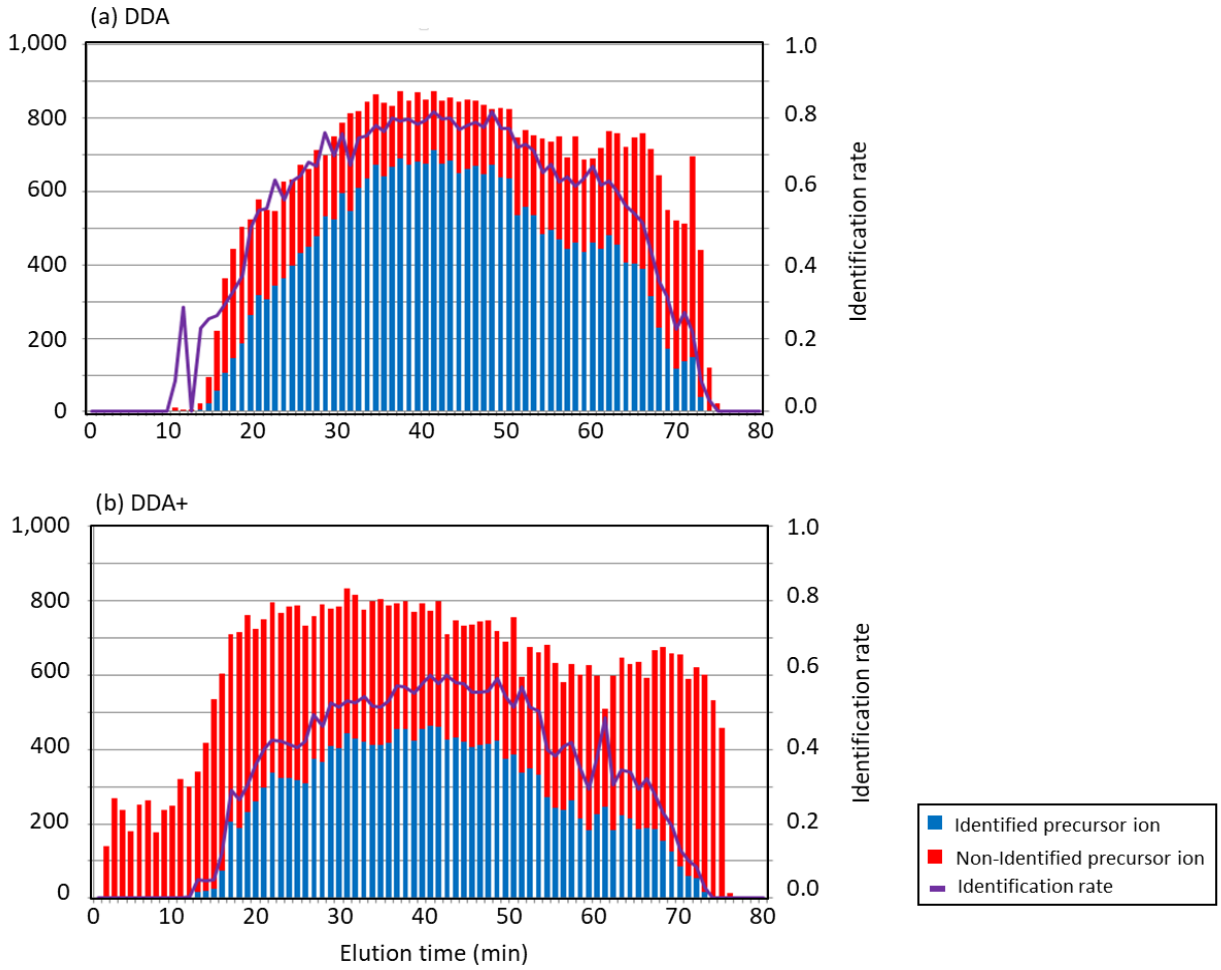

Figure 2-9 Number of identified and non-identified MS/MS scans over elution time. The histograms colored in blue and red are the numbers of identified and non-identified MS/MS scans over time, respectively. Purple line is the identification rate plot along the secondary y-axis. The identification rate is the number of identified precursor ion divided by total number of precursor ions.

2.3.4.2 The precursor monoisotopic peak relative to the center of the isolation window

One of the major differences between the DDA and DDA+ methods is the isolation window placement when generating MS/MS spectra. DDA centers isolation window on targeted precursor ions, while DDA+ isolates the precursor ions with predefined isolation windows (Figure 2-10). To understand the effect of the precursor ion's relative location on identification, I plotted the observed frequency of the monoisotopic peaks of precursors at different locations within an isolation window, as well as the identification rates of the resulting MS/MS spectra from those precursor ions. The results from doubly charged precursor ions are shown in Figure 2-11. As expected, the results show that the majority of precursor ions are in the center of the isolation window in DDA spectral data, and evenly (relative to DDA) distributed within the isolation window in DDA+. Interestingly, the multiple peaks of precursor ion distribution reflects the fact that *m/z* values of peptide ions are clustered and separated by 1 Da due to specific masses of 20 amino acids[64]. The decrease in identification rate toward the right side of isolation windows may be due to a smaller proportion of precursor isotopic envelopes were isolated for fragmentation, thus less fragment information for sequencing in resulting MS/MS spectra.

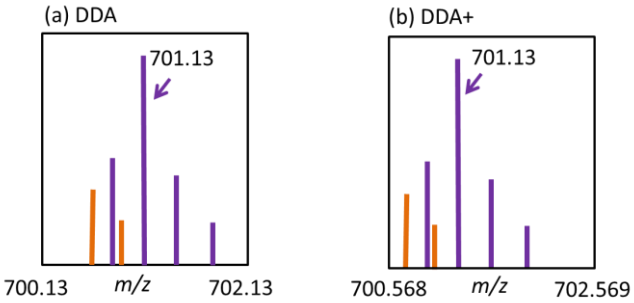

Figure 2-10 Isolation window placement for the targeted *m/z* of 701.13 Th.

Purple lines indicate an isotopic envelope of precursor ion. (a) Isolation window is center to the targeted *m/z* in DDA, (b) In DDA+, the isolation window is predefined, thus the precursor ion is often not in the middle of isolation window

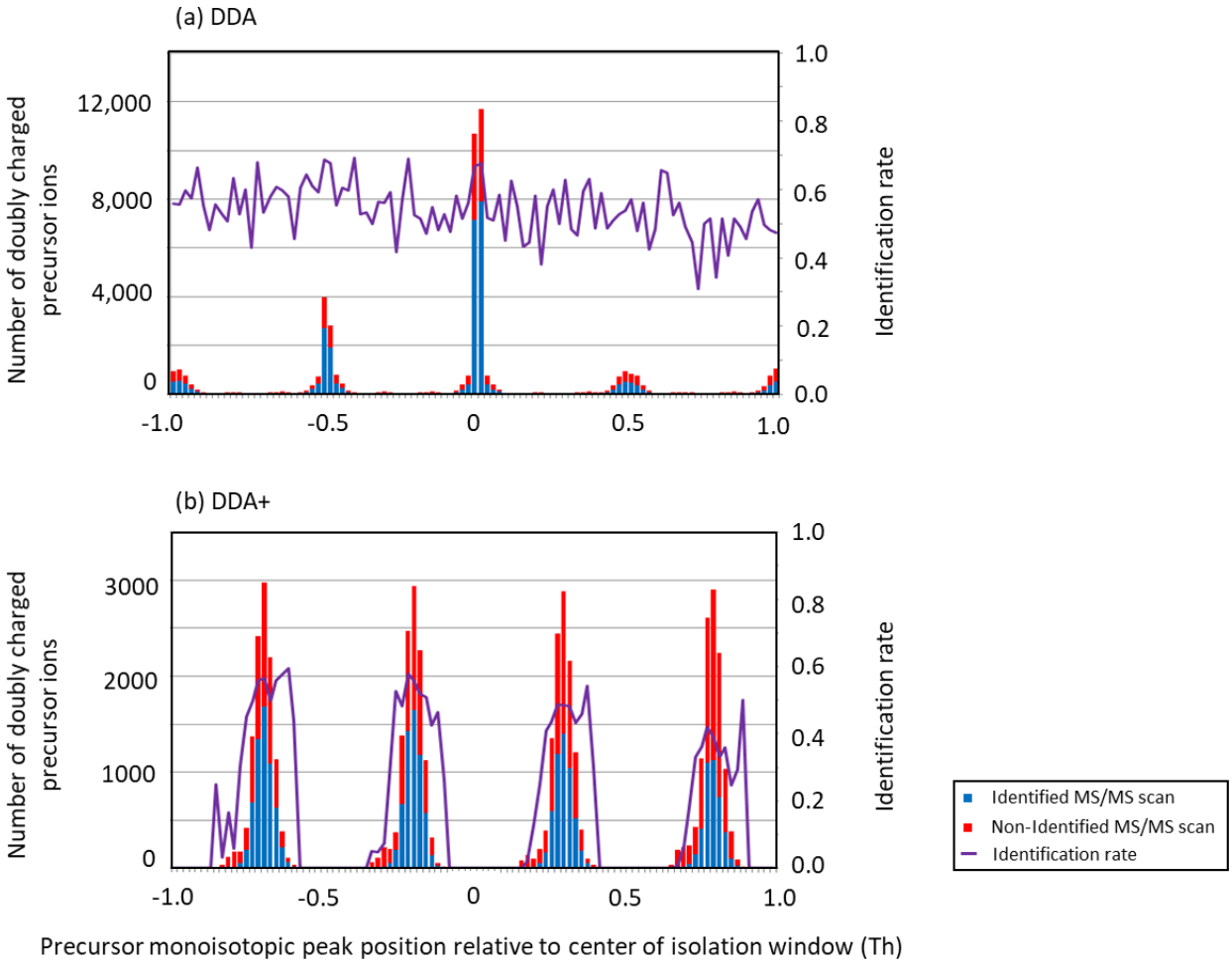

Figure 2-11 The effect of precursor monoisotopic peak position on identification rate. (a) DDA, and (b) DDA+. X-axis: precursor ion position is relative to the center of the 2 *m/z*-wide isolation window

2.3.4.3 Sampling position on a chromatographic peak

Peptides elute off the liquid chromatography column over time, ideally forming a Gaussian shaped chromatographic profile. To know whether the relative sampling position on the

chromatographic peak affects identification, I counted the frequency and identification rate of peptide-spectrum match over a chromatographic peak. (Figure 2-12). In DDA data, the number of PSMs is highest towards the beginning of the elution profile in DDA data, which indicates that DDA tends to sample peptide features early in chromatographic program. In contrast to DDA, DDA+ appears to sample more frequently at the end of eluting. The identification rate in DDA+ is lower than DDA regardless of the sampling position on a chromatographic peak.

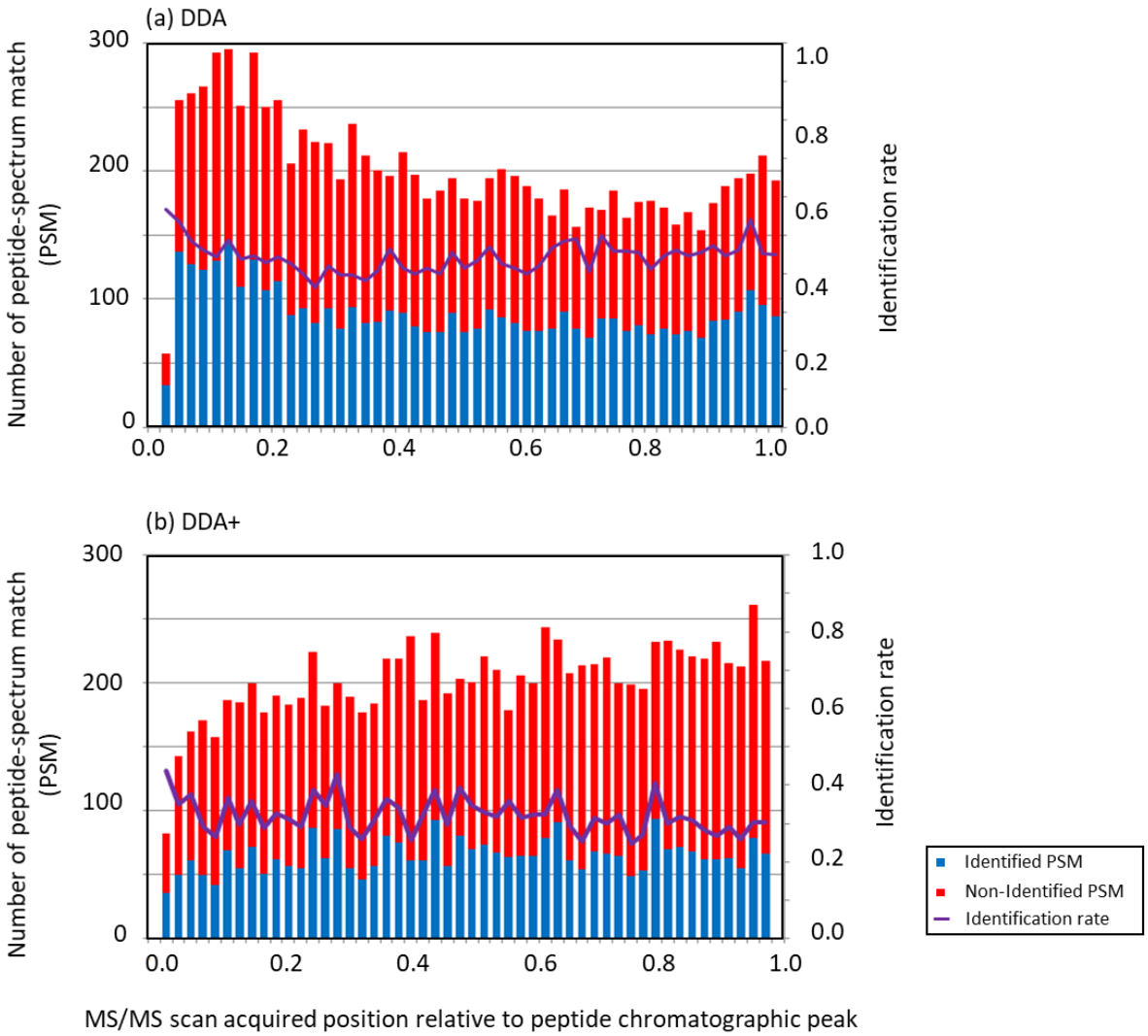

Figure 2-12 The effect of MS/MS sampled position on identification rate

(a) DDA, and (b) DDA+. X-axis is the normalized chromatographic peak width, 0 and 1indicate the beginning and end of chromatographic peak respectively.

2.3.4.4 Precursor ion intensity

Theoretically, precursor ions with higher intensity will produce more peptide fragment ions in the resulting MS/MS spectra, therefore, higher intensity precursor ions preferred for identification. The relationship of precursor intensity and identification in the two datasets are shown in Figure 2-13. The resulting trend reveals show that DDA+ sampled precursor ions from

a wider dynamic range than DDA, and that there is a larger portion of low intensity precursor ions were sampled in DDA+. Additionally, I show that the identification rates increases with increasing precursor intensity, and that spectral data from DDA gave higher identification rates than spectral data from DDA+ over all.

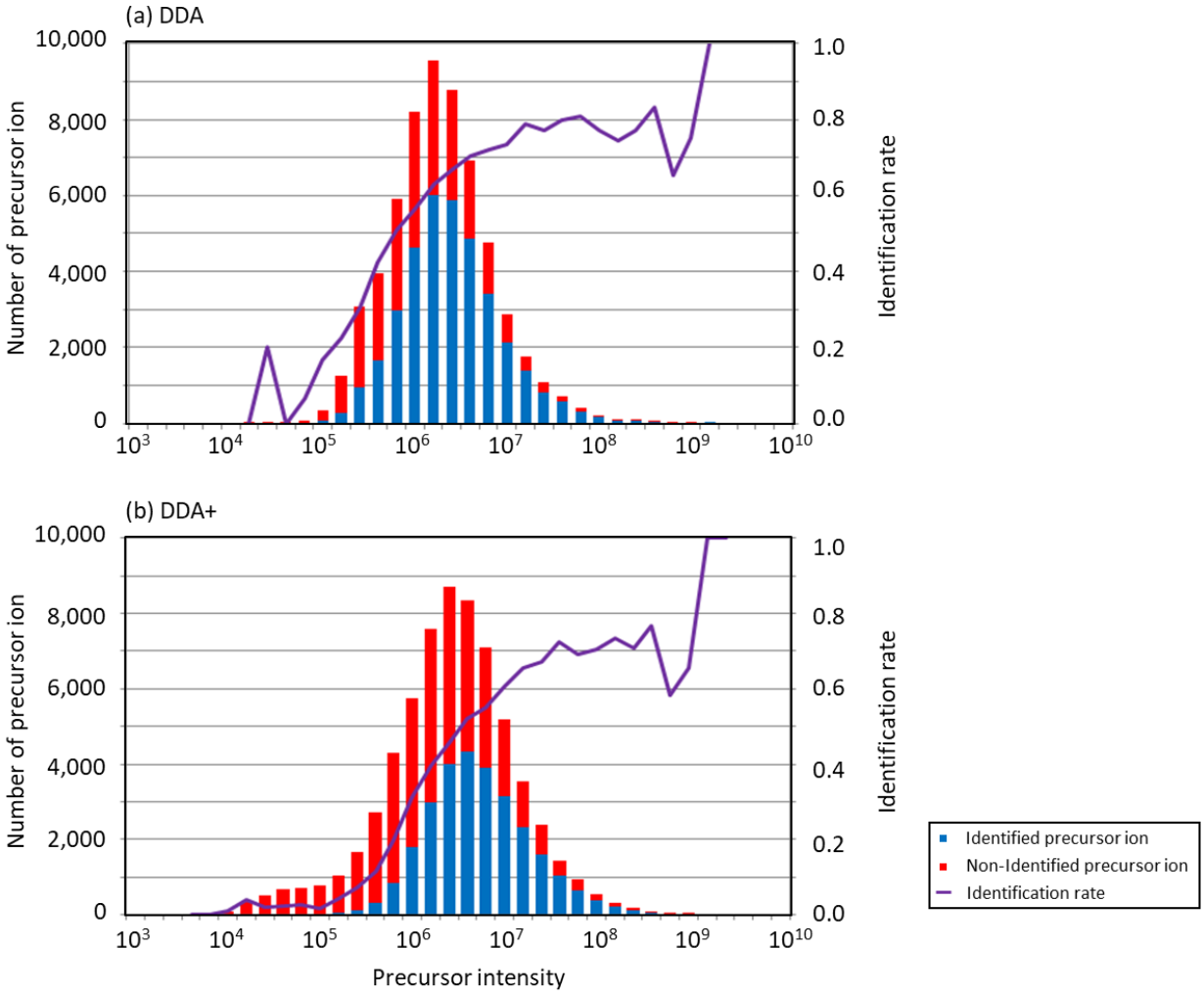

Figure 2-13 The effect of precursor ion intensity on identification rate.

The precursor ion intensity distributions from (a) DDA, and (b) DDA+ data acquired on first HeLa replicate. The identified and non-identified precursor ions are colored in blue and red, respectively. Purple line (plot along the secondary y-axis) represents the identification rate of precursor ion under different intensity range.

2.4 DISCUSSION

2.4.1 *Large portion of "No-precursor" MS/MS scans in DDA+*

DDA+ performs peptide feature detection on the fly and selects isolation windows for MS/MS acquisition if a peptide feature is detected within isolation window m/z region, thus it is expected that a majority of MS/MS scans in DDA+ can be associated back to peptide features during post-acquisition analysis. However, there is a significant amount (13,000; 22%) of acquired MS/MS spectra that cannot be matched to any peptide feature in post-acquisition analysis (Figure 2-5), and more than 65% (8,800) non-precursor MS/MS spectra were acquired during early elution time (first 15 minutes, Figure 2-6). Compared with the number of peptide features detected over the entire elution time (Figure 2-7), it is clear that there are only handful of peptide features detected in early elution time from the post-acquisition analysis. Since there is a minutes-long time delay (specifically, at least five minutes in the system used for this work) between gradient changing in the nano-LC system and peptide de-solvation from the column emitter, it is possible that only a few (or no) peptide features eluted off the LC system. Considering the low identification rate (almost 0) in MS/MS acquired in early elution time (Figure 2-9) and the nano-LC system set up, I hypothesize that MS/MS events triggered in early elution time were caused by false positives in on-the-fly peptide detection.

In contrast to DDA+, DDA acquired a very small number of MS/MS scans early in the chromatographic program (10 MS/MS scans in 10 minutes, and 500 MS/MS scans in first 15 minutes), which indicates almost all ions detected in a MS survey scan do not pass the criteria for triggering a MS/MS scan in DDA. To trigger a MS/MS event, DDA requires a minimum of 10 % targeted ion amounts (AGC target) observed in a MS survey scan. While a detected peptide ion isotope distribution is preferred when selecting ions for MS/MS analysis, it is not required in

DDA. Therefore, MS/MS scans were not acquired in elution time by DDA due to an in-sufficient amount of ions based on the information in the MS survey scan. In DDA+, a minimum intensity (ion per second) threshold is also required for triggering a MS/MS event, however, the ion under fill rate/intensity threshold parameter is one of the parameters that I could not access directly via the instrument API. Based on the differences between the two datasets (Figure 2-6), it is clear that the intensity threshold used in DDA+ is less restrictive than the ion under fill rate used in DDA. The result in Figure 2-13 also shows the same phenomenon that more low-abundant precursor ion species were analyzed in DDA+.

In addition to early elution time, DDA+ acquired a bigger portion of "no precursor" MS/MS scan across entire gradient, which may also be due to the less restrictive on-the-fly peptide detection algorithm and intensity thresholding in DDA+.

2.4.2 *Sampled peptide feature in DDA and DDA+*

The detected and sampled peptide feature analysis is based on post-acquisition peptide feature detection results. Since DDA+ sequentially acquires MS/MS scans on selected isolation window from low m/z to high m/z (Figure 2-3), one concern is that, peptides with high *m/z* values may have already eluted by the time the isolation window's spectrum request reaches the top of the acquisition queue. To avoid sampling bias due to precursor *m/z*, a real-time acquisition-updating strategy was implemented in DDA+ (see section 2.2.1.3). However, figure 2.8 shows that the DDA+ sampling rate (defined as the ratio of analyzed peptide feature to total peptide features) is biased toward peptide features with smaller *m/z* value. To further investigate the possible factors leading to sampling bias across *m/z*, I acquired data in the reverse order (from high *m/z* to low *m/z*) with another sample set (data not shown). In this experiment, the sampling bias was improved but not much. The sampling rate was still biased toward small *m/z* values despite the

fact that acquisition started with high *m/z* values, which indicates there are other factors causing the observed sampling bias. Sampling rates across the entire elution time is plotted in Figure 2-7. The result shows that DDA+ has a higher sampling rate at around 20 minutes than around 65 minutes and that the number of detected features is about the same at the two time periods. Since peptide with higher m/z values usually elute later in a reverse LC set up, the observed sampling bias under different elution time may still be due to low sampling rates at high m/z regions. I hypothesize that two peptide detection algorithms (on-the-fly vs. post-acquisition) agree with one another less when detecting peptide features with higher *m/z*. The underlying factors for this hypothesis need to be further investigated.

2.4.3 *Identification of MS/MS scans and peptide features*

The resulting MS/MS in both DDA and DDA+ were identified with database search pipeline using the same parameters, there is similar number of MS/MS scans were subjected for search pipeline, but there are fewer identified scans, PSM and peptides in DDA+ (Table 2-2). In addition, there are less commonly identified peptide sequences between two replicates in DDA+. Considering number of none-precursor MS/MS scans were acquired in DDA+, it is possible that DDA+ spent too much time on low quality or low intensity peptide features (Figure 2-6) The identification results are also associated with peptide feature level (Table 2-3), which shows that, even if the number of sampled peptide features are comparable in DDA and DDA+, there are less peptide features identified in DDA+ data.

Of all the parameters that may affect identification of acquired MS/MS (see section 2.3.3), the precursor location within the isolation window seems to play an important role compared to the other explored parameters, since the identification rate of DDA+ is lower than

DDA regardless of how other parameters change (Figure 2-9,11, and 13). Figure 2-11 shows that the identification rate decreases from the left to the right side of the isolation window in both DDA and DDA+. DDA and DDA+ have similar identification rates across three fourths of the isolation window (relative positions from -1 to 0.5), and DDA+ has a slightly lower identification rate on one fourth of the isolation window (0.5 to 1). There are two factors that could cause low identification rates on the right side of the isolation window: 1) Isolation efficiency on the edge of isolation window[65] being lower due to the properties of the quadrupole used in Q-Exactive HF, or 2) the majority of peptide ions have one or two heavy isotope atoms in their composition, which results in peptide ions with mass of the monoisotopic (M) plus one or two Dalton (M+1 or M+2). In Figure 2-11, the relative position of the monoisotopic peak to the center of the isolation window was used; therefore, for precursors located one the right side of the isolation window (0.5 to 1), their M+1 or M+2 isotope ions (0.5 or 1 Th more for doubly charged peptide) will be outside of the window, which does not get isolated for fragmentation. As the result, less fragment ion information is captured in the acquired MS/MS scan, which makes identification difficult for precursor ions located on the right side of the isolation window.

Since the majority of the precursor ion population in DDA is located on the center of the isolation window and about 25% of the precursor ion population in DDA+ are located in the low identification rate region of the isolation window, the average identification rate of DDA+ will be lower than DDA.

2.4.4 *The differences between on-the-fly and post-acquisition peptide feature detection.*

Peptide isotopic envelope detection in a MS survey scan is the first step in peptide feature detection. In both on-the-fly and post-acquisition peptide feature detection algorithms, peptide isotopic envelope is detected by the same version of Hardklor[60]; therefore, the detected peptide

isotopic envelope should be the same in both peptide detection algorithms. There is a small chance that the spectra I obtained on the fly with DDA+ are different from the spectra written in the raw file, in which case the detected isotopic envelope could be different. The difference in peptide detection is more likely due to the filter criteria applied on post-acquisition peptide feature detections. For example, in post-acquisition detection, a peptide feature with extremely long (2 minutes was used) or short elution time (3 seconds) is removed since it is more likely to be chemical noise or a false positive match. When detecting peptide features on the fly, the detection algorithm I used could not utilize the complete spectral information in the data due to time constraints; thus, it reasonably may report more false positive peptide feature matches. Therefore, additional filters may be required for on-the-fly detection to improve peptide feature detection.

2.5 CONCLUSION

Here I have described a novel acquisition method, DDA+, that aims to sample the majority of on-the-fly detected peptide features in a complex sample by selecting predefined isolation windows. Instead of only selecting a subset of high intense peptide precursors for MS/MS analysis like DDA, all the detected peptide features are considered for MS/MS analysis in DDA+. The performance of DDA+ was evaluated, and was compared to DDA with replicated Hela digest. From the acquired spectral data, I observed a large portion of MS/MS spectra were triggered on low quality peptide features when those features where determined by the DDA+ on-the-fly; therefore, those MS/MS spectra were not subjected to database searching, which also indicates that DDA+ could spend instrument time more efficiently by implementing a restricted on-the-fly peptide feature detection algorithm or by raising the intensity threshold for triggering

MS/MS events. The reason for this peptide feature sampling bias in DDA+ is not clear but it may due to the difference between the peptide feature detection algorithm on-the-fly detection versus post-acquisition, and due not entirely to the acquisition ordering scheme in DDA+.

The low identification rates were investigated, and indicate that the identification rate is not affected by where the MS/MS is acquired on a chromatographic elution profile (Figure 2-12). The MS/MS scans from precursor ions with higher intensity have higher identification rate (Figure 2-13). The identification rate of MS/MS scans is correlated with the number of peptide features eluting off the LC column (Figure 2-9, 7). However, those factors do not explain why the MS/MS scans in DDA+ give a lower identification rate than DDA. The precursor ion position within an isolation window could be one of the reasons leading to lower identification rate in DDA+. Although the number of chimeric MS/MS spectra is about the same in DDA and DDA+, the complexity of the MS/MS spectra in DDA and DDA+ may be different. How the MS/MS complexity effect on identification may worth to be further investigated.

In conclusion, several parameters in the DDA+ mode described here could be optimized in future work, specifically improving peptide feature detection on the fly for better isolation window selection and more efficient instrument time usage.

Chapter 3. CHAPTER III – DEVELOPMENT OF AMYLOIDOSIS TYPING METHODS USING DATA-INDEPENDENT ACQUISITION MASS SPECTROMETRY

3.1 INTRODUCTION

Amyloidosis is a group of diseases characterized by abnormal protein folding and aggression in extracellular regions, which then leads to progressive organ dysfunction and failure. To date, more than thirty functional and structurally diverse amyloidogenic proteins can transform into characteristic beta-sheet structured insoluble amyloid fibrils and cause disease. Amyloidosis is classified into different types based on the major amyloid fibril species in the affected tissue. Treatments are available but vary between types. Accurate amyloidosis typing is essential and paramount for an effective treatment strategy[47]. Traditionally, amyloidosis is first diagnosed by observing apple-green birefringence on a Congo red stained formalin-fixed and paraffin-embedded (FFPE) tissue section under polarized light. Subsequent typing is usually inferred with immunohistochemistry (IHC) on adjacent tissue sections. However, the typing result of IHC can be ambiguous due to the complexity of tissue background, conformational changes of amyloid fibril, and lack of specific antibodies. To remedy the limitations of IHC, laser capture microdissection mass spectrometry (LMD-MS) based typing methods have been proposed and have become the gold standard for amyloidosis typing in the clinical lab[45,48,49]. Specifically, Congo red positive regions are isolated by LMD which enriches amyloid fibrils in the sample mixture. Proteins in LMD region are digested and then analyzed with liquid chromatography tandem mass spectrometry where the tandem mass spectra (MS/MS spectra) are sampled using

data dependent acquisition (DDA). Amyloidosis typing is determined by comparing spectral counts of amyloidogenic proteins in the sample. Indented technique has shown high specificity and sensitivity for several amyloidosis typing studies[6665]. In addition to typing known amyloidosis types, this pipeline has also shown the capability of finding novel amyloidogenic proteins in recent study[67].

Although LMD-MS based methods are superior alternative to IHC and have shown promising results, there are potential issues of current LMD-MS pipeline need to be addressed. First, the data-dependent acquisition (DDA) mass spectrometry method is used to select subset of peptides for analysis in current amyloidosis pipeline. DDA was originally designed to maximize number of peptides analyzed in an assay in a way that compromises reproducibility, however, the ability of reporting consistent results of repeated assays is critical in clinical usage since the conclusion of the assay may direct effect on treatment decision and patient care. Therefore, low variability between assays is desired in a clinic assay and DDA may not be the most suitable method in clinical application. Second, current LMD-MS based methods compare spectral counts between different proteins in a sample in order to find the most dominant amyloidogenic protein in the affected tissue. However, it is known that spectral counting is biased to the proteins generate more proteotypic peptides, and the differences between signal abundance of proteins in the mass spectrometry is usually not reflect the real relationship of protein concentrations in a sample due to many factors, such as differences between enzymatic cleavage efficiencies of proteins and ionization efficiencies of peptides. Therefore, proteins/peptides are usually compared to its counterpart in different samples or conditions when using mass spectrometry in proteomics studies.

As an alternative to DDA, several acquisition methods have been developed to fit different purposes. Targeted acquisition methods are used to improve reproducibility and sensitivity of peptide detection and quantification over DDA by reducing multiplexing capability of the assay, in which only analyzing targeted peptide list during acquisition. This type of method such as selected reaction monitoring (SRM), has shown great success in clinical protein assays[30,68], and have been applied in amyloidogenic proteins for different purposes in recent studies[66].

Proteins of interest are known before developing the assay, the fully proteotypic peptides with low polymorphism frequencies are usually selected when developing the assay. However, it's not clear if fully proteotypic peptides are best indicators for monitoring in amyloidosis cases, since amyloidosis proteins could be in a truncated form and the causative sequence regions may not produce standard proteotypic peptides. In addition, several amyloidogenic proteins have shown a large number of sequence variants, which increases number of possible peptide sequences substantially and may not be able to fit in a single targeted assay when considering special sequence variants for amyloidosis. In addition, using targeted approaches will not be able to discover other potential amyloidogenic proteins in the sample.

In addition to DDA and targeted approaches, data-independent acquisition (DIA) has emerged as a new alternative for better quantification measurements than DDA and higher multiplex capability than targeted acquisitions. DIA systematically analyzes desired precursor *m/z* with relatively wide isolation windows regardless peptide information in the MS survey scans. The peptide chromatographic peak at the MS/MS level is used to estimate peptide quantity, which provides better quantitative sensitivity and accuracy than DDA spectral counting. Considering the amyloidosis typing is based on the comparison between quantities,

DIA with chromatographic peak might be a better alternative to DDA spectral counting used in the current clinical assay. In addition, DIA samples ions over a wide precursor *m/z* range with high scan cycling frequency, thus the acquired spectral data can be considered as a comprehensive digital record of a sample. New hypotheses can be tested with previously acquired DIA samples, which is especially attractive when it comes to samples that are difficult to obtain such as patient samples.

In this study, I proposed a new amyloidosis typing workflow based on quantification measurements from DIA. The performance of the DIA-based workflow is evaluated and compared to DDA spectral counting with LMD FFPE tissue samples across three major amyloidosis types and control cases. The DDA-based spectral counting pipeline was implemented based on descriptions in previous studies. A voting system for amyloidosis typing was developed in this study, in which multiple Naïve Bayes classifiers predict amyloidosis type based on integrated MS/MS level peak area from DIA data. Instead of aggregating quantitative values of all peptides to protein level, a subset of informative peptides was used as a proxy of amyloidosis-related proteins in a new typing method. Interestingly, I found some specific regions of amyloidogenic proteins that are enriched and the non-standard proteotypic peptides from those regions give better discriminatory power for typing. The typing decision is based on the relative quantitative value of peptides in different samples.

With developed majority voting system, I demonstrated a way of using DIA and normalized intensity value for amyloidosis typing, which shows better typing performance compare with current DDA based spectral counting pipeline for AL lambda, AL kappa and AA types, and better reproducibility between duplicated samples from same FFPE tissue block.

3.2 MATERIAL AND METHOD

3.2.1 *Study Subjects and laser capture microdissection*

Eighty-four samples were acquired from 42 cases which include 6 amyloid A (AA), 14 amyloid light chain kappa (AL-kappa), 14 amyloid light chain lambda (AL-lambda) amyloidosis cases, and 8 time zero allograft control cases. All the samples were from formalin-fixed paraffin-embedded (FFPE) renal biopsy specimens collected from 2003 through 2015. This study was approved by University of Washington institutional review board (IRB #48306). For each case, 10µm-thick tissue slices were sectioned from FFPE renal specimens and placed on DIRECTOR laser microdissection slides (Expression Pathology). The tissue sections were de-paraffinized and then stained with Congo-Red to confirm amyloid fibril deposition regions. We isolated Congo-Red positive glomeruli from amyloidosis cases with a Leica LMD6500 laser capture microdissection system, and similarly isolated representative glomeruli from control cases. For each case, two replicates were collected from tissue sections with at least 50,000 – 60,000 mm^2 area for each replicate. Dissected tissue samples were placed in 0.5 ml Eppendorf vials with 25 µl 0.1% RapiGest (Waters) in 50mM ammonium bicarbonate. The samples were stored at -80 degree Celsius until the day of protein digestion.

To develop the DIA typing algorithm for amyloidosis typing, 34 Congo-red positive cases (68 samples) were divided into training and testing sets for building and evaluating the models respectively.

The training set consisted of 4 AA, 10 AL-kappa, and 10 AL-lambda cases, whereas the testing set had 2 AA, 4 AL-kappa, and 4 AL-lambda cases. The cases were randomly assigned into the two sets. Another 12 samples from 4 cases (2 AA, 1 AL-kappa, and 1 AL-lambda; triplicate) were also used as testing set for the final model evaluation. This set of samples were

acquired, processed and analyzed with same experimental procedures but in different batches. All the 4 cases were from FFPE autopsy samples collected from 2001 to 2012. This study was also approved by University of Washington institutional review board (IRB # 48306).

3.2.2 *Protein extraction and digestion*

Dissected tissue samples were thawed at 4 degree Celsius and spun down for a minute. An additional 35 µl 0.1% RepiGest was then added to make a final volume of 60 µl. Proteins were extracted and resolubilized by heating at 98 degrees Celsius for two hours and followed by water bath sonication for an hour. Five microliters of 15 ng/µl 15N-Apolipoprotein was added as internal standard for post-acquisition digestion efficiency correction. Proteins were reduced by 5mM dithiothreitol (DTT) for 30 min at 60 degree Celsius, alkylated with 15 mM iodoacetamide (IAA) for 30 min at room temperature in the dark, and digested with 1ug of trypsin for 18 hours at 37 degree Celsius, 600 rpm. Digestion was quenched by adding 5 M HCl to make a final concentration of 50mM and incubating at 37 degree Celsius for 60 min. The supernatants containing peptides were removed after spinning at 14K, 4 degree Celsius for 10 min. The samples were stored in -80 degree Celsius until the day of mass spectrometry acquisition.

3.2.3 *Nanoflow liquid chromatography and tandem mass spectrometry*

Protein digests were separated using a reversed-phase liquid chromatography system (Waters nanoACQUITY) and analyzed on a Q-Exactive HF mass spectrometer (Thermo Scientific). An injection of 3 µl peptides was made with 8 µl trapping column wash with flow rate of 2 µl/min. The peptides were eluted from analytical column using a 60 min linear gradient of 2-40% buffer

B and a 5-min gradient of 40-60% buffer B. Mobile phase A consisted water with 0.1% formic acid. Mobile phase B consisted acetonitrile with 0.1% formic acid. The homemade trapping and analytical columns were made using 3cm (150 μm inner diameter) and 15 cm (75μm inner diameter) fused silica capillary column (Polymicro Technologies, Phoenix, AZ), packed with 90Å 4μm C12 (Jupiter Proteo; Phenomenex, Ventura, CA) and 120Å 3μm C18-AQ resin (Dr. Maisch; GmbH, Germany), respectively.

Mass spectra were acquired using both data-dependent acquisition (DDA) and data-independent acquisition (DIA). In DDA mode, spectra were acquired with a cycle of one high-resolution MS scan (400-1400*m/z*, 120,000 resolution at 200 *m/z*) followed by twenty MS/MS scans (30,000 resolution at 400 *m/z*). The precursor ion of MS/MS scan was selected based on information in MS scan and isolated with 2-*m/z* wide window. The mass-to-charge ratios of selected ions would be excluded for 20 second. The DIA data was acquired with a cycle of one mass scan (445-885 *m/z*, 120,000 resolution at 400 *m/z*) followed by twenty-one MS/MS scans isolating from predefined *m/z* regions across 450-870 *m/z*. Forty-two 20 *m/z*-wide with alternative 10 *m/z* overlap. The isolation window edge placement was optimized[1].

3.2.4 *Protein database preparation*

The standard human protein sequence database was downloaded from Swiss-Port. To consider amyloidosis related sequence variations in analysis, we collected known mutations and polymorphisms from Uniprot (the version released on 20141123). 511 additional sequence entries from 30 amyloidogenic proteins were incorporated into the standard protein database. Common contaminants and trypsin sequence were also appended to the protein database, resulting in 42,469 total entries in the database. To estimate peptide identification confidence by

target-decoy approach[11,14], we generated a decoy database with reversed protein sequences and appended decoys to the target protein database. The final concatenated protein database was used for database search.

3.2.5 *DDA analysis*

3.2.5.1 DDA data interpretation and protein quantification

The Raw files acquired with the DDA method were converted into ms1 and ms2 peak list format using MSConvert[61]. The monoisotopic mass of ms2 precursors were corrected using Hardklor[60] (v2.3.0) and Bullseye[2] (v1.3.0). The resulting ms2 files were searched against a concatenated protein sequence database using the search engine Comet, X!Tandem, and Mascot. The search parameters were configured to derive semi-tryptic peptides with maximum of two miss cleavage and static carbamidomethyl modification (+57.02 Da) of cysteine, variable oxidation on methionine. In addition, X!Tandem also considered acetylation (+42.01 Da) on protein N-terminal, water loss (-18.01 Da) of N-terminal glutamates, and ammonia loss (-17.02 Da) on glutamine and carbamidomethylated cysteine by default. The mass tolerance for precursor and MS/MS matching were 10 ppm and 0.2 Da respectively. The spectra identification results from the three search engines were further integrated and analyzed using Scaffold,[69] which calculates probabilities for identified peptides and proteins. Peptides and proteins with probabilities higher than or equal to 0.9 were considered for further analysis. Qualified peptides and proteins were quantified with total spectral counts and quantification tables were exported for amyloidosis diagnosis and typing.

3.2.5.2 Amyloidosis typing with DDA spectral counting table

We performed amyloidosis typing on all Congo-red positive samples using methods described in previous studies[3–5], in which the amylodosis typing decision was made by comparing spectral counts between all amyloidogenic proteins within a sample. Typing was then assigned according to the amyloidogenic protein associated with the highest spectral counts in a sample. A minimum of five counts was required for typing; otherwise the sample was considered a failed typing case. For immunoglobulin proteins, we treated constant and variable regions as different protein groups for typing.

3.2.6 *DIA analysis*

3.2.6.1 DIA precursor *m/z* range

To have a desired scanning duty cycle, isolation window width, and number of MS/MS events across a peptide chromatographic peak over elution time, 400 *m/z*-wide precursor *m/z* range were chosen and covered by 20 *m/z*-wide isolation window for DIA experiment. To find the optimal precursor *m/z* region that covers the majority of peptides of amyloidosis related proteins (30 known amyloidogenic protein and 2 amyloidosis signature proteins), the frequency distribution of observed and predicted tryptic amyloid peptides were plotted across different *m/z* values (Figure 3-9 a, b). Observed amyloid peptides were identified in preliminary DDA studies on biopsy samples from AL-lambda, AL-kappa, ATTHY, and AA types.

3.2.6.2 DIA data interpretation and protein quantification

DIA Raw files were converted into mzML with MSConvert and then searched against a spectral library using EncyclopeDIA (version 0.6.0) with the following parameters: Mass tolerance of 10 ppm for precursor, fragment and library matching, tryptic peptides, overlapping DIA as

acquisition type, normal target and decoy search mode, percolator version of 3.01, accept both b and y ions, and minimum quantitative ions of 5. The spectral library used in the EncyclopeDIA search was built from integrated peptide identifications from Scaffold. EncyclopeDIA computes scores for a peptide chromatographic peak based on the fragmentation and retention time information from spectral library. Among all confidentially detected peptide chromatographic peaks (FDR of 0.01) in EncyclopeDIA, I kept peptides associated with known amyloidosis related proteins (Appendix A1), and quantified peptides by the integrated peak areas of curated transitions. The transitions were selected and ranked by EncyclopeDIA based on the signal intensity and correlation over time. The top six transitions were used following manual curation. Only peptides with three or more (up to six) transitions were included for further analysis. Skyline was used to extract integrated peak area under selected transitions of peptides and then generated the peptide quantification report used for amyloidosis typing.

3.2.6.3 Normalization for peptide quantification in DIA.

Two normalization strategies were applied to correct for variabilities introduced during sample preparation. To normalize total protein amount between samples, protein concentration is usually measured in the beginning of the experiment; however, the protein amount of the LCM samples was too low to be reliably measured or to be detected by conventional protein assays. Therefore, I used the total ion current (TIC) of all multiply charged, peptide-like features in the mass spectral raw data as a proxy of total peptide amount in the sample. The difference of peptide loading amounts between samples is then normalized by dividing by the TIC. I used a HeLa digest standard (Pierce, #88328) to demonstrate the linearity between peptide concentration and TIC, described in detail in the next section. In addition, I spiked ^{15}N-heavy labeled

apolipoprotein A1 (APOA1) in each sample, then used the integrated peak area of the ^{15}N version of APOA1 peptide VQPYLDDFQK (VQP; amino acid position from 120 to 129) to capture the differences of the digestion process between samples. Specifically, to account for the digestion difference, I divided peptide intensity by the intensity of ^{15}N VQP. The final ratio value of each peptide was used in analysis.

3.2.7 *Linearity between peptide concentration and TIC*

20 µg Hela digest (Pierce, #88328) was reconstituted and diluted with 50mM ammonia bicarbonate into dilutions of 0.05, 0.025, 0.01, 0.005, 0.0025, and 0.001 µg /µL. The diluted Hela digest samples were eluted from analytical column using a 20 min linear gradient of 2-40% buffer B and a 3-min gradient of 40-60% buffer B. Mobile phase A consisted water with 0.1% formic acid. Mobile phase B consisted acetonitrile with 0.1% formic acid. The homemade trapping and analytical columns were made using 3cm (150 µm inner diameter) and 10 cm (75µm inner diameter) fused silica capillary column (Polymicro Technologies, Phoenix, AZ), packed with 90Å 4µm C12 (Jupiter Proteo; Phenomenex, Ventura, CA) and 120Å 3µm C18-AQ resin (Dr. Maisch; GmbH, Germany), respectively. Since the goal of this experiment was to evaluate the relationship between peptide concentration and TIC value, only high-resolution (400-1400*m/z*, 120,000 resolution at 200 *m/z*) full MS survey spectra were acquired throughout the entire experiment. Bullseye was used to detect peptide-like features. Bullseye uses the averaged isotope distribution of peptides to predict isotope distribution at given *m/z* value, and then compares predicted isotope distribution with observed ones within a tolerance. Since the detection was based on an approximated isotope distribution, there could be false-positive detections reported by Bullseye. Considering the majority of tryptic peptides carry multiple

charges, and the metabolic or chemical ions are more likely to be singly charged ions, only the multiply charged peptide-like features were used to represent total mass spectral signal of peptides in the sample.

3.2.8 *Peptide selection for naïve Bayes classifiers*

Due to divergent biochemical properties of peptides and digestion efficiencies with a protein, different peptides from the same protein could behave very differently during sample processing and ionization, therefore, not every peptide from a protein is equally informative for amyloidosis typing. In addition, it has been suggested that the amyloid fibrils in affected region may mainly come from specific regions of proteins, therefore, instead of using proteins level quantification, peptides were considered as indicators in this study. To select peptides as an indicator for specific amyloidosis types, a training set was used for peptide selection process and model development. The first step for peptide selection was to curate the peptide chromatographic peak, where the peptides that show more than one chromatographic peaks are removed. The peptides with missed cleavages were also removed from consideration.

The next step transformed all the amyloidosis cases of the training set into multiple binary class groups using one-vs.-rest strategy (AA vs. others; AL-lambda vs. others; AL-kappa vs. others). The goal here is to find a subset of peptides that are enriched and have better capability to distinguish specific amyloidosis type from others. A reliable peptide indicator should be observed in majority of samples that from specific amyloidosis type, therefore, only peptide that has a chromatographic peak presents in at least 70% of specific type amyloidosis samples are selected for further consideration. For following computational process, the intensity of a missing peptide was set to an arbitrary number that was lowest normalized intensity value

among all the samples divided by 2. Since all the peptide intensity was transformed into log2 space during computation, the arbitrary number was the lowest normalized intensity minus 1 in log2 space. The idea here was to use a very small number to represent the peptide quantity below the limit of detection in a MS.

The discriminatory power of each peptide for different binary amyloidosis type group is estimated with area under curve (AUC) of the receiver operating characteristic curve (ROC). The ROC curve is a plot of classification sensitivity against one minus classification specificity, as the intensity cut-off value is increased from minimum to maximum observed intensity value. The higher AUC of an ROC means better separation between intensity distributions of two classes (For example, AA vs. others). The AUC values are normalized score with minimum of 0 and maximum of 1. The top three peptides with AUC score equal to or above 0.85 was subjected to build Naïve Bayes models.

Each selected peptide is used to build a binary naïve based classifier, in which two Gaussian probability density models are created based on intensity distributions of specific amyloidosis type and other types in training set. For each Naïve Bayes classifier, it reports an odds ratio based on the given normalized peptide intensity. The odd ratio is the likelihood of "being a specific amyloidosis type" divided by likelihood of "not being a specific amyloidosis type" (For example, the odd ratio of AA/non-AA). The odds ratio above 1 means this given sample is typed to specific amyloidosis type based on the intensity of selected peptide indicator, and vice versa. To evaluate whether a peptide Naïve Bayes model is biased to training set or has a very high variance during typing differently sized training samples were used to build Naïve Bayes models, and plot the relationship between the prediction accuracy against training size. A three-fold cross validation was used to select a different subset of samples to train and validate

the model. Peptides were selected as an indicator if the curve showed an increased and converged prediction accuracy between cross-validation set and training set as size of training set increased.

3.2.9 *Voting system for Amyloidosis typing*

For AA, AL-lambda, and AL-kappa, a different subset of peptides were selected to build three groups of binary Naïve Bayes classifiers to distinguish AA vs. others, AL-lambda vs. other, and AL-kappa vs. others, respectively. For each amyloidosis type, several indicator peptides and associated Naïve Bayes classifier were selected and built. A simple voting system is used to merge and compare decisions from three groups of classifiers, in which the percentage of positive typing decisions within each amyloidosis typing group is computed. The final type is assigned to the group gives highest positive typing percentage over all the peptides. If none of three classifier groups give positive typing decision, then the sample will be reported as ambiguous.

3.3 RESULT

3.3.1 *Amyloidosis peptide is detected by MS but not quantified by spectral counting*

To give an overview of the data acquired using DDA and DIA, the spectral counting values and integrated peak areas for identified amyloidosis related proteins across all samples are shown in a heatmap (Figure 3-1). The Figure 3-1 shows DIA has better quantitative sensitivity than DDA where a large portion of the data matrix are missing values (colored with grey). Figure 3-2 shows an example of a peptide that has spectral counts of 0, but a chromatographic peak indicative of peptide truly quantified in the data.

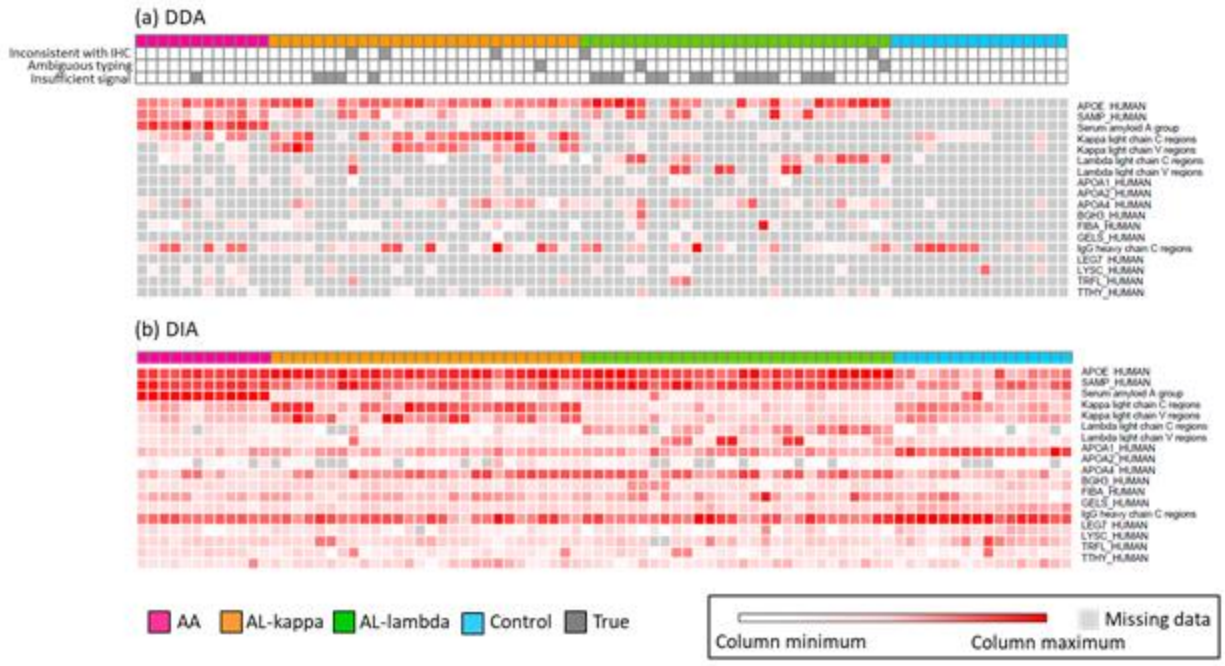

Figure 3-1 Quantification profiles of samples from two quantification methods

Protein is quantified with a) spectra counts from DDA results or b) integrated peak area extracted from DIA. Each column indicates a sample, rows are identified amyloidosis related proteins. The replicates from same sample are adjacent columns. Majority of proteins have one or more peptides detected by MS and quantified in DIA dataset. The dark grey blocks above DDA data matrix indicate the samples has 1) inconsistent typing result with IHC, 2) ambiguous typing results where two amyloidogenic proteins have same spectral counts, and 3) insufficient spectral counts for typing

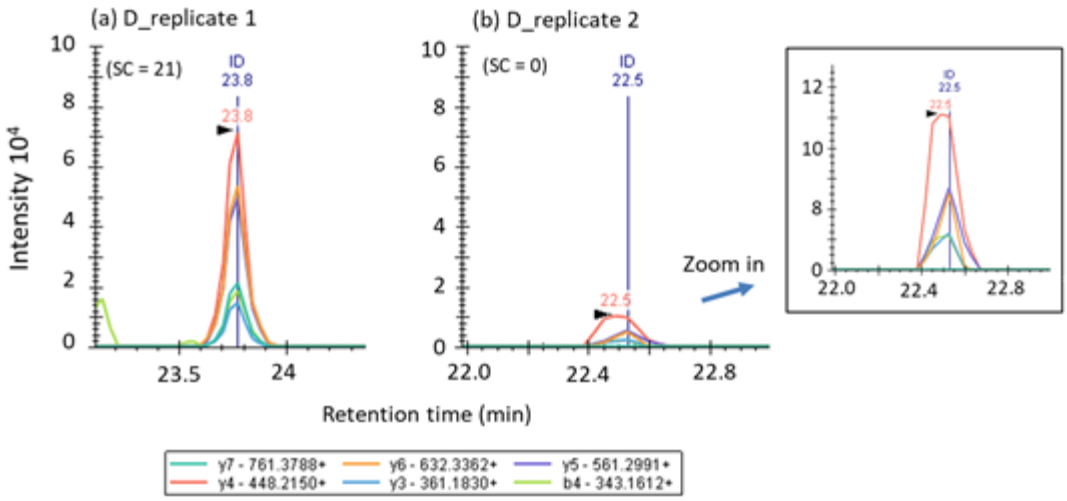

Figure 3-2 Example of a peptide can be reproducible observed in DIA but not DDA.

A doubly charged serum amyloid A peptide AAEAISDAR are observed in both replicates from same biopsy. The SC indicates the spectral counts of the peptide

3.3.2 *Typing with DDA spectral counts*

Fifty-seven (83.8%) out of 68 Congo-red positive amyloidosis samples were confirmed by spectral counting through observation of two amyloidosis signature proteins: apolipoprotein E (APOE) and serum amyloid P component (SAMP). None of the control cases had spectral evidence for APOE or SAMP. I performed typing by spectral counting profile on all Congo-red positive samples (Figure 3-1, Appendix A2). Eleven (91.7%) AA, twenty (71.4%) AL-kappa and ten (35.7 %) AL-lambda samples were typed in agreement with the IHC report. Five samples had typing results inconsistent with IHC, specifically one was typed as AL-lambda with DDA data, but as AL-kappa by IHC. Another 4 samples were typed as amyloid heavy due to significant amounts of immunoglobulin heavy proteins, whereas IHC reports them as AL-lambda

or AL-kappa. Three samples gave ambiguous results where more than one amyloidogenic proteins had equal amounts of spectral counts. Finally, nineteen (27.9%) of 68 samples had insufficient spectral counts (less than five counts) for any of amyloidogenic proteins. The typing consistency between spectral counting profile and IHC would be 100% (11 out of 11), 83.3% (20 out 24), and 71.4 % (10 out 14) for AA, AL-kappa, and AL lambda if only samples with sufficient spectral counts were considered.

Among all 34 duplicated Congo-red positive sample pairs, 22 pairs had sufficient spectral counting results in both replicates, where 5 (out of 5) AA, 8 (out of 11) AL-kappa, and 4 (out of 6) AL-lambda sample pairs had consistent typing results between replicates and IHC (Figure 3-1, Appendix A2)

3.3.3 *Different dynamic ranges in three causative amyloidogenic proteins*

The intensity of all detected peptides in three causative amyloidogenic proteins were summed to the protein level, and the intensity distribution of proteins in three typing groups plotted in Figure 3-3. The figure shows that causative amyloidogenic proteins have higher intensity in samples from associated amyloidosis types than samples from other types. The intensity distributions of serum amyloid A separated the AA and non-AA groups well; however, the intensity distributions of immunoglobulin light chain kappa constant (IGKC) and immunoglobulin light chain lambda constant (IGLC) proteins had significant overlap between the two populations. In addition, Figure 3-3 also shows that the intensities of different proteins are distributed on different scales.

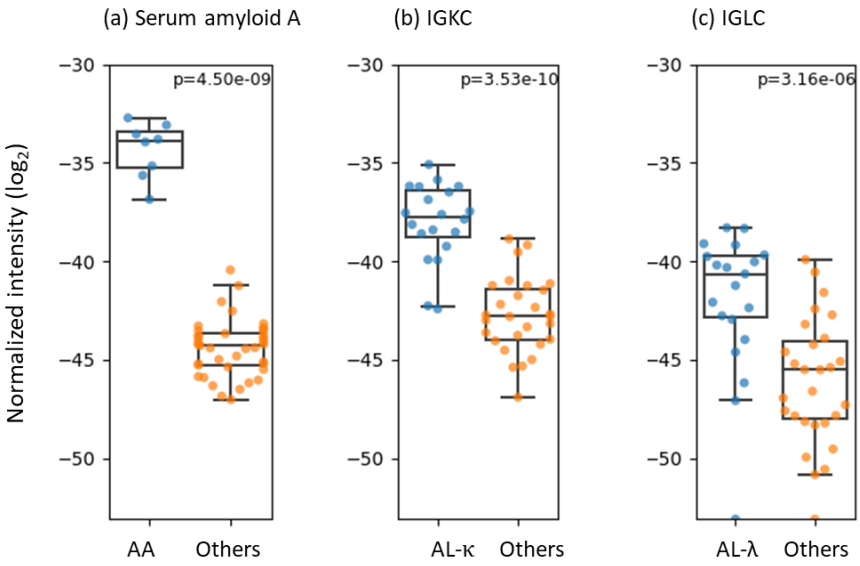

Figure 3-3 Intensity distributions of three causative amyloidogenic proteins in 68 Congo-red positive samples

3.3.4 *Intensity distribution of peptides from IGLC proteins*

The intensities of peptides from the immunoglobulin light chain lambda constant (IGLC) were distributed very differently in AL-lambda vs. “other types” of patient populations in this study (Figure 3-4). Some peptides separated well between the two populations, which suggests those peptides could be better indicators for AL-lambda typing.

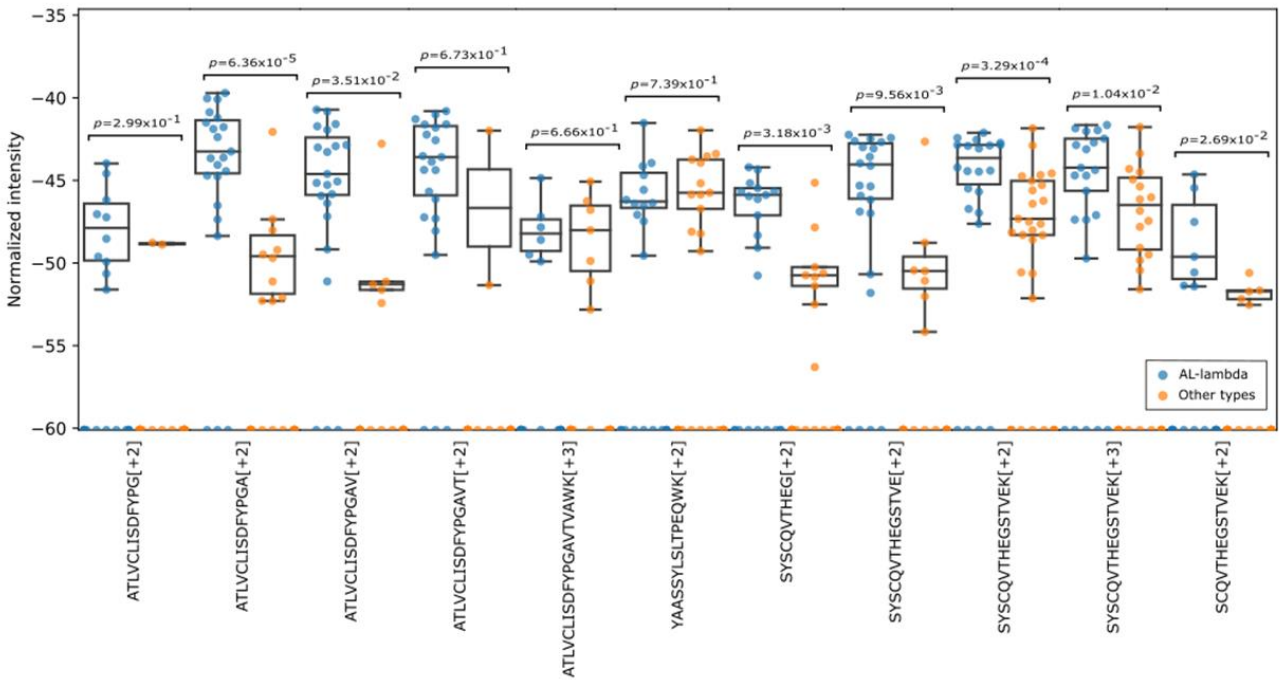

Figure 3-4 Intensity distributions of peptides from immunoglobulin light chain lambda constant regions.

The missing values are aligned to the bottom of y-axis. The p-value calculation is based on measurable data points in two populations (AL-lambda vs. others)

3.3.5 *Selected peptide indicators for amyloidosis typing*

Three serum amyloid A peptides, immunoglobulin light chain kappa constant (IGKC) region peptides, and immunoglobulin light chain lambda constant (IGLC) region peptides are selected to build a binary Naïve Bayes classifier to classified AA vs. others, AL-kappa vs. others, and AL-lambda vs. others respectively. The peptides were selected based on multiple criteria described in section 3.2.8, and three characteristics (intensity distribution, discriminatory power, and peak shape) of the peptides in the training set (48 samples from 24 cases) are shown in Figure 3-4,5,6. The three best peptides were selected for each binary amyloidosis typing task. Although the chromatographic elution profiles of the selected IGKC peptides are less ideal

(Figure 3-5), other characteristics of the selected peptides are better than the rest of the peptides from the IGKC protein.

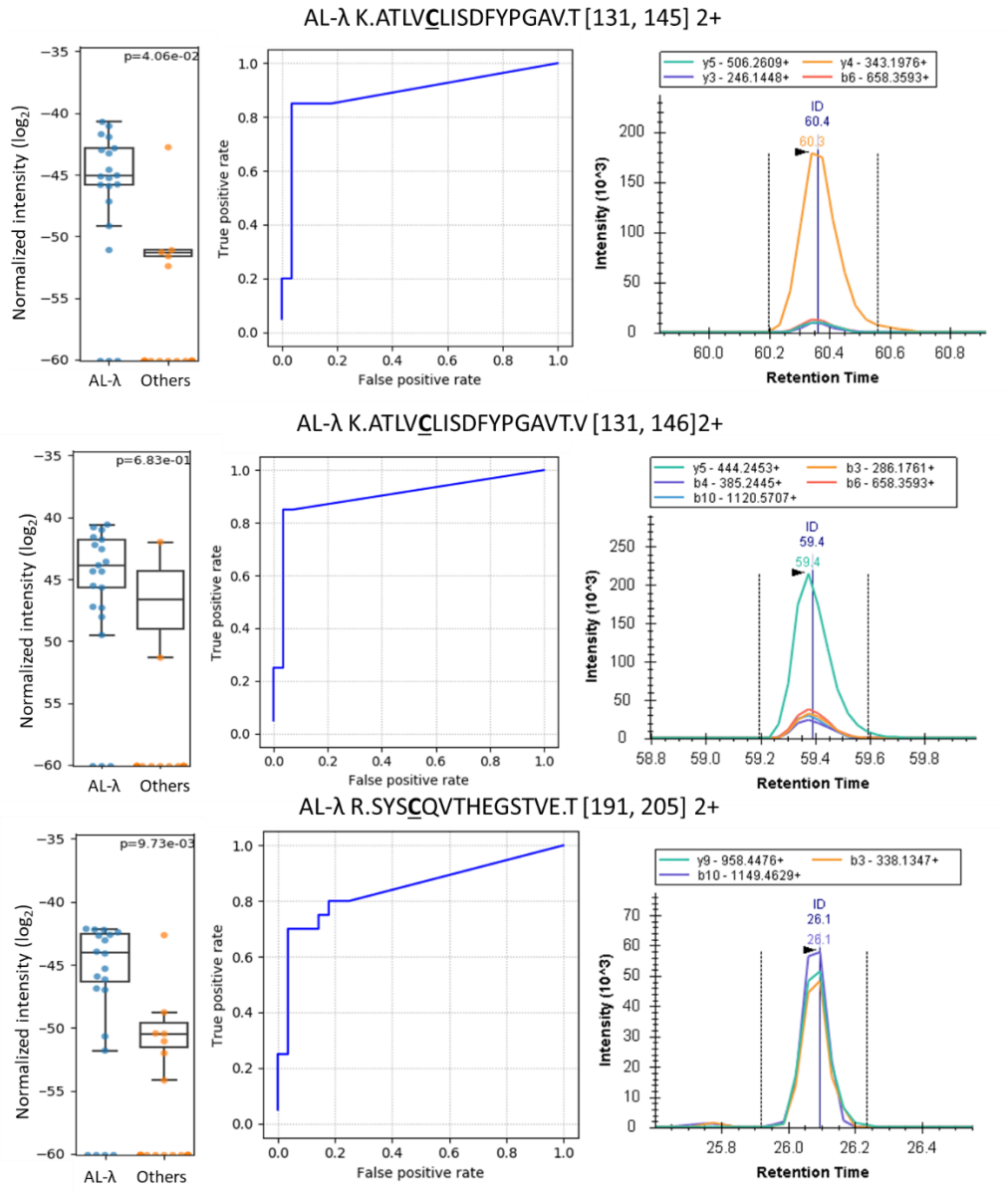

Figure 3-5 The peptide indicators used for AL-lambda binary classifiers.

The left panel is the intensity distributions of peptides in AL-lambda and "other types" populations. The middle panel is the sensitivity and 1-specificity, as intensity cut-off increases, in which sample with peptide intensity above cut-off was assigned to positive for AL-lambda. The area under the curve is used to represent discriminatory power of peptide. The right panel is peptide chromatographic peak at MS/MS level extracted from DIA data

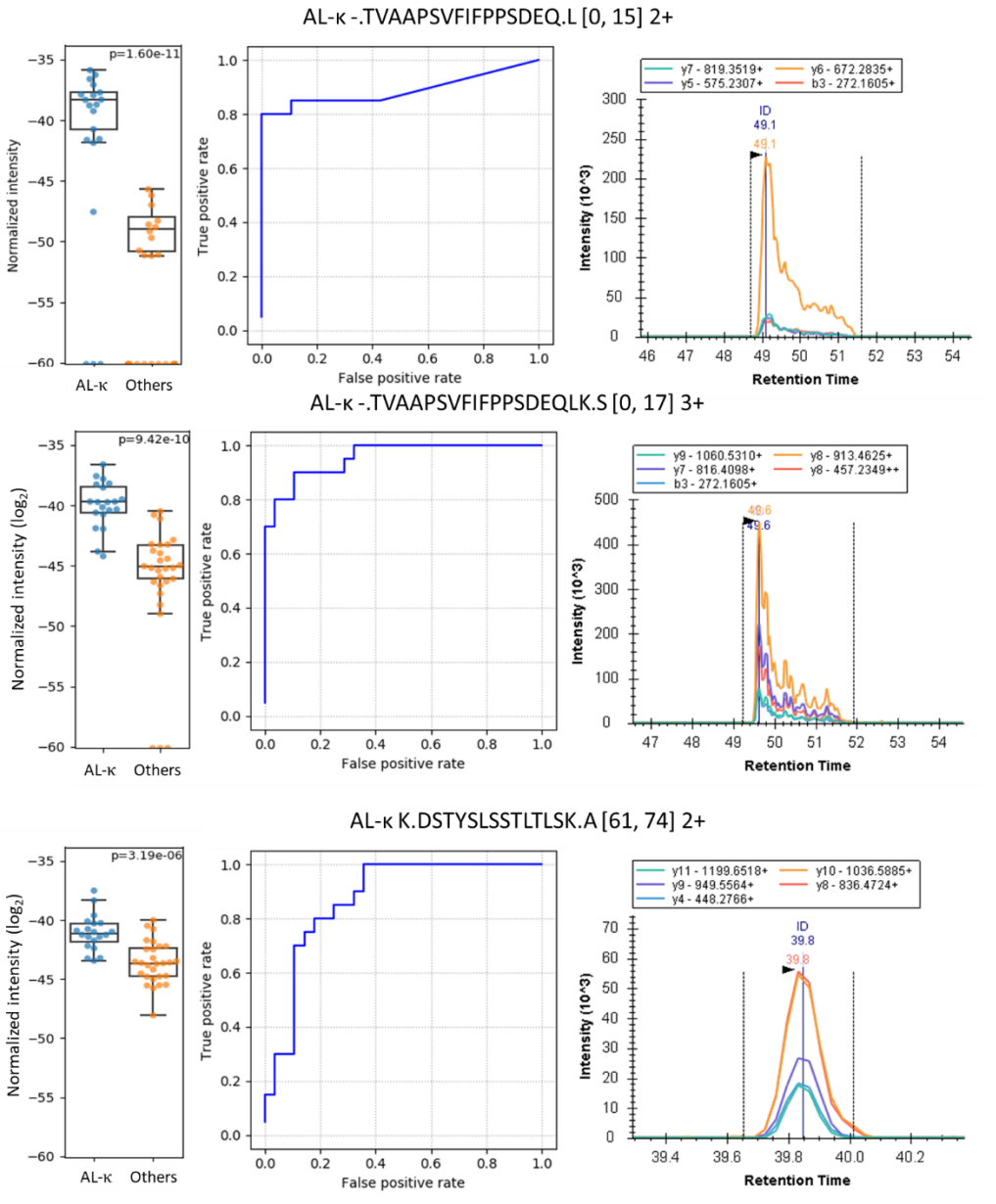

Figure 3-6 The peptide indicators used for AL-kappa binary classifiers

The left panel is the intensity distributions of peptides in AL- kappa and "other types" populations. The middle panel is the sensitivity and 1-specificity, as intensity cut-off increases, in which sample with peptide intensity above cut-off was assigned to positive for AL- kappa. The area under the curve is used to represent discriminatory power of peptide. The right panel is peptide chromatographic peak at MS/MS level extracted from DIA data

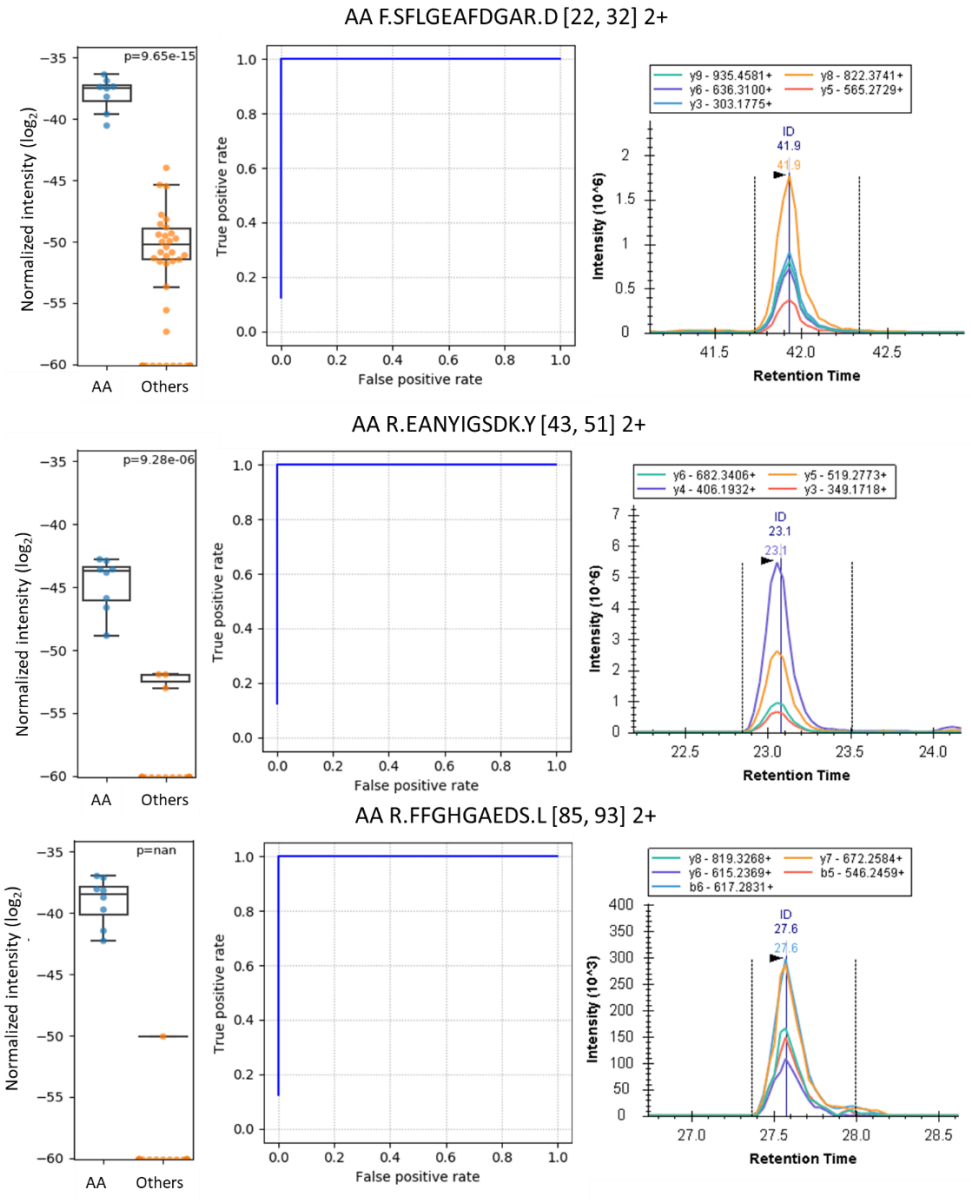

Figure 3-7 The peptide indicators used for AA binary classifiers

The left panel is the intensity distributions of peptides in AA and “other types” populations. The middle panel is the sensitivity and 1-specificity, as intensity cut-off increases, in which sample with peptide intensity above cut-off was assigned to positive for AA. The area under the curve is used to represent discriminatory power of peptide. The right panel is peptide chromatographic peak at MS/MS level extracted from DIA data.

3.3.6 *Evaluation of voting system for amyloidosis typing*

A voting system for amyloidosis typing using DIA data was developed. The model was developed using a training set consisting of 4 AA, 10 AL-kappa, and 10 AL-lambda cases (Figure 3-8 a). The result shows that 6 samples in the training set were either typed differently than the IHC method, or could not be assigned to a specific type. The developed model was further tested with two testing data sets. The first testing data set was composed of 24 samples from 12 biopsy cases, and the second set had 12 samples from 4 autopsy cases (in triplicate). All the samples in the first testing set were typed in agreement with the IHC reported type. One (three replicated samples) of four autopsy cases in the second testing set could not be assigned to a single specific type due to positive results for both AL-kappa and AL-lambda typing; therefore, it was considered as ambiguous typing (Figure 3-8 b)

Table 3-1 Amyloidosis typing performance of developed voting system in training and testing sets

	DIA training set		***DIA testing set 1***		***DIA testing set 2***	
	Sensitivity	1-Specificity	Sensitivity	1-Specificity	Sensitivity	1-Specificity
AA	1.00	0.00	1.00	0.00	1.00	0.00
AL-kappa	0.85	0.00	1.00	0.00	1.00	0.00
AL-lambda	0.85	0.02	1.00	0.00	0.00	0.00

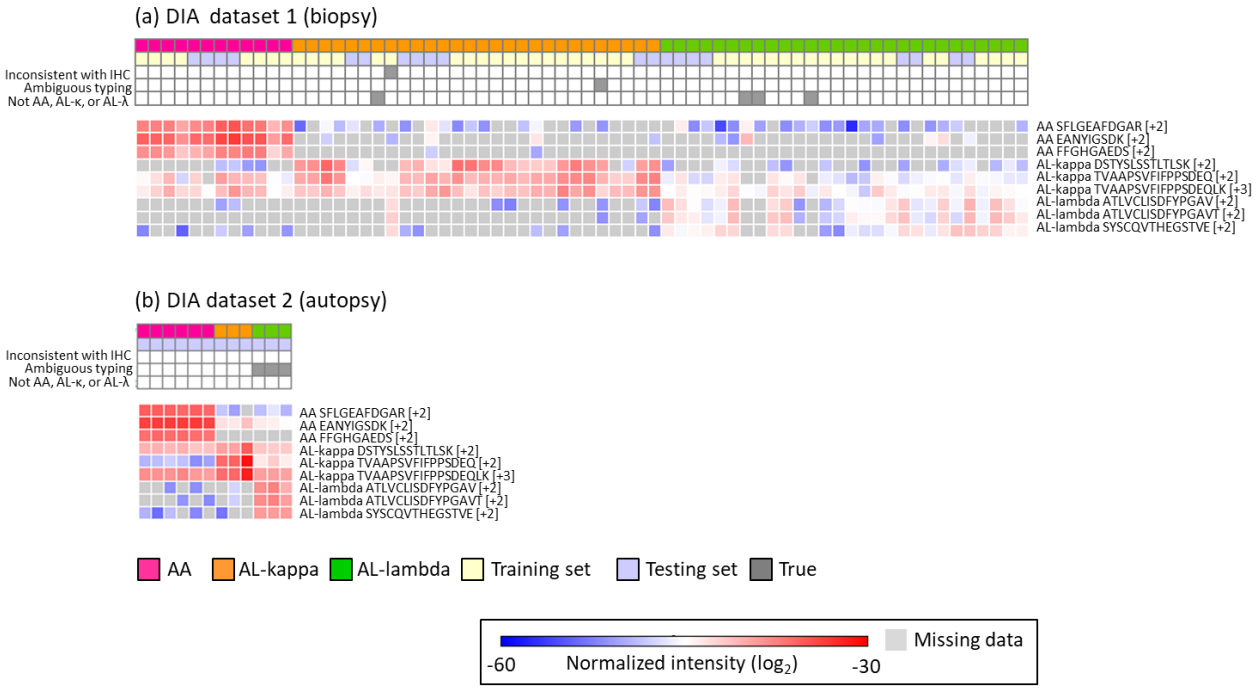

Figure 3-8 Quantification profile with selected peptide indicators.

Each column indicates a sample, rows are selected peptide indicator. The replicates from same sample are adjacent columns. The data matrix are shown in the absolute quantification value of each peptide in samples. The quantification values is normalized integrated peak area of peptide chromatographic peak at MS/MS level extracted from DIA data. The dark grey blocks above data matrix indicate the samples 1) give inconsistent typing result with IHC, 2) give ambiguous typing results, and 3) was not typed as AA, AL-kappa, or AL-lambda. The light purple and light yellow blocks indicate the samples used as training or testing set during typing model development.

3.3.7 *Amyloidosis typing performance in spectral counting method and developed voting system*

The amyloidosis typing performance for all 68 Congo-red positive samples is summarized in Table 3-2, in which samples were typed with the spectral counting method and also the voting system developed above. There was one AL-kappa sample that was typed as AL-lambda by both

mass spectrometry-based typing methods, which leads to an overall AL-lambda typing specificity of 0.98. The typing reproducibility between replicates from the sample biopsy is listed in Table 3-3. Many of the inconsistent typing cases from the DDA spectral counting method were due to insufficient spectral counts in one/both replicates for typing, including 1 AA, 3 AL-kappa, and 8 AL-lambda cases.

Table 3-2 Performance on typing using spectral counting and developed voting system

	Spectral counting (DDA)		***Voting system (DIA)***	
	Sensitivity	1-Specificity	Sensitivity	1-Specificity
AA	0.92	0.00	1.00	0.00
AL-kappa	0.71	0.00	0.93	0.00
AL-lambda	0.36	0.02	0.89	0.02

Table 3-3 Typing reproducibility between replicates from same biopsy

	Reproducibility	
	Spectral counting (DDA)	*Voting system (DIA)*
AA	5/6	6/6
AL-kappa	8/14	12/14
AL-lambda	4/14	12/14

3.3.8 *Precursor m/z range for the DIA experiment*

The *m/z* frequency of the predicted and identified amyloid peptides (from the pilot experiment) are shown in (Figure 3-9 a, b). The percentage of peptide species covered by different 400 m/z regions is shown in Figure 3-9 c. Although the region from *m/z* 400 to 800 covered most of the

identified amyloid peptides, the coverage of predicted amyloid peptides in this region is relative lower than neighboring regions. Considering the coverage of identified and predicted amyloid peptides, the region from m/z 460 to 860 was chosen for analysis in the DIA experiment.

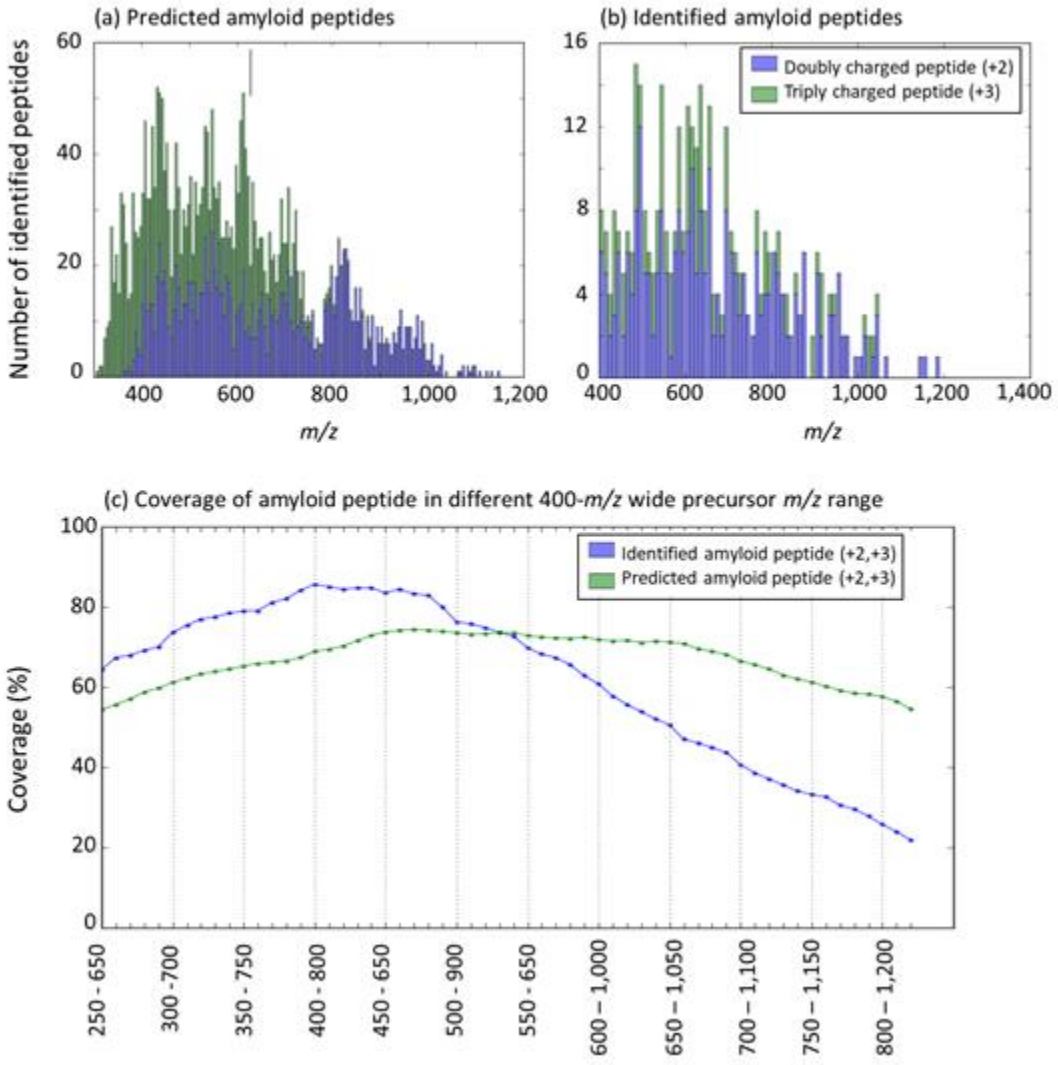

Figure 3-9 amyloid peptide m/z distribution and coverage under different precursor m/z range.

The *m/z* distribution of (a) predicted and (b) identified doubly and triply amyloid peptides. (c) The coverage of amyloid peptide species under different 400-m/z wide precursor m/z range. The m/z range of 460-860 was chosen based on both identified and predicted coverage of amyloid peptides.

3.3.9 *Linearity between peptide concentration and TIC*

The total mass spectral signal from multiply charged peptide-like features is correlated with the amount of peptide loaded on the column. The correlations from three replicates agreed with each other; therefore, I used the total signal from +2, +3, and +4 peptide-like features to normalize the loading amount between the different LMD samples.

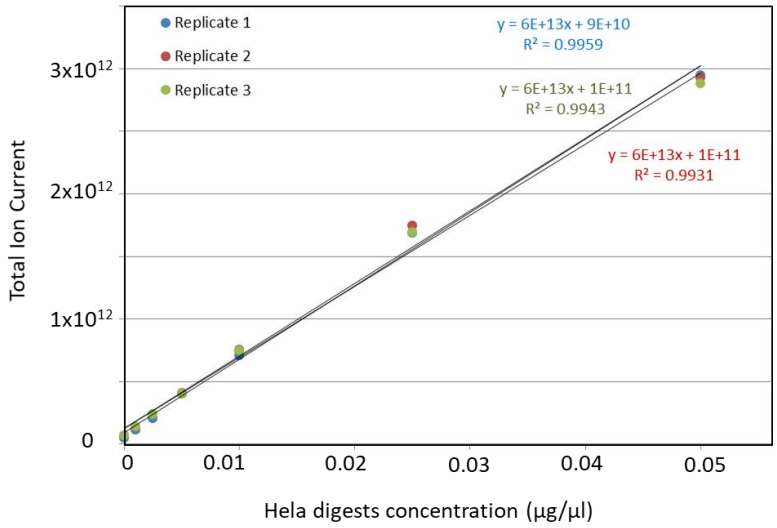

Figure 3-10 Linearity between mass spectral signal from multiply charge peptide-like species and peptide concentration

3.4 DISCUSSION AND CONCLUSION

The recent development of an MS-based amyloidosis typing assay has solved several limitations of the conventional IHC assay and has been successfully applied in the clinic. Unlike other clinical protein assays that use targeted MS methods, the current amyloidosis typing assay uses a shotgun proteomics approach, which has the advantage of finding novel amyloidosis types and also has good multiplexing capacity. However, the stochastic nature of DDA sampling has certain limitations, most notably the quantification inaccuracy of spectral counting and irreproducibility between assays. Furthermore, comparing the MS signals of different proteins against each other might not be the best way for determining amyloidosis types since the difference in MS signal may not reflect the real concentration relationship of proteins in the sample due to several factors. The best example of this is that the observed signal of peptides from a given protein can span many orders of magnitude even they should be equimolar in theory (Figure 3-4). In this study, I developed a voting system by combining multiple probability-based binary classifiers for amyloidosis typing using normalized integrated peak areas extracted from DIA data. Typing was determined based on comparison of the same peptide species across different samples.

To perform sample-wise comparisons, correcting for variability introduced during sample preparation is critical and required. I used two normalization factors to minimize differences in the protein amount of the samples and the trypsin digestion efficiencies between samples. Protein amount of samples is usually estimated with protein assays or approximated by surface area of dissected tissue sections. However, those methods were not suitable in this study, because the protein concentration of LCM samples is too low to be measured by conventional protein assays, and sample losses could be substantial when microdissected tissue pieces flying from tissue

slides to eppendorf tubes. Therefore, I used the total ion current from peptides in each LC-MS run as a surrogate for the total amount of peptide loaded on column. To distinguish signal from peptide and non-peptide molecules in the mass spectral data, I used Bullseye to do peptide feature detection by matching theoretical isotope distributions of peptides with observed signal patterns over time. I only included multiply charged peptide-like features so as to reduce the possibility of including false positive matches to other non-peptide chemicals in the sample. I have shown that the total ion signal of multiply charged peptide-like features is highly correlated with the peptide concentration of the sample, which demonstrates that the total signal of peptide-like feature could be a surrogate for peptide concentration (Figure 3-10). I then used this approach to normalize the difference between samples. To normalize digestion differences between samples, ^{15}N-labeled apolipoprotein A1 was spiked into samples before digestion.

One of the major differences between the method developed here and the current spectral counting based method is that I performed typing using the peptide level quantification instead of the protein level. As shown in the results of this study and previous studies, peptides even from the same protein have different quantitative properties and further not all of them carry equal information for typing. Merging quantification values for all identified peptides into the protein level and ignoring the differences between informative and uninformative peptides will introduce noise to the protein quantification (Figure 3-3 c and 3-4), and lead to a loss in the power to distinguish different amyloidosis types. This situation might be even worse if the signal from uninformative peptides is more intense than the signal from informative peptides in a protein. Therefore, I considered the peptides themselves as markers and found a subset of informative peptides in causative amyloidogenic proteins for typing. Interestingly, many semi-tryptic peptides are detected in amyloidosis samples and many of these semi-tryptic peptides are from

same region of a protein (Figure 3-4). Moreover, based on our selection scheme (Figure 3.5-7) six of the semi-tryptic peptides passed all the selection criteria and were considered better indicators than tryptic peptides for typing. This phenomenon could be due to some artificial factors during the experiment such as LMD process or/and due to the underlying biological understanding that amyloid fibrils are from specific region of proteins.

The voting system was evaluated with two testing datasets consisting of 4 AA, 5 AL-kappa, and 5 AL-lambda cases. Almost all the cases and replicated samples were successfully typed in agreement with their IHC report, except for one AL-lambda case in the second testing set. There were three replicated samples from this AL-lambda sample. All three of the triplicated samples were unable to type as AL-lambda successfully due to significant signal from IGKC protein (Figure 3-8 b). Although all AL-lambda peptide indicators reported very high odds ratios for being the AL-lambda type, all three AL-kappa peptide indicators also gave positive results for AL-kappa typing. Therefore, the positive typing frequency was 1 (3/3) in both typing groups, and the final typing could not be determined. The ambiguous typing result could be due to insufficient sample size leading to inability to fully characterize the relationship between peptide intensity values and the probability of being disease; however, it is also possible that this case does indeed have both abnormal abundant lambda and kappa light chain, a situation which has been reported in a previous study. The performance of the voting system developed here was about 90 % (29 out of 32 samples) on the two testing datasets and 79 % on training set.

The typing performance on 64 Congo-red positive regions using the voting system developed here and the spectral counting based method are 90% and 64%, respectively. The low typing performance of spectral counting was mainly due to low quantitative sensitivity If typing was only performed on samples with sufficient spectral counts, the overall performance would be

83.7% (41 out 49). Nevertheless, the voting system developed here still outperformed the spectral counting method. One concern is that 48 (75%) out of 64 Congo-red positive regions were used to develop the voting system, which may bias the result. However, considering the voting system also gave high typing performance (90%) on the testing datasets, this approach could potentially be a better alternative to the current spectral counting method. More testing datasets are needed for further evaluation.

BIBLIOGRAPHY

1. Yarmush, M. L. & Jayaraman, A. Advances in Proteomic Technologies. *Annu. Rev. Biomed. Eng.* **4,** 349–373 (2002).
2. Yates, J. R. Mass Spectrometry and the Age of the Proteome. *J. Mass Spectrom.* **33,** 1–19 (1998).
3. Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S. & Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci.* **83,** 6233–6237 (1986).
4. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, G. M. Electrospray ionization for mass-spectrometry of large biomolecules. *Science* **246,** 64–71 (1989).
5. Hunt, D. F. *et al.* Characterization of peptides bound to the class I MHC molecule HLA-A2. 1 by mass spectrometry. *Science* **255,** 1261–1263 (1992).
6. Gale, D. C. & Smith, R. D. Small volume and low flow-rate electrospray lonization mass spectrometry of aqueous samples. *Rapid Commun. Mass Spectrom.* **7,** (1993).
7. Emmett, M. R. & Caprioli, R. M. Micro-electrospray mass spectrometry: ultra-high-sensitivity analysis of peptides and proteins. *J. Am. Soc. Mass Spectrom.* **5,** 605–613 (1994).
8. Wilm, M. S. & Mann, M. Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? *Int. J. Mass Spectrom. Ion Process.* **136,** 167–180 (1994).
9. Hunt, D. F., Bone, W. M., Shabanowitz, J., Rhodes, J. & Ballard, J. M. Sequence analysis of oligopeptides by secondary ion/collision activated dissociation mass spectrometry. *Anal. Chem.* **53,** 1704–1706 (1981).
10. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989 (1994).

11. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4,** 207–214 (2007).

12. Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *J. Proteome Res.* **7,** 29–34 (2008).

13. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100,** 9440–9445 (2003).

14. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4,** 923–925 (2007).

15. Fenyö, D. & Beavis, R. C. A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes. *Anal. Chem.* **75,** 768–774 (2003).

16. Cottrell, J. S. & London, U. Probability-based protein identification by searching sequence databases using mass spectrometry data. *electrophoresis* **20,** 3551–3567 (1999).

17. Geer, L. Y. *et al.* Open Mass Spectrometry Search Algorithm. *J. Proteome Res.* **3,** 958–964 (2004).

18. Cox, J. *et al.* Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **10,** 1794–1805 (2011).

19. Stahl, D. C., Swiderek, K. M., Davis, M. T. & Lee, T. D. Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. *J. Am. Soc. Mass Spectrom.* **7,** 532–540 (1996).

20. Michalski, A., Cox, J. & Mann, M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC−MS/MS. *J. Proteome Res.* **10,** 1785–1793 (2011).

21. Hebert, A. S. *et al.* The One Hour Yeast Proteome* DS. *Mol. Cell. Proteomics* **13,** 339

22. Tabb, D. L. *et al.* Repeatability and Reproducibility in Proteomic Identifications by Liquid Chromatography−Tandem Mass Spectrometry. *J. Proteome Res.* **9,** 761–776 (2010).

23. Liu, H., Sadygov, R. G. & Yates, J. R. A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Anal. Chem.* **76,** 4193–4201 (2004).

24. Purvine, S., Eppel*, J.-T., Yi, E. C. & Goodlett, D. R. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *PROTEOMICS* **3,** 847–850 (2003).

25. 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).

26. Chapman, J. D., Goodlett, D. R. & Masselon, C. D. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling: MULTIPLEXED AND DATA-INDEPENDENT TANDEM MS. *Mass Spectrom. Rev.* **33,** 452–470 (2014).

27. Makawita, S. & Diamandis, E. P. The Bottleneck in the Cancer Biomarker Pipeline and Protein Quantification through Mass Spectrometry-Based Approaches: Current Strategies for Candidate Verification. *Clin. Chem.* **56,** 212–222 (2010).

28. Anderson, L. & Hunter, C. L. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell. Proteomics* **5,** 573–588 (2006).

29. 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).

30. Henderson, C. M. *et al.* Measurement by a Novel LC-MS/MS Methodology Reveals Similar Serum Concentrations of Vitamin D-Binding Protein in Blacks and Whites. *Clin. Chem.* **62,** 179–187 (2016).

31. Shi, T. *et al.* Advancing the sensitivity of selected reaction monitoring-based targeted quantitative proteomics. *PROTEOMICS* **12,** 1074–1092 (2012).

32. Shi, T. *et al.* Advances in targeted proteomics and applications to biomedical research. *PROTEOMICS* **16,** 2160–2182 (2016).

33. Gallien, S. *et al.* Targeted Proteomic Quantification on Quadrupole-Orbitrap Mass Spectrometer. *Mol. Cell. Proteomics* **11,** 1709–1723 (2012).

34. Schilling, B. *et al.* Multiplexed, Scheduled, High-Resolution Parallel Reaction Monitoring on a Full Scan QqTOF Instrument with Integrated Data-Dependent and Targeted Mass Spectrometric Workflows. *Anal. Chem.* **87,** 10222–10229 (2015).

35. Meyer, J. G. & Schilling, B. Clinical applications of quantitative proteomics using targeted and untargeted data-independent acquisition techniques. *Expert Rev. Proteomics* **14,** 419–429 (2017).

36. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **389,** 1017–1031 (2007).

37. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **404,** 939–965 (2012).

38. Colinge, J., Chiappe, D., Lagache, S., Moniatte, M. & Bougueleret, L. Differential Proteomics via Probabilistic Peptide Identification Scores. *Anal. Chem.* **77,** 596–606 (2005).

39. Lundgren, D. H., Hwang, S.-I., Wu, L. & Han, D. K. Role of spectral counting in quantitative proteomics. *Expert Rev. Proteomics* **7,** 39–53 (2010).

40. Zybailov, B. *et al.* Statistical Analysis of Membrane Proteome Expression Changes in *Saccharomyces c erevisiae*. *J. Proteome Res.* **5,** 2339–2347 (2006).

41. Griffin, N. M. *et al.* Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat. Biotechnol.* **28,** 83–89 (2010).

42. Ishihama, Y. *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4,** 1265–1272 (2005).

43. Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **25,** 125–131 (2007).

44. Kyle, R. A. Amyloidosis: a convoluted story. *Br. J. Haematol.* **114,** 529–538 (2001).

45. Dogan, A. Amyloidosis: Insights from Proteomics. *Annu. Rev. Pathol. Mech. Dis.* **12,** 277–304 (2017).

46. Dember, L. M. Amyloidosis-Associated Kidney Disease. *J. Am. Soc. Nephrol.* **17,** 3458–3471 (2006).

47. Leung, N., Nasr, S. H. & Sethi, S. How I treat amyloidosis: the importance of accurate diagnosis and amyloid typing. *Blood* **120,** 3206–3213 (2012).

48. Vrana, J. A. *et al.* Classification of amyloidosis by laser microdissection and mass spectrometry-based proteomic analysis in clinical biopsy specimens. *Blood* **114,** 4957–4959 (2009).

49. Theis, J. D., Dasari, S., Vrana, J. A., Kurtin, P. J. & Dogan, A. Shotgun-proteomics-based clinical testing for diagnosis and classification of amyloidosis: Shotgun-proteomics-based clinical testing. *J. Mass Spectrom.* **48,** 1067–1077 (2013).

50. Sipe, J. D. *et al.* Nomenclature 2014: Amyloid fibril proteins and clinical classification of the amyloidosis. *Amyloid* **21,** 221–224 (2014).

51. Muchtar, E., Buadi, F. K., Dispenzieri, A. & Gertz, M. A. Immunoglobulin Light-Chain Amyloidosis: From Basics to New Developments in Diagnosis, Prognosis and Therapy. *Acta Haematol.* **135,** 172–190 (2016).

52. Ramirez-Alvarado, M. Amyloid formation in light chain amyloidosis. *Curr. Top. Med. Chem.* **12,** 2523–2533 (2012).

53. Lavatelli, F. & Vrana, J. A. Proteomic typing of amyloid deposits in systemic amyloidoses. *Amyloid* **18,** 177–182 (2011).

54. Obici, L., Raimondi, S., Lavatelli, F., Bellotti, V. & Merlini, G. Susceptibility to AA amyloidosis in rheumatic diseases: A critical overview. *Arthritis Rheum.* **61,** 1435–1440 (2009).

55. Takase, H., Tanaka, M., Miyagawa, S., Yamada, T. & Mukai, T. Effect of amino acid variations in the central region of human serum amyloid A on the amyloidogenic properties. *Biochem. Biophys. Res. Commun.* **444,** 92–97 (2014).

56. Real de Asua, D. *et al.* Systemic AA amyloidosis: epidemiology, diagnosis, and management. *Clin. Epidemiol.* 369 (2014). doi:10.2147/CLEP.S39981

57. van der Hilst, J. C. H. *et al.* Increased susceptibility of serum amyloid A 1.1 to degradation by MMP-1: potential explanation for higher risk of type AA amyloidosis. *Rheumatology* **47,** 1651–1654 (2008).

58. Rennegarbe, M., Lenter, I., Schierhorn, A., Sawilla, R. & Haupt, C. Influence of C-terminal truncation of murine Serum amyloid A on fibril structure. *Sci. Rep.* **7,** (2017).

59. Egertson, J. D. *et al.* De Novo Correction of Mass Measurement Error in Low Resolution Tandem MS Spectra for Shotgun Proteomics. *J. Am. Soc. Mass Spectrom.* **23,** 2075–2082 (2012).

60. Hoopmann, M. R., Finney, G. L. & MacCoss, M. J. High-Speed Data Reduction, Feature Detection, and MS/MS Spectrum Quality Assessment of Shotgun Proteomics Data Sets Using High-Resolution Mass Spectrometry. *Anal. Chem.* **79,** 5620–5632 (2007).

61. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30,** 918–920 (2012).

62. Hsieh, E. J., Hoopmann, M. R., MacLean, B. & MacCoss, M. J. Comparison of Database Search Strategies for High Precursor Mass Accuracy MS/MS Data. *J. Proteome Res.* **9,** 1138–1143 (2010).

63. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: An open-source MS/MS sequence database search tool. *PROTEOMICS* **13,** 22–24 (2013).

64. Nefedov, A. V., Mitra, I., Brasier, A. R. & Sadygov, R. G. Examining Troughs in the Mass Distribution of All Theoretically Possible Tryptic Peptides. *J. Proteome Res.* **10,** 4150–4157 (2011).

65. Scheltema, R. A. *et al.* The Q Exactive HF, a Benchtop Mass Spectrometer with a Pre-filter, High-performance Quadrupole and an Ultra-high-field Orbitrap Analyzer. *Mol. Cell. Proteomics* **13,** 3698–3708 (2014).

66. Kim, Y. J. *et al.* Quantification of SAA1 and SAA2 in lung cancer plasma using the isotype-specific PRM assays. *PROTEOMICS* **15,** 3116–3125 (2015).

67. Benson, M. D., James, S., Scott, K., Liepnieks, J. J. & Kluve-Beckerman, B. Leukocyte chemotactic factor 2: a novel renal amyloid protein. *Kidney Int.* **74,** 218–222 (2008).

68. Hoofnagle, A. N., Becker, J. O., Wener, M. H. & Heinecke, J. W. Quantification of Thyroglobulin, a Low-Abundance Serum Protein, by Immunoaffinity Peptide Enrichment and Tandem Mass Spectrometry. *Clin. Chem.* **54,** 1796–1804 (2008).

69. Searle, B. C. Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies. *PROTEOMICS* **10,** 1265–1269 (2010).

APPENDIX A

Table A-1

AMYLOIDOSIS SIGNATURE PROTEIN	PROTEIN NAME	PROTEIN ID PATTERN IN UNIPORT
APOLIPOPROTEIN E	Apolipoprotein E	APOE
SERUM AMYLOID P-COMPONENT	Serum amyloid P-component	SAMP
AMYLOIDOSIS TYPE	**PRECURSOR PROTEIN NAME**	**PROTEIN ID PATTERN IN UNITPORT**
AL-KAPPA	Immunoglobulin Light Chain Kappa	IGKC; KV[0-9]+
AL-LAMBDA	Immunoglobulin Light Chain Lambda	LAC[0-9]+; LV[0-9]+; IGLL5
AH	Immunoglobulin Heavy Chain	HV[0-9]+; IGHG1
AA	(Apo) Serum Amyloid A	SAA1;SAA2
ATTR	Transthyretin	TTHY
ABETA2M	b2-Microglobulin	B2MG
AAPOAI	Apolipoprotein A I	APOA1
AAPOAII	Apolipoprotein A II	APOA2
AAPOAIV	Apolipoprotein A IV	APOA4
AGEL	Gelsolin	GELS
ALYS	Lysozyme	LYSC
ALECT2	Leukocyte Chemotactic Factor-2	LECT2
AFIB	Fibrinogen a	FIBA
ACYS	Cystatin C	CYTC
ABRI	ABriPP	ITM2B
ADAN*	ADanPP	ITM2B
ABETA	Ab protein precursor	A4
APRP	Prion protein	PRNP
ACAL	(Pro)calcitonin	CALC
AIAPP	Islet Amyloid Polypeptide	IAPP
AANF	Atrial Natriuretic Factor	ANF
APRO	Prolactin	PRL
AINS	Insulin	INS
ASPC	Lung Surfactant Protein	SFTP [.]+
AGAL7	Galectin 7	LEG7
ACOR	Corneodesmosin	CDSN
AMED	Lactadherin	MFGM
AKER	Kerato-epithelin	BGH3
ALAC	Lactoferrin	TRFL
AOAAP	Odontogenic Ameloblast-Associated Proteins	ODAM
ASEM1	Semenogelin 1	SEMG1
AAPOCII	Apolipoprotein CII	APOC2

	Immunohistochemistry (IHC) or Congo-red staining result	*Amyloid A (AA)*											
	Typing consistent with IHC												
	Typing inconsistent with IHC												
	Ambiguous typing												
	Insufficient spectral counting												
protein id	Protein description	6BB_replicate1	6BB_replicate2	6R_replicate1	6R_replicate2	D_replicate1	D_replicate2	E_replicate1	E_replicate2	MM_replicate1	MM_replicate2	NN_replicate1	NN_replicate2
APOE_HUMAN	Cluster of Apolipoprotein E	**19**	**16**	**21**	**20**	**20**	**2**	**20**	**15**	**19**	**22**	**5**	**8**
SAMP_HUMAN	Cluster of Serum amyloid P-component	**23**	**16**	**17**	**19**	**5**	**1**	**8**	**8**	**15**	**20**	**3**	**6**
APOA4_HUMAN	Cluster of Apolipoprotein A-IV	**1**	**4**			**7**		**1**					
SAA1_HUMAN	Cluster of Serum amyloid A-1 protein	**43**	**46**	**43**	**56**	**45**	**2**	**58**	**12**	**42**	**52**	**16**	**12**
IGKC_HUMAN	Cluster of Ig kappa chain C region	**1**	**4**	**14**	**19**	**7**		**13**	**2**	**17**	**15**		
KV101_HUMAN	Cluster of Ig kappa chain V-I region AG												
KV105_HUMAN	Ig kappa chain V-I region DEE												
KV106_HUMAN	Ig kappa chain V-I region EU												
KV112_HUMAN	Ig kappa chain V-I region Kue												
KV119_HUMAN	Ig kappa chain V-I region Wes												
KV120_HUMAN	Ig kappa chain V-I region Mev												
KV121_HUMAN	Ig kappa chain V-I region Ni												
KV206_HUMAN	Cluster of Ig kappa chain V-II region RPMI 6410			**1**	**2**						**1**		
KV303_HUMAN	Cluster of Ig kappa chain V-III region NG9 (Fragment)			**4**	**5**	**1**		**1**		**1**			
KV306_HUMAN	Ig kappa chain V-III region POM												
KV308_HUMAN	Ig kappa chain V-III region CLL												
KV309_HUMAN	Ig kappa chain V-III region VG (Fragment)				**2**								
KV401_HUMAN	Cluster of Ig kappa chain V-IV region (Fragment)												
IGLL5_HUMAN	Cluster of Immunoglobulin lambda-like polypeptide 5			**1**	**3**	**1**				**2**	**2**		
LAC7_HUMAN	Ig lambda-7 chain C region				**1**								
LV101_HUMAN	Ig lambda chain V-I region VOR										**1**		
LV102_HUMAN	Ig lambda chain V-I region HA												
LV301_HUMAN	Ig lambda chain V-III region SH												
LV302_HUMAN	Ig lambda chain V-III region LOI												
LV403_HUMAN	Ig lambda chain V-IV region Hil												
LV601_HUMAN	Cluster of Ig lambda chain V-VI region AR												
LV605_HUMAN	Cluster of Ig lambda chain V-VI region EB4												
APOA1_HUMAN	Cluster of Apolipoprotein A-I					**3**		**1**			**2**		**1**
APOA2_HUMAN	Apolipoprotein A-II												
BGH3_HUMAN	Cluster of Transforming growth factor-beta-induced protein ig-h3							**2**					
FIBA_HUMAN	Cluster of Fibrinogen alpha chain		**1**	**3**	**1**	**4**				**3**	**2**		**1**
GELS_HUMAN	Cluster of Gelsolin												
IGHA1_HUMAN	Cluster of Ig alpha-1 chain C region			**9**	**14**			**5**		**3**	**2**		
IGHG1_HUMAN	Cluster of Ig gamma-1 chain C region	**1**	**3**	**8**	**17**	**2**		**10**	**3**	**12**	**9**		**1**
IGHG4_HUMAN	Ig gamma-4 chain C region		**1**	**1**	**7**			**1**			**3**		
IGHM_HUMAN	Cluster of Ig mu chain C region		**2**	**12**	**18**	**2**		**6**	**2**	**8**	**4**		
LEG7_HUMAN	Galectin-7												
LYSC_HUMAN	Cluster of Lysozyme C		**1**			**1**		**1**		**1**	**3**		
TRFL_HUMAN	Cluster of Lactotransferrin					**1**		**4**		**4**	**1**		
TTHY_HUMAN	Cluster of Transthyretin							**1**				**1**	

Figure A-1 Spectral counts of identified amyloidosis related proteins. Green color indicates the protein has the most spectral counts within a sample. The orange color indicates the mistyping or ambiguous cases where more than one proteins has the most abundant spectral counts

	Amyloid light chain kappa (AL-kappa)																												
Typing consistent with IHC																													
Typing inconsistent with IHC																													
Ambiguous typing																													
Insufficient spectral counting																													
protein id	AA_replicate1	AA_replicate2	C_replicate1	C_replicate2	G_replicate1	G_replicate2	HH_replicate1	HH_replicate2	JJ_replicate1	JJ_replicate2	K_replicate1	K_replicate2	LL_replicate1	LL_replicate2	OO_replicate1	OO_replicate2	PP_replicate1	PP_replicate2	QQ_replicate1	QQ_replicate2	RR_replicate1	RR_replicate2	S_replicate1	S_replicate2	SS_replicate1	SS_replicate2	T_replicate1	T_replicate2
APOE_HUMAN	8	9	34	20		1	5	8	10	10	12	15	17	9	5	6	12	7	13	18	12	11	17	15	10	13	9	17
SAMP_HUMAN	1	3	5	3			5	11	4	9	1	6	3	1	2	6	4	2	2	3	3	3	6	1	5	4	2	
APOA4_HUMAN			7	4				4			2		2	1	2	5	1		7	11	10	7	9	10	5	1	7	4
SAA1_HUMAN												1						2										
IGKC_HUMAN	5	4	15	19			1		9	4	5	14	11	2	7	6	10	8	13	15	42	20	32	10	5	7	29	27
KV101_HUMAN	5	6	29	6									7	6			1	1			8	8	25	18			15	30
KV105_HUMAN			22	3																	2							
KV106_HUMAN		3															1						1	9				
KV112_HUMAN			2																									
KV119_HUMAN																					1							
KV120_HUMAN			10	3									3										1					
KV121_HUMAN			9														9	5			1							
KV206_HUMAN																					2							
KV303_HUMAN												1			5	8					5							
KV306_HUMAN											1																	
KV308_HUMAN											3	26									1							
KV309_HUMAN																												
KV401_HUMAN																				3								
IGLL5_HUMAN								6							1	1					8					2	1	
LAC7_HUMAN								1													6							
LV101_HUMAN																												
LV102_HUMAN																					1							
LV301_HUMAN																					1							
LV302_HUMAN			1																									
LV403_HUMAN																					1							
LV601_HUMAN								14																				
LV605_HUMAN								5																				
APOA1_HUMAN			2	1				1				2									10		2					
APOA2_HUMAN																					7							
BGH3_HUMAN			1																									
FIBA_HUMAN	1	1	1	2				1	1	1		5					1			2	10	2	1					1
GELS_HUMAN																					1							
IGHA1_HUMAN						1	1								3	1				2	15						1	
IGHG1_HUMAN	2	2	1	1				4	2	2	3	5	2		8	1			1	2	69	7	1	1	5	1	1	
IGHG4_HUMAN																					11				3	2		
IGHM_HUMAN											9	20									5				5	6		
LEG7_HUMAN																												
LYSC_HUMAN			1			1						1			1													
TRFL_HUMAN			1					8																			6	
TTHY_HUMAN			1	1								1	1				2			5	10		1					

Figure A-1 Spectral counts of identified amyloidosis related proteins. Green color indicates the protein has the most spectral counts within a sample. The orange color indicates the mistyping or ambiguous cases where more than one proteins has the most abundant spectral counts

	Amyloid light chain lambda (AL-lambda)																												
Typing consistent with IHC																													
Typing inconsistent with IHC																													
Ambiguous typing																													
Insufficient spectral counting																													
protein id	B_replicate1	B_replicate2	BB_replicate1	BB_replicate2	EE_replicate1	EE_replicate2	F_replicate1	F_replicate2	GG_replicate1	GG_replicate2	H_replicate1	H_replicate2	II_replicate1	II_replicate2	KK_replicate1	KK_replicate2	U_replicate1	U_replicate2	V_replicate1	V_replicate1	W_replicate1	W_replicate2	X_replicate1	X_replicate2	Y_replicate1	Y_replicate2	Z_replicate1	Z_replicate2
APOE_HUMAN	14	15	15	15	13	12			16	9	3				4	2	4	7	5	4		8	11	12	18	29	19	15
SAMP_HUMAN	8	4		3	9	9		1	13	16	1		1	5	1	1		8	4	3	2	2	5	5	3	4	3	6
APOA4_HUMAN	6	1		3	2	7			9	7			1	2	2	3			5	1		2		7	2	1	5	3
SAA1_HUMAN		1																										
IGKC_HUMAN		1			2	1			2	3	1						1	1		1								
KV101_HUMAN																												
KV105_HUMAN																												
KV106_HUMAN																												
KV112_HUMAN																												
KV119_HUMAN																												
KV120_HUMAN																												
KV121_HUMAN																												
KV206_HUMAN																												
KV303_HUMAN						1																						
KV306_HUMAN																												
KV308_HUMAN																												
KV309_HUMAN																												
KV401_HUMAN																												
IGLL5_HUMAN	2		1	3	6	10			5	6	1							1	7	3	1	4	1	12	7	12	5	8
LAC7_HUMAN				1	2	3			1	3	1								3	2			4	5	5	4		2
LV101_HUMAN								1					13	8														2
LV102_HUMAN				2										1														
LV301_HUMAN																												
LV302_HUMAN																												
LV403_HUMAN																												
LV601_HUMAN			1						18	14			1	1					1	4								
LV605_HUMAN									6	4								1	22	16								
APOA1_HUMAN		1			1	1			1	1	1				1			1	1									1
APOA2_HUMAN																												
BGH3_HUMAN					2	10																						
FIBA_HUMAN					1	5				1							14	1				1			5	2	1	2
GELS_HUMAN						1			1		1																	
IGHA1_HUMAN	10	6			1	2					1							3										
IGHG1_HUMAN					2	2			6	3	5		3	1			2				1	1	1	1	2	1	6	8
IGHG4_HUMAN											1																	
IGHM_HUMAN																												
LEG7_HUMAN																	2											
LYSC_HUMAN		1			1	1			1	1							1							1				
TRFL_HUMAN									9	13																		
TTHY_HUMAN						4												2						2		3		1

Figure A-1 Spectral counts of identified amyloidosis related proteins. Green color indicates the protein has the most spectral counts within a sample. The orange color indicates the mistyping or ambiguous cases where more than one proteins has the most abundant spectral counts

	Congo-red negative (Control)															
Typing consistent with IHC																
Typing inconsistent with IHC																
Ambiguous typing																
Insufficient spectral counting																
protein id	6J _replicate1	6J _replicate2	6K_replicate1	6K_replicate2	6L_replicate1	6L_replicate2	6N_replicate1	6N_replicate2	6O_replicate1	6O_replicate2	6Q_replicate1	6Q_replicate2	CC_replicate1	CC_replicate2	DD_replicate1	DD_replicate2
APOE_HUMAN										**1**						
SAMP_HUMAN																
APOA4_HUMAN				**1**										**1**		
SAA1_HUMAN																
IGKC_HUMAN			**4**	**3**	**1**	**1**	**1**	**1**	**1**					**1**		
KV101_HUMAN																
KV105_HUMAN																
KV106_HUMAN																
KV112_HUMAN																
KV119_HUMAN																
KV120_HUMAN																
KV121_HUMAN																
KV206_HUMAN																
KV303_HUMAN				**2**										**1**		
KV306_HUMAN																
KV308_HUMAN																
KV309_HUMAN																
KV401_HUMAN																
IGLL5_HUMAN			**1**													
LAC7_HUMAN																
LV101_HUMAN																
LV102_HUMAN																
LV301_HUMAN																
LV302_HUMAN																
LV403_HUMAN																
LV601_HUMAN																
LV605_HUMAN																
APOA1_HUMAN																
APOA2_HUMAN																
BGH3_HUMAN																
FIBA_HUMAN					**1**											
GELS_HUMAN																
IGHA1_HUMAN			**2**	**1**												
IGHG1_HUMAN	**2**		**5**	**5**	**2**	**2**	**2**	**3**			**1**		**1**	**1**		
IGHG4_HUMAN						**1**		**1**					**1**			
IGHM_HUMAN																
LEG7_HUMAN																
LYSC_HUMAN									**2**					**1**		
TRFL_HUMAN																
TTHY_HUMAN																

Figure A-1 Spectral counts of identified amyloidosis related proteins. Green color indicates the protein has the most spectral counts within a sample. The orange color indicates the mistyping or ambiguous cases where more than one proteins has the most abundant spectral counts

VITA

Han-Yin Yang was born in a small island nation, Taiwan, on the Pacific Ocean. She earned a Bachelor of Science degree from National Taiwan Normal University in 2006, majoring in Life Science with certificate in bioinformatics. Driven by her enthusiasm to learn bioinformatics, she earned Master of Science in bioinformatics under the mentorship of Dr. Wailap Victor Ng in National Yang-Ming University, Taiwan in 2008. During Han-Yin's master training, she found mass spectrometry-based proteomics interesting and challenging. After receiving Master degree, she worked as research assistant in Dr. Ting-Yi Sung's lab at Institute of Information Science, Academia Sinica where she worked closely with computational scientist and mass spectrometry chemist to develop analysis tool for proteomics data. To fulfill her curiosity in mass spectrometry, she moved to Seattle and joined Dr. Michael J. MacCoss's lab for graduate training in 2012. After immersing herself in coffee, beer, and mass spectrometers for more than five years, she earned Ph.D. degree from Department of Genome Sciences at University of Washington in 2018.