Shifting the Paradigm: Peptide-Centric Analysis of Systematically Sampled Mass Spectrometry Data

Ying Sonia Ting

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

Doctor of Philosophy

University of Washington
2016

Reading Committee:
Michael J. MacCoss, Chair
William Stafford Noble
Judit Villén

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Shifting the Paradigm: Peptide-Centric Analysis of Systematically Sampled Mass Spectrometry Data

Ying Sonia Ting

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

In mass spectrometry-based shotgun proteomics, data-independent acquisition (DIA) is an emerging technique due to its systematic and unbiased sampling of precursor ions. However, current DIA methods often use wide precursor isolation windows, resulting in co-fragmentation and complex mixture spectra, posing various challenges in DIA data analysis.

In this dissertation, I aim to shift the paradigm of data analysis in DIA tandem mass spectrometry with peptide-centric analysis. I first describe the analytical advantages of peptide-centric analysis over the conventional spectrum-centric analysis for DIA data analysis. Next, I describe a new tool I developed, named PECAN, that enables peptide-centric analysis for robust peptide detection for DIA data without a

prerequisite library. I demonstrate rigorous validation of PECAN performance, including validation with synthesized analytical standards using an *in vitro* translation system. I further discuss the impact on the selectivity of DIA data to peptide detection.

All DIA methods balance between acquisition parameters, such as precursor selectivity, fragmentation method, ion injection time/dwell time, and resolving power of MS/MS analysis. I discuss using PECAN detection as a quantitative metric to characterize the effects of these acquisition parameters in the context of DIA analysis. Along with two other applications, I demonstrate the versatility of PECAN and peptide-centric analysis of DIA data.

Finally, I apply peptide-centric analysis to characterize the murine heart proteome with SS-31 treatment. Here, I first show improvement in cardiac function of SS-31-treated old mice. Next, I interrogate the mitochondrial and global proteome changes impacted by the SS-31 treatment in the murine heart proteome with DIA MS/MS and PECAN. I show the proteome dynamics of several pathways hypothesized to be involved with the SS-31 mechanisms, shedding light on the underlying mechanisms of the SS-31 cardiac protective effects.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am deeply grateful to the many individuals who have offered me guidance, criticisms, opportunities, and support during my academic career.

I first thank my Ph.D. supervisor, Michael J. MacCoss, for being a fantastic mentor to me. Mike's passion in science and technology is truly inspiring. Chatting with Mike about new and exciting projects were some of my favorite moments in grad school. In addition, he is always approachable, supportive, and understanding when things take longer than expected. I truly appreciate the genuine care he has shown for his students, not only about their academic achievement but also their success in life.

I also thank the remaining members of my committee, William Stafford Noble, Joshua Akey, Judit Villén, and Peter Rabinovitch, for their guidance during my Ph.D. training. I am very grateful to Bill for co-mentoring me during the development of the work described in Chapter 2 and 3, and for frequently providing constructive criticism and suggestions for my work. Josh steered me away from unrealistic expectations of myself and my projects, and challenged me to think from a different perspective multiple times. I am also grateful for the encouragements and insightful discussions from Judit which helped me through my frustration at times. Peter supported and generously shared his expertise regarding the work described in Chapter 5.

It was my great privilege to carry out my doctoral research in the Department of Genome Sciences at the University of Washington, where I was surrounded by a fantastic group of faculty and trainees. I thank the department for creating such a positive and collaborative environment and thank all the talented and hardworking trainees for giving me valuable interdisciplinary perspectives and keeping me on my toes. I thank the faculty

members for their feedback in research reports and journal clubs, which in turn greatly improved my skills in presentation and communication. I thank the amazing GSIT and administrative staffs for providing reliable support that keeps the department running. I would also like to acknowledge the 2011 GS class, particularly Elyse Hope, Rachel Gittelman, Stephanie Battle, and Alex Mason, for their amazing friendship and emotional support through this journey.

I benefited enormously from interactions with members of the MacCoss lab. I am now competent at operating the instruments thanks to Michael Bereman, Jarrett Egertson, and Richard Johnson. Gennifer Merrihew taught me many wet-lab skills, including how to not do unnecessary experiments that your advisor casually brought up. I also thank my collaborators, particularly Ying Ann Chiao, Jimmy Eng, and Romain Huguet, for generously sharing their time and expertise.

I would not be here without these three important professors in my early academic career. Dr. Hsueh-I Lu's class inspired my interest in algorithms. Victor Ng introduced me to the fascinating world of proteomics. David Goodlett gave me an opportunity and made me believe that I have what it takes to be a real scientist. I am very lucky and grateful to have encountered them as great mentors and teachers.

Finally, I would like to thank my family. My mother, to whom this work is dedicated, always believed that I can achieve what I set out for. She made me a strong and determined woman and gave me the strength I needed to complete this journey. I thank my husband, Jarrett Egertson, for his patience, understanding for many late nights, and unconditional love and support.

DEDICATION

This work is dedicated to my mother, Pei-Shan Chen (1957-2011),

the person who taught me to love, to live, and to never give up.

Chapter 1

INTRODUCTION

In the post-genomic era, proteomics has emerged to become an indispensable domain of biological, clinical, and pharmaceutical research. Proteomics is the study of the protein complement of a biological system. The goal is to gain insights into the cellular functions and biological mechanisms at the molecular level. Unlike genomics, proteomics is challenged by the extremely dynamic nature of protein expression, degradation, and post-translational modification. For instance, the dynamic range of human plasma protein expression is larger than 10 orders of magnitude. Understanding the protein composition and relative abundance of the plasma proteome is essential for developing a plasma assay or a biomarker. To deliver such information, many high-throughput approaches using mass spectrometry have been developed, among which I will focus on "shotgun proteomics".

In this chapter, I briefly review the paradigm of shotgun proteomics, describe the technologies used for data acquisition, and discuss the challenges in data analysis. I further detail the objectives and organization of this dissertation.

Some content of this chapter has been adapted with changes from: Ting, Y.S., Egertson, J.D., Payne, S.H., Kim, S., MacLean, B., Käll, L., Aebersold, R., Smith, R.D., Noble, W.S., and MacCoss, M.J. (2015). Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Mol Cell Proteomics* 14, 2301–2307.

1.1 Mass spectrometry-based shotgun proteomics

Mass spectrometry (MS)-based shotgun proteomics is a powerful and versatile tool in modern life sciences. Shotgun proteomics, also known as bottom-up proteomics, refers to the characterization of proteins by analysis of peptides, short chains of amino acid monomers linked by peptide bounds. The name is derived from the analogy to shotgun sequencing of DNA, named after the quasi-random firing pattern of a shotgun. In shotgun proteomics, a mixture of proteins is first digested into a mixture of peptides, most commonly with proteases that exhibit strong specificity of proteolytic cleavages, such as trypsin and lysyl-endopeptidase. The resulting mixture of proteolytic peptides is typically separated by liquid chromatography (LC), ionized with electrospray ionization, and then analyzed by tandem mass spectrometry. During tandem mass spectrometry, MS analysis surveys the intact precursor ions, and MS/MS analysis characterizes the product ions generated from isolation and fragmentation of the selected precursor ions (Figure 1.1 a). Any one of several modes of data acquisition may be used to select precursor ions for MS/MS analysis.

Compared to intact protein analysis, the physical and chemical properties of peptides circumvent the poor solubility, modest ionization efficiency, separation efficiency, and the limitation in the size of analyte detectable in mass spectrometers. The downside is that much information of proteoforms, including protein isoforms and post-translation modifications (PTMs), is degenerated along with the denaturation of proteins. Nonetheless, shotgun proteomics allows global protein detection as well as the ability to systematically profile dynamic proteomes (Wu and MacCoss, 2002).

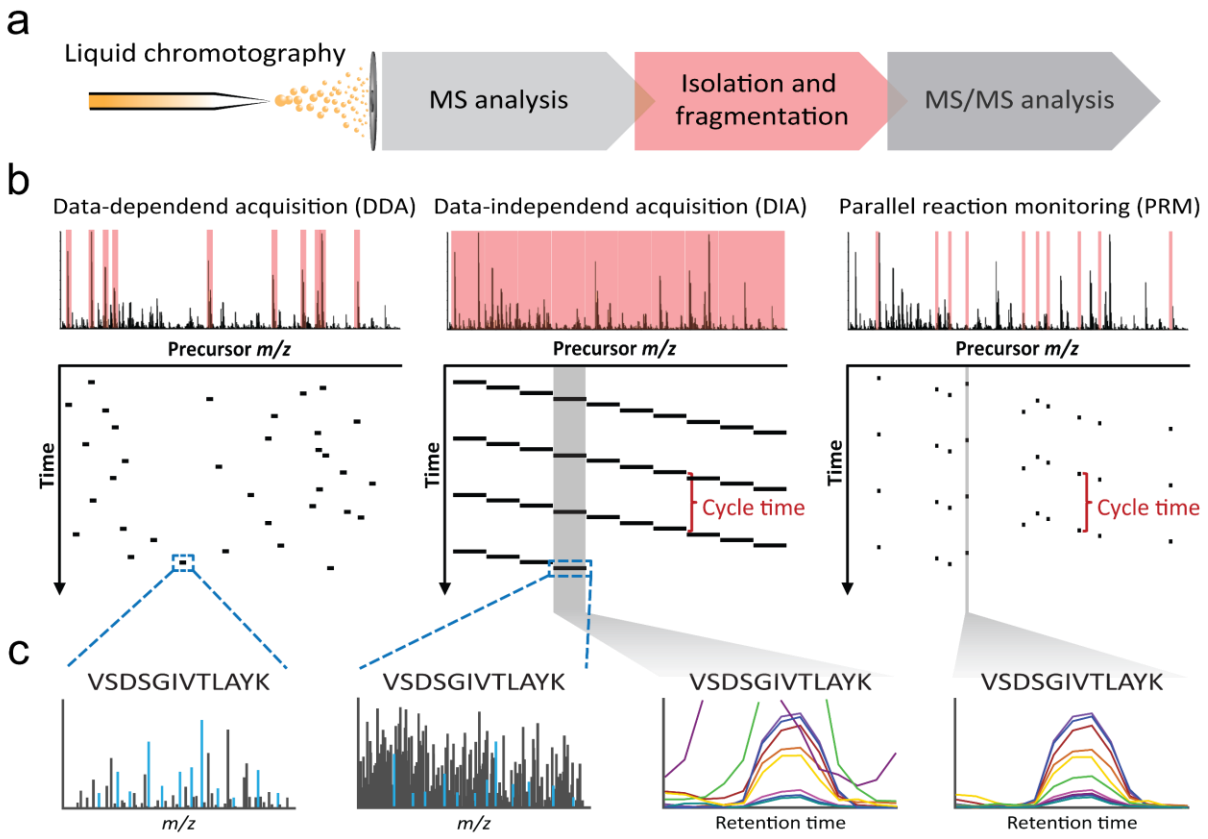

Figure 1.1 Mass spectrometry-based shotgun proteomics

(a) In shotgun-proteomics, peptides are typically separated by a liquid chromatography, and then ionized and introduced to tandem mass spectrometry for MS and MS/MS analyses. (b) The precursor regions sampled for MS/MS analysis over time in data-dependent acquisition, data-independent acquisition, and parallel reaction monitoring (PRM). Red shades represent the sampling of 10 precursor regions with a given MS spectrum. Cycle time for DIA and PRM indicates the time instrument takes to cycle through the predetermined precursor regions. (c) The left two panels show the peptide spectrum matches (PSMs) of a peptide to a MS/MS spectrum from DDA and DIA. The right two panels show the extracted ion chromatograms of a peptide from DIA or PRM data.

1.2 Acquisition methods for tandem mass spectrometry

1.2.1 *Data-dependent acquisition*

The most commonly applied mode uses data-dependent acquisition (DDA), in which MS/MS spectra are acquired from the dissociation of precursor ions selected from an MS survey spectrum (Figure 1.1). Constrained by the speed of instrumentation, DDA can sample only a subset of precursor ions for MS/MS characterization, generally targeting the top-*N* most abundant ions detected in the most recent survey spectrum. In addition, DDA is typically coupled with a method referred to as "dynamic exclusion" (Gatlin et al., 2000) that attempts to prevent reselection of the same *m/z* for some specified period of time. These acquisition strategies greatly increase proteome coverage and extend the dynamic range of detection for shotgun proteomics. DDA is a powerful and well-established technique for LC-MS/MS data acquisition. By targeting precursor ions observed in MS survey scans with highly selective MS/MS scans, DDA generates a large set of high quality MS/MS spectra which can be automatically interpreted by database searching to identify thousands of proteins in a complex sample. When DDA was introduced, instrumentation was not fast enough to sample every observed precursor in the survey scan; thus, high-intensity precursors were preferentially targeted because they tend to generate higher quality MS/MS spectra that lead to peptide identification. While this sampling approach results in a large amount of peptide identification in a single sample run, it comes at the cost of reproducibility of MS/MS acquisition between sample analyses and an inherent bias against low abundance analytes that are less likely to be sampled (Michalski et al., 2011). On modern instrumentation, the speed of MS/MS

acquisition has dramatically improved to the point where the majority of MS precursors that are not already in the dynamic exclusion list can be sampled for MS/MS analysis. However, even if every precursor observed in each survey MS scan is sampled, DDA is still biased against low abundance analytes that fall below the limit of detection in the MS analysis and will never be sampled. This bias is a practical limitation in the analysis of a complex mixture with high dynamic range where many analytes will be below the limit of detection in MS analysis, but remain detectable by the more selective and more sensitive MS/MS analysis (Panchaud et al., 2009).

1.2.2 *Targeted acquisition*

DDA remains a powerful method for identifying a large number of proteins in a sample. However, due to the incomplete sampling, when a peptide is not identified in a conventional shotgun experiment using DDA, it is incorrect to conclude that the peptide is missing from the sample, or even below the limit of detection of MS/MS, because the peptide ions may have never been sampled for MS/MS analysis. To overcome such limitations, targeted acquisition approaches such as selected reaction monitoring (SRM) and parallel reaction monitoring (PRM) are often the methods of choice. In targeted acquisition approaches, a set of predetermined precursor ions are systematically subjected to MS/MS characterization throughout the LC time domain (Figure 1.1b). In general, the collision energy for each targeted ion can be optimized for fragmentation efficiency. With systematic MS/MS sampling and the combined specificity of chromatographic retention time, precursor ion mass, and the distribution of product ions, targeted acquisition allows highly sensitive and reproducible detection of the targeted

analytes within a complex mixture.

Modern targeted acquisition approaches are the gold standard for sensitively and reproducibly measuring hundreds of peptides in a single LC-MS/MS run (Escher et al., 2012; Marx, 2013; Burgess et al., 2014). However, data acquired in this manner is only informative for the set of peptides targeted for analysis. Due to this narrow focus, iterative testing of different hypotheses (i.e. a different set of target peptides) also requires iterative acquisition of additional data. Moreover, assay development often requires retention time scheduling and/or refinement steps to find the optimal peptides and transitions, pairs of precursor-product ions, for testing a particular hypothesis.

1.2.3 *Data-independent acquisition*

With the existence of two complementary but distinct approaches – DDA for broad sample characterization and targeted acquisition for interrogation of a specific hypothesis – the natural question is if the benefits of both techniques may be combined in a single technique. A potential solution is an alternative mode of bottom-up proteomics referred to as data-independent acquisition (DIA) that has been described and realized with various implementations (Panchaud et al., 2009; Chapman et al., 2014; Gillet et al., 2012; Weisbrod et al., 2012; Purvine et al., 2003; Venable et al., 2004; Silva et al., 2006; Plumb et al., 2006; Bern et al., 2010; Carvalho et al., 2010; Egertson et al., 2013). In DIA, the instrument acquires MS/MS spectra systematically and independently from the content of MS survey spectra (Figure 1.1). These DIA approaches differ from DDA methods, targeted acquisition methods, and from each other in MS/MS isolation window width, total range of precursor *m/z* covered, duration of completing one cycle of isolation scheme

(called the *cycle time*), single or multiple isolation windows per MS/MS analysis, and instrument platform. Due to the benefits of systematic sampling of the precursor *m/z* range by MS/MS, data from a single DIA experiment can be useful for both peptide detection and quantification in a complex mixture. Similar to DDA approaches, DIA data is broadly informative because the MS/MS characterization is not specific to a pre-defined set of peptide targets. Similar to targeted approaches, MS/MS information about peptides across the entire LC time domain can be extracted from DIA data to test a particular hypothesis. As the acquisition speed of modern instrumentation continues to increase, DIA has become more popular because of its comprehensive and unbiased sampling.

Although DIA circumvents the problem of biased or incomplete MS/MS sampling, current DIA methods come with compromises (Egertson et al., 2015), where the most common compromise is precursor selectivity. Constrained by the speed and accuracy of instrumentation, DIA methods typically use five- to ten-fold wider isolation windows compared to DDA to achieve the breadth and depth of a single LC-MS/MS run. Because of this reduction in precursor selectivity, MS/MS spectra from DIA are noisier than DDA spectra. In particular, DIA by design generates mixture spectra, each containing product ions from multiple analytes with various abundance and different charge states. Fragmenting multiple analytes together also precludes DIA from tailoring collision energy for every analyte, a standard optimization in DDA and targeted acquisition.

1.3 Challenges in tandem mass spectrometry data analysis

1.3.1 *General challenges*

It is essential to remember that in shotgun proteomics, the mass spectrometer measures peptides and peptide fragments, not proteins. Using the detected peptides to infer which proteins were present in the original biological sample is a major challenge in shotgun proteomics, largely because of the degenerate nature of peptides, of which one peptide could be derived from multiple proteins or protein isoforms (Nesvizhskii and Aebersold, 2005). Guided by the parsimony principle, protein inference approaches often create a minimal list of proteins that can explain the detected peptides. However, the mere presence of parsimonious proteins is far from reality where a proteome is a composite of proteofroms resulting from all sources of combinational variation including splicing variants, single nucleotide polymorphisms, and post-translational modifications (Smith et al., 2013a).

To further complicate the problem, most data analysis approaches for shotgun proteomics rely heavily on protein sequence databases, which commonly contain only canonical sequences and not sequence variants. Thus, many peptides have never been detected because they were never interrogated. This challenge could be addressed by customizing the protein database with exome sequencing. Approaches like this that use genomic techniques to improve proteomics, or vice versa, have become an emerging field called proteogenomics (Jaffe et al., 2004; Nesvizhskii, 2014).

Another major challenge is peptide positional isomers, of which multiple peptides have the same amino acid composition but differ in the sequences and/or the location of

modifications. Tandem mass spectrometry data does not always provide enough information to differentiate positional isomers. Depending on the position in the peptide, positional isomers could share a majority of fragment ions from a given fragmentation method. A common example for differentiating positional isomers is the localization of phosphorylation sites, with scores designed for this purpose, such as the A-score (Beausoleil et al., 2006). Varying fragmentation methods is sometimes required to differentiate positional isomers (Aguiar et al., 2010; Swaney et al., 2009).

1.3.2 *DDA data analysis*

For DDA data, database searching is the dominant, high-throughput approach for peptide identification from MS/MS spectra. SEQUEST is the first and one of the most widely used database searching approaches (Eng et al., 1994, 2008). SEQUST correlates MS/MS spectra of peptides with amino acid sequences in a protein database. Database searching is designed to identify DDA spectra, where the precursor ion for a spectrum is known and used as a filter for candidate peptides in the database. For the past two decades, the proteomics community has developed many algorithms with various scoring schemes for database searching, including Mascot, X!Tandem, MaxQuant, Comet, MS-GF+, and OMSSA (Cox and Mann, 2008; Craig and Beavis, 2004; Eng et al., 2013; Geer et al., 2004; Kim and Pevzner, 2014; Koenig et al., 2008). Other approaches for analyzing DDA data include *de novo* sequencing (Frank et al., 2007; Ma et al., 2003; Taylor and Johnson, 1997), and searching against a spectrum library (Frewen et al., 2006; Lam et al., 2008; Yen et al., 2011). Regardless of the approach used, the match between a MS/MS spectrum to a candidate peptide is called a peptide-spectrum match (PSM) (Figure 1.1 c).

Because most tools report best matches for every spectrum, PSMs from a dataset are typically a mixture of correct and incorrect matches. These PSMs can be subjected to a statistical model / classifier, such as PeptideProphet (Keller et al., 2002), to estimate the accuracy of peptide identifications, or Percolator (Käll et al., 2007), to distinguish between correct and incorrect matches and estimate false discovery rates (FDRs) using the target-decoy strategy (Elias and Gygi, 2007).

Most database searching approaches aim to identify one peptide in a spectrum. However, it is estimated that the majority of DDA spectra are generated from more than one analyte (Houel et al., 2010; Hsieh et al., 2010). The major challenges in interpreting these mixture spectra lie in allowing for multiple contributing precursor ions, assessing the dynamic range of mixture peptides, distributing intensities of product ions shared by contributing peptides, and adjusting statistical confidence estimates. Some sophisticated approaches, such as ProbIDtree, MixDB, MixGF, and ChimeraCounter (Zhang et al., 2005; Houel et al., 2010; Wang et al., 2011, 2014; Zhang et al., 2014), address these challenges by deconvolving mixture spectra into pseudo spectra or by matching mixture spectra to combinations of product ions from multiple candidate peptides. However, identification of low abundance analytes from mixture spectra is inherently difficult because the MS/MS signals from low abundance analytes are naturally overwhelmed by the signals from high abundance ones.

The major challenge for DDA data analysis results from the stochastic nature of DDA sampling. For a peptide that is not identified in a DDA experiment, it is incorrect to conclude that this peptide is not detectable from the sample because the peptide ions may have never been sampled for MS/MS analysis. In addition, because of the stochastic

sampling pattern, DDA data show low reproducibility between two acquisitions. Thus, the quantitative matrix, such as the relative abundance of peptides or proteins, derived from DDA data for between sample comparisons could contain up to 40% missing data. Many statistical approaches have been developed to impute missing values in peptide quantitative matrix for DDA data; however, strategies to guide users in the selection of the suitable imputation for their dataset and analysis objectives are much needed (Webb-Robertson et al., 2015).

1.3.3 *Targeted acquisition data analysis*

Data from targeted acquisition are typically analyzed using "targeted data analysis" (MacLean et al., 2010; Picotti and Aebersold, 2012; Prakash et al., 2009). In targeted analysis, the tools typically look for the co-eluting patterns from a group of predetermined pairs of precursor-product ions (called *transitions*) or extracted ion chromatograms, created by plotting the intensity of the signal observed at a chosen *m/z* value or set of values in a series of mass spectra recorded as a function of retention time (Murray et al., 2013). Following, a series of signal processing such as peak boundaries detection and peak background subtraction are performed. Eventually, the quantity of a detected peptide is represented by the area under the curve of the intensities from a handful of transitions.

Chromatographic peak picking and peak boundaries determination are major challenges in targeted data analysis. Typically, to confirm a peptide detection, measurements of 3–5 intense, co-eluting transitions and, if present, heavy isotope–labeled standards for each endogenous peptide are required. In a complex sample,

analytes with precursor and product ion masses similar to that of a target peptide can result in ambiguity of chromatographic peak picking and interference for peak boundaries determination, which can lead to false positive identifications or imprecise quantification (Picotti and Aebersold, 2012). While PRM data can be analyzed by database searching, this approach circumvents the challenge of chromatographic peak picking but not peak boundaries determination (Peterson et al., 2012). Although there is no need for extensive transition-assay optimization prior to acquiring PRM data, one still needs to select 3–5 co-eluting transitions that are not interfered by the fragmentation of other co-isolated analytes for quantification, posing a different challenge in targeted data analysis.

The peptide detections from targeted acquisition data typically require manual validation, a process that is labor-intensive and error-prone. Several bioinformatics solutions have incorporate some empirical criteria and scores to improve this process (MacLean et al., 2010; Prakash et al., 2009). In addition, mProphet combines the empirical criteria and scores into a statistical model that is designed to estimate false positives in SRM peptide detection (Reiter et al., 2011).

1.3.4 *DIA data analysis*

The low precursor selectivity and resulting complexity of DIA spectra severely challenges the performance of traditional database searching which generally assumes that the detected product ions were derived from a single, isolated precursor. Because almost every spectrum is mixed in DIA data, such data is poorly suited for analysis by classic database searching approaches initially designed for DDA data. Approaches deigned to demultiplex mixture spectra from DDA data often forgo the systematically

sampled nature of DIA and thus are also poorly suited for analyzing DIA data.

In recent years, new strategies have been development to enable effective analysis of DIA data. One example is OpenSWATH (Röst et al., 2014), which leverages the fragmentation patterns and relative retention times from a library of previously identified MS/MS spectra to detect peptides from DIA data. Another tool is MSPLIT-DIA (Wang et al., 2015) that iteratively extracts pseudo spectra from DIA data by matching library spectra to DIA spectra. These library-based approaches successfully facilitate DIA data analysis, making the most of the rich knowledge accumulated from previous studies. However, while library-based approaches are sensitive, relying on libraries limits the data interrogation to analytes that have been observed and identified previously. Such limitation confines the discovery potential for DIA data. Thus, tools designed to detect peptides from DIA data without libraries are necessary.

These library-free tools can be broken into two categories. The first category of library-free tools includes DIA-Umpire (Tsou et al., 2015) and Group-DIA (Li et al., 2015), which generate pseudo spectra from DIA data with methods like detecting covarying precursor-product ion groups or deconvolving the multiplexed spectra. Naturally, the quality of each pseudo spectrum is often determined by the quality and interpretability of the precursor signal in MS analysis. These pseudo spectra are then sent to conventional database searching pipelines designed for DDA identification, where precursor signal is a key filtering criterion for candidate peptides. As a result, pseudo spectra with poor precursor signal are less successful in yielding confident identifications. Such precursor dependency in database searching, a legacy from analyzing DDA data, hinders the detection of analytes with detectable product signal but highly interfered or not detectable

precursor signal in DIA data (Panchaud et al., 2009). The lack of detectable precursor signal for some detectable analytes is a common phenomenon resulting from intra-scan dynamic range limitation. This happens when the dynamic range of a mass analyzer is exceeded by the dynamic range of analytes in a single MS analysis, but not exceeded by the dynamic range of product ions from the selected analytes in a single MS/MS analysis. The phenomenon is commonly observed when acquiring DIA data from complex samples separated with hour-long chromatography.

The second category of library-free approaches looks for the best supporting evidence of detection from the data for each query peptide. FT-ARM (Weisbrod et al., 2012) was the first demonstration of a library-free peptide-centric approach for DIA data. The method works by querying simple theoretical spectra of peptides against high mass accuracy (<5 ppm) DIA data using a dot product score function. As a proof of principle, the algorithm from FT-ARM was novel and straightforward, but with much to improve, specifically in its sensitivity and false discovery rate.

Another major challenge for DIA data analysis is for peptide quantification. In theory, DIA provides MS/MS-based quantification for all detectable peptides. In reality, the high interference resulting from low precursor selectivity makes peak boundary determination and transition selection much more difficult, even when the chromatographic peaks center has been assigned. Despite the challenges in data analysis, DIA remains attractive not only because of its systematic measurements of the proteome, but also because DIA enables subsequent reanalysis of the data with different hypotheses (Gillet et al., 2012) – effectively providing a permanent digital record of the content of a sample.

1.4 Objective of this dissertation

In this dissertation, I aim to shift the paradigm of data analysis in DIA tandem mass spectrometry with peptide-centric analysis. I also describe a new tool, PECAN, that enables peptide-centric analysis for robust peptide detection from DIA data.

In Chapter 2, I describe peptide-centric analysis and spectrum-centric analysis that generalize the data analysis approaches for tandem mass spectrometry data. I describe my perspective of the inherited analytical advantages of peptide-centric analysis for DIA data, specifically in handling the challenges of mixture spectra and precursor information (Ting et al., 2015). I further discuss the extensibility of the peptide-centric framework to analyte-centric framework for general mass spectrometry data analysis.

In Chapter 3, I describe a new peptide-centric tool "PECAN" (PEptide Centric Analysis) that detects peptides directly from DIA data without prerequisite spectral or retention time libraries. I demonstrate several rigorous assessments of key parameters in the PECAN algorithm, followed by the performance validation of PECAN including accuracies in chromatographic peak picking and peptide detection. I further show the comparison of PECAN to a spectrum-centric DIA data analysis workflow.

In Chapter 4, I demonstrate three applications of PECAN. The first application uses PECAN as a quantitative metric for evaluating DIA methods varied with several acquisition parameters. The second application uses PECAN to build a chromatogram library from human plasma. The third application uses PECAN to query sequence variants against DIA data. These applications demonstrate the unique features of PECAN and also reveal some limitation of DIA data.

In Chapter 5, I apply peptide-centric analysis to characterize the murine heart

proteome with SS-31 treatment. I first show improvement in cardiac function of SS-31-treated old mice. Next, I interrogate the mitochondrial and global proteome changes impacted by the SS-31 treatment in the murine heart proteome with DIA MS/MS and PECAN. I show the proteome dynamics of several pathways hypothesized to be involved with the SS-31 mechanisms, shedding light on the underlying mechanisms of the SS-31 cardiac protective effects.

Chapter 2

PEPTIDE-CENTRIC PROTEOME ANALYSIS

This chapter has been adapted with changes from: Ting, Y.S., Egertson, J.D., Payne, S.H., Kim, S., MacLean, B., Käll, L., Aebersold, R., Smith, R.D., Noble, W.S., and MacCoss, M.J. (2015). Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Mol Cell Proteomics* 14, 2301–2307.

Abstract

In mass spectrometry-based bottom-up proteomics, data-independent acquisition (DIA) is an emerging technique due to its comprehensive and unbiased sampling of precursor ions. However, current DIA methods use wide precursor isolation windows, resulting in co-fragmentation and complex mixture spectra. Thus, conventional database searching tools that identify peptides by interpreting individual MS/MS spectra are inherently limited in analyzing DIA data. Here we discuss an alternative approach, peptide-centric analysis, which tests directly for the presence and absence of query peptides. We discuss how peptide-centric analysis circumvents some limitations of traditional spectrum-centric analysis, and we outline the unique characteristics of peptide-centric analysis in general.

2.1 Introduction

As discussed in Chapter 1, several modes of data acquisition have been developed for bottom-up proteomics. The most commonly applied mode uses data-dependent acquisition (DDA), in which MS/MS spectra are acquired from the dissociation of precursor ions selected from an MS survey spectrum. DDA greatly increase proteome coverage and extend the dynamic range of detection for shotgun proteomics. The resulting MS/MS spectra are typically analyzed using sequence database searching software such as SEQUEST, Mascot, X!Tandem, MaxQuant, Comet, MS-GF+, or OMSSA (Cox and Mann, 2008; Craig and Beavis, 2004; Eng et al., 2008, 2013; Geer et al., 2004; Kim and Pevzner, 2014; Koenig et al., 2008). Because these algorithms identify peptides by first associating each individual spectrum with a matching peptide sequence and then aggregating the thus matched spectra into a list of identified peptides, we refer to them as "**spectrum-centric analyses**". In spectrum-centric analysis, spectra are most commonly interpreted using database searching, but can also be interpreted using *de novo* sequencing (Frank et al., 2007; Ma et al., 2003; Taylor and Johnson, 1997), or by searching against a spectrum library (Frewen et al., 2006; Lam et al., 2008; Yen et al., 2011). For the past two decades, spectrum-centric analysis has been an essential driving force for the development of large-scale shotgun proteomics using DDA.

Another emerging data acquisition approach is data-independent acquisition (DIA), where the instrument acquires MS/MS spectra systematically and independently from the content of MS survey spectra. DIA circumvents the problem of biased and

Table 2.1 Spectrum-centric analysis vs. peptide-centric analysis

	Spectrum-centric analysis	Peptide-centric analysis
Query unit	MS/MS spectrum	Peptide
Assumption	Each spectrum is generated from at least one peptide	Each peptide elutes once (for a short period of time) during liquid chromatography
Goal	Identify peptide(s) from each spectrum	Find evidence of detection for each peptide
Scoring	Candidate peptides from the sequence database compete for the best scoring PSM	Candidate spectra from the acquired data compete for the best scoring evidence of detection
Example tools	SEQUEST, Comet, MASCOT, X!Tandem, OMSSA, ProbID, MS-GF+, MaxQuant, DIA-Umpire, GroupDIA	SALSA, FT-ARM, OpenSWATH, Skyline

stochastic MS/MS sampling in DDA; however, DIA methods often compromise precursor selectivity, making peptide detection from DIA data much more challenging (Egertson et al., 2015).

Recently, Gillet *et al.* demonstrated an alternative approach that analyzes DIA data in a targeted fashion (Gillet et al., 2012), opening a new door for the investigation of tandem mass spectrometry data. Much like targeted analysis of transitions used in targeted acquisition methods, Gillet *et al.* use extracted ion chromatograms (XICs) to detect and quantify query peptides. Similarly, Weisbrod *et al.* identify peptides by searching peptide fragmentation patterns against DIA data (Weisbrod et al., 2012).

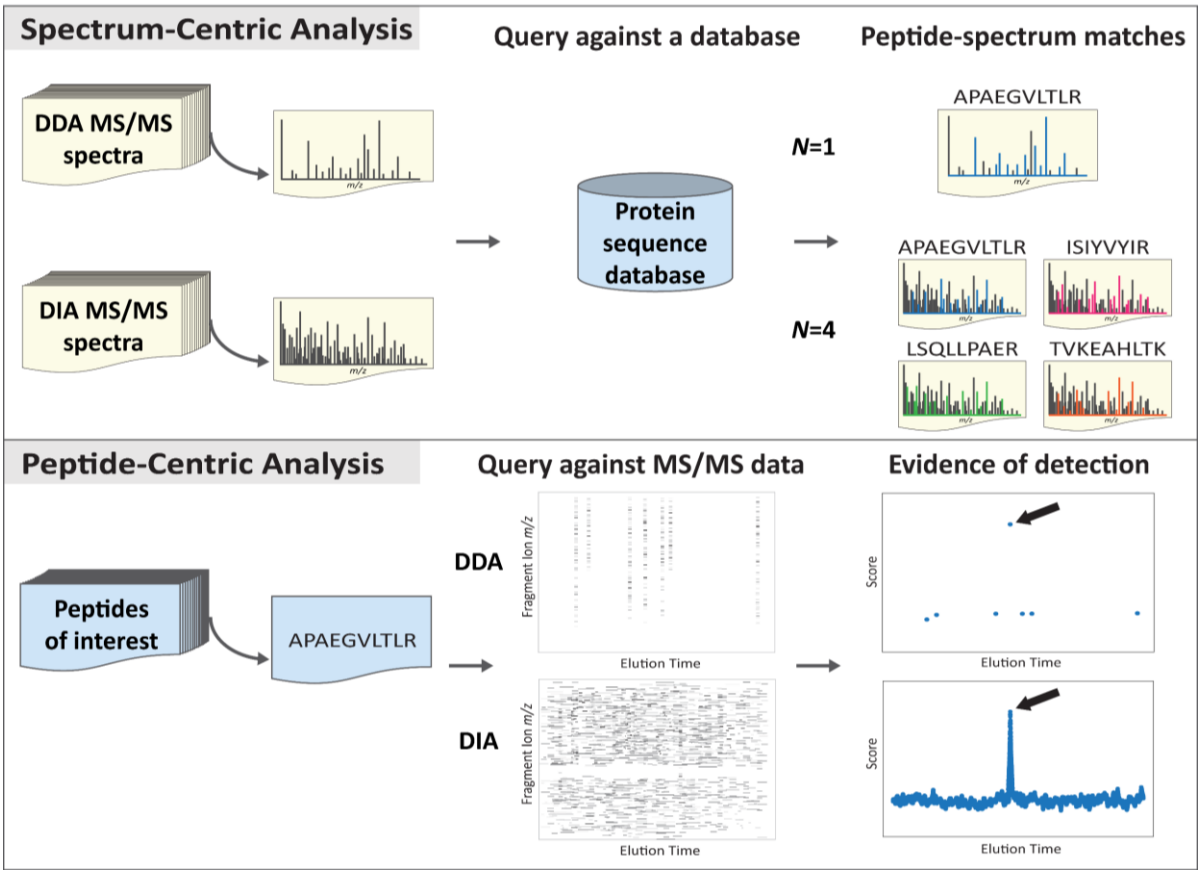

Figure 2.1 Spectrum-centric analysis and peptide-centric analysis

In spectrum-centric analysis, each MS/MS spectrum from either a DDA or DIA experiment is queried against a protein sequence database. The peptides that yield the best scoring N statistically significant PSMs are assigned to the corresponding MS/MS spectrum. Typically, *N* is one for a DDA spectrum and multiple for a DIA spectrum (showing *N*=4). In peptide-centric analysis, every peptide of interest is queried against the acquired MS/MS data. The bottom-middle panel shows the extracted MS/MS signal of the query peptide over time in which the signal is extracted from any MS/MS spectrum generated from isolating the query precursor *m/z*. The extraction window width corresponds to the acquisition method, showing here 2 *m/z* for DDA and 10 *m/z* for DIA. The precursor *m/z* of the query peptide is sampled stochastically and sparsely in DDA but systematically in DIA. The MS/MS signal that provides the best scoring evidence of detection is assigned to the query peptide (indicated by the arrows).

Instead of interpreting individual spectra in a spectrum-centric fashion, these alternative approaches take each peptide of interest and ask: "Is this peptide detected in the data?" We refer to this approach as "**peptide-centric analysis**" in contrast with "**spectrum-centric analysis**." In peptide-centric analysis, each peptide is detected by searching the MS and MS/MS data for signals selective for the query peptide. Peptide-centric analysis covers all methods that use peptides as an independent query unit, including but not limited to the targeted analysis. Peptide-centric analysis is intrinsically very different from spectrum-centric analysis (Table 2.1, Figure 2.1) and better suited for addressing many biological problems. This perspective discusses the analytical advantages of peptide-centric analysis and how they could translate to improvements in protein inference, and the analysis of DIA data.

2.2 Unique characteristics of peptide-centric analysis

2.2.1 *Direct statistical measurements of query peptides*

A drawback of spectrum-centric analysis is that the confidence estimates for peptides are indirect. In spectrum-centric analysis, each MS/MS spectrum is first assigned at least one peptide identity, yielding a large set of peptide-spectrum matches (PSMs). These PSMs are classified into accepted or not accepted by methods (Elias and Gygi, 2007; Käll et al., 2007; Keller et al., 2002; Tabb et al., 2002) that assign to each PSM statistical confidence estimates, indicating the confidence of either a set of PSMs being correct (e.g. FDR) or an individual PSM being correct (e.g. p-values and E-values). Subsequently, peptide-level confidence estimates can be assigned by aggregating the best

PSM per peptide in a post-processing step (Käll et al., 2007; Shteynberg et al., 2011). Because the query unit for spectrum-centric analysis is an MS/MS spectrum, only the peptides that are matched to at least one spectrum are subject to the peptide level statistical tests. As a result, only this subset of peptides is assigned statistical confidence estimates, and the remaining peptides are implicitly considered missing.

Peptide-centric analysis, on the other hand, tests every peptide queried, providing direct and complete statistical measurements. The goal of peptide-centric analysis is to ascertain whether a query peptide was detected in an experiment. Thus, in a given data set, all of the query peptides can be separated into those with or without evidence of detection (i.e. detected or not detected). An empirical null can be estimated by generating decoy query peptides with shuffled sequences, measuring the null score distribution, and calculating p-values and q-values for every query peptide using common statistical methods (Käll et al., 2007; Keller et al., 2002; Reiter et al., 2011). With peptide-centric analysis, direct peptide-level testing makes answering biological questions more straightforward, and the completeness of statistical measurements makes subsequent comparison and quantification much easier.

Peptide-centric analysis could be very useful when considering the protein inference problem, which involves estimating the set of detected proteins from the set of detected peptides (Nesvizhskii and Aebersold, 2005). Protein inference is heavily affected by the observed peptides. The value of peptide-centric analysis is that each peptide in a database can be directly assigned a confidence estimate of being detected/not detected because each peptide is directly investigated. In contrast, spectrum-centric analysis implicitly assigns all “missing” peptides equal, very low

confidence estimates. These imputed confidence estimates could lead to biases in the inferred set of detected proteins. This includes peptides that distinguish splice isoforms or paralogs. Therefore, when comparing the result from a peptide-centric analysis to the detectability of such a peptide (Li et al., 2010; Mallick et al., 2007), it is possible to begin to probabilistically evaluate the presence/absence of a protein isoform. With directly tested peptide probabilities, peptide-centric analysis makes the input of protein inference more straightforward and transparent.

2.2.2 *Considerations for mixture spectra*

When investigating a complex proteome with shotgun proteomics, mixture spectra are a common occurrence. Although conventional DDA uses narrow isolation windows (typically ~2 *m/z*-wide) targeting single precursor ion species for fragmentation, as many as 50% of the MS/MS spectra are mixed (Houel et al., 2010; Zhang et al., 2014; Luethy et al., 2008). The frequency and impact of mixture spectra in a DDA experiment vary with the sample complexity, LC separation, acquisition parameters, and instrumentation. Some studies used isolation windows as narrow as 0.7 *m/z*-wide to minimize unwanted precursor ions from being co-isolated and co-fragmented (Hebert et al., 2014; Michalski et al., 2011). In the context of DIA, all spectra are essentially mixture spectra because DIA isolates and fragments all precursor ions within a wide *m/z* range. As discussed previously, identification of multiple components in a mixture spectrum is challenging: most spectrum-centric software is designed to identify a single component from each spectrum.

Peptide-centric analysis excels in handling mixture spectra because it does not

interpret individual spectra. Rather than deconvolving each individual spectrum, peptide-centric analysis searches for evidence of detection for individual peptides, explicitly tolerating co-fragmentation. While spectrum- centric analysis struggles to identify multiple components with wide dynamic range from each mixture spectrum, peptide-centric analysis queries each peptide independently from other peptides (Table 2.1). This subtle but significant change of query unit shifts the problem from "peptides competing with each other to explain the mixture spectrum" to "spectra competing with each other to represent the query peptide". With peptide-centric analysis, a single spectrum can be the top-scoring evidence of detection for multiple distinct peptides, as expected in the case of mixture spectra. In addition, peptide-centric analysis readily benefits from the systematic sampling of DIA when each analyte is sampled multiple time across its chromatographic peak. Conversely, even if the product ions of the query peptide comprise the minority of the mixture spectra, peptide detection can still be achieved using peptide-centric analysis.

2.2.3 *Roles of precursor ion signals*

Precursor information is a powerful component of MS/MS data analysis. Inherently designed to identify DDA spectra, spectrum-centric approaches typically use precursor information as a "filter" to constrain peptide candidates for PSMs (Cox and Mann, 2008; Craig and Beavis, 2004; Eng et al., 2008, 2013; Geer et al., 2004; Kim and Pevzner, 2014; Koenig et al., 2008). These approaches assign precursor ion(s) to each spectrum in various ways spanning from using the un-processed precursor ion target, considering multiple monoisotopic ions in the isolation window, to detecting peptide

features in the MS space. With high mass measurement accuracy and high resolution instruments, spectrum-centric searches could allow for only ±10 ppm of monoisotopic mass tolerance, thus greatly reducing the number of peptide candidates for PSMs and reducing the false discovery rate.

In the context of analyzing DIA data there is no clear consensus on how to use precursor information. Recent DIA methods emphasize the systematic measurement of both precursor and product ions, allowing for the detection of precursor and product ions that co-vary over elution time and likely are derived from the same analyte (Purvine et al., 2003). This concept of detecting co-varying precursor-product ion groups has been used to generate deconvolved spectra from DIA spectra. Each deconvolved spectrum contains precursor and product ions ostensibly derived from a single analyte and are thus more compatible with spectrum-centric analysis (Li et al., 2009; Tsou et al., 2015).

Peptide-centric approaches could also use precursor information as evidence of detection. Rardin *et al.* recently demonstrated improved quantification from DIA data using Skyline with precursor ion filtering and transition filtering by correlation analysis (MacLean et al., 2010; Rardin et al., 2015). Although filtering with precursor ions and precursor-product groups improves selectivity and specificity, the detection process could reduce sensitivity because analytes may have no MS signal, or an MS signal with substantial chemical noise despite having an MS/MS signal amenable for quantification. One way to incorporate precursor information without reducing sensitivity is to use it as a scoring feature rather than a filter, which is employed in some peptide-centric approaches such as the algorithms used in Skyline (MacLean et al., 2010). When analyzing complex mixtures, incorporating precursor information without filtering may

provide greater confidence in peptide detection for analytes with a signal in MS spectra without compromising sensitivity by eliminating analytes which may have an MS/MS signal but not detectable MS signal.

2.3 Applications of peptide-centric analysis

Peptide-centric analysis is particularly suited for DIA experiments given its advantages in handling mixture spectra. In addition, peptide-centric analysis can easily incorporate valuable properties from DIA data, such as retention time and elution profile, that are commonly ignored by spectrum-centric analysis. For example, Gillet *et al.* demonstrated peptide detection and quantification by extracting peptide-specific product ion chromatograms, or XICs, from DIA data using 26-*m/z* SWATH acquisition (Gillet et al., 2012). Weisbrod *et al.* demonstrated peptide detection and quantification by searching theoretical or empirical peptide fragmentation patterns against the DIA data acquired using high mass accuracy Fourier transform-all reaction monitoring (FT-ARM) of 100- *m/z* wide isolation windows (Weisbrod et al., 2012). With low precursor selectivity and high intra-scan dynamic range in both cases, correctly interpreting the spectra using spectrum-centric analysis is extremely challenging.

Peptide-centric analysis can also be applied to DDA data. For example, Liebler *et al.* used a pattern recognition algorithm (SALSA) to search for peptide-specific ion series against the DDA MS/MS spectra (Liebler et al., 2002). Due to the stochastic nature of the DDA data, the evidence for peptide detection appears sparse and scattered compared to analyzing DIA data (Figure 1). Nonetheless, peptide-centric

analysis provides statistical measure for every query peptide regardless of whether the data is sparse or dense. In addition, given that many DDA spectra are mixed, peptide-centric analysis retains the benefits of handling of mixture spectra when analyzing DDA data.

2.4 Extensible framework for mass spectrometry

This concept of defining peptides as analytes and directly searching for their evidence of detection generalizes into a broader paradigm, which we call "**analyte-centric analysis**." Analyte-centric analysis comprises any method that uses the analyte as the query unit to ask whether the analyte is detected or not. It includes the traditional targeted data analysis, but is not limited to the methods that scores based on transitions or XICs. The analyte of interest can be naturally extended from peptides to include small molecules, peptides with modifications, intact proteins, lipids and metabolites. In this analyte-centric paradigm, any properties of an analyte can be naturally incorporated into the score that summarizes the evidence supporting an analyte being detected. For example, "Does the discovered fragmentation evidence coincide with chromatographic expectations?" Also, as mass spectrometer resolution continues to improve toward fine-scale isotope resolution (Rose et al., 2013), the analyte-centric approach can discriminate an isotopic profile based on the elemental composition of the analyte.

One of the subtle but significant benefits of analyte-centric analysis is the change in the query unit and null hypothesis. In the spectrum-centric approach, validation

programs that modeled a false distribution of decoy hits were in reality posing the null hypothesis as, “This spectrum is made up of a random analyte.” For analyte-centric analysis, the null hypothesis is, “The analyte is not detected in the data.” This more direct hypothesis is better suited for answering most biological problems.

2.5 Conclusions

In this perspective, we discuss the analytically unique characteristics of peptide-centric analysis compared to traditional spectrum-centric analysis in analyzing shotgun proteomics data. Specifically, peptide-centric analysis provides direct statistical measurements for every peptide, and could improve the analysis of mixture spectra common in DIA data. We also discussed how peptide-centric approaches could use precursor signals as essential or supporting evidence of detection. As mass spectrometry instruments continue to improve in acquisition speed, DDA will be able to sample deeper for lower abundance analytes and DIA will be able to systematically acquire MS/MS spectra with improved precursor selectivity or a shorter cycle time. Analysis of the resulting large collections of data could benefit from the alternative peptide-centric approaches. Specifically, changing the perspective from identifying as many spectra as possible to confidently detecting peptides from an experiment greatly benefits protein inference and quantitative comparison. The fact that the same peptide is fragmented in DIA datasets multiple times generating a chromatographic elution profile for each product ion further increases the achievable quantitative accuracy. Furthermore, a peptide-centric perspective can be naturally extended to other analytes such as intact

proteins, lipids, and metabolites. We hope the analytical advantages of analyte-centric analysis over spectrum-centric analysis will incite the field to further advance bioinformatics and statistical solutions for analyzing mass spectrometry data.

Chapter 3

PECAN

In Chapter 2, I have discussed the inherent analytical advantages of peptide-centric approaches over traditional spectrum-centric approaches when analyzing DIA data. Accordingly, I developed a new peptide-centric algorithm called PECAN (PEptide Centric Analysis) that detects peptides directly from DIA data without prerequisite spectral or retention time libraries. PECAN is a library-free, peptide-centric tool for peptide detection from data-independent acquisition (DIA) tandem mass spectrometry data. In this chapter, I first illustrate the core algorithm of PECAN and demonstrate the assessment of its key parameters. I further validate PECAN detection and evaluate its performance with various datasets. Finally, I discuss the impact of precursor selectivity on PECAN and compare the performance with DIA-Umpire workflow, a library-free, spectrum-centric approach for peptide identification from DIA data. Together, I demonstrate that PECAN detects peptides robustly and accurately from DIA data without using a library.

3.1 Introduction

As illustrated in Chapter 1, data independent acquisition (DIA) is an emerging technique in shotgun proteomics which systematically selects a mixture of precursor ions

for MS/MS analysis in an unbiased fashion. To achieve unbiased sampling while maintaining a comprehensive measure of the proteome, most DIA methods use wide isolation windows that sacrifice precursor selectivity in favor of comprehensiveness. Such trade-off makes the detection of peptides from DIA data very challenging because each DIA MS/MS spectrum contains product ions from multiple analytes.

PECAN offers three primary advances relative to existing approaches for peptide detection from DIA data. First, PECAN scoring weights a theoretical fragment ion based on its specificity to the query peptide relative to the rest of the proteome. This strategy boosts the score contribution of peptide specific fragment ions even when they are low intensity ions. Second, PECAN incorporates a novel background score subtraction to correct for the scoring bias caused by uneven distribution of the proteome in retention time and precursor space. PECAN background scores represent how high on average a score can be achieved by chance at a given point of time. Subtracting background scores significantly reduces random matches in the peptide “hot zones”, where a large number of peptides with similar precursor ions are introduced to the instrument simultaneously. Last, PECAN scoring is primarily based on fragment ions. In PECAN, precursor information is used as an auxiliary feature, independent from the determination of evidence of detection. By not depending on the observation of precursor ions, PECAN takes full advantages of the better sensitivity in MS/MS analysis resulting from better selectivity compared to MS analysis. Here, we present in detail the PECAN algorithm, validation, and performance evaluated with various data sets validation, and performance evaluated with various data sets.

3.2 The workflow and algorithm of PECAN

PECAN takes as input centroided DIA data, a list of query peptides that contains the sequences of proteolytic peptides of interest and the names of the proteins that could generate the peptide, and a background proteome database that typically is the species protein sequence database (Figure 3.1). PECAN outputs 19 auxiliary scores describing the assigned evidence of detection with an associated retention time for every query (target) peptide and PECAN generated decoy peptide. These evidence of detection and associated auxiliary scores are used by Percolator (Käll et al., 2007) to estimate false discovery rate and report confident peptide and protein detection.

PECAN uses the open source application programming interface pymzML (Bald et al., 2012) and supports the HUPO Proteomics Standard Initiative standard file format—mzML (Martens et al., 2011). PECAN search results can be imported into Skyline (MacLean et al., 2010), an open source platform for mass spectrometry data visualization, quantification, interactive analyses, and report generation.

The PECAN workflow comprises four steps: generate peptide vectors, subtract background scores, report evidence of detection, and estimate detection FDRs. Here, we first describe PECAN's primary score function, decoy generation, and then each of the four steps.

3.2.1 *PECAN primary scoring*

PECAN uses matrix multiplication to score each peptide relative to its fragment extracted ion chromatograms (XICs), created by plotting the intensity of the signal observed at a chosen *m/z* value or set of values in a series of mass spectra recorded as a

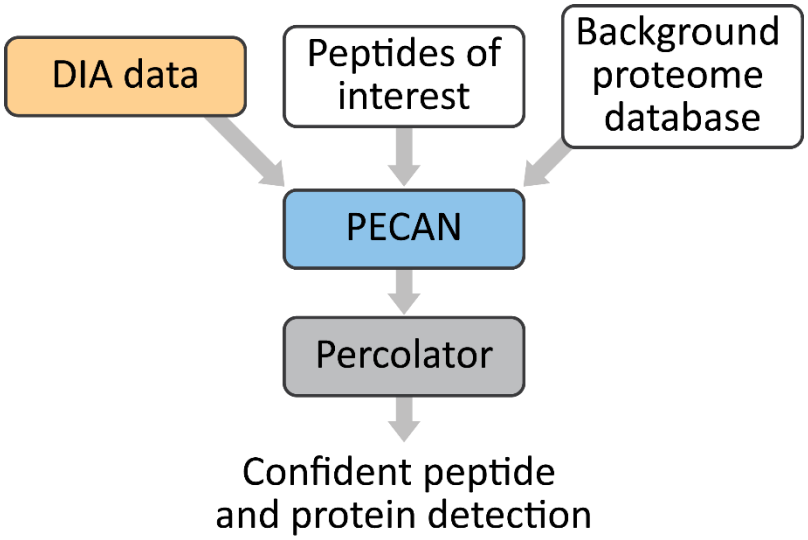

Figure 3.1 Overview of PECAN workflow.

PECAN takes DIA data, peptides of interest, and a background proteome database as inputs, and outputs evidence of detection with auxiliary scores for every query peptide and PECAN generated decoy peptide. Percolator uses PECAN output to train a classifier to distinguish correct and incorrect matches, and then outputs confident peptide and protein detection with estimated FDR.

function of retention time (Murray et al., 2013). For a DIA dataset where m MS/MS spectra are generated from the isolation window that contains the precursor ion of peptide p, the fragment XICs of peptide p can be represented as

$$XIC_p = \begin{bmatrix} I_{b_1,t_1} & I_{b_1,t_2} & \cdots & I_{b_1,t_m} \\ \vdots & \vdots & & \vdots \\ I_{b_{n-1},t_1} & I_{b_{n-1},t_2} & \cdots & I_{b_{n-1},t_m} \\ I_{y_1,t_1} & I_{y_1,t_2} & \cdots & I_{y_1,t_m} \\ \vdots & \vdots & & \vdots \\ I_{y_{n-1},t_1} & I_{y_{n-1},t_2} & \cdots & I_{y_{n-1},t_m} \end{bmatrix}$$

where $I_{x,t}$ is the extracted intensity of an expected fragment ion with m/z value x at

retention time t. The extracted intensity is the sum of the square root of the intensities of ions with m/z values within the extraction mass error tolerance (default ±10 ppm) of x. Let the peptide vector (see 3.2.2 below) corresponding to peptide p be V_p. Then the peptide score matrix is calculated as

$$S_p = V_p \cdot XIC_p = [s_{t_1}, s_{t_2} \dots, s_{t_m}]$$

where each s_t is mathematically equivalent to the scalar projection of O_t, the observed MS/MS spectrum at retention time t, onto the peptide scoring vector V_p. Because the scalar value s_t represents the magnitude of the spectrum at retention time t supporting a peptide with V_p, the vector S_p represents the evidence of detection for peptide p over time.

3.2.2 *Decoy generation*

PECAN uses two types of decoys: one for background score estimation and the other for FDR estimation. Decoy peptides in PECAN are generated by random shuffling a reference proteolytic peptide while keeping the proteolytic site (e.g. C-terminal R and K for trypsin digestion). In all cases, a decoy is invalid if it is present in either the list of query (target) peptides or the background proteome.

For background score estimation, the background proteome is used as a reference to seed for decoy generation (described in 3.3.1). A new decoy will be generated with the same reference until either a valid decoy has been generated from the reference or three attempts has been made. In case of no valid decoy after three attempts, PECAN will shuffle the reference sequence without maintaining the proteolytic site.

For use of target-decoy paradigm, the list of query (target) peptides is used as a

reference to seed for decoy generation. Here, a decoy is further validated by the fragment similarity to its reference. Upon generation, a decoy is invalid if it shares more than 40 % of the theoretical fragment ion *m/z* values with its reference. A new decoy will be generated with the same reference until either a valid decoy has been generated from the reference or ten attempts has been made. In case of no valid decoy after ten attempts, the decoy with the least shared theoretical fragment ion *m/z* values will be used.

3.2.3 *Generate peptide vectors*

For each query peptide, PECAN generates a normalized scoring vector called a "peptide vector". A peptide vector is a unit vector that represents the theoretical fragmentation pattern of the peptide. For a peptide p with n amino acids, let $p = [b_2, \dots, b_{n-1}, y_1, \dots, y_{n-1}]$ where b_i and y_i are the theoretical m/z values of the corresponding fragment ions at position i. By default, PECAN considers only +1 fragment ions for precursor ions with less than or equal to +2 charges, and includes +2 fragment ions for precursor ions with +3 charges and above. The peptide vector for peptide p is then

$$V_p = \frac{[w_{b_1},\dots,w_{b_{n-1}},w_{y_1},\dots,w_{y_{n-1}}]}{\|[w_{b_1},\dots,w_{b_{n-1}},w_{y_1},\dots,w_{y_{n-1}}]\|} = [w'_{b_1}, \dots, w'_{b_{n-1}}, w'_{y_1}, \dots, w'_{y_{n-1}}],$$

where w_x is the "raw weight" of a fragment ion with m/z value x, and w'_x is the weight normalized to the magnitude of the vector containing raw weights. The raw weight w_x is calculated as the multiplicative inverse of the frequency of observing fragment ions with m/z value x (plus or minus a given mass accuracy, such as 10 ppm), generated by *in silico* fragmentation of proteolytic (e.g. tryptic) peptides from the background proteome

database.

The w_x is calculated in a window-by-window fashion. For each distinct isolation window in a DIA experiment, only proteolytic peptides that could generate precursor ions within the m/z range of the isolation window, and therefore could contribute to product ion interference for the query peptide, are used to calculate the w_x of the window. As a result, fragment ions with high frequency m/z values, such as 147.113 (y_1-Lysine) and 175.119 (y_1-Arginine) for trypsin digestion, are weighted less than those with low frequency m/z values. While w_x represents the specificity of observing a fragment ion with m/z value x in an isolation window with a given species database, w'_x represents the relative specificity for such observation to the peptide p.

3.2.4 *Subtract background scores*

In DIA, multiple precursor ions within an isolation window are fragmented together, resulting in highly multiplexed MS/MS spectra. Because these spectra typically contain so many fragment ions, the expected score for a typical peptide against such spectra is non-zero. To estimate how high a peptide score can be achieved by chance, PECAN calculates "background scores" represented by the means of thousands of decoy peptides. In addition, within the same isolation windows, higher charged precursor ions are assigned more fragment ions and hence exhibit a different score distribution compared to lower charged precursor ions. To account for these differences, the background scores are calculated in a window-by-window and charge-by-charge fashion. Peptides with precursor ions in different isolation windows, or in the same window but of different charge states, have different calibrating backgrounds (more discussion in 3.3.1).

To calculate background scores, PECAN generates thousands of decoys by shuffling proteolytic peptides from the background proteome database and scores each decoy against the data. Let z be a charge state of interest. The background score $B_{y,z}$ for isolation window y at charge state z is calculated as the average score of the thousands of decoys generated within window y with charge state z. With the background scores, PECAN calibrates each peptide score by

$$S'_p = S_p - B_{y,z}$$

Here, the isolation window y and charge state z are selected by the precursor ions of query peptide p. The calibrated score S'_p is then subjected to a simple moving average smoothing with a factor u. One of the strengths of DIA is the systematic measurement of the product ions. Depending on the liquid chromatography separation and DIA cycle time, PECAN uses the smoothing to capture the continuous scoring patterns and smooth out the scattering noise caused by random incidence. PECAN considers the average score at every time point as an "evidence of detection" centered at this time point. The evidence of detection E for peptide p, at center time t is:

$$E_p(t) = \frac{1}{u} \sum_{k=t-\frac{u}{2}}^{t+\frac{u}{2}-1} S'_{p_k}$$

The smoothing factor u is an estimate of the number of times a peptide is analyzed by MS/MS at its full width at half maximum (FWHM) on average. This factor is calculated by dividing the user input minimum peptide elution time (in seconds) to the averaged cycle time of the first one hundred cycles. For example, with a 90-minute linear gradient liquid chromatography on a 30-cm long 3 μm C18 column, most peptides elute for 12 to

20 seconds at FWHM. If a DIA method has a cycle time of 2 seconds, then a peptide would be measured by MS/MS at least 6 times. In this case, PECAN would then use $u = 6$ for the moving average calculation.

3.2.5 *Report evidence of detection*

For every peptide, PECAN by default reports the best scoring evidence of detection and its associated center time t from all evidence that pass empirical criteria of the evidence qualifying procedure (Figure 3.2). The goal of these empirical criteria is to disqualify evidence whose scores are predominantly contributed by a small number of fragment ions, suggesting that the score could result from interference of a few high abundance ions rather than a collaboration of multiple fragment ions. To this end, two hyperparameters, α and β, are used to set the criteria. Let peptide p contain N components (i.e. number of theoretical fragment ions) in the peptide vector V_p. For the candidate $E_p(t)$, the evidence of detection for peptide p at time t, the component score threshold is set as

$$T_p(t) = \frac{1}{N^{\alpha}} \sum_{k=t-\frac{u}{2}}^{t+\frac{u}{2}-1} S_{p_k}$$

The score contribution of a fragment ion component with m/z value x to the $E_p(t)$ is:

$$ionS_p(x,t) = w'_x \cdot \sum_{k=t-\frac{u}{2}}^{t+\frac{u}{2}-1} I_{x,k}$$

We call the fragment ion components that score no less than the threshold $T_p(t)$

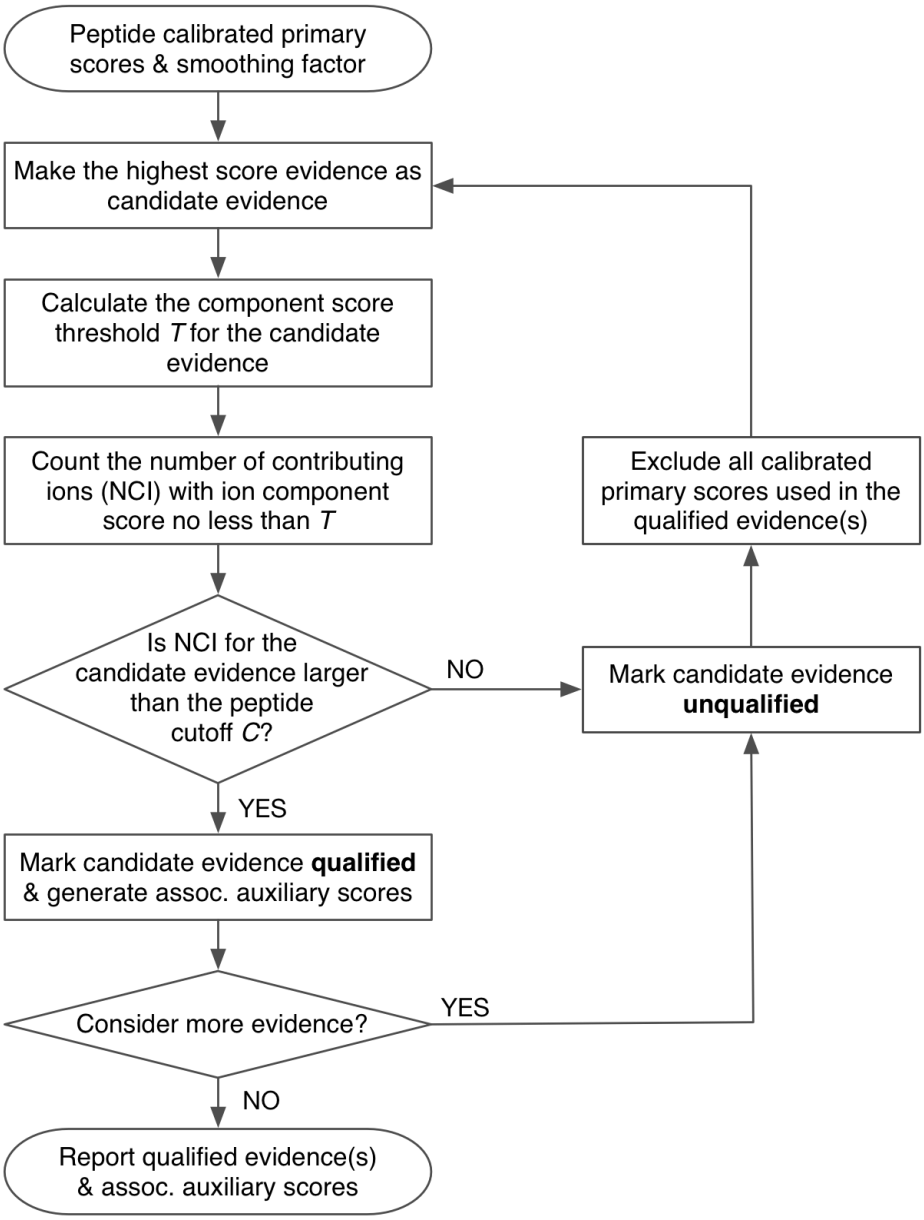

Figure 3.2 Evidence qualifying procedure in PECAN

An evidence of detection (abbr. evidence) for a query peptide p at the time t is the average of the calibrated primary scores from a short period of retention time, centered at the time t. Following this flowchart, PECAN reports a user-defined number of qualified evidence(s) that are calculated from primary scores which have never been used to calculate other qualified evidence(s).

Table 3.1 Auxiliary scores for qualified evidence of detection

Feature	Level	Description
peak score	fragment	Average of pre-calibrated primary scores from a short period of time centered at the retention time *t* for the evidence
peak calibrated score	fragment	Average of calibrated primary scores from a short period of time centered at the retention time *t* for the evidence (i.e. $E_p(t)$, the evidence of detection for peptide *p* at time *t*)
peak weighted score	fragment	Average weighted score of pre-calibrated primary scores from a short period of time centered at *t*, where each fragment ion contribution is weighted by multiplied with its m/z value
peak Z score	fragment	Average of standardized calibrated primary scores from a short period of time centered at the retention time *t* for the evidence, where each calibrated primary score is standardized with the mean and standard deviation of the 2,000 decoy scores of the same precursor charge state
spectra norm	fragment	Average of magnitudes of MS2 spectrum within a short period of time centered at the evidence retention time, where each magnitude is calculated as the Euclidean length of spectrum with square root of the intensities.
NCI	fragment	Number of contributing ions (CIs)
rank	fragment	Rank of the evidence relative to other qualified evidence (if any) for the query peptide
delta Sn	fragment	Normalized delta "peak calibrated score" of the evidence to the next qualified evidence
CI mass error mean	fragment	Mean of the weighted mass errors in ppm from the contributing ions (CI), where the mass error of each CI is weighted by the observed intensity
CI mass error variance	fragment	Variance of the weighted mass errors in ppm from the contributing ions (CI), where the mass error of each CI is weighted by the observed intensity
similarity	fragment	Average cosine similarity of the observed spectra to the peptide scoring vector, where the observed spectra are MS/MS spectra from a short period of time centered at the evidence time *t*
sampled times	fragment	Number of MS/MS spectra from a short period of time centered at the retention time *t* of the evidence
retention time	fragment	Midpoint retention time *t* of the evidence
Average idotp	precursor	Average isotopic dot product score between expected and observed isotopic envelope distributions from MS1 spectra of a short period of time centered at the evidence time *t*
Midpoint idotp	precursor	Isotopic dot product score between expected and observed isotopic envelope distributions from MS1 spectrum at the center time *t* of the evidence
precursor mass error mean	precursor	Mean of the weighted mass errors in ppm from the precursor ions, where the mass error of each precursor ion is weighted by the observed intensity
precursor mass error variance	precursor	Variance of the weighted mass errors in ppm from the precursor ions, where the mass error of each precursor ion is weighted by the observed intensity
peptide length	peptide	Numbers of amino acid from the query peptide
precursor charge state	peptide	Charge state of the query peptide precursor

"contributing ions." Let the number of contributing ions (NCI) of the evidence $E_p(t)$ be the number of ion components with score contribution at time t no less than the threshold $T_p(t)$. If the number of contributing ions of $E_p(t)$ is larger than the threshold $C_p = \beta N$, the evidence of detection $Ep(t)$ is marked "qualified" and will be reported. If the candidate evidence of detection is disqualified, the next highest scoring evidence will be considered until one qualified candidate evidence arise (more discussion in 3.3.2).

3.2.6 *Estimating detection FDR*

PECAN employs Percolator (Käll et al., 2007), a semi-supervised support vector machine algorithm, to rank peptides and estimate FDR of the reported evidence of detection. PECAN generates one decoy peptide for every query (target) peptide by shuffling the target sequence (more description in 3.2.2). These decoys undergo the same scoring processes as the targets, including subtraction with the same background scores. For each reported evidence of detection, whether for a target or a decoy peptide, PECAN calculates 19 auxiliary scores (Table 3.1). The auxiliary scores of the reported evidence from the target and decoy queries are used by Percolator to train a classifier to distinguish correct from incorrect matches and then estimate FDRs. In this target-decoy paradigm, the set of targets contains a mixture of detectable and undetectable peptides, whereas decoys by design consist only of undetectable peptides. Thus, PECAN reported evidence of detection for targets are a mixture of correct and incorrect, whereas all evidence for decoys are incorrect by design. We combined all PECAN reported evidence of detection from different isolation windows of one experiment so that Percolator could use the auxiliary scores to separate correct from incorrect evidence. We refer to the PECAN

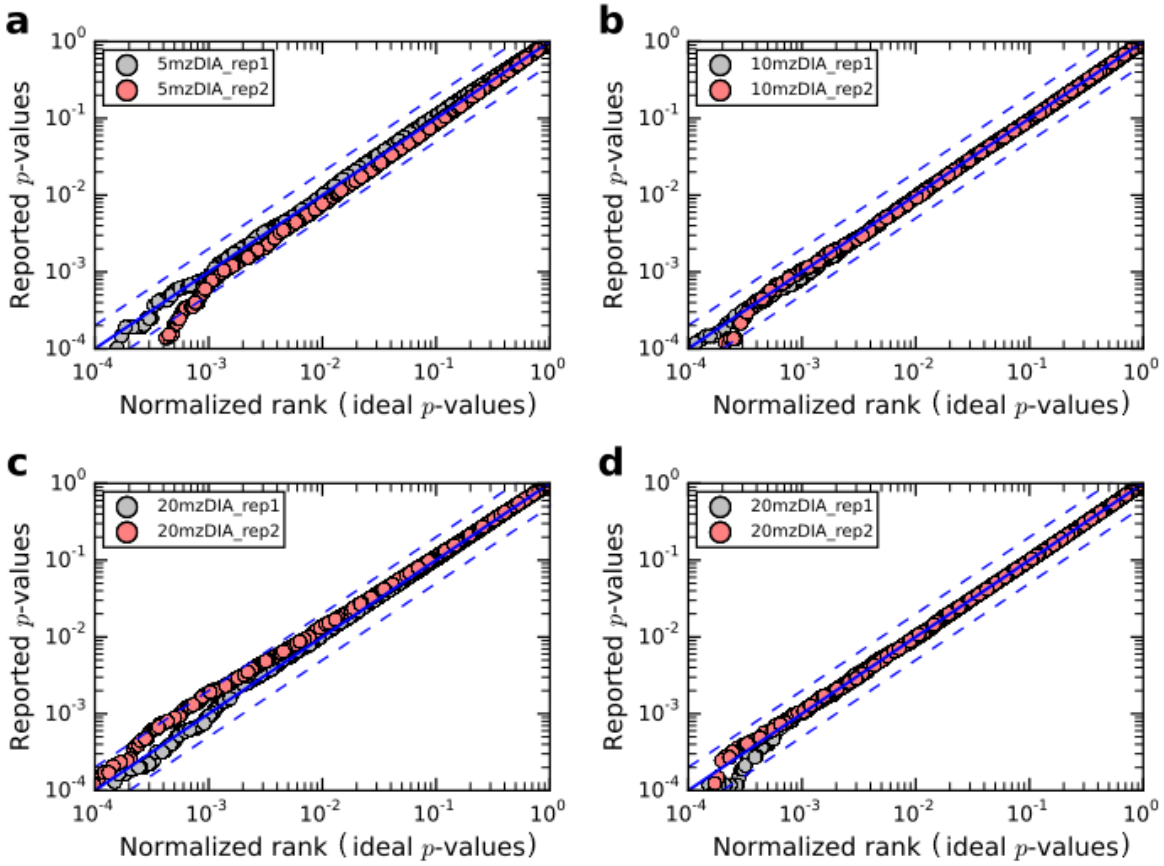

Figure 3.3 Q-Q plots of reported and ideal p-values with various DIA datasets

(a) The Reported p-values are plotted relative to an ideal, uniform distribution or p-values. All p-values were estimated using the Percolator score. The y=x diagonal is indicated by a blue line, and both y=2x and y=x/2 are shown in blue dashed lines. Three HeLa DIA datasets, each containing two technical replicates, were tested: 4xGFP 5mz DIA (a), 2xGFP 10mz DIA (b), and 1xGFP 20mz DIA (c, d). During PECAN analysis, the background proteome used was either the E. coli Swiss-Prot protein sequence database (a, b, c), or the human Swiss-Prot protein sequence database (d).

reported evidence of detection with q-value < 0.01 after Percolator as “PECAN detection”.

To test if the auxiliary scores, single or combined, incorrectly differentiated targets from decoys when used by Percolator, we queried ~100,000 tryptic peptides from the E. coli proteome against HeLa DIA datasets with various DIA isolation schemes. By design, no query peptides are supposed to be detected from the datasets, and thus the target p-values should be uniformly distributed (Granholm et al., 2013). We generated quantile-quantile (Q-Q) plots to compare the p-values reported by Percolator with the normalized rank p-values that represent the uniform distribution (Figure 3.3). The results showed that Percolator could not differentiate the targets from decoys in this test, indicating that the auxiliary scores from PECAN did not introduce undesired separation of targets from decoys. Furthermore, tests of the same dataset with peptide vectors generated from either an E. coli or a human protein sequence database showed that different origins of peptide vectors did not introduce undesired separation of targeted from decoys (Figure 3.3 c and d).

3.3 Assessment of PECAN parameters

3.3.1 *Assessment of background scores estimation*

Background scores estimation is a key component to PECAN scoring. Because MS/MS spectra acquired with DIA contain many peptide-like fragment ions, any peptide could score non-zero against the same MS/MS spectrum. To estimate how high on average a peptide score can be achieved merely by chance with a dataset, PECAN calculates estimated background scores represented by the arithmetic means of

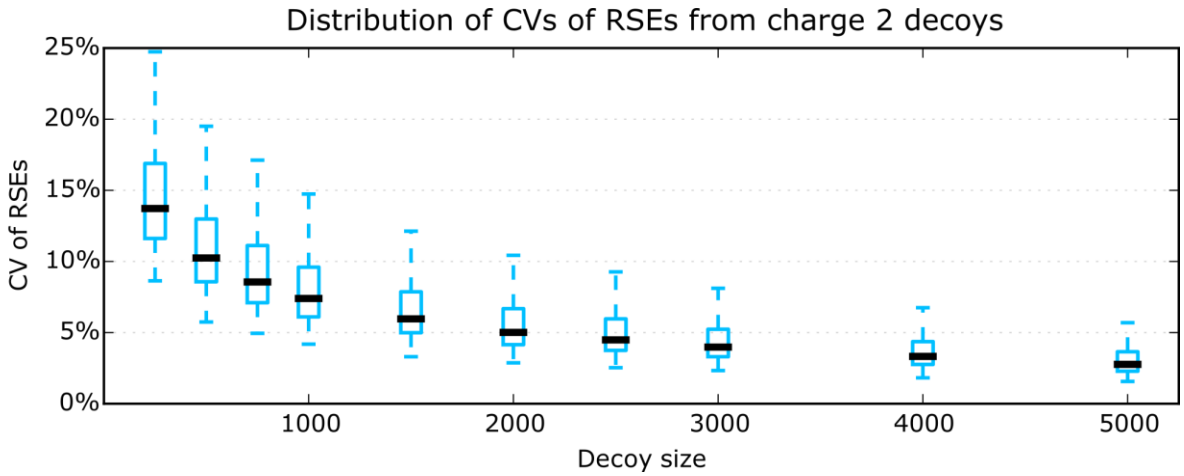

Figure 3.4 Various decoy sizes for 1,000 random sampling estimations

Boxplot shows the distribution of 2,185 CVs of the RSEs from 1,000 random sampling at each decoy size.

thousands of decoy peptides over time. These decoys are generated from shuffling a random selection of proteolytic peptides from the background proteome databases, typically the protein sequence database of the targeted species when analyzing complex sample.

The approach PECAN uses to estimate a background score for individual spectra is analogous to estimating the population mean using a random sample. Such estimation is used because even with a strict proteolytic rule (e.g. fully trypsin digestion), calculating the population mean from all possible proteolytic peptides and precursor ions for every MS/MS spectrum is computationally expensive. For example, there are ~10^{10} possible unmodified peptides with C-terminal arginine or lysine that could generate charge 2

precursor ions between 500-505 *m/z*. In light of this, we adopted the standard practice of estimating the population mean using a random sample.

To determine the sample size *N* (i.e. number of decoys) for background scores estimation, we selected ten different sizes and evaluate the resulting estimate with relative standard error of the mean (RSE), a standard metric indicating how far the estimate is likely to be from the true population mean expressed as a fraction of the estimate. In addition, to account for the sampling effect, for each sample size we performed 1,000 estimations, resulting 1,000 RSEs for every spectrum (Figure 3.4). In this experiment, we used data from one isolation window (500-505 *m/z*) of a mouse DIA dataset that contains 2,185 MS/MS spectra between retention time 20-50 minutes, where most of the peptides were eluted. Charge 2 decoys were generated by randomly sampling the corresponding number of possible tryptic peptides without replacement from the mouse Swiss-Prot database. In one estimation, a set of *N* decoys was generated to calculate 2,185 sample means for 2,185 spectra, followed by 2,185 RSEs. According to the central limit theorem, both the sample means and the RSEs from 1,000 random sampling should be normally distributed. In light of this, to demonstrate sampling effect and evaluate the robustness of the estimation, we calculated the coefficient of variation (CV) of the 1,000 RSEs for individual spectra. Overall, the CVs of the 1,000 RSEs across the data decreased as the sample size increased (Figure 3.4). At sample size 2,000, the RSEs of more than 75% of the 2,185 spectra varied less than 7% CV. Thus, we chose decoy size 2,000 for background score estimation throughout the current study.

Next, we wanted to determine if background scores should be charge state dependent. We used the Wilcoxon rank-sum test with the null hypothesis that the

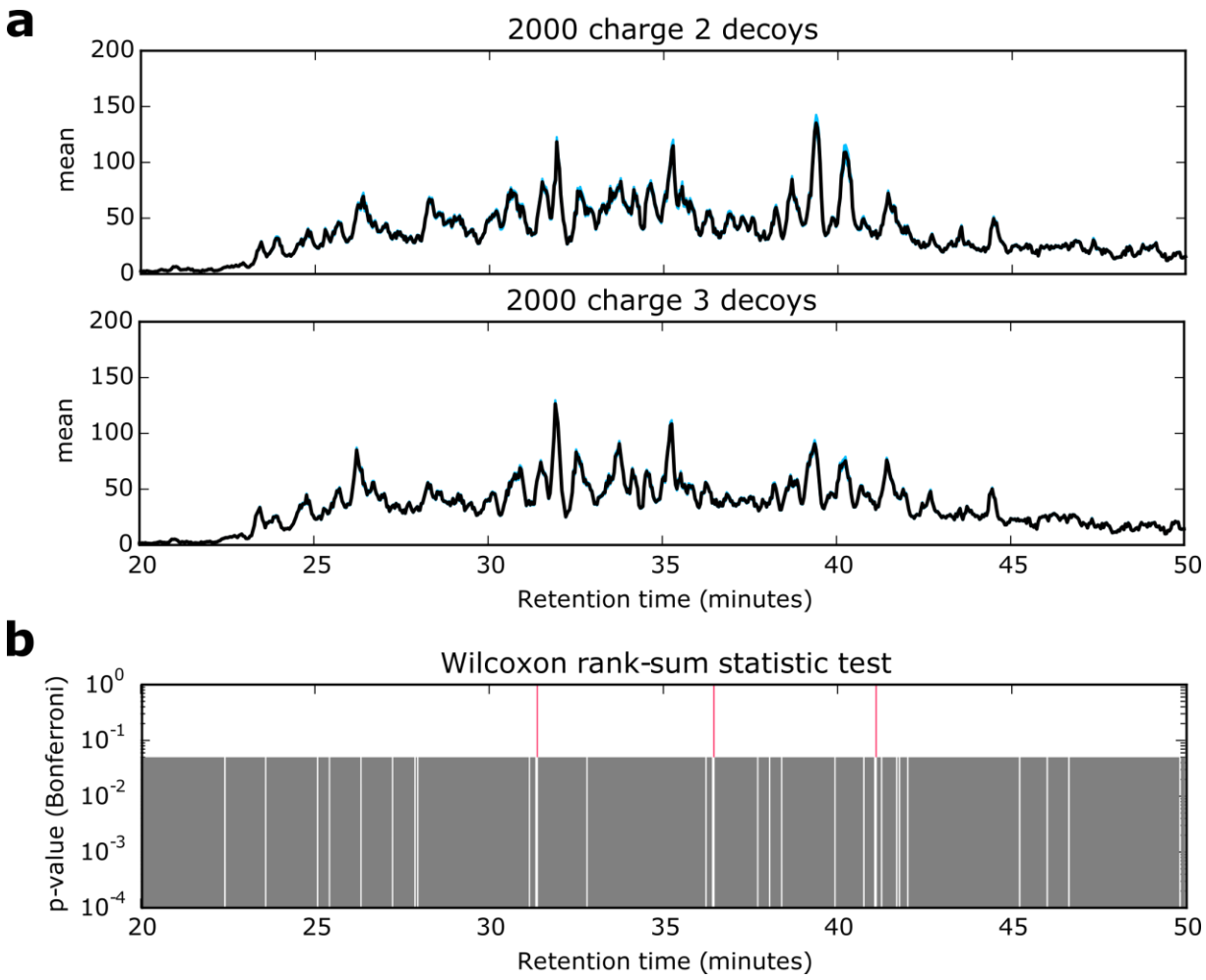

Figure 3.5 Charge state dependency for background score estimations
(a) The estimated background scores with 2,000 charge 2 and 2,000 charge 3 decoys for 2,185 MS/MS spectra presented over retention time. Black lines trace the median of the decoy means from 1,000 estimations by random sampling and the blue shades are segments between the 25th and 75th percentiles. (b) 2,185 Bonferroni corrected p-values from the Wilcoxon rank-sum test for each spectrum between the 1,000 estimations from 2,000 randomly generated charge 2 and charge 3 decoys. For each spectrum, either a grey or red line is drawn below of above the *p*-value cutoff 0.05 if the Bonferroni corrected *p*-value from the Wilcoxon rank-sum test is smaller than the cutoff, or larger than or equal to the cutoff.

underlying score distribution for each MS/MS spectrum from charge 2 and charge 3 decoys are identical (Figure 3.5). At decoy size of 2,000, only 30 of 2,185 spectra tested with Bonferroni corrected p-value >= 0.05 and therefore failed to reject the null hypothesis. This number was further reduced when we increased the decoy size (data not shown). This results demonstrated that the majority of the underlying background score distribution from charge 2 and charge 3 decoys are not identical, and in cases where the two distributions appeared to be identical it was likely an effect of sample size. Thus, PECAN estimates background scores in a charge state dependent fashion.

3.3.2 *Hyperparameters for the evidence qualifying procedure*

PECAN uses empirical criteria during the evidence qualifying procedure to disqualify evidence of detection whose scores are predominantly contributed by a small number of fragment ions, suggesting that the score could result from interference of a few high abundance ions rather than a collaboration of multiple fragment ions. Two hyperparameters α and β are used to set the criteria as described in 3.2.5.

To determine the hyperparameter α and β, I used a S. cerevisiae lysate DIA dataset, acquired on a Q-Exactive using a 10-m/z-wide isolation window DIA approach in which the mass range from 500 to 700 m/z is analyzed with twenty non-overlapping 10-m/z wide isolation window targeted MS/MS scans. This dataset contained 6 biological replicates; each included manually curated boundaries of chromatographic peaks from 204 peptides verified by DDA identification. A total of 1,224 peak boundaries were used as reference for the following test (available at Panorama Public, see 3.7.6).

We first looked at the NCI distribution of PECAN reported evidence resulted from

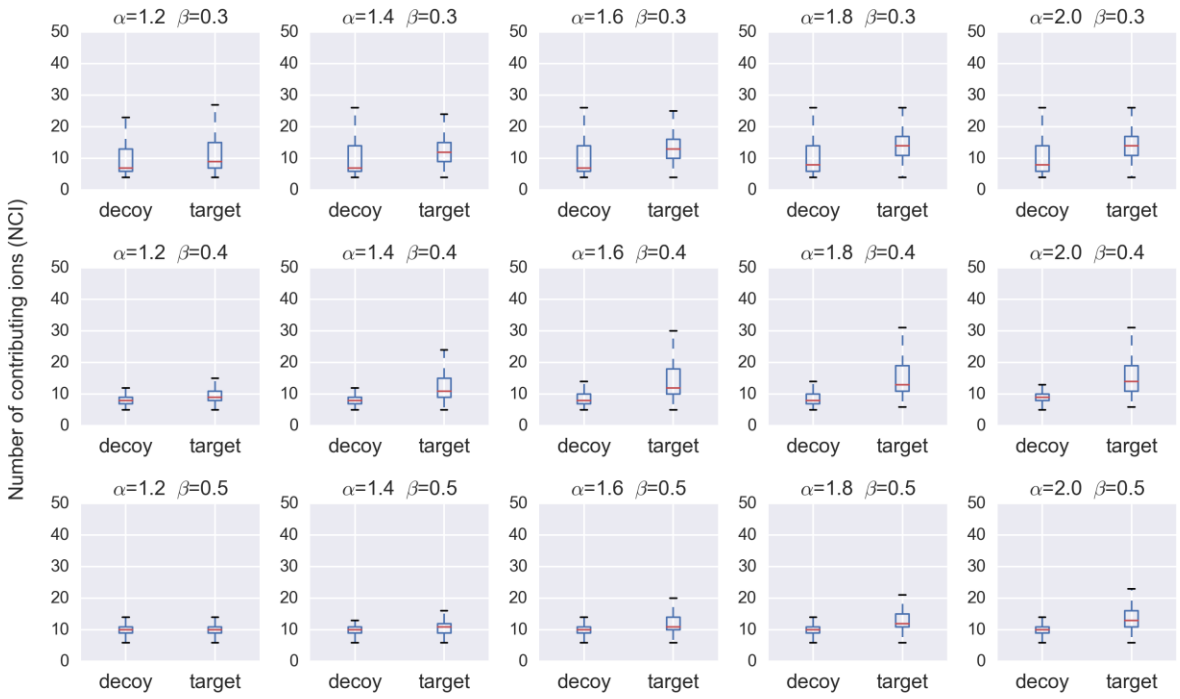

Figure 3.6 NCI distributions with various hyperparameter combinations

Box plots show the NCI distribution of PECAN reported top 1 evidence of detection from 12 representative sets of α and β. Both target and decoy evidence reported by PECAN are included without any FDR control.

various combinations of α and β (Figure 3.6). Overall, as the α increased, the median of the NCI distribution also increased. This is expected because the increase of α decreased the component score threshold of each evidence. As a result, with a lower component score threshold, more fragment ions were considered "contributing ions" for passing the threshold. On the other hand, as the β increased, the range of NCI distribution became tighter, especially for decoys. Because β controls the threshold of NCI required for an evidence to be qualified, larger β favors evidence with more uniformly distributed

component contributions. However, the larger the β is, the less sensitive the evidence qualifying procedure is. Finding the balance between α and β is key to the sensitivity and specificity of the procedure.

With different α and β, we evaluated the performance of the evidence qualifying procedure by comparing the reference peak boundaries to the retention time of PECAN reported evidence when considering the top 1, top 2, or top 3 evidence(s) for each query peptide. A reported evidence was classified as correct if the reported retention time (i.e. center time of the evidence) had fallen between the reference peak boundaries of the query peptide. We defined sensitivity to be the number of peptides with one correct evidence over the total number of query peptides, and specificity to be the number of correct evidence over the total number of reported evidence. As expected by these definitions, we observed that at a given set of α and β, the specificity dropped significantly when PECAN reported top 2 or top 3 evidence per peptide with minimum sensitivity gains compared to reported only the top 1 evidence (Figure 3.7). This result indicates that the calibrated primary score PECAN used to rank the candidate evidence of detection for each peptide was effective so that rarely the second or third best evidence were correct. Together, with α =0.4 and β =1.8 PECAN resulted the best balance between sensitivity and specificity determined by area under the curve when consider only the top 1 evidence (Figure 3.7). This set of α and β values were set to be the default values for PECAN, and used throughout this dissertation.

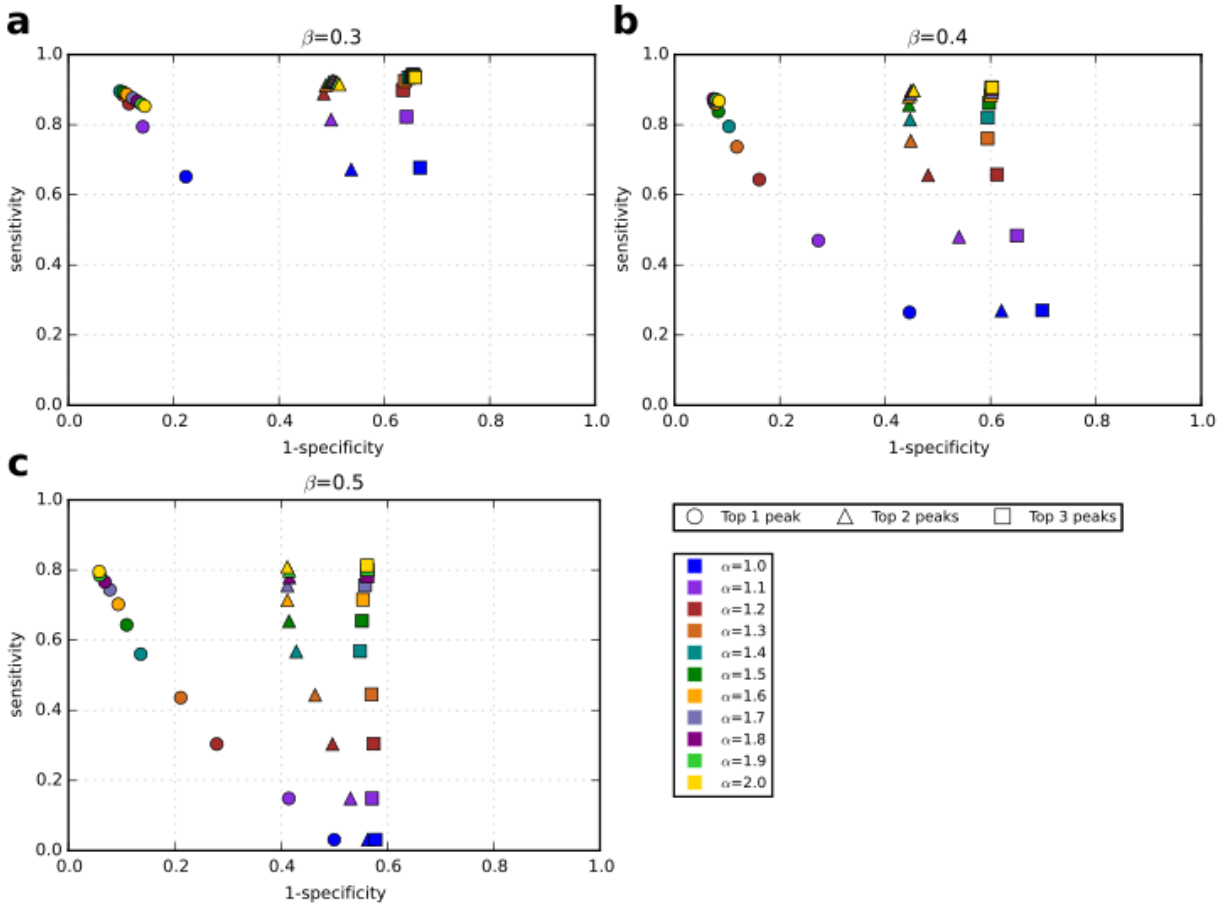

Figure 3.7 Performance of the evidence qualifying procedure with different hyperparameters

Sensitivity and specificity of the qualifying procedure when reported top 1, top 2, or top3 qualified evidence(s) of detection with $1.0 \leq \alpha \leq 2.0$ and $\beta = 0.3$ (a), $\beta = 0.4$ (b), or $\beta = 0.5$ (c). At any given set of α and β, the sensitivity gains were minimum when reporting top 2 or top 3 qualified evidence compared to only reporting the top 1 qualified evidence, indicating that the primary score used to rank the qualified evidence of detection for query peptides were effective.

3.3.3 *Discriminatory power of auxiliary scores*

Auxiliary scores generated by PECAN play an important role in the Percolator SVM. However, there is no measurement of statistical importance for individual auxiliary scores in an SVM. This is because, unlike a method such as logistic regression, which assumes that the underlying data is normally distributed, the SVM is a non-parametric method that makes no assumption about the form or the distribution that generates the data. Without such assumptions, a null model for confidence estimation cannot be analytically derived.

In light of this, practitioners frequently resort to empirical methods to estimate the relative importance of SVM features, in this case the auxiliary scores. This can be done, for example, by deleting one feature from the input and measuring the extent to which this removal affects the performance of the trained classifier. Such methods are admittedly imperfect, both because they do not discriminate between uninformative versus redundant features and because they are conditional on the particular data used for the evaluation. But this type of approach can still provide valuable information.

Unfortunately, this empirical approach still requires a gold standard set of labels against which to evaluate performance. In the case of proteomics, such a gold standard is not easily obtained. We therefore adopted the standard practice of using an empirical null model based on decoy peptides. With this assumption, we can estimate the importance of an auxiliary score by leaving it out of the SVM. For each auxiliary score, we counted the number of peptides detected by the SVM without the score relative to the number of peptides detected with the score in three GPF datasets (Figure 3.8a). In this leave-one-out analysis, the absence of score NCI had the largest impact on the overall

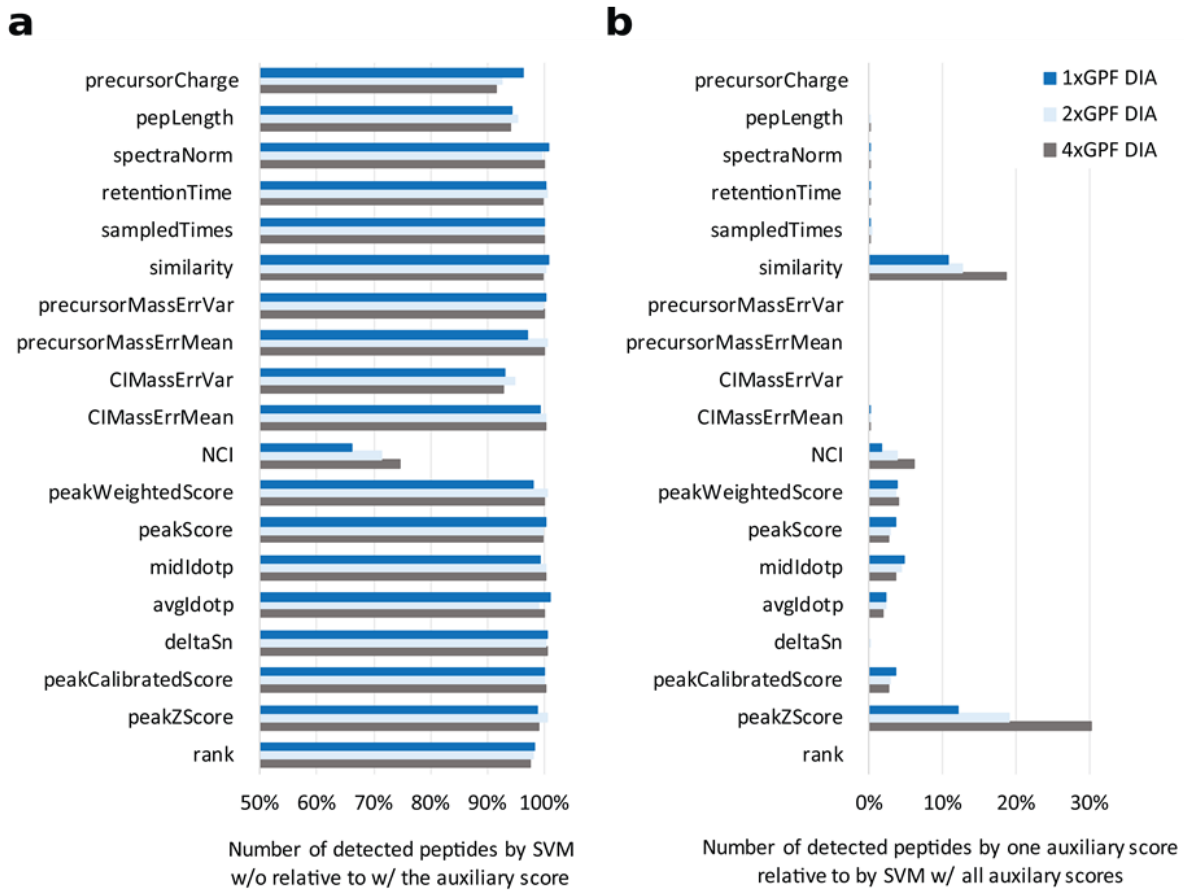

Figure 3.8 Discriminatory power analysis of auxiliary scores

(a) Leave-one-out analysis shows the number of peptides detected with q-value < 0.01 by SVM without the corresponding auxiliary score relative to with the corresponding auxiliary score. (b) Leave-one-in analysis demonstrates the discriminatory power of each auxiliary score on its own with q-value < 0.01, relative to the power of SVM with all auxiliary scores.

discriminatory power of the SVM.

In addition, we investigated how discriminative each auxiliary score is on its own, independent of the SVM. With the empirical null model based on decoy peptides, we

counted the number of peptides detected at q-value < 0.01 by each auxiliary score relative to the number of peptides detect by the SVM with all auxiliary scores (Figure 3.8b). In this leave-one-in analysis, the auxiliary score "peakZscore" had the highest discriminatory power by itself, averaged out to around 20% of the number of peptides detected by the SVM.

3.4 Performance validation of PECAN

3.4.1 *PECAN peak picking performance*

For every query peptide, PECAN reports the best evidence of detection with an associated center time by scoring through the entire DIA data. In many ways, this process is analogous to picking one chromatographic peak from the full retention time range extracted ion chromatograms (XICs) of the query peptide. To evaluate the "peak picking" performance of PECAN, we analyzed a published SWATH-MS dataset that contains 422 synthetic stable isotope-labeled standard (SIS) peptides spiked into a human proteome with 10 two-fold dilution steps, each measured in triplicate (Röst et al., 2014). The dataset was published with a manually curated reference specifying the boundaries of chromatographic peaks of 387 detectable SIS peptides in each dilution step.

Based on the published reference, we calculated the percentage of correct to total SIS peaks reported by PECAN. To do so, the PECAN reported evidence of detection for SIS peptides were classified as correct or incorrect by whether their retention times fall in between or outside of the reference peak boundaries. Before subjecting to the

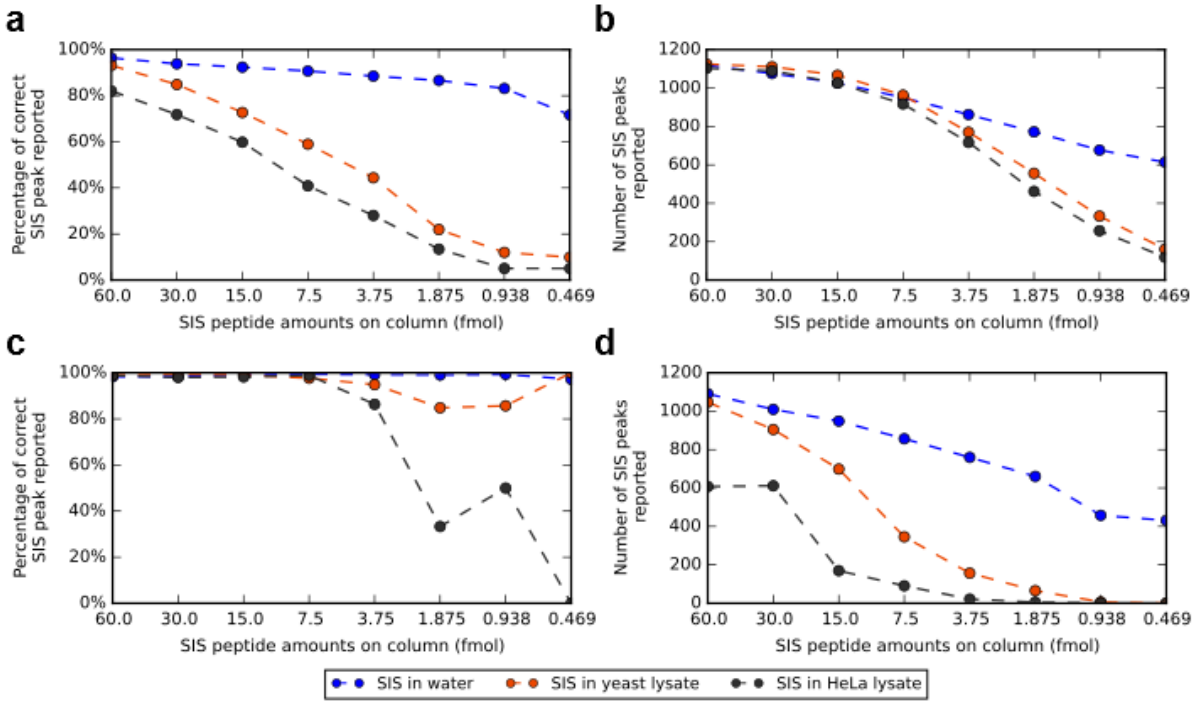

Figure 3.9 PECAN peak picking performance on SIS dataset

The percentage of total correct SIS peaks (a) and the number of SIS peaks (b) reported by PECAN prior to FDR control from three replicates combined. Same figures (c) and (d) respectively after the PECAN reported evidence of detection were subjected to peptide level FDR control per measurement at q-value < 0.01 by Percolator.

discriminative classifier of Percolator, the percentage of correct SIS peaks reported decreased as the SIS spiked-in concentration decreased and as the complexity of the matrix increased (Figure 3.9). After subjecting to the discriminative classifier of Percolator and FDR control at the peptide-level *q*-value < 0.01, the percentage of correct SIS peaks reported greatly improved, even at low SIS spiked-in concentration and high

interference from the background human proteome (Figure 3.9c). This improvement resulted from the success of FDR control. While PECAN reports the best evidence of detection for every query peptide, not every reported evidence is correct just as not every query peptide is detectable from the data. These results show that with PECAN reported decoy evidence, Percolator was able to successfully reject most of the incorrect evidence of detection (Figure 3.9d), and thus greatly improved the performance of PECAN peak picking.

3.4.2 *PECAN detection validation*

We validated the quality of the resulting set of detected peptides in two ways. First, we employed the 8,207 GST-fusion-protein database to interrogate the 90-min deep gas phase fractionation (4xGPF, see 3.7.4 for more description) HeLa datasets. At 1% FDR by Percolator, we compared peptides detected from DIA by PECAN with those from DDA by Comet, yielding 12,767 and 6,221 unique peptides, respectively, with an overlap of 5,182 peptides (Figure 3.10 a). In other words, 83% of Comet-DDA peptides were detected by PECAN directly from the DIA data without assistance from any libraries. Of the 1,039 peptides only identified in Comet-DDA, 179 were identified with charge 4 precursors that was not considered in this PECAN analysis; 428 had PECAN reported evidence but these evidence did not pass the FDR control; 96 were identified with precursors that fall in the gaps between the DIA isolation windows edge; and 336 had no qualified evidence that passed the evidence qualifying procedure. The PECAN-DIA and Comet-DDA peptides then led to detection of 2,613 and 1,759 protein groups respectively, with an overlap of 1,510

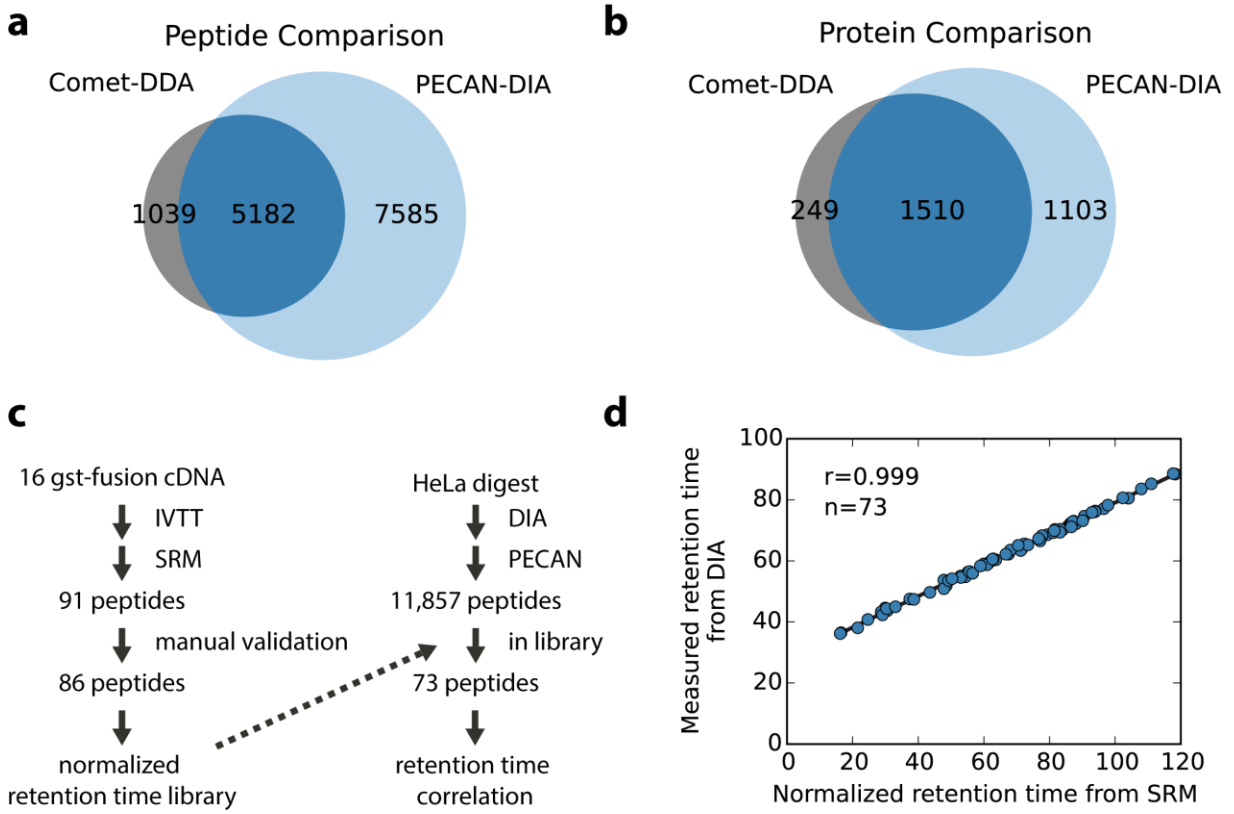

Figure 3.10 Validate PECAN detection with GST-fusion proteins

Comparative analysis of peptide detection from DIA and DDA data from HeLa protein digest. Peptide (a) and protein (b) comparison of PECAN-DIA detection and Comet-DDA identification. (c) SRM validation workflow for a set of analytical standards synthesized using in vitro transcription translation (IVTT). (d) Comparative analysis of retention time of HeLa peptides detected by PECAN from DIA data and IVTT peptides detected from SRM.

proteins (

Figure 3.10 b). This result shows that many of the distinct peptides from Comet-DDA and PECAN-DIA were in fact derived from the same proteins.

To verify the PECAN-DIA specific detections, we selected 16 GST-fusion proteins and expressed them using the in vitro transcription translation (IVTT) protocol for individual cDNA constructs (Figure 3.10 c). We then measured the corresponding 91 peptides individually using SRM from the corresponding trypsin digestion of the GST-enriched proteins. Of the 91 peptides monitored in SRM, we manually assigned boundaries of chromatographic peaks for 86 peptides without ambiguity of detection, and created a normalized retention time library referenced to the spiked-in stable-isotope labeled peptides. We then compared the measured retention time of the 73 peptides detected in PECAN-DIA to the normalized retention time from the SRM and received a 0.999 correlation coefficient (Figure 3.10d). With a threshold of < 0.1% difference in total normalized retention time, all 73 peptides were correct, suggesting that the large majority of the PECAN-DIA specific set of detected peptides were correct.

3.4.3 *Deep gas phase fractionation DDA*

As a reference for deep gas phase fractionation (GPF) DIA analysis, we also analyzed the DDA data acquired with matching GPF settings. We searched the 1xGPF, 2xGPF and 4xGPF DDA data with Comet and used Percolator and Fido to report peptide and protein identification at q-value < 0.01, respectively. With different GPF settings, DDA should sample in various depths using the same top-20 method because each fractionation focused on various widths of precursor m/z range (Yi et al., 2002). From the 1xGPF, 2xGPF and 4xGPF DDA data, we identified 5,934, 5,915, and 6,221 unique peptides, and 1,504, 1,678, and 1,759 protein groups, respectively (Figure 3.11 a).

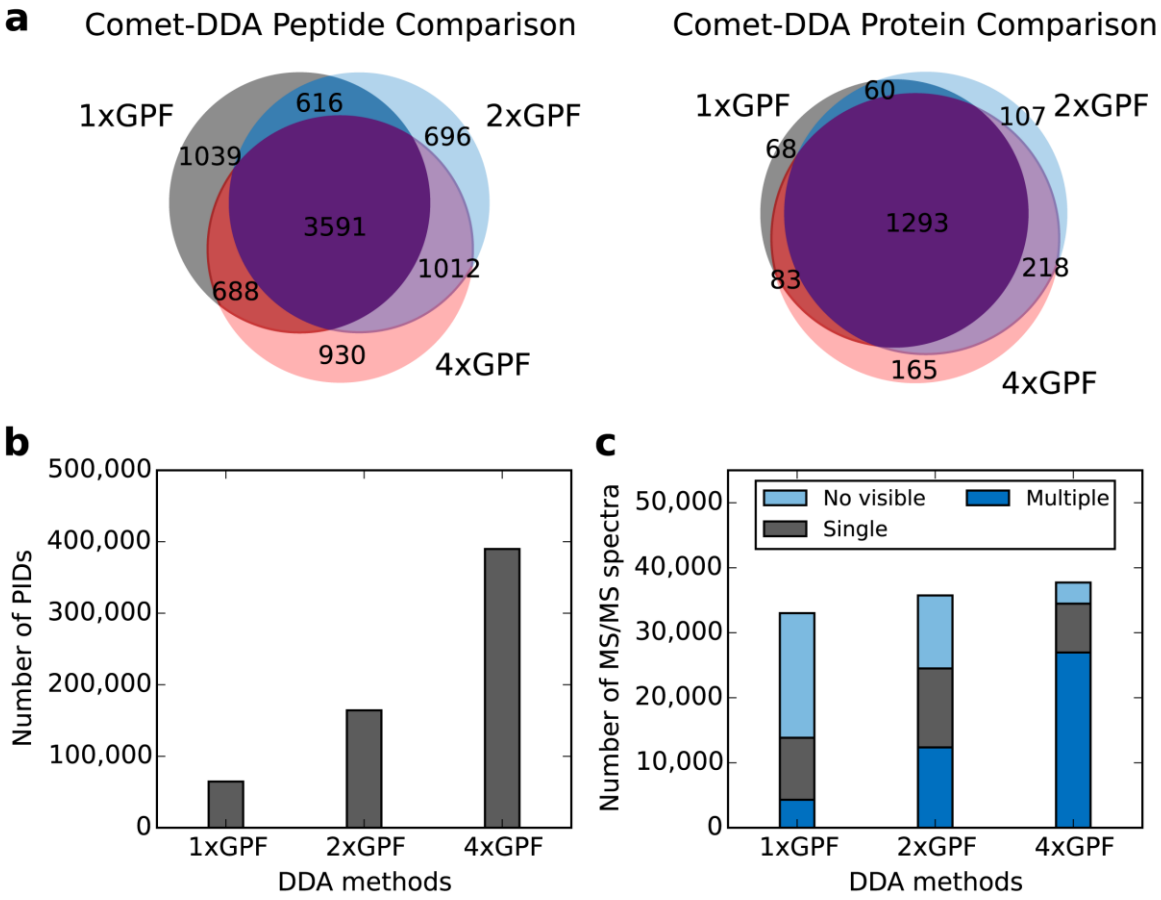

Figure 3.11 Gas phase fractionation DDA

Deeper gas phase fractionation revealed more precursor isotope distributions but failed to improve DDA identification due to unoptimized acquisition parameters. (a) Comparison of peptides and proteins identified by Comet from 1xGPF, 2xGPF, and 4xGPF DDA data. (b) Number of peptide isotope distributions (PIDs) identified. (c) Number of MS/MS spectra assigned with no visible, single, or multiple PIDs.

Surprisingly, when we compared 4xGPF to 1xGPF DDA data, only 14 % more MS/MS spectra were acquired with 4xGPF (Figure 3.11 c). This was unexpected because 4xGPF cost four times the sample and instrument time of what 1xGPF cost so that in each

fractionation DDA only needed to sample from one quarter of the precursor range. In addition, with an Orbitrap mass analyzer, reducing the ion variety for MS1 analysis (i.e. improved MS1 selectivity) by deep GPF should improve the MS1 sensitivity. To test if MS1 sensitivity was improved in deep GPF data, we used Hardklör (v2.16) (Hoopmann et al., 2012) to identify peptide isotopic distributions (PIDs) in the MS1 spectra. As expected, five times more PIDs were identified in 4xGPF than in 1xGPF, indicating that the MS1 sensitivity was greatly improved with deep GPF (Figure 3.11 b). Next, we used Bullseye (v1.26) (Hsieh et al., 2010) to assign these PIDs within ± 3 seconds in retention time to each MS/MS spectrum. As MS1 signal got more selective from 1xGPF to 4xGPF, a significantly higher percentage of MS/MS spectra were assigned with multiple PIDs (Figure 3.11 c). These results indicate that while the DDA method used here was not optimized for the corresponding GPF settings, the sensitivity of MS1 signal was successfully improved by deep GPF.

3.4.4 *Querying modified peptides*

PECAN can be used to query modified forms of the peptides. For fixed modifications, such as carbamidomethyl cysteine, the delta mass of the modification is applied globally to the modified residue, including target peptides, peptides in the background proteome, and every decoy generated. For querying peptides with variable modifications, PECAN treats each peptide query independently. PECAN leverages precursor information in the form of auxiliary scores. In the case where multiple modified forms of one peptide have different intact masses, the evidence reported by PECAN for each modified from will have different auxiliary scores, including precursor isotopic dot

products, and means and variances of precursor mass error, even if the same group of spectra provides the best evidence for more than one modified forms of the peptide. In case of positional isomers, PECAN treats each peptide query independently. Thus, it is possible that multiple isomers could be scored equally highly with the same group of spectra.

To demonstrate how PECAN performs when considering modifications, we queried the modified peptides from protein *N*-terminal acetylation (i.e. *N*-acetylation) in addition to the unmodified peptides of the human UniProt Swiss-Prot database against the 4xGPF DIA data. PECAN detected 34,958 unique peptides, including 267 peptides from protein *N*-acetylation. In addition, we used Comet to search the 4xGPF DDA data allowing for variable modification of protein *N*-terminal acetylation. Comet identified 15,656 unique peptides including 120 peptides from protein *N*-acetylation. 91 modified peptides were detected in both methods (Figure 3.12 a). The measured retention time of these 91 peptides from DDA and DIA data aligned nicely and thus further confirmed the detection with modification made by PECAN (Figure 3.12 b).

Differentiating modifications from DIA data is a lot more challenging than from DDA data. Depending on the DIA isolation scheme, multiple modification forms of the same peptide could all fall in the same MS2 isolation window. For example, oxidation on a 2+ peptide only has a precursor shift of 8 m/z. The oxidation form and the non-modified form of one peptide could share most of the fragment ions and be measured in the same MS2 scans in DIA with isolation windows larger than 8 m/z-wide. In this case, one could only distinguish the detection if the MS1 provides strong support preferring one

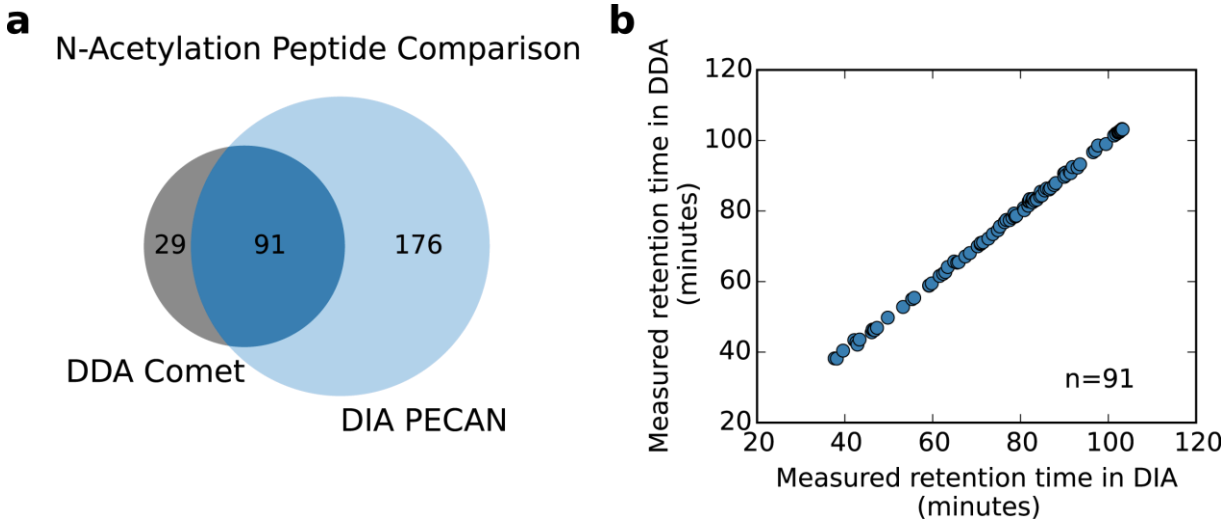

Figure 3.12 Detection of modified peptides from protein *N*-acetylation (a) Comparison of modified peptides of protein N-terminal acetylation (N-acetylation) detected by Comet from 4xGPF DDA data and by PECAN from 4xGPF DIA data. (b) Retention time analysis of 91 modified peptides detected by both methods.

precursor, or if the distinguishing product ions were observed. For this reason, PECAN leverages precursor information when it is available to improve search results and distinguish between the modifications.

Currently, PECAN does not further filter detections if the same group of spectra provide evidence for multiple forms of one peptide. By design, PECAN treats the detection of every peptide independently from others. Keep in mind that DIA data could provide enough evidence for some modifications, but may not have enough evidence to differentiate one form from the others. This is also a challenge that traditional database searching approaches have faced, with scores designed for this purpose, such as the A-

score for site localization (Beausoleil et al., 2006). Thus, while it is possible to query for variable modifications with the current implementation of PECAN, users are strongly urged to further scrutinize the results, especially if the goal is site-localization of modified forms.

3.5 Impact of precursor selectivity on PECAN detection

Current DIA methods often use 5-10 times wider isolation windows compared to conventional DDA (typically 1.5 to 2 *m/z*-wide) to sample a desired precursor range. The wide precursor range ensures that each analyte is measured multiple times. However, using wide isolation windows (i.e. low precursor selectivity) dramatically increases the complexity of the resulting MS/MS spectra because of the interference from multiple co-fragmenting analytes (Gillet et al., 2012). To test how precursor selectivity impacts PECAN's performance in detection, we used a technique called gas-phase fractionation (GPF) to vary precursor selectivity of DIA while maintaining the same cycle time and precursor *m/z* range measured. Three GPF settings, named 1xGPF, 2xGPF and 4xGPF, were used to collect three DIA HeLa datasets with 20, 10, and 5 *m/z*-wide isolation windows, respectively. We used the human UniProt Swiss-Prot database to interrogate the three DIA datasets. From the 1xGPF (20 m/z), 2xGPF (10 m/z), and 4xGPF (5 m/z) DIA datasets, PECAN detected 14,135, 23,398, and 34,813 unique peptides, and 1,834, 5,191, and 9,132 protein groups, respectively (Figure 3.13 a and b). This result shows that better precursor selectivity (i.e. narrower isolation windows) dramatically improves PECAN's performance. In addition, the majority of peptide and protein detections from

DIA data with lower precursor selectivity were successfully captured by data with higher precursor selectivity. Furthermore, we projected the retention time of 12,952 peptides detected in all three datasets from 1xGPF and 2xGPF to 4xGPF, and found only 17 peptides that were detected with disagreement in retention time after projecting the retention times to the 4xGPF dataset (Figure 3.13 c), indicating the robustness of PECAN detection.

As a benchmark, we processed the GPF DIA datasets with DIA-Umpire followed by Comet database searching. From the 1xGPF, 2xGPF and 4xGPF DIA datasets processed by DIA-Umpire, Comet identified 13,978, 20,266, and 24,721 unique peptides at Percolator peptide-level q-value < 0.01. Compared to the results from this DIA-Umpire workflow, PECAN detected 1%, 15%, and 41% more unique peptides from 1xGPF, 2xGPF and 4xGPF, respectively, while 9,919, 15,369, and 20,015 peptides were detected by both tools (Figure 3.13 d). Overall, PECAN outperformed DIA-Umpire more when the precursor selectivity of DIA is better.

Both PECAN and DIA-Umpire showed significant improvements in peptide detection from 1xGPF to 4xGPF data. Compared to 1xGPF, the 4xGPF setting not only improves the precursor sensitivity for MS analysis by reducing the isolation range of precursor ions in each fractionation, but also improves the precursor selectivity for MS/MS analysis. For DIA-Umpire, the process that groups and extracts covarying product signals to precursor signals prior to database searching, the improvement in precursor sensitivity from 4xGPF compared to 1xGPF was particularly beneficial as it revealed more precursor ions that were undetectable in 1xGPF. However, 4xGPF improves precursor sensitivity but not the resolving power in MS analysis. Thus, if a

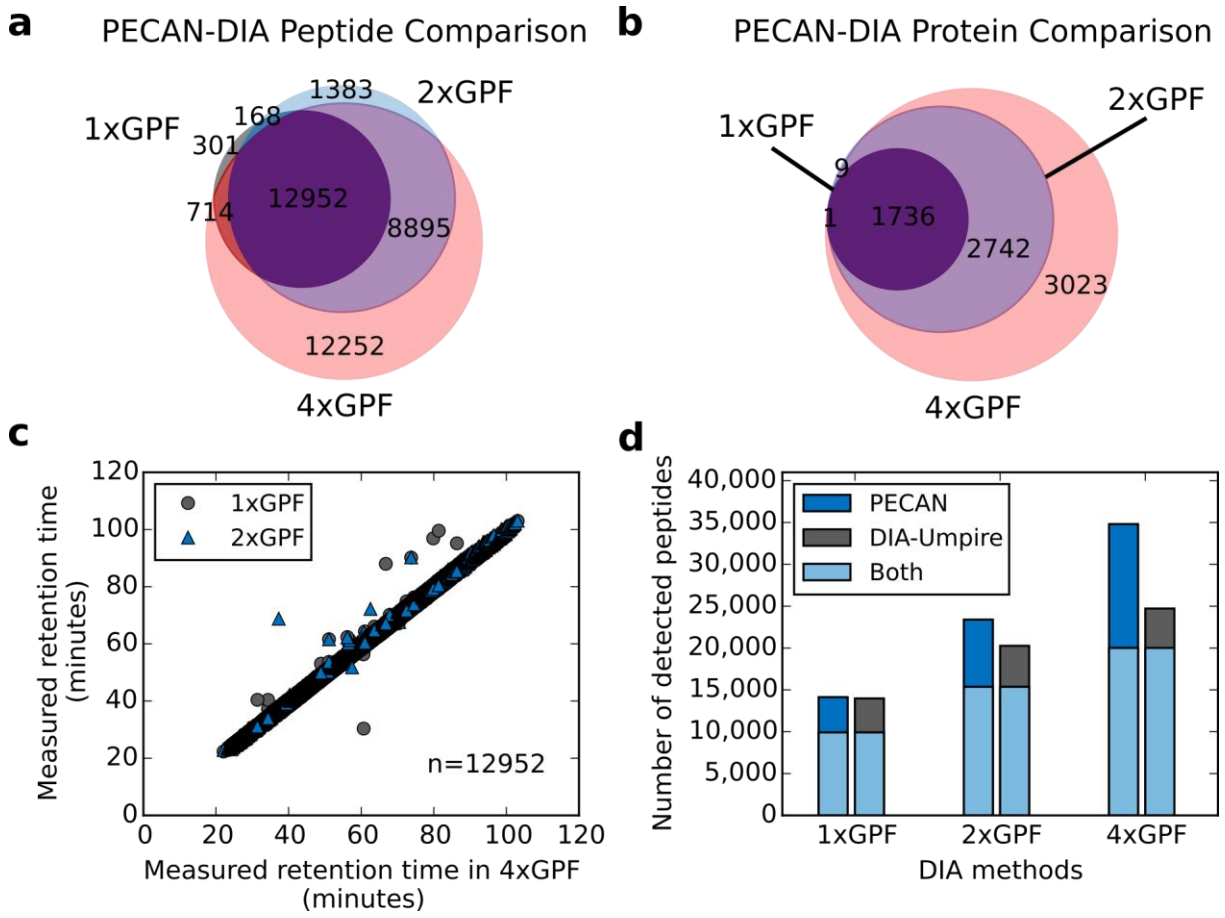

Figure 3.13 Altering precursor selectivity with gas phase fractionation Comparison of peptides (a) and proteins (b) detected by PECAN from 1xGPF, 2xGPF, and 4xGPF DIA data. (c) Retention time comparison of 12,952 PECAN detected peptides form 1xGPF and 2xGPF relative to 4xGPF DIA. (d) Number of peptides detected by either, or both PECAN and DIA-Umpire from the three GPF DIA datasets.

precursor ion shows interference because of the chemical noise in 1xGPF, it is likely to show interference in the 4xGPF because the resolution of MS analysis in both GPF settings remains constant. Moreover, the low intensity precursors that are only detectable

in 4xGPF but not in 1xGPF are more likely to have interference and more susceptible to stochastic sources of noise such as spray instability. Thus, detecting the co-variation with product ions for low intensity precursors is generally more challenging than for high intensity precursors. Nonetheless, DIA- Umpire demonstrated a 77% improvement in peptide detection from 1xGPF to 4xGPF. On the other hand, PECAN reports evidence of detection based on product scoring, taking advantage of the fact that improved precursor selectivity generates less product interference. In addition, PECAN also benefits from the improvement in precursor sensitivity in the forms of the auxiliary scores for precursors. Together, PECAN demonstrated a 146% improvement in peptide detection from 1xGPF to 4xGFP.

3.6 Conclusions

We have demonstrated the ability of PECAN to detect peptides robustly and accurately from DIA data without using a library. Because the detection of peptides improves as the precursor isolation window is decreased, PECAN enables a workflow where libraries can be built directly from DIA data collected using narrow isolation windows and applied to wide isolation data later.

PECAN may not be as sensitive as library-based tools for detecting some peptides. This is because, by design, PECAN does not use experimental retention time or relative fragment ion intensities. To further improve the sensitivity of PECAN, we expect that training the hyperparameters, α and β, for the evidence qualifying procedure with DIA data of various precursor selectivity will be effective. We also expect that incorporating a

retention time predictor, such as SSRCalc (Krokhin et al., 2004) or BioLCCC (Gorshkov et al., 2006), to filter unexpected time regions of signals will improve the sensitivity of PECAN detection.

3.7 Materials and methods

3.7.1 *Liquid chromatography*

All chromatography was performed using a nanoACQUITY (Waters) system set to a flow rate of 250 nl/min during linear gradient. Buffer A was 2% ACN, 0.1% formic acid and 97.9% water. Buffer B was 99.9% ACN and 0.1% formic acid.

Homemade 3-cm-long 100-µm inner diameter (I.D.) trapping columns were used prior to the homemade 75-µm I.D. resolving column that is either 15 or 30-cm-long for a 27.5-min or 90-min linear gradient from 2% to 32% Buffer B respectively. Both trapping and resolving columns were packed with 3-µm ReproSil-Pur C18 AQ (Dr. Maisch GmbH). The gradient was followed by a wash at 80% Buffer B and a column re-equilibration at 2% Buffer B.

3.7.2 *Select proteins and peptides for IVTT SRM*

Ninety-one peptides were selected for the 16 GST-fusion proteins based on a preliminary analysis of PECAN during its early development. The proteins and peptides were selected based on the preliminary PECAN results from the 4xGPF HeLa DIA data acquired with 5*m/z*-wide isolation windows, and on the Comet results from 4xGPF HeLa DDA data.

First, tryptic peptides with up to one missed cleavage from the 8,207 GST-fusion-protein database were queried against DIA data by PECAN. DDA data was analyzed by Comet using the same database and up to one missed cleavage was allowed. The detected peptides, both reported at Percolator q-value<0.01, were compared and mapped to the proteins in the GST-fusion-protein database. From a random order, the first 16 proteins[1] with at least more than 3 additional peptides detected by PECAN-DIA compared to Comet-DDA, and with at most 1 peptide identified by Comet-DDA were selected for IVTT synthesis.

As mentioned in 3.4.2, the SRM assay development described above was done with peptides detected by an earlier version of PECAN. Since then, minor adjustments were made and additional features, such as hyperparameters alpha and beta, were added to PECAN. The earlier version of PECAN was only used in selecting the peptides for IVTT and SRM. All the validation and comparison of PECAN detection in this manuscript and supplementary were performed with PECAN (v 0.9.9).

3.7.3 *SRM validation of IVTT proteins*

Full-length cDNA clones for the 16 selected proteins were obtained from the pANT7_cGST clone collection distributed by the Arizona State University Biodesign Institute plasmid repository. Each bacterial stock clone was grown independently overnight in 5 ml of Luria-Bertani broth with 100 μg ml^{-1} ampicillin (LB-amp). Plasmid DNA was extracted using the manufacturer's spin mini-prep protocol (QIAGEN).

[1] During the culturing step, one of the 16 clones (library well ID: HsxXG003443-A06) did not grow to the desired O.D. We replaced that protein with one that passed all of the aforementioned criteria, but had already been synthesized in the lab (HsxXG006208-E04).

Proteins were then synthesized from plasmid DNA using the Pierce Human *in vitro* Protein Expression kit (Thermo) according to the manufacturer's protocol with GFP control. We then enriched the GST-fusion proteins using glutathione sepharose 4B beads (GE) with a published method (Stergachis et al., 2011). Finally, these enriched GST-fusion proteins were reduced, alkylated, and digested for 2 h with trypsin individually.

Ninety-one peptides were selected for the 16 proteins based on a preliminary analysis of PECAN during its early development (more description in 3.7.2). Each protein digestion was injected separately and analyzed with a TSQ-Vantage triple-quadrupole instrument (Thermo) using a nanoACQUITY UPLC (Waters). A 3-µl aliquot of sample was loaded for a 27.5-min LC setting. Ions were isolated in both Q1 and Q3 using 0.7 FWHM resolution. Peptide fragmentation was performed at 1.5 mTorr in Q2 without peptide specific collision energies. Data was acquired using a scan width of 0.002 mass to charge ratio (*m/z*) and a dwell time of 10 ms.

3.7.4 *HeLa datasets*

HeLa protein digest (Thermo) spiked-in with stable-isotope labeled peptides (PRTC, Thermo) was analyzed on a Q-Exactive HF mass spectrometer. One µg of HeLa peptides and 40 fmol of PRTC were loaded in each injection and separated with a 90-min linear gradient LC. Three gas-phase fractionation (GPF) (Davis et al., 2001) settings were used to cover the precursor *m/z* range of 500 to 900: one injection (1xGPF), two injections covering 500-700 and 700-900 *m/z* (2xGPF), and four injections covering 500-600, 600-700, 700-800, and 800-900 *m/z* (4xGPF). The isolation ranges of MS analysis for all GPF settings correspond to the precursor range covered in each injection. For example,

the third injection of 4xGPF contains MS analysis with scanning ranges of 700 to 800 *m/z*, and MS/MS analysis of selected (either by DDA or DIA) precursor ions within precursor ranges of 700 to 800 m/z. Thus, the costs of sample amount and instrument time are double of the costs in 1xGPF for 2xGPF, and quadruple for 4xGPF.

Both DDA and DIA data were acquired with three GPF settings. A standard, top-20 DDA method (MS analysis with 120,000 resolution and MS/MS analysis with 15,000 resolution) with 1.5 m/z-wide isolation windows was used in data collection of 1xGPF DDA, 2xGPF DDA, and 4xGPF DDA. A standard (one MS analysis with 60,000 resolution followed by twenty MS/MS with 30,000 resolution) DIA method with 20, 10, or 5 m/z-wide isolation windows was used to acquire 1xGPF DIA, 2xGPF DIA, or 4xGPF DIA respectively. For FDR control, data from multiple injections were analyzed together as if they were from one instrument run.

3.7.5 *Databases and data analysis*

Three sequence databases were used in this chapter: the GST-fusion-protein database containing 8,207 proteins translated from the DNASU human cDNA plasmid library, the human UniProt Swiss-Prot database containing 42,128 protein isoforms, and the UniProt Swiss-Prot human natural variant database containing 74,733 single amino acid variants. For validating PECAN detection, we queried the GST-fusion-protein database using PECAN against DIA data, and searched DDA data using Comet against the same database instead of the UniProt Swiss-Prot database, guided by the principle of searching only the peptides we are interested in (Noble, 2015). For the comparison with the DIA-Umpire workflow, we targeted the human UniProt Swiss-Prot database. In all

cases, the human UniProt Swiss-Prot database was used as the background proteome for PECAN.

In all analysis, only fully tryptic peptides with up to one missed cleavage sites were considered, and only a fixed modification of carbamidomethyl cysteine was considered. For the PECAN workflow, PECAN (v. 0.9.9) was used to query peptides from the target database, allowing for 2+ or 3+ precursor ion charge states. All PECAN analysis was done in y-ion mode where only product y-ion series were considered. For DDA data, Comet (v 2016.01 rev. 0) was used to search the MS/MS spectra against the target database allowing for up to +4 precursor ions. For the DIA-Umpire workflow, DIA-Umpire (v. 1.4) was used to extract signal and generate pseudo spectra from DIA data, allowing for 2+ to 4+ precursor ions. The resulting pseudo spectra were searched by Comet (v 2016.01 rev. 0) with corresponding charge states. A $\pm$10 ppm mass error tolerance is used for precursor ions (in PECAN and Comet) and fragment ions (in PECAN only). A 0.02 *m/z*-bin-width for fragment ions is used in Comet. Both PECAN and Comet results are processed by Percolator (Käll et al., 2007) (v. 2.08.01) to separate targets and decoys. All peptides are reported by Percolator at the peptide level with *q*-value < 0.01, and proteins are reported by Percolator's built-in Fido algorithm (Serang et al., 2010) with *q*-value < 0.01, unless indicated otherwise. For protein comparison, protein groups reported by Fido are represented by the first protein sorted by accession number.

3.7.6 *Data and software access*

PECAN is open-source, freely available at http://pecan.maccosslab.org. All raw data acquired for this manuscript are publicly available at Chorus Project, project number

Table 3.2 Direct links for downloading the raw files

Dataset Name	Chorus ID	Link
SRM validation of IVTT proteins	2427	https://chorusproject.org/anonymous/download/experiment/4846597907291871276
HeLa datasets Part I: DDA	2448	https://chorusproject.org/anonymous/download/experiment/-2822210361803919543
HeLa datasets Part II: DIA	2449	https://chorusproject.org/anonymous/download/experiment/1929128726775705417
DIA plasma library	2655	https://chorusproject.org/anonymous/download/experiment/-3803766532162238398

1105 (Table 3.2). Skyline documents and libraries are publicly available at Panorama Public (https://panoramaweb.org/labkey/pecan-manuscript.url).

Chapter 4

APPLICATIONS OF PECAN

PECAN detects peptides from DIA data without a prerequisite library. Without the limitation of a library, one can fully leverage the discovery potential of DIA data using PECAN. In this chapter, I show various applications of PECAN, demonstrating its unique features. I also discuss the limitations of DIA data in some applications.

Much of this work is done in collaboration with Romain Huguet, Philip M. Remeš, and Vlad Zabrouskov from Thermo Fisher Scientific, San Jose, California.

4.1 Exploring the limits of DIA

DIA attempts to combine the benefits of DDA and targeted acquisitions. Depending on the purpose and design, DIA experiments balance between precursor selectivity, cycle time, and precursor range. These parameters are further influenced by the fragmentation method, ion injection time or dwell time, and resolving power of MS/MS analysis. The Orbitrap Fusion series mass spectrometer is a unique instrument because it contains a dual-pressure linear ion trap, a mass filtering quadrupole, and an Orbitrap detector. In addition, the Orbitrap Fusion Lumos is equipped with an electrodynamic ion funnel and a high capacity transfer tube that

together provide a bright ion source. The Orbitrap Fusion Lumos is versatile and enables diverse DIA methods.

Here, we use PECAN detection as a quantitative metric to evaluate peptide detection from DIA data acquired with various parameters. We demonstrate that resonance excitation CID improves peptide detection in DIA. Better precursor selectivity also improves peptide detection, even when it compromises the cycle time. Furthermore, we show that longer filling time and high resolution improves the MS/MS sensitivity, leading to more peptide detection from DIA.

4.1.1 *Charge state dependency of fragmentation methods*

Fragmenting multiple analytes together precludes DIA from tailoring collision energy for every analyte, a standard optimization in DDA and SRM. Therefore, charge state independent fragmentation is an attractive feature for DIA. Currently, the most used instruments for DIA are the TripleTOF series and Q-Exactive series, which are both only capable of fragmenting ions by beam-type collision-induced dissociation (beam CID) (Olsen et al., 2007). With beam CID, the activation energy needs to be tuned based on the precursor ion charge state for efficient fragmentation. In addition, it is harder to prevent multiple fragmentation events with beam CID as the ions must transit through the entire length of the fragmentation device. When acquiring DIA data with beam CID, a charge state must be assumed, and thus peptides of other charge states are likely to be fragmented with a non-optimal energy.

The Orbitrap Fusion series can fragment ions with beam CID or resonance excitation collision-induced dissociation (resonance CID) in the linear ion trap. Unlike

beam CID, resonance CID activates only a specific *m/z* range for fragmentation by resonance excitation. Once a fragment is produced it is no longer in resonance and will not be fragmented again. Thus, compared to the beam CID, the efficiency of resonance CID fragmentation is less dependent on precursor charge state, and the multiple fragmentation events produced by resonance CID are unlikely unless the fragment ions fall in the fragmentation target *m/z* range. In addition, resonance CID activation preserves more of the b-ion fragment series, which are particularly prone to multiple bond breakage events. However, resonance CID obliterates fragment ions that fall in the fragmentation target *m/z* range, which is as wide as the isolation window for DIA.

To determine whether beam CID or resonance CID is the better fragmentation method for DIA, we acquired 25 x 4 *m/z* DIA data on a HeLa digest using either beam CID or resonance CID, at activation target (AT) of charge state 1, 2 or 3. We used PECAN to query the Hela protein database against the data with y-ion mode for beam CID data and b- and y- ion mode for resonance CID. Overall, PECAN detected more peptides from data acquired with resonance CID (Figure 4.1 a). In addition, the coefficient of variances (CVs) of product over precursor area ratio is much smaller for resonance CID than beam CID (Figure 4.1 b). This result indicates that the fragmenting peptides with non-optimal AT charge states has a higher impact when using beam CID than resonance CID. In this application, PECAN provides a quantitative metric to compare the two fragmentation methods in the context of DIA acquisition. Together, resonance CID demonstrates reproducible fragmentation regardless of the activation targeted (AT) charge state, resulting in reproducible peptide detection.

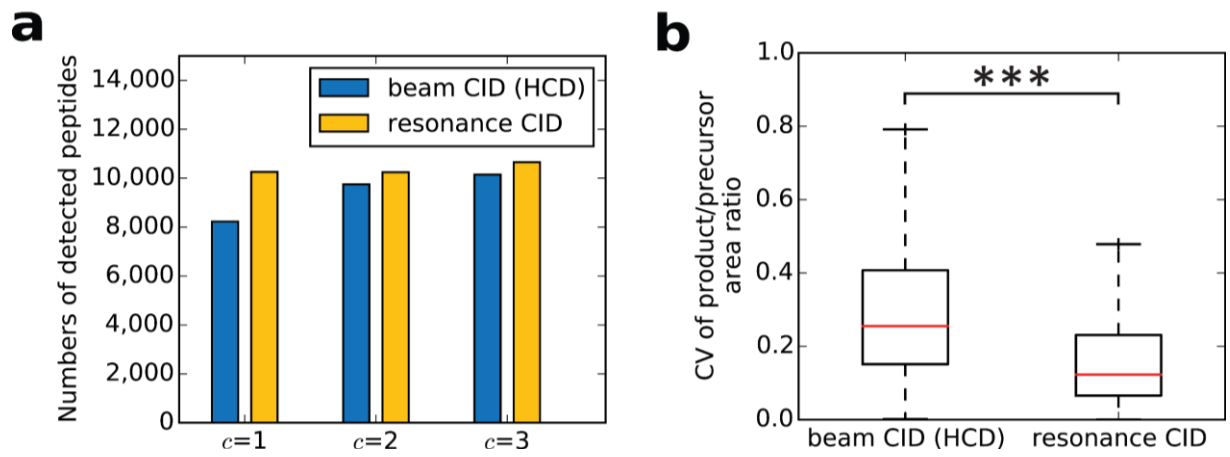

Figure 4.1 Charge state independent resonance excitation CID provides reproducible fragmentation for DIA

(a) Number of peptides detected from 25 x 4 m/z DIA data, acquired with either beam-type CID (i.e. HCD) or resonance CID, at activation target (AT) of either charge state c=1, c=2, or c=3. (b) Charge state dependency estimated with the reproducibility of fragmentation from different AT charge states. The coefficient of variances (CVs) of product over precursor area ratio are calculated per transition within the same fragmentation method between data collected at default AT of c=1, c=2, and c=3, from peptides detected in all three data. *** marks significant difference in two distributions with p-value < 0.01

4.1.2 *The tradeoff between precursor selectivity and cycle time*

Previously in 3.5, we have demonstrated that precursor selectivity has a direct impact on the sensitivity of PECAN detection. In addition to precursor selectivity, the cycle time of DIA MS/MS analysis could be an important factor in achieving high rates of peptide detection. DIA cycle time directly influences how many times an analyte is measured and how likely an analyte is missed from being sampled. While the rule of

thumb for peptide quantification is at least 7 measurements, it is unclear how the number of measurements impacts PECAN detection. Here, we discuss the tradeoff between precursor selectivity and cycle time.

Pecan provides an easy metric to compare these fragmentation approaches in the context of DIA acquisition. To compare peptide detection between DIA using resonance CID and beam CID, a series of DIA methods are tested on a HeLa digest. These methods use either 3, 4, 5, 8, or 10 *m/z*-wide isolation windows with 1 or 2 LC-MS/MS runs (i.e. injection) to cover 500-700 *m/z* precursor region. The resulting data is analyzed using PECAN for peptide detection (Figure 4.2). With every DIA method, resonance CID increased the number of peptide detections compared to beam CID. The improvement in peptide detection from resonance CID increases as the width of the isolation window increases. These improvements are despite an increase in cycle time when using resonance CID, which is more time-consuming than beam CID fragmentation.

There is a clear trend toward increasing peptide detections with increasing precursor selectivity. In the acquisitions using 2 LC-MS/MS runs per analysis, peptide detections increase by 36% moving from 5 *m/z* wide MS/MS scans to 2 *m/z* wide scans using resonance CID despite the associated ~150% increase in cycle time. The same trend is also observed in the data acquired using a single LC-MS/MS run per analysis, despite the extremely long cycle time (5.7 seconds) required to cover this wider precursor m/z range with 4 m/z wide MS/MS scans. We further compared the pair DIA methods that halved in cycle time with either 4, or 5 m/z-wide isolation windows. For 4 *m/z* and 5 *m/z* DIA, halving the cycle time increases by 10-20% the number of peptide detections for both fragmentation methods.

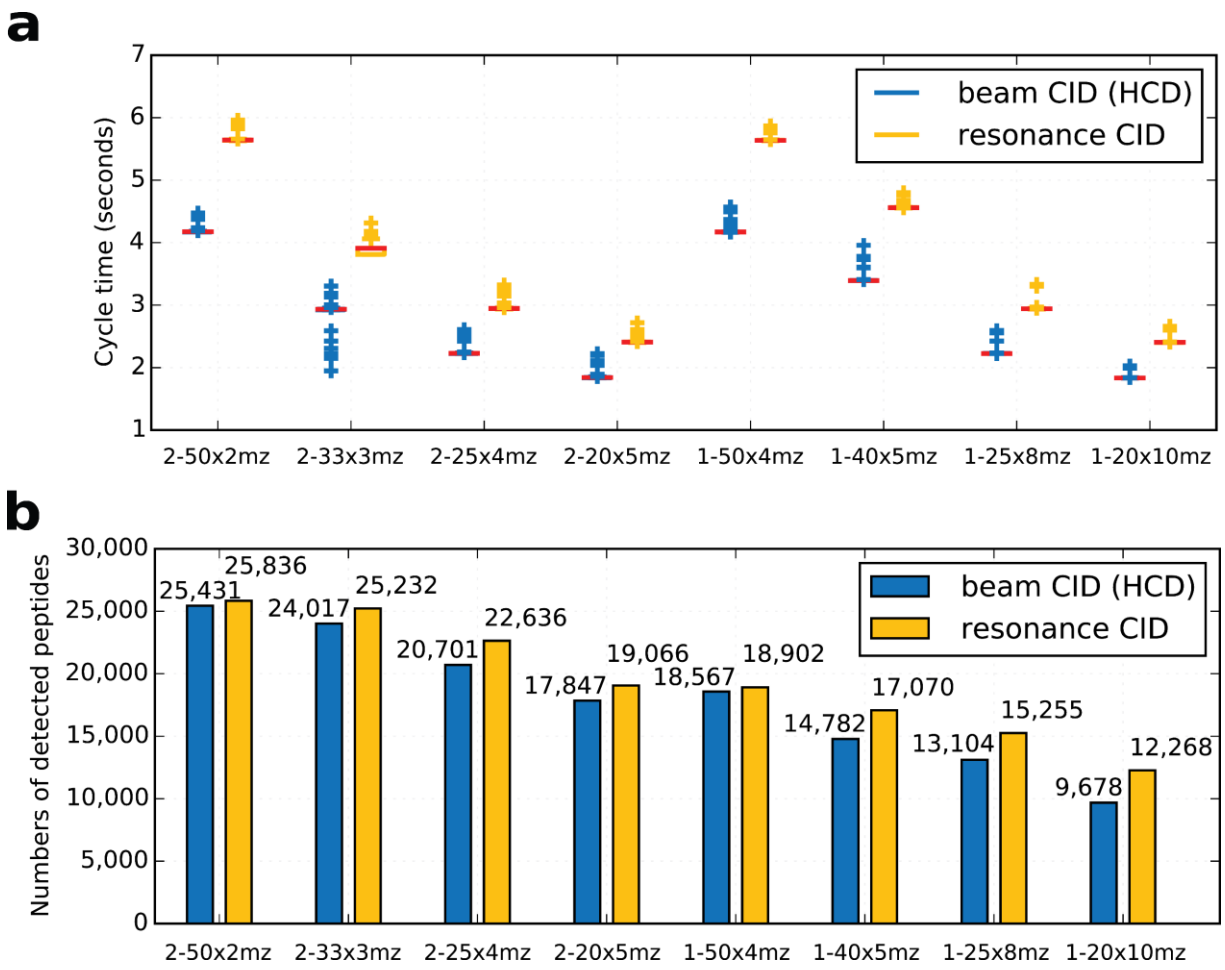

Figure 4.2 Increased precursor selectivity improves peptide detection
HeLa digest is analyzed with 8 DIA MS/MS methods, with various isolation window widths. DIA methods is labeled by the number of injections (1: one injection covers 500-*700* precursor *m/z;* or 2: two injections cover 500-600 and 600-700 precursor *m/z*), followed by the number of MS/MS events per DIA cycle x the width of the isolation windows. Color of the bar and box indicates the fragmentation method. (a) Box plot showing cycle time of each DIA method calculated from the average of the first 100 cycles of the data. (b) Number of peptides detected by PECAN from each DIA method.

4.1.3 *Ion filling time and resolving power for MS/MS analysis*

On Oribtrap instrumentation, the resolution of an MS/MS can be increased by increasing the transient acquisition time in the Orbitrap. For example, the Orbitrap Fusion can scan at roughly 20 Hz when acquiring MS/MS scans at 15,000 resolving power (RP), but at 10 Hz acquiring 30,000 RP MS/MS scans. While there is a trade-off in MS/MS scans, there is potential gain due to increased resolving power and fill times. On both the Q-Exactive and Orbitrap Fusion series, precursor ions are accumulated for the next MS/MS scan in parallel with mass analysis in the Orbitrap. Therefore, as resolving power increases, the amount of time that can be spent accumulating ions for MS/MS also increases, thus further improving sensitivity. PECAN was used to assess the increase in peptide detection sensitivity with 30,000 RP acquisition and a 55 ms maximum injection time (MIT) compared to 15,000 RP and a 20 ms MIT. Using multiple DIA methods, the higher resolving power accompanied with longer fill time MS/MS acquisition increased peptide detection by ~30-50% despite the associated increase in acquisition cycle time (Figure 4.3). Furthermore, analysis of MS spectra indicates that the precursor intensity of the detected peptides using the higher RP / MIT scans are significantly higher than the lower RP / MIT scans.

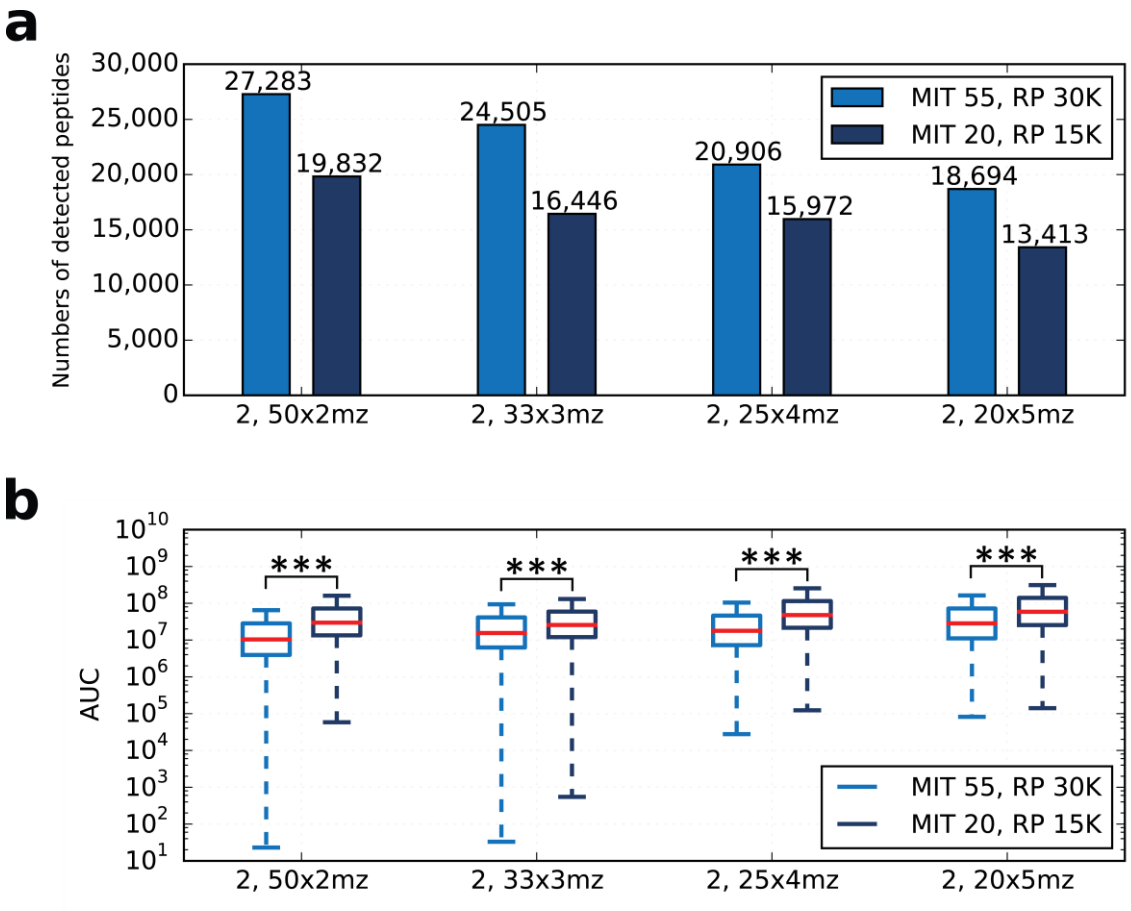

Figure 4.3 Longer filling time and higher resolving power increases peptide detection and MS/MS sensitivity

(a) Number of peptides detected by PECAN from each DIA method. The DIA methods acquired data with 50x2mz, 33x3mz, 25x4mz or 20x5mz isolation windows. (b) Box plots show the precursor abundance distribution of PECAN detected peptides in each DIA method. Precursor abundance of each detected peptide is calculated using the integrated area under the curve (AUC) of precursor ions [M], [M+1], and [M+2] extracted by Skyline. The integrated region is +/- 15 seconds of the PECAN detection. For each method two injections were used, covering 500-600, and 600-700 precursor m/z. Color of the bar indicates the maximum injection time (MIT) and resolving power (RP) settings for MS/MS analysis. *** marks significant difference in two distributions with p-value < 0.01

4.2 Building libraries from DIA data with PECAN

A common workflow for analyzing DIA data is to use a library to improve the sensitivity of detection. Typically, these libraries are generated from previous identifications derived from DDA data, which are inherently dependent on detectable precursors. One key disadvantage of relying on only DDA identification is the dependency on detectable precursor signal. Not only does DDA depend on detectable precursor signal to trigger MS/MS analysis, but most database searching tools also depend on precursor information to narrow down candidate peptides. Such dependency hinders the detection of low abundant analyte with detectable product signal but no detectable or highly interfered precursor signal. When using an Orbitrap mass analyzer, this phenomenon is particularly common as the MS1 intra scan dynamic range is limited by the ion capacity of the Orbitrap.

With PECAN, one can generate a library directly from DIA data. We demonstrated this capability by acquiring library data using twelve gas phase fractionation DIA runs, each with twenty-five 2-m/z-wide isolation windows, on a non-depleted, pooled plasma sample. We queried the data with the human UniProt Swiss-Prot database using PECAN and generated a detection library. From the library data, PECAN detected 3,689 peptides and 520 protein groups. We looked at the relative concentration of 248 plasma proteins gathered from the literature (Source: Leigh Anderson, The Plasma Proteome Institute, Washington, DC, USA) (Anderson and Anderson, 2002). Based on this reference, the DIA plasma library spans > 5 orders of magnitude of protein concentration (Figure 4.4). Note that some literature values are measurements for protein complex or specific fragments of the protein (e.g. values for Prothrombin and Fibrinogen alpha chain), of which the

intact protein concentration could be higher. Of the 3,689 detected peptides, 379 were not present in the PeptideAtlas Human Plasma spectral library (2012-08 release) constructed from 177 public datasets. We put the PECAN detections from the plasma library data on Panorama Public for the benefit of the community. Additionally, the fragmentation patterns and retention time information resulting from PECAN detection can be incorporated into existing libraries.

From the observation that the sensitivity of PECAN detection scales well with precursor selectivity (discussed in 3.5) follows a practical application that leverages gas phase fractionation to build libraries directly from DIA data. This approach can augment existing DDA-based libraries as evidenced by detection of hundreds of novel peptides from 12 LC-MS/MS runs that were not detected in over 100 DDA experiments in plasma. The detection directly from DIA data, and the ability to detect peptides with weak or undetectable MS1 signal could be drivers of these novel detections. Existing libraries may be extended even further by combining the DIA library approach with sample fractionation and/or depletion.

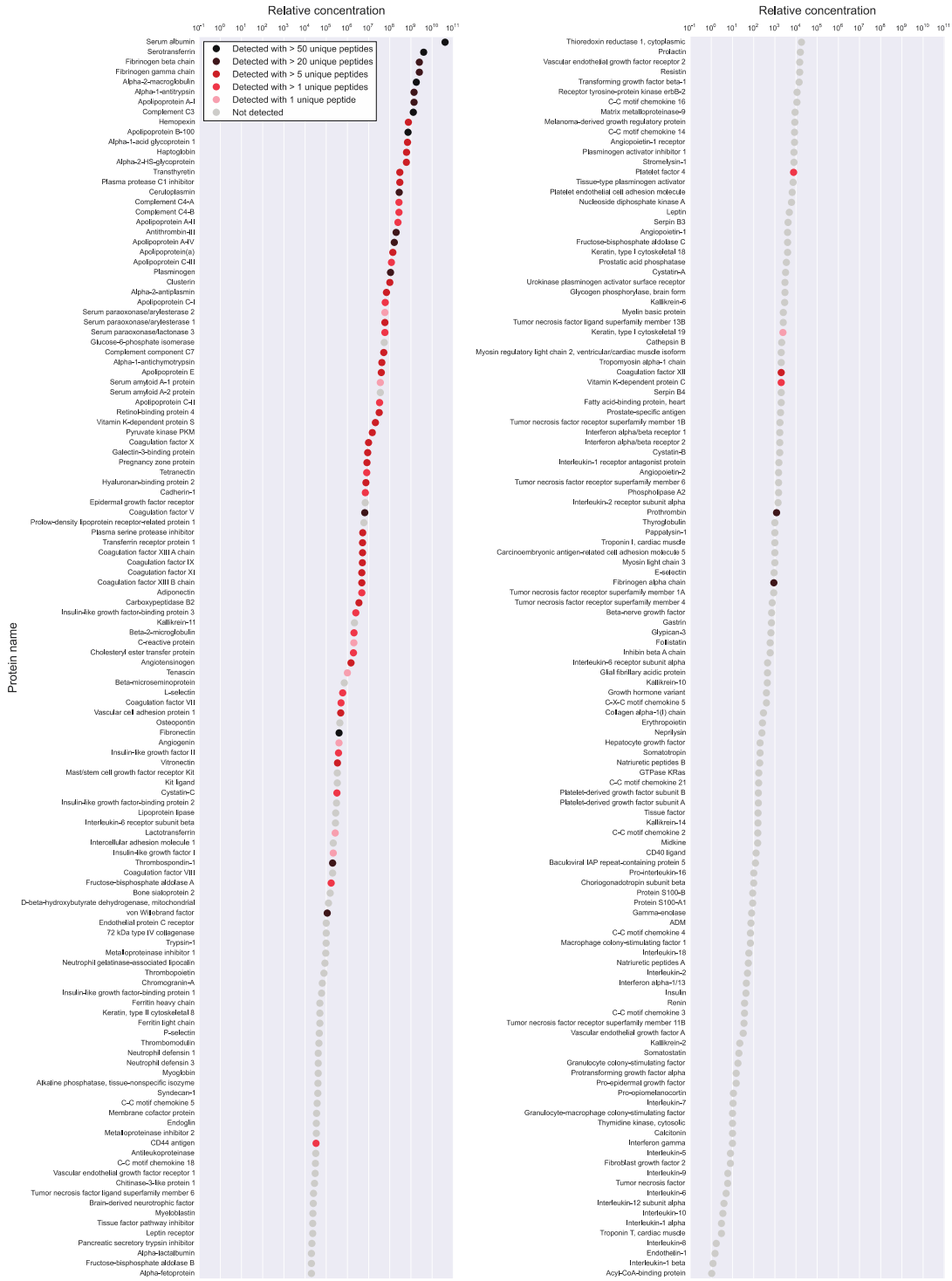

Figure 4.4 Dynamic range of DIA plasma library

Relative concentration values of 248 plasma proteins are taken from the literature. Color of the dot represents the number of peptides unique to the protein or only shared by its isoforms in the DIA plasma library.

4.3 Querying sequence variants with PECAN

For decades, genetics and genomics have focused on studying sequence variation and its influence on phenotype. Modern large-scale exome and genome sequencing projects have done much to expand the catalog of known sequence variation. Recently, the rapid growth of measured sequence variation as well as great improvements in mass spectrometry instrumentation and proteomic techniques have re-stimulated the development of proteogenomics (Na et al., 2016; Nesvizhskii, 2014; Payne et al., 2015). With PECAN, one can easily query novel peptides, such as variant containing peptides, against proteomics data. It is important to point out that PECAN makes proteomics data more intuitive to general researchers. This is because as a peptide-centric tool (see definitions in Chapter 2), PECAN practices direct hypothesis testing by asking if a peptide of interest is detected in the data. The combination of PECAN and data-independent acquisition represents an elegant approach to leverage known sequence variation for proteogenomics studies.

Here we demonstrate querying sequence variants with PECAN. 342 of the proteins detected by PECAN in the DIA plasma library are represented in the UniProt Swiss-Prot human natural variant database, and collectively contain a total of 4,264 single amino acid variants. Of the 4,264 variants, 3,714 resulted in at least one theoretical tryptic peptide specific to the variant and not present in the reference human UniProt Swiss-Prot database in the mass range of 600 to 4000 Da. We used PECAN to query these variant-specific, tryptic peptides from 3,714 variants against the plasma library data. PECAN detected 133 variant-specific peptides, corresponding to 115 variants (Appendix A). In some cases, PECAN detected multiple variant-specific peptides resulting from the same

sequence variant. For instance, in Serotransferrin (Figure 4.5 a), two variant-specific peptides were detected for the variant Ile448Val while no canonical peptide spanning Ile448 was detected. In addition, three variant-specific peptides were detected for Pro589Ser, of which two were from the introduction of a trypsin cleavage site by the variant.

It is important to know that mass spectrometry data itself may not be sufficient to conclusively demonstrate the presence of some sequence variants. For example, some variants, such as leucine to isoleucine, are identical in mass and indistinguishable by the mass spectrometer. Other variants, such as glutamic acid to lysine, shift the peptide mass so little that the canonical and variant peptide ions will likely be isolated and fragmented together, resulting in similar MS/MS spectra. In this case, depending on the variant position relative to the peptide N-terminus, two peptide ions may share most of the y-ions. For example, among the three glutamic acid to lysine variants in Apolipoprotein A-1, only Glu160Lys has the definitive peptide resulting from cleavage at the tryptic site introduced by the variant. (Figure 4.5 b). In addition, PECAN treats each peptide and peptide variant independently. Thus, when the data provides strong evidence, PECAN does not further filter detections even if multiple forms of the same peptide are detected. In the example of Glu134Lys and Gly222Lys of Apolipoprotein A-1, the same MS/MS spectra provided evidence of detection with q-value < 0.01 for both canonical and variant peptides.

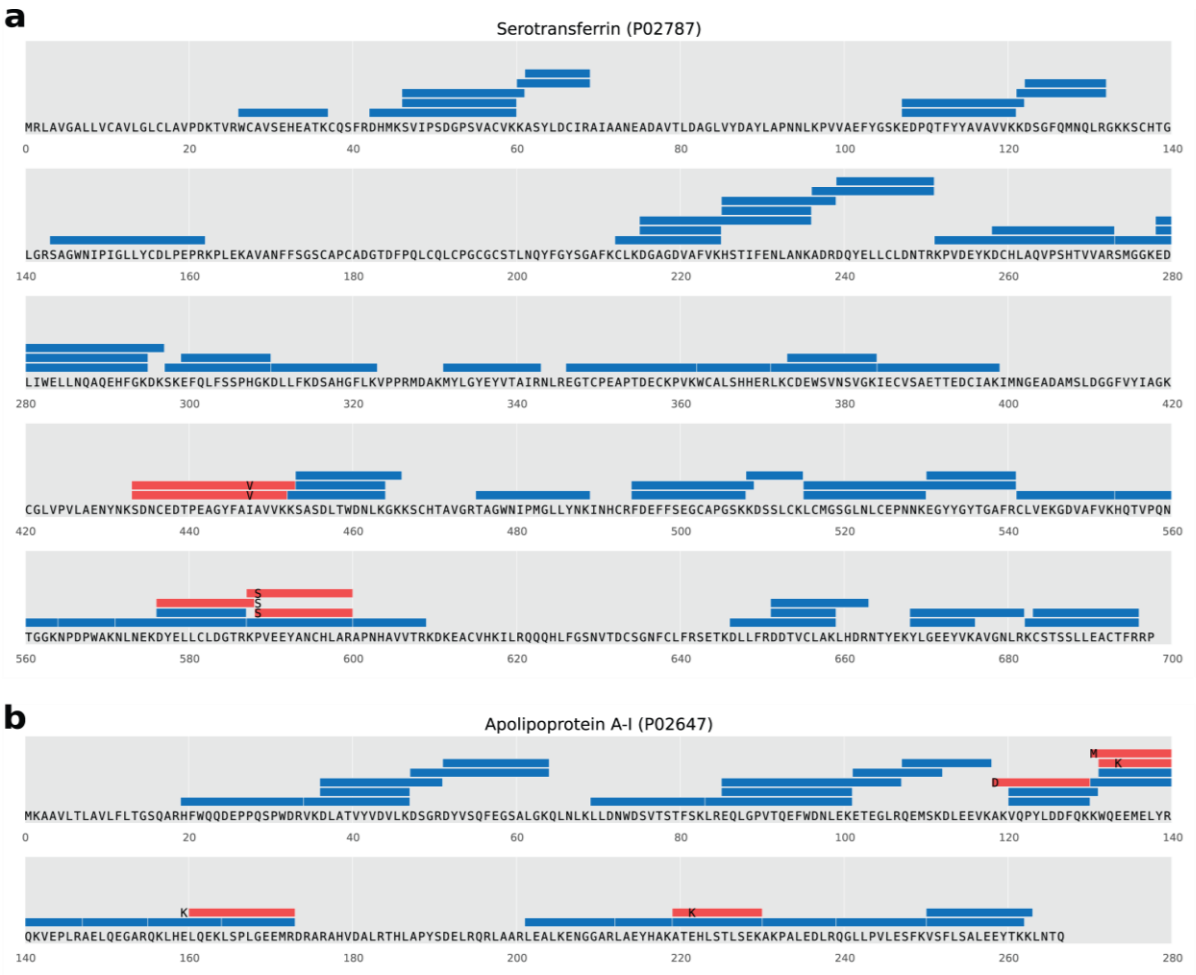

Figure 4.5 Natural variants in the plasma library data

Full-length canonical sequences of Serotransferrin (a) and Apolipoprotein A-1 (b) are obtained from the human UniProt Swiss-Prot database, accession number P02647 and P02787, respectively. Blue boxes represent PECAN detected peptides from the plasma library data when queried with canonical sequences. Red boxes represent PECAN detected variant-specific peptides from the plasma library data when queried with variant-specific tryptic peptides from 3,714 variants.

4.4 Materials and methods

4.4.1 *HeLa datasets*

HeLa protein digest (Thermo) spiked-in with stable-isotope labeled peptides (PRTC, Thermo) was analyzed on an Orbitrap Fusion™ Lumos (Thermo) mass spectrometer. For each injection, one μg of HeLa peptides and 40 fmol of PRTC were loaded to a 25-cm-long reversed phase EASY-Spray™ C18 column (Thermo) prepacked with 2-μm Spherical, and separated with a 60-min linear gradient liquid chromatography (LC). LC was performed using an EASY-nLC™ 1200 system (Thermo) set to a flow rate of 250 nl/min during linear gradient with Buffer A (2% ACN, 0.1% formic acid and 97.9% water) and Buffer B (99.9% ACN and 0.1% formic acid). The gradient was followed by a wash at 80% Buffer B and a column re-equilibration at 2% Buffer B.

Two gas-phase fractionation (GPF) settings were used to cover the precursor *m/z* range of 500 to 700: one injection, or two injections covering 500-600 and 600-700 *m/z*. The isolation ranges of MS analysis for all GPF settings correspond to the precursor range covered in each injection with SIM scans. MS scans were acquired by the Orbitrap mass analyzer with 60,000 resolution and 20 ms of maximum injection time (MIT). We used date independent acquisition (DIA) MS/MS methods with various isolation window width w, each contains n non-overlapping isolation windows so that $n \times w$ covers either the 200 or 100 precursor *m/z* range. MS/MS scans were acquired by the Orbitrap mass analyzer with 30,000 resolution and 55 ms of MIT unless specified otherwise. For FDR control, data from multiple injections were analyzed together as if they were from one instrument run.

4.4.2 *Plasma library data*

Non-depleted plasma samples from five deidentified donors and a normal female plasma standard (Lampire Biological Laboratories) were individually digested. Plasma samples were diluted 200-fold prior to digestion with a diluent containing heavy labeled protein and peptide standards and PPS silent surfactant in 50 mM ammonium bicarbonate. Post dilution, the sample contained 1 ng/uL 15N-labeled human Apolipoprotein A-1 (Cambridge Isotope Laboratories), 2.5nM heavy lysine labeled GST peptides, and 0.1% PPS silent surfactant (Protein Discovery). Each diluted plasma sample was boiled at 95°C for 5 minutes to denature proteins. After denaturing, dithiothreitol (DTT, Sigma Aldrich # D0632) was added to a final concentration of 5 mM and samples incubated at 60 °C for 30 minutes to reduce disulfide bonds. Iodoacetamide (Sigma Aldrich # I1149) was then added to a concentration of 15 mM followed by a 30-minute room temperature incubation in the dark to alkylate reduced cysteine. The alkylation reaction was quenched by addition of DTT to a final concentration of 10mM added. Sequencing grade trypsin (Pierce # 1862748) was added to a 1:10 trypsin to protein ratio followed by sample incubation at 37 °C / 1200 RPM for 4 hours to digest proteins. The digestion reaction was quenched by addition of hydrochloric acid to a final concentration of 9.4 mM. The resulting digests of equal volume were pooled to make the plasma library sample.

Chromatography was performed using a nanoACQUITY (Waters) system set to a flow rate of 250 nl/min during linear gradient with Buffer A (2% ACN, 0.1% formic acid and 97.9% water) and Buffer B (99.9% ACN and 0.1% formic acid). A homemade 2-cm-long 150-µm I.D. trapping column was used prior to a self-packed 30-cm-long 75-µm I.D.

PicoFrit resolving column (New Objective) for a 90-min linear gradient from 2% to 35% Buffer B. Both trapping and resolving columns were packed with 3-μm ReproSil-Pur C18 AQ (Dr. Maisch GmbH). The gradient was followed by a wash at 80% Buffer B and a column re-equilibration at 2% Buffer B.

Twelve gas phase fractionations were used in acquiring the DIA plasma library data. Together, the precursor range of 400-1000 *m/z* was analyzed, where each fractionation covered a 50 *m/z*-wide portion of the precursor range: 400-450, 450-500, 500-550, 550-600, 600-650, 650-700, 700-750, 750-800, 800-850, 850-900, 900-950, or 950-1000. One μg of plasma sample, 50 fmol of PRTC, and 2.8 ng N15-APO-A1 were loaded in each injection, separated with a 90-min linear gradient LC, and analyzed on a Q-Exactive HF mass spectrometer. For each fractionation, DIA method cycled with 25 non-overlapping 2 *m/z*-wide isolation MS/MS scans (at 30,000 resolution), one 50 *m/z*-wide MS scan (at 30,000 resolution), and one 600 *m/z*-wide MS scan that covers 400-1000 *m/z* (at 15,000 resolution). The MS spectra with 400-1000 *m/z* precursor range were stripped from mzML files prior to PECAN analysis.

4.4.3 *Peptide detection*

Two sequence databases were used in this chapter: the Hela protein database that contains 18,926 previously identified protein isoforms, and the human UniProt Swiss-Prot database that contains 42,128 protein isoforms. All raw files were converted to mzML format by ProteoWizard (Chambers et al., 2012). PECAN (v. 0.9.9) was used to query peptides from the target database. Only fully tryptic peptides with up to one missed cleavage sites were queried, with a fixed modification of carbamidomethyl cysteine.

PECAN was set to query peptides with 2+ or 3+ precursor ion charge states, and in y-ion mode where only product y-ion series were considered. A ±10 ppm mass error tolerance was used for both precursor ions and fragment ions. Percolator (Käll et al., 2007) (v. 2.08.01) was used to separate targets and decoys. All detected peptides are reported by Percolator at the peptide level with q-value < 0.01, and detected proteins are reported by Percolator's built-in Fido algorithm (Serang et al., 2010) with q-value < 0.01, unless indicated otherwise.

Chapter 5

CHARACTERIZATION OF MURINE HEART PROTEOME WITH SS-31 TREATMENT

This work is done in collaboration with Ying Ann Chiao and Peter Rabinovitch from the Department of Pathology, University of Washington, Seattle.

Abstract

Aging results in decline in cardiac function and is an independent risk factor for cardiovascular diseases, including heart failure. One of the key components in the development and progression of heart failure is mitochondrial oxidative damage. Previously we have demonstrated attenuated cardiac aging phenotypes by reducing mitochondrial oxidative damage with two methods: transgenic overexpression of mitochondrial catalase (mCAT) and treatment of mitochondrial protective SS-31 peptide. Although SS-31 treatment shows similar benefits as mCAT overexpression, our preliminary proteomics analysis showed distinct proteome changes affecting multiple pathways. To further understand the mechanisms of SS-31 intervention on cardiac aging, we designed a comprehensive proteome study using data independent acquisition (DIA) on an old mouse cohort.

5.1 Introduction

Aging results in decline in cardiac function and is an independent risk factor for cardiovascular diseases, including heart failure, one of the leading causes of morbidity and mortality worldwide. According to the American Heart Association, Americans 60-79 years of age have a 72% prevalence rate of cardiovascular disease and Americans over 80 years of age have over 80% prevalence rate (Members et al., 2012). In 2016, heart failure incidence approached 10 per 1000 population after 65 years of age (Mozaffarian et al., 2015). Over three million Americans suffer from congestive heart failure, the most common reason for hospital admission in older individuals (Tecce et al., 1999). Aging has a remarkable effect on the heart and arterial system, leading to an increase in prevalence of cardiovascular diseases including atherosclerosis, hypertension, myocardial infarction, and stroke (Lakatta and Levy, 2003). As the average lifespan of humans increases with modern medicine, the percentage of people entering the 65 and older group will continue to grow. In all likelihood, cardiovascular disease will remain the leading cause of death, and the cost associated with treatment will continue to increase.

No intervention has yet been developed to treat or prevent cardiac aging. In addition, the underlying mechanisms of cardiac aging are not well established even though the phenotypes of cardiac aging, such as hypertrophy, altered left ventricular (LV) diastolic function, diminished LV systolic reverse capacity, increased arterial stiffness, and impaired endothelial function, are well-characterized (North and Sinclair, 2012). Many studies have demonstrated that oxidative stress caused by accumulation of reactive oxygen species (ROS) and mitochondrial dysfunction are two important factors contributing to cardiac aging and the development of heart failure (Murdoch et al., 2006;

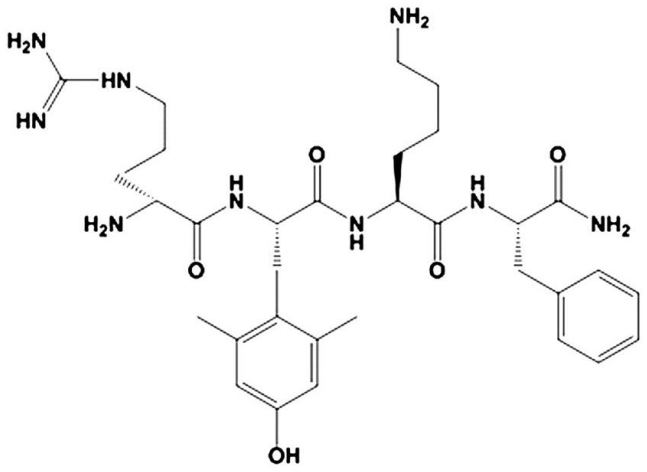

Figure 5.1 Structure of mitochondria targeted peptide SS-31
The Szeto–Schiller tetrapeptide SS-31 (D-Arg-2',6'-dimethyl-tyrosine-Lys-Phe-NH2) contains an aromatic–cationic sequence motif that specifically enables them to be delivered to mitochondria.

Rosca and Hoppel, 2013; Sawyer et al., 2002). In mice, transgenic overexpression of catalase, a ROS scavenger targeted to the mitochondria (mCAT), attenuate the cardiac aging in mice (Dai et al., 2009). The protective effects of mCAT from cardiac aging result in substantial changes in the cardiac proteome, including mitochondrial proteins (Dai et al., 2012). Although genetically engineered mouse models, such as mCAT, provide great systems to study cardiac aging, they offer very little potential to translate into treatments.

Recently, a mitochondria-targeted tetrapeptide, SS-31 (Figure 5.1), has been shown to improve skeletal muscle and cardiac function in aged mice (Dai et al., 2011; Siegel et al., 2013; Szeto, 2014). SS-31 is a member of the Szeto–Schiller (SS) family of tetrapeptides that contain an aromatic–cationic sequence motif that specifically enables them to be delivered to mitochondria, where they localize to the inner mitochondrial

membrane (IMM) with an approximate 1000–5000-fold accumulation (Zhao et al., 2004). Despite the functional improvements associated with this treatment, the underlying mechanisms of SS-31 in this intervention are not well understood. Originally thought to be an antioxidant attributed to the dimethyltyrosine residue, SS-31 was thought to scavenge hydrogen peroxide and peroxynitrite and inhibit lipid peroxidation (Szeto, 2006). Most recently, SS-31 has been shown to interact with cardiolipin and improve the electron carrying function of cytochrome c, while reducing its peroxidase activity (Szeto, 2014). Deeper understanding of the mechanisms behind the cardiac function improvement produced by SS-31 and the impacts of SS-31 on mitochondrial dysfunction is much needed.

Although SS-31 treatment shows similar benefits to cardiac function as mCAT overexpression, our preliminary proteomics analysis shows distinct proteome changes affecting multiple pathways. To further understand the mechanisms of SS-31 intervention on cardiac aging and mitochondrial dysfunction, we designed a comprehensive proteome study using data independent acquisition (DIA) on an old mouse cohort.

5.2 SS-31 improves cardiac functions in old mice

We used osmotic minipumps to deliver SS-31 or saline to 24-month-old mice for eight weeks. Cardiac function was measured longitudinally at 0-, 4- and 8-week time points by echocardiography. At the 8-week endpoint, hearts were harvested for biochemical assays and proteome analysis.

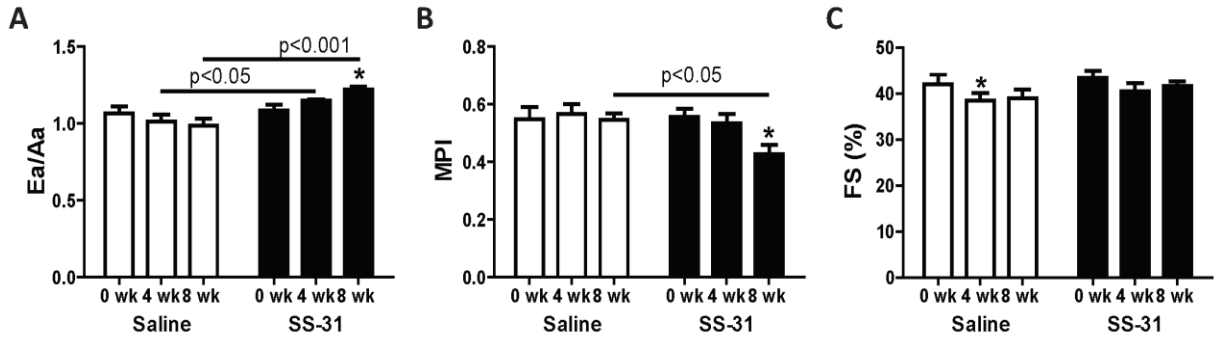

Figure 5.2 SS-31 treatment reverses age-related decline in cardiac function

Echocardiography of 14 mice form 0-, 4-, and 8-weeks of saline or SS-31 treatment (n=7/group). Treatment related changes in (A) diastolic function (Ea/Aa) measured by tissue Doppler imaging of the mitral annulus, (B) myocardial performance (myocardial performance index, MPI), and (C) fractional shortening (FS) percentage. For Ea/Aa and MPI, SS-31 treated mice show significant ($p<0.05$) linear regression over time. * marks significant ($p<0.05$) change to the corresponding 0-week baseline.

To determine the effect of SS-31 treatment on cardiac aging, we compared the cardiac function measured from fourteen age-matched, wild-type (WT) mice, of which seven were treated with saline and seven with SS-31. We found significant SS-31 treatment-dependent linear trends for improved diastolic function with improved Ea/Aa (Figure 5.2 a), and enhanced myocardial performance with decreased myocardial performance index (MPI) (Figure 5.2 b). The fractional shortening (FS) was not altered between two treatment groups, indicating no deleterious effects to global left ventricular

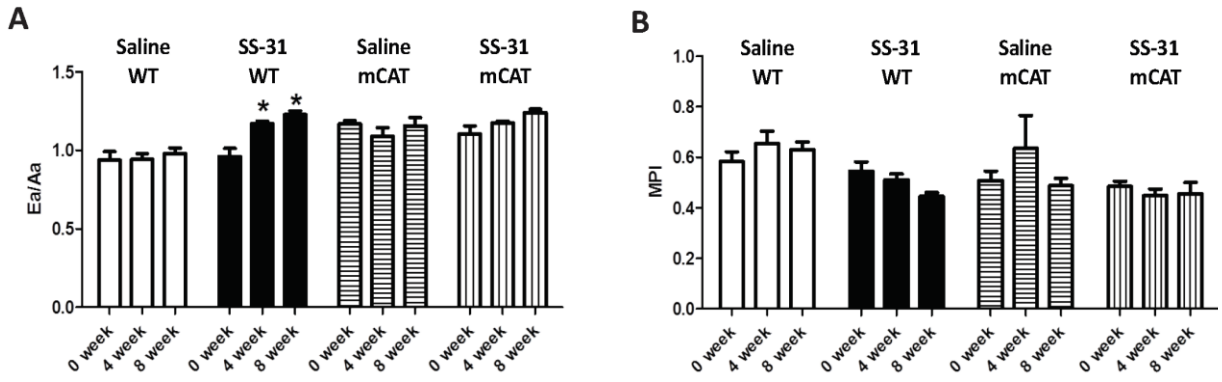

Figure 5.3 SS-31 treatment and mCAT overexpression are not additive in cardiac function improvement

Doppler echocardiography showed that SS-31 treatment was not additive to the (A) diastolic function (Ea/Aa) and (B) myocardial performance (MPI) improvement conferred by mCAT overexpression. * marks significant ($p<0.05$) change to the corresponding 0-week baseline.

(LV) systolic function from the SS-31 treatment. (Figure 5.2 c). These results show that SS-31 reverses the age-related decline in cardiac function in old WT mice.

To compare the effects of SS-31 and mCAT overexpression on cardiac aging, a separate cohort of old, age-matched mCAT and WT mice were treated with SS-31 for eight weeks. Cardiac function of the saline-treated or the SS-31-treated, mCAT or WT age-matched mice was monitored at 0-, 4-, and 8-weeks of treatment. While SS-31 treatment showed significant improvement in diastolic function (Ea/Aa) for WT mice at 4-, and 8-week time point, it did not show further improvement for mCAT mice (Figure 5.3). In other words, the 8-week SS-31 treatment and mCAT overexpression are not additive in cardiac function improvement.

These results indicate that the effects of SS-31 and mCAT in cardiac aging could be overlapping. Such overlap could result from shared pathways, common downstream substrates, or from different mechanisms that share the same target organelle – mitochondria. Alternatively, it is possible that the targets for the mechanisms of SS-31 are absent from the mCAT mice to begin with, considering that mCAT overexpression has a life-long effect on mice. While the primary role of mCAT overexpression is to scavenge ROS from mitochondria, SS-31 could act as an antioxidant that directly scavenges ROS , or as a stabilizer that interacts with cardiolipin (Zhao et al., 2004). To further understand the role of SS-31 in cardiac aging, below we show in-depth proteome analysis on the proteome in the heart.

5.3 Mitochondrial proteome profiling in heart

Because both mCAT and SS-31 are known to selectively target the inner mitochondrial membrane (IMM) and interact with the mitochondrial function, we first profile the mitochondrial proteome. We interrogated the whole heart proteome data with 1,265 mitochondrial proteins from the mouse MitoP2 database. From the 6,333 unique peptides detected by PECAN, 5,560 peptides were quantified across all 22 samples, representing 758 mitochondrial proteins (see details in 5.7.5).

Differential quantification analyses showed 22 peptides with significant fold changes (with p-value < 0.05 and larger than a two-fold change) from the saline-treated to the SS-31-treated WT mice, five peptides from the saline-treated to the SS-31-treated mCAT mice, 101 peptides from the saline-treated mCAT to the WT mice, and 48 peptides

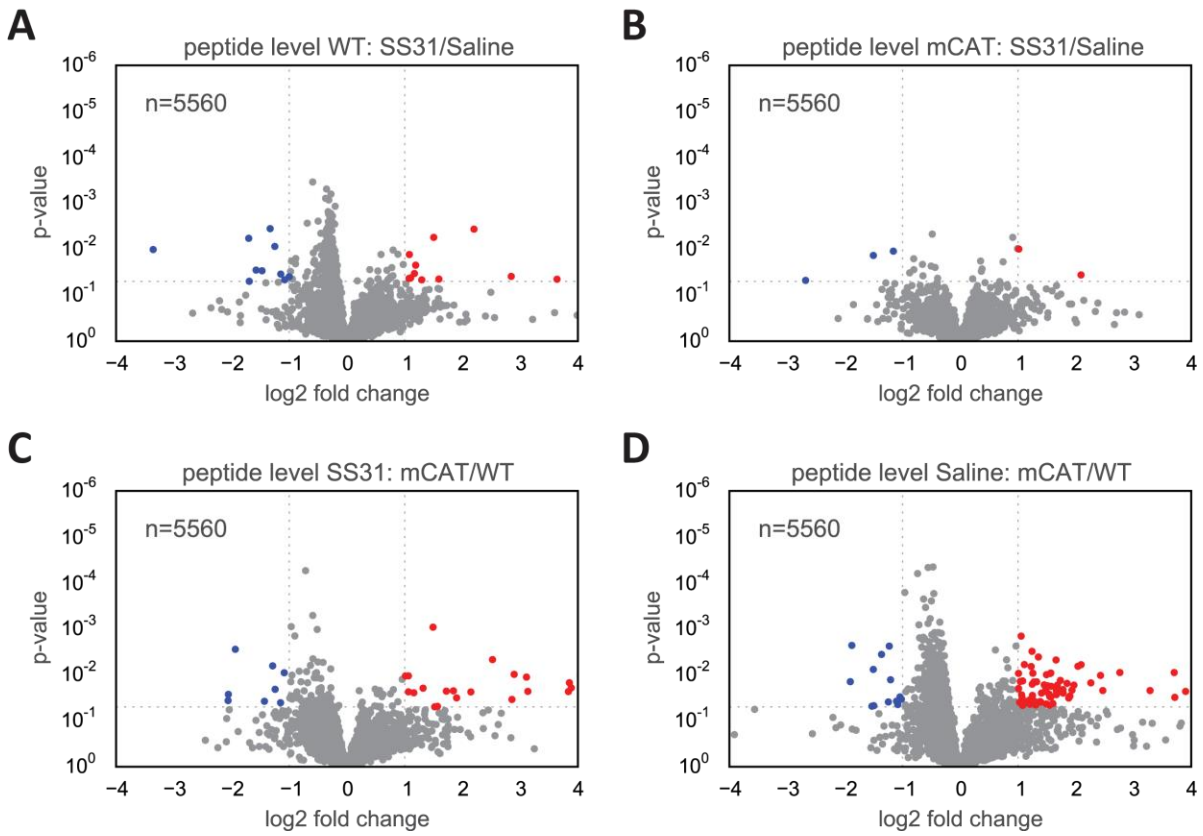

Figure 5.4 Differential quantification of peptides from mitochondrial proteins

Volcano plot representation of the fold changes from 5,560 peptides between (A) SS-31 and saline treated WT, (B) SS-31 and saline treated mCAT, (C) SS-31 treated mCAT and WT, and (D) saline treated mCAT and WT. Peptides with significant group difference (p-value < 0.05) and with log2 fold-change > 1 or < -1 are colored in red or blue, respectively.

from the SS-31-treated mCAT to the WT mice (Figure 5.4). With normalization to the median of the detected peptide area, these fold changes of peptides between two groups were calculated relative to the mitochondrial proteome. As expected, life-long mCAT

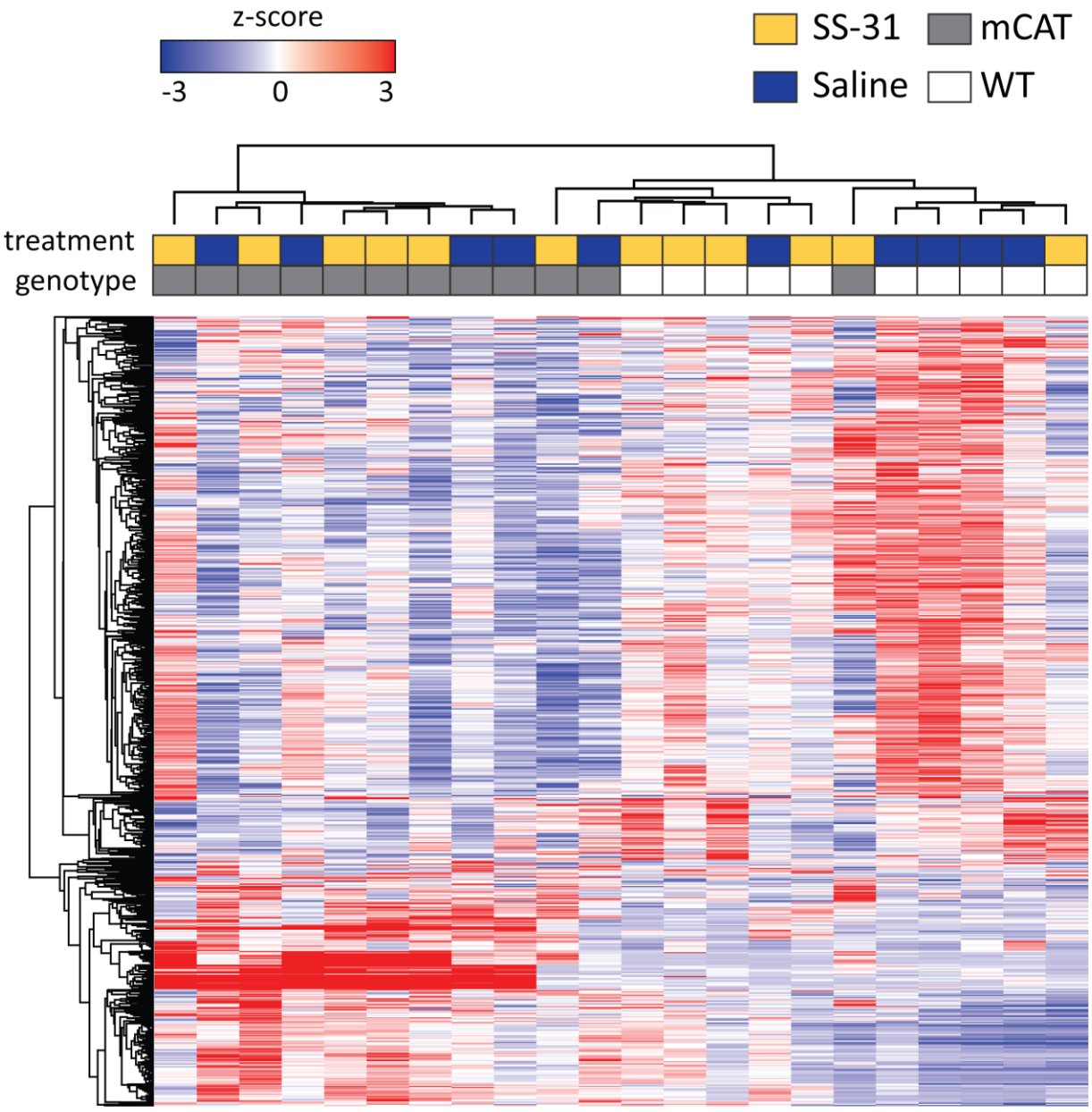

Figure 5.5 Proteome signature of the interactions of the treatment and the genotype

Hierarchical clustering of 703 peptides with significant interactions (p-value < 0.05) of the treatment and the genotype determined by two-way ANOVA. Each row is a peptide and each column is a sample, labeled with treatment (SS-31 or saline) and genotype (mCAT or WT). The values are normalized peptide area with z-score transformation using row median and median absolute deviation.

showed broader changes in the mitochondrial proteome than the 8-week treatment of SS-31. In addition, the 8-week treatment of SS-31 resulted in more changes to the WT mitochondrial proteome than to the mCAT one, consistent with the observation that the improvements produced by SS-31 and mCAT in cardiac function were not additive (Figure 5.3).

We used two-way analysis of variance (ANOVA) to investigate the relative abundance of peptides related to the interaction of the treatment (i.e. SS-31 or saline) and the genotype (i.e. mCAT or WT). Of 5,845 peptides tested, 703 peptides showed significant interactions (p-value < 0.05) of the two independent variables, meaning that the effect of the treatment differs between mCAT and WT. These 703 peptides showed visibly distinct signature for mCAT and WT, and a strong signature for different treatments in WT mice (Figure 5.5). Within the mCAT group, no signature was observed that could differentiate the SS-31 from the hearts of the saline-treated mice. Because all mice were old-aged in this study, the inter-individual variance was much higher than that of a cohort of young mice with an identical genetic background. Thus, the "noise" in the proteome signature is expected even after normalization (see details in 5.7.5). However, the mitochondrial proteome from some hearts (e.g. the leftmost one in Figure 5.5) exhibited signatures that resembled that of the other genotype and the opposite treatment. Further investigation is needed to understand these phenomena.

5.4 Global proteome profiling in heart

To understand the global proteome changes in the heart, we interrogated the whole heart proteome data with the 16,831 proteins from the mouse UniProt Swiss-Prot database. From the 29,971 unique peptides detected by PECAN, 28,247 peptides were quantified across 22 samples, representing 6,267 proteins. Differential quantification analyses showed 133 peptides and 30 proteins with significant fold changes (with p-value < 0.05 and larger than a two-fold change) from the saline-treated to the SS-31-treated WT mice, 72 peptides and 19 proteins from the saline-treated to the SS31-treated mCAT mice (Figure 5.6).

The number of proteins with significant fold changes was much lower than the number of proteins represented by the peptides with significant fold changes. This is largely because of the challenges in determining the proteoform of a protein with shotgun proteomics (Smith et al., 2013). A single protein coding gene can give rise to hundreds of forms of protein when considering endogenous proteolysis, RNA splicing, nonsynonymous single-nucleotide polymorphisms (SNPs), and post-translational modifications (PTMs). With shotgun proteomics, the information about proteoforms is degenerate because a peptide could come from multiple proteoforms (as discussed in 1.3.1). While measuring the changes of a peptide is straightforward, inferring the changes of a protein is much more complicated. Here, the relative abundance of a protein was represented by the sum of the peptide areas from the detected peptides that are unique to the protein (see details in 5.7.5). This conventional approach excluded the peptides that could be derived from more than one protein coding gene, even when there was no

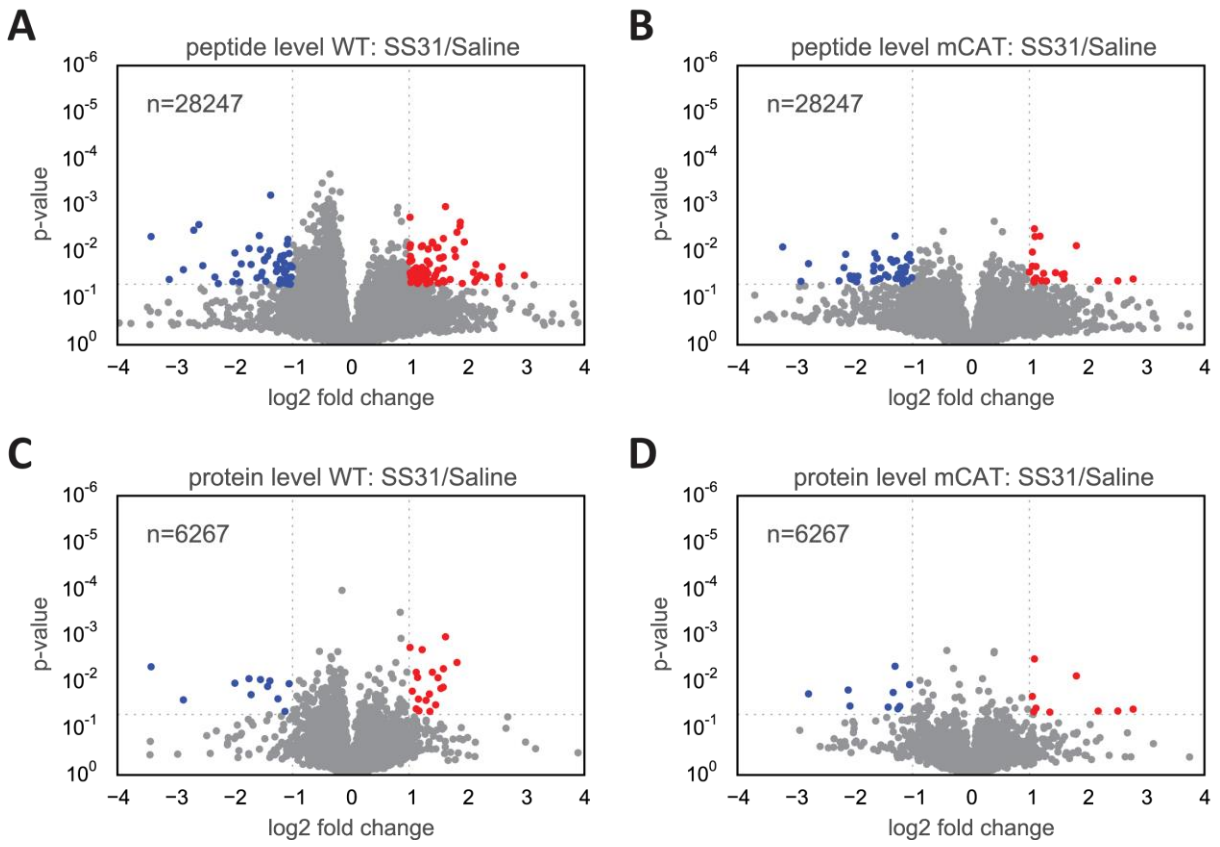

Figure 5.6 Differential quantification from the global proteome profiling in heart

Volcano plot representation of the fold changes from 28,247 peptides and between (A) SS-31 and saline treated WT, (B) SS-31 and saline treated mCAT, and the fold changes from 6,267 proteins between (C) SS-31 and saline treated WT, (D) SS-31 and saline treated mCAT. Peptides or proteins with significant group difference (*p*-value < 0.05) and with log2 fold-change > 1 or < -1 are colored in red or blue, respectively.

evidence of expression for the other gene(s), e.g. no unique peptide was detected. In addition, it is possible that two peptides of one protein changed in different directions as

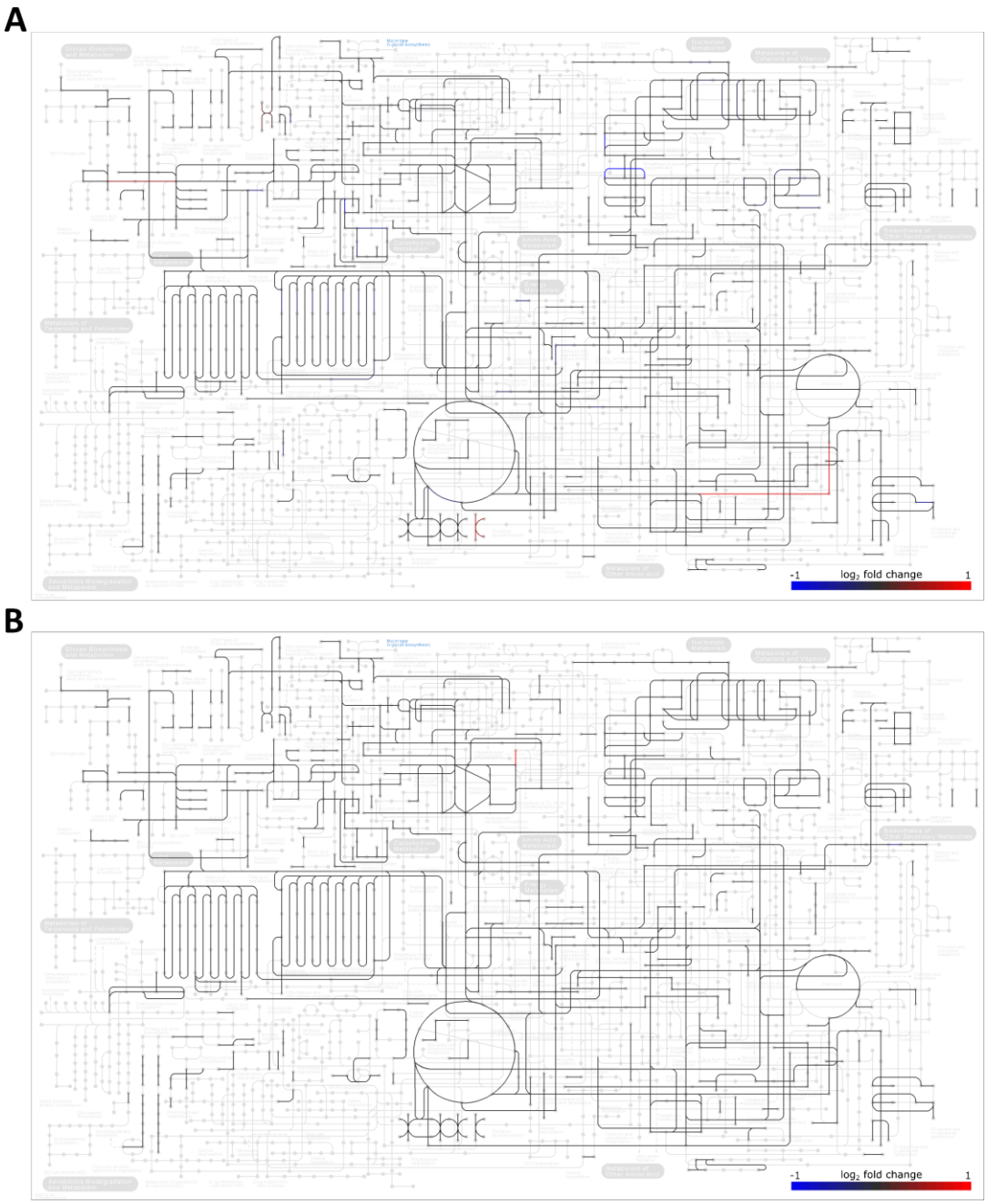

Figure 5.7 The effects of SS-31 treatment to the metabolic pathways

641 of the 1,294 KEGG genes in the mouse metabolic pathways were mapped by the proteins from global proteome profiling. The fold change from the saline-treated to the SS-31-treated of each protein was calculated from the normalized protein area of the (A) WT mice or (B) mCAT mice. Proteins with significant (p-value < 0.05) log2 fold-change are colored, and those with log2 fold-change ≥ 1 or ≤ -1 are colored in red or blue, respectively. Protein with insignificant (p-value ≥ 0.05) log2 fold-change are shown in black.

a result of one peptide being post-translationally modified. In other words, the dynamic of the unmodified peptides within a protein could average out the changes observed in individual peptides. In this case, selecting peptides with consistent changes to represent a protein should reduce the interference caused by PTMs or other factors, such as accessibility of the peptides from membrane vs. soluble domains.

While most of the proteome in the heart remained unchanged, the impact of the SS-31 treatment on the proteome was different between WT and mCAT mice. For instance, the proteome changes resulting from the SS-31 treatment in the metabolic pathways were noticeably different in WT and mCAT mice (Figure 5.7). Under the life-long influence of mCAT, the molecular and physical phenotypes of the heart are very different from the WT mice. Considering the broad changes in the mitochondrial proteome resulting from the life-long mCAT itself (Figure 5.4 D), the effects of SS-31 on the heart proteome is subsequent to the effects of mCAT. Thus, such effects are expected to be different in the WT and mCAT mice. Additionally, the improvements in cardiac function from SS-31 and from mCAT were not additive (Figure 5.3). Thus, to establish a better understanding in the mechanisms of SS-31 for improving cardiac function, we focus on the proteome changes from the SS-31 treatment in the WT mice hereafter.

5.5 Impacts of SS-31 on the whole heart proteome in WT mice

5.5.1 *SS-31 impacts metabolic pathways*

In WT, the SS-31 treatment resulted in a significant decrease of dCTP pyrophosphatase 1 (dCTPase 1), the protein that hydrolyzes and converts deoxycytidine

triphosphate (dCTP) and deoxycytidine diphosphate (dCDP) to deoxycytidine monophosphate (dCMP) (Figure 5.7). dCMP is used as a substrate to form dCDP which upon phosphorylation to dCTP supports DNA biosynthesis. A significant decrease in dCTPase 1 compared to the saline-treated group should result in increased dCTP, suggesting that the SS-31 treatment may increase the DNA biosynthesis. In addition, the SS-31 treatment resulted in a significant increase of delta-1-pyrroline-5-carboxylate synthase (P5CS), a protein from the aldehyde dehydrogenase family that converts glutamate to glutamate 5-semialdehyde (Figure 5.7). Glutamate 5-semialdehydan is an intermediate in the biosynthesis of proline and arginine, two key amino acids in peptide biosynthesis. A significant increase in P5CS compared to the saline-treated group should result in increased glutamate 5-semialdehydan, suggesting that the SS-31 treatment may increase the peptide biosynthesis.

5.5.2 *SS-31 increases production of phosphatidic acids*

In SS-31-treated WT mice, we observed a significant increase of lysophosphatidic acid acyltransferase (LPPAT) and diacylglycerol kinase (DAGK), and a significant decrease of cytosolic phospholipase A2 (PLA2) compared to the saline-treated WT mice (Figure 5.8). These enzymes are essential in the formation and degradation of phosphatidic acid (PA). First, LPPAT converts lysophosphatidic acid (LPA) into PA by incorporating an acyl moiety at the sn-2 position of the glycerol backbone, whereas PLA2 catalytically hydrolyzes the sn-2 acyl bond of PA and convers PA into LPA. Second, DAGK converts diacylglycerol (DAG) into PA with ATP as a source of the phosphate, whereas

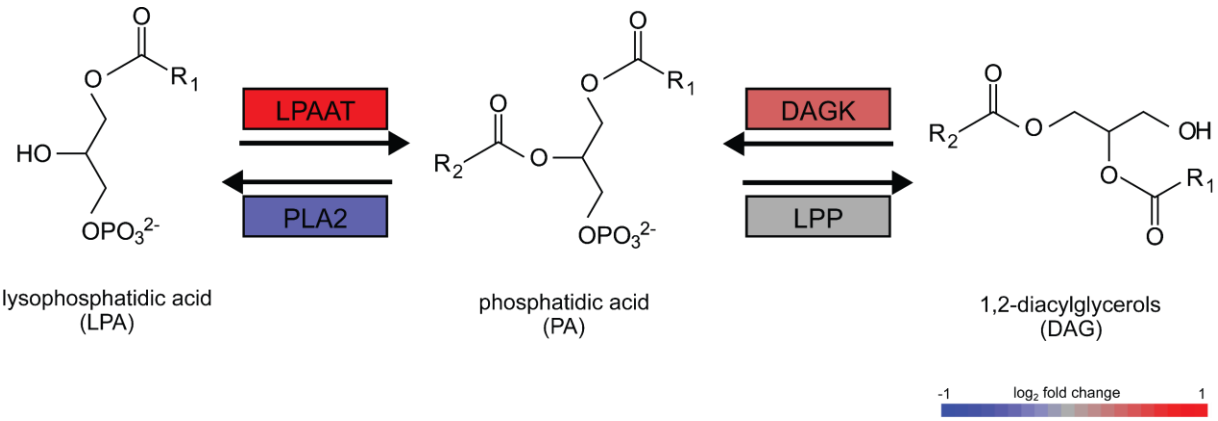

Figure 5.8 SS-31 increases synthesis of phosphatidic acid

Color indicates the log2 fold change of an enzyme category, defined by the enzyme commission (EC) number from the KEGG Ligand Database, from the saline-treated to the SS-31-treated WT mice. When more than one protein is associated with an enzyme category, the color indicates the largest absolute log2 fold change among these proteins. Lysophosphatidic acid acyltransferase (LPPAT) and diacylglycerol kinase (DAGK) both catalyze the synthesis of phosphatidic acid (PA), and show a significant (p-value < 0.05) positive fold-change. Cytosolic phospholipase A2 (PLA2) catalyzes the conversion of PA to lysophosphatidic acid (LPA) and shows a significant (p-value < 0.05) negative fold-change. Phosphatidate phosphatase (LPP) catalyzes the conversion of PA to 1,2-diacylglycerols (DAG), and shows no significant fold-change.

phosphatidate phosphatase (LPP) hydrolyzes PA and converts it into DAG. Together, the changes we observed in these enzymes should significantly promote the production of PA.

PA is a major component of the cell membrane and an essential precursor for the synthesis of many other phospholipids, including cardiolipin. PA is also involved in a

wide range of biological processes including several signaling pathways. SS-31 is known to selectively bind to cardiolipin via electrostatic and hydrophobic interactions (Birk et al., 2013; Szeto, 2014). The increase in production of PA may be a compensating mechanism for the cell to synthesize more cardiolipin. Further analyses of the metabolome are required to determine whether the proteome changes from the SS-31 treatment truly result in accumulation of PA or an increase of cardiolipin. PA also plays an important role in the phospholipase D signaling pathway as a single lipid generated from the phospholipase D (PLD)-mediated hydrolysis of other phospholipids. Although we did not observe significant changes in PLD and PLD activators, such as PKCα, resulting from the SS-31 treatment, the increased production of PA may impact the downstream pathways that PA interacts with.

5.5.3 *Involvement of cardiac mast cells*

Mast cells store and release a variety of biologically active mediators, including histamine, basic fibroblast growth factor (bFGF), and mast cell-specific proteases such as chymase, tryptase, and carboxypeptidase A (MC-CPA) type (Krishnaswamy et al., 2006). While low numbers of mast cells are present in the heart of healthy individuals, patients with cardiometabolic diseases and animals with experimentally-induced cardiometabolic pathologies have high numbers of mast cells with increased activity in the affected tissues (Shi et al., 2015). Studies have demonstrated that cardiac mast cells and their mediators are involved in cardiac remodeling and may directly contribute to the pathogenesis of cardiovascular diseases. (Janicki et al., 2015; Levick et al., 2011). We observed protein changes indicating the cardiac mast cells may be involved in the mechanisms of SS-31. In

SS-31-treated WT mice, we observed a significant increase of mast cell carboxypeptidase A (MC-CPA) and chymase, a major secreted protease of mast cells. The increase of these mast cell specific proteins suggests an increase in the population or the activation of cardiac mast cells resulting from SS-31. However, this observation might contradict the current understanding that the increased numbers and activity of mast cells are associated with reduced cardiac function. For example, active chymase increases angiotensin II formation in damaged tissues and induces the development of cardiovascular remodeling (Takai and Jin, 2016). Further analysis of chymase activity may be needed to determine if the SS-31 treatment cause increased chymase-mediated cardiovascular remodeling.

5.6 Conclusions

Treatment with SS-31 for eight weeks partially reversed the cardiac aging phenotypes and improved the cardiac function. Such improvement in the cardiac function was not additive to the improvement of life-long mCAT overexpression. As indicated by the functional studies, proteome profiling in heart showed that SS-31 treatment had larger impacts on WT mice than mCAT mice. Signatures observed in the mitochondrial proteins suggested some interactions between the treatment and the genotype. Differential quantification analysis also demonstrated that SS-31 resulted in distinct proteome changes between the WT and mCAT mice. These results indicated that the effects of SS-31 on the heart proteome could be overwhelmed by the life-long influence of mCAT.

Significant changes in several key enzymes involved in the metabolic pathways were observed in the analysis of the SS-31 impacts on the whole heart proteome in WT. The changes of these enzymes may promote the biogenesis of DNA and peptide, and the production of phosphatidic acids, a major component of the cell membrane. Together, these results suggested that SS-31 may increase mitochondrial biogenesis where cells repair their damaged mitochondrial DNA. Furthermore, SS-31 increased some mast-cell specific proteases in WT, suggesting involvement of cardiac mast cells in the mechanisms of SS-31.

5.7 Materials and methods

5.7.1 *Old mouse cohort*

24-month-old C57BL/6 wild-type (WT) mice were obtained from the National Institute of Aging Charles River colony. 26-month-old C57BL/6.mCAT mice (MCAT transgenic mice) were obtained from the Rabinovitch laboratory at the University of Washington (Schriner et al., 2005). Mice were housed at 20°C in an AAALAC accredited facility under Institutional Animal Care and Use Committee (IACUC) supervision. Both WT and mCAT mice were randomly assigned to two groups and injected with saline (as a vehicle control) or SS-31 via osmotic pumps. At 8 weeks, hearts were harvested for biochemical assays or proteome analysis. In this study, 12 mCAT male mice (7 treated with SS-31 and 5 with saline) and 10 WT male mice (5 treated with SS-31 and 5 with saline) were processed for proteome analysis.

5.7.2 *Whole heart proteome data*

Upon harvest, individual whole heart was flash frozen in liquid nitrogen, added to the blender, and then disrupted by pulsing the blender for several minutes until the tissue resembles a fine powder. The pulverized tissues were individually stored at -80°C. During preparation for mass spectrometry analysis, pulverized tissues were homogenized in 0.1% RapiGest™ SF surfactant (Waters) in 50mM ammonium bicarbonate with Halt™ phosphatase inhibitor (a mixture of sodium fluoride, sodium orthovanadate, sodium pyrophosphate and beta-glycerophosphate) (Thermo) and deacetylase inhibitors (Trichostatin A and Nicotinamide). Protein concentration for each sample was then determined by BCA assay (Thermo).

The sample was diluted in 50mM ammonium bicarbonate with unlabeled human Apolipoprotein A-1 as the internal standard for digestion. Post dilution, each sample contained 0.2 ug/uL of heart protein extract, 1 ng/uL human Apolipoprotein A-1, and 0.1% RapiGest™ SF surfactant. Dithiothreitol (DTT, Sigma Aldrich # D0632) was added to a final concentration of 5 mM and samples incubated at 50 °C for 30 minutes to reduce disulfide bonds. Iodoacetamide (Sigma Aldrich # I1149) was then added to a concentration of 15 mM followed by a 30-minute room temperature incubation in the dark to alkylate reduced cysteine. Next, sequencing grade trypsin (Promega) was added to a 1:50 enzyme to protein ratio followed by sample incubation at 37 °C shaking for 2 hours to digest proteins. The digestion reaction was quenched by addition of hydrochloric acid to a final concentration of 200 mM followed by 45 mins of incubation at 37 °C shaking. After a 20-min spin at 4 °C / 14,000 RPM, the supernatant was kept for MCX column (Waters) clean up.

Chromatography was performed using a nanoACQUITY (Waters) system set to a flow rate of 250 nl/min during linear gradient with Buffer A (2% ACN, 0.1% formic acid and 97.9% water) and Buffer B (99.9% ACN and 0.1% formic acid). A homemade 3-cm-long 150-µm I.D. trapping column was used prior to a self-packed 30-cm-long 75-µm I.D. PicoFrit resolving column (New Objective) for a 60-min linear gradient from 2% to 35% Buffer B. Both trapping and resolving columns were packed with 3-µm ReproSil-Pur C18 AQ (Dr. Maisch GmbH). The gradient was followed by a wash at 80% Buffer B and a column re-equilibration at 2% Buffer B.

For each sample, four gas phase fractionations were used in acquiring the mouse heart proteome data. Together, the precursor range of 500-900 *m/z* was analyzed, where each fractionation covered a 100 *m/*z-wide portion of the precursor range: 500-600, 600-700, 700-800, or 800-900. One µg of plasma sample, 50 fmol of PRTC, and 2.8 ng APO-A1 were loaded in each injection, separated with a 60-min linear gradient LC, and analyzed on a Q-Exactive HF mass spectrometer. For each fractionation, data-independent acquisition (DIA) method cycled with 25 non-overlapping 4 *m/z*-wide isolation MS/MS scans (at 30,000 resolution), and one 100 *m/*z-wide MS scan (at 30,000 resolution).

5.7.3 *Peptide detection*

Two sequence databases were used in this study: the mouse UniProt Swiss-Prot database that contains 16,831 proteins, and the mouse mitochondrial proteome database MitoP2 that contains 1,265 proteins known to be associated with mitochondria (Prokisch et al., 2006).

PECAN (v. 0.9.9) was used to query target peptides, and to generate and query decoy peptides, from the target database against the DIA data. Only fully tryptic peptides with up to one missed cleavage sites were queried, with a fixed modification of carbamidomethyl cysteine. PECAN was set to query peptides with 2+ or 3+ precursor ion charge states, and in y-ion mode where only product y-ion series were considered. A ±10 ppm mass error tolerance was used for both precursor ions and fragment ions. As described in 3.2, PECAN reported nineteen auxiliary scores (Table 3.1) for every target or decoy evidence of detection. These auxiliary scores are used by Percolator (Käll et al., 2007) (v. 2.08.01) to train a classifier to separate correct from incorrect matches and estimate FDR. All detected peptides are reported by Percolator at the peptide level with *q*-value < 0.01, and detected proteins are reported by Percolator's built-in Fido algorithm (Serang et al., 2010) with *q*-value < 0.01, unless indicated otherwise.

5.7.4 *Detection synchronization*

Not every peptide was detected in every sample. Thus, for peptide quantification, we synchronized the detection with the following four steps: retention time alignment, outlier removal, peak imputation, and peak boundary adjustment.

For retention time alignment, an anchor sample was first chosen based on the lowest average retention time difference compared to the other samples. The average retention time difference was calculated from the peptides detected in all samples. The retention time of each sample was aligned to the anchor sample by a support vector regression. Next, we determined the detection outliers by the delta retention time. After retention time alignment, if the observed delta retention time of a detection was not

within three standard deviations of the mean of delta retention times, the detection was considered an outlier and would be removed. With the alignment information, the peak boundaries of empty detections for each sample were imputed base on the retention time from the anchor sample. The empty detections could result from not detected by PECAN or outlier removal.

Finally, with the product ion chromatograms, the boundaries of every peak were adjusted individually. In this process, a detection was represented by the trace of median normalized intensities, of which the intensities of each non-zero transition was individually normalized to the sum of its intensities from the ± 30 seconds of the retention time of detection. The apex of the trace of median normalized intensities was considered the peak apex. From the peak apex walking out, the boundaries were drawn when the trace of median intensities showed continuous increases or reached 1% of the apex intensities. With peak boundaries from all sample, we calculated the median peak duration for each peptide. If the difference between a peak duration and the median duration of the peptide was more than 20% of the median duration, the peak was further adjusted equally on both boundaries so that the difference to median was 20%. After enforcing the peak durations of every peptide, the peak boundaries were finalized and used for quantification.

5.7.5 *Quantification and normalization*

The relative abundance of a peptide, called the “peptide area” hereafter, was represented by the sum of the area under the curve (AUC) of the product signal from all detected charge states. For each detected charge state of a peptide, four product ion

transitions, extracted with a ± 10 ppm mass error tolerance, were selected to calculated the AUC across all samples. The transitions selection was based on a voting mechanism. Each sample casted one vote to every non-interfered transition determined by the following two criteria. First, within a given peak boundary, if the intensity of a transition at the left boundary or the right boundary is more than 60% of the intensity at the peak apex, the transition was considered interfered. In this process, the peak apex was determined by the apex of median intensities from all non-zero transitions. Second, at a given peak duration determined by the peak boundaries, if the sum of intensities of a transition from the same duration of either right before or right after the peak is more than 60% of the sum of intensities within peak, the transition was considered interfered. After casting from all samples, the four transitions of the charge state of the peptide with the most votes were used for quantification.

To account for the run-to-run variance introduced by instrumentation, the peptide areas were first normalized by the sum of the total ion current (TIC) from MS1 spectra for each run. The data exhibited high variance among individual samples, likely contributed by the diversity of aging process in the mice. To account for the sample-to-sample variance, the peptide areas were normalized to the median area of detected peptides from each sample with the database interrogated.

The relative abundance of a protein, called the “protein area”, was represented by the sum of the peptide area from the detected peptides that were unique to the protein or the isoforms of the protein and not shared with any other proteins in the database interrogated.

5.7.6 *Hierarchical clustering*

Hierarchical clustering was performed with GENE-E, a matrix visualization and analysis platform (http://software.broadinstitute.org/GENE-E). Modified *z*-score transformation with row median and median absolute deviation (MAD) was used prior to clustering. One minus the Spearman's rank correlation coefficient was used as the distance metric for both column (sample) and row (peptide or protein).

BIBLIOGRAPHY

Aguiar, M., Haas, W., Beausoleil, S.A., Rush, J., and Gygi, S.P. (2010). Gas-phase rearrangements do not affect site localization reliability in phosphoproteomics data sets. J Proteome Res *9*, 3103–3107.

Anderson, N.L., and Anderson, N.G. (2002). The Human Plasma Proteome History, Character, and Diagnostic Prospects. Mol Cell Proteomics *1*, 845–867.

Bald, T., Barth, J., Niehues, A., Specht, M., Hippler, M., and Fufezan, C. (2012). pymzML - Python module for high throughput bioinformatics on mass spectrometry data. Bioinformatics bts066.

Beausoleil, S.A., Villén, J., Gerber, S.A., Rush, J., and Gygi, S.P. (2006). A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotech *24*, 1285–1292.

Bern, M., Finney, G., Hoopmann, M.R., Merrihew, G., Toth, M.J., and MacCoss, M.J. (2010). Deconvolution of Mixture Spectra from Ion-Trap Data-Independent-Acquisition Tandem Mass Spectrometry. Anal. Chem. *82*, 833–841.

Birk, A.V., Liu, S., Soong, Y., Mills, W., Singh, P., Warren, J.D., Seshan, S.V., Pardee, J.D., and Szeto, H.H. (2013). The Mitochondrial-Targeted Compound SS-31 Re-Energizes Ischemic Mitochondria by Interacting with Cardiolipin. JASN *24*, 1250–1261.

Burgess, M.W., Keshishian, H., Mani, D.R., Gillette, M.A., and Carr, S.A. (2014). Simplified and Efficient Quantification of Low-abundance Proteins at Very High Multiplex via Targeted Mass Spectrometry. Mol Cell Proteomics *13*, 1137–1149.

Carvalho, P.C., Han, X., Xu, T., Cociorva, D., Carvalho, M. da G., Barbosa, V.C., and Yates, J.R. (2010). XDIA: improving on the label-free data-independent analysis. Bioinformatics *26*, 847–848.

Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., et al. (2012). A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotech *30*, 918–920.

Chapman, J.D., Goodlett, D.R., and Masselon, C.D. (2014). Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. Mass Spec Rev *33*, 452–470.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotech *26*, 1367–1372.

Craig, R., and Beavis, R.C. (2004). TANDEM: matching proteins with tandem mass spectra. Bioinformatics *20*, 1466–1467.

Dai, D.-F., Santana, L.F., Vermulst, M., Tomazela, D.M., Emond, M.J., MacCoss, M.J., Gollahon, K., Martin, G.M., Loeb, L.A., Ladiges, W.C., et al. (2009). Overexpression of Catalase Targeted to Mitochondria Attenuates Murine Cardiac Aging. Circulation *119*, 2789–2797.

Dai, D.-F., Chen, T., Szeto, H., Nieves-Cintrón, M., Kutyavin, V., Santana, L.F., and Rabinovitch, P.S. (2011). Mitochondrial targeted antioxidant peptide ameliorates hypertensive cardiomyopathy. J Am Coll Cardiol *58*, 73–82.

Dai, D.-F., Hsieh, E.J., Liu, Y., Chen, T., Beyer, R.P., Chin, M.T., MacCoss, M.J., and Rabinovitch, P.S. (2012). Mitochondrial proteome remodelling in pressure overload-induced heart failure: the role of mitochondrial oxidative stress. Cardiovascular Research *93*, 79–88.

Davis, M.T., Spahr, C.S., McGinley, M.D., Robinson, J.H., Bures, E.J., Beierle, J., Mort, J., Yu, W., Luethy, R., and Patterson, S.D. (2001). Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry II.Limitations of complex mixture analyses. Proteomics *1*, 108–117.

Egertson, J.D., Kuehn, A., Merrihew, G.E., Bateman, N.W., MacLean, B.X., Ting, Y.S., Canterbury, J.D., Marsh, D.M., Kellmann, M., Zabrouskov, V., et al. (2013). Multiplexed MS/MS for improved data-independent acquisition. Nat Meth *10*, 744–746.

Egertson, J.D., MacLean, B., Johnson, R., Xuan, Y., and MacCoss, M.J. (2015). Multiplexed peptide analysis using data-independent acquisition and Skyline. Nat. Protocols *10*, 887–903.

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

Eng, J.K., McCormack, A.L., and Yates III, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry *5*, 976–989.

Eng, J.K., Fischer, B., Grossmann, J., and MacCoss, M.J. (2008). A Fast SEQUEST Cross Correlation Algorithm. J. Proteome Res. *7*, 4598–4602.

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

Escher, C., Reiter, L., MacLean, B., Ossola, R., Herzog, F., Chilton, J., MacCoss, M.J., and Rinner, O. (2012). Using iRT, a normalized retention time for more targeted measurement of peptides. Proteomics *12*, 1111–1121.

Frank, A.M., Savitski, M.M., Nielsen, M.L., Zubarev, R.A., and Pevzner, P.A. (2007). De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry. J. Proteome Res. *6*, 114–123.

Frewen, B.E., Merrihew, G.E., Wu, C.C., Noble, W.S., and MacCoss, M.J. (2006). Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries. Anal. Chem. *78*, 5678–5684.

Gatlin, C.L., Eng, J.K., Cross, S.T., Detter, J.C., and Yates, J.R. (2000). Automated Identification of Amino Acid Sequence Variations in Proteins by HPLC/Microspray Tandem Mass Spectrometry. Anal. Chem. *72*, 757–763.

Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., and Bryant,

S.H. (2004). Open Mass Spectrometry Search Algorithm. J. Proteome Res. *3*, 958–964.

Gillet, L.C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. Mol Cell Proteomics *11*, O111.016717.

Gorshkov, A.V., Tarasova, I.A., Evreinov, V.V., Savitski, M.M., Nielsen, M.L., Zubarev, R.A., and Gorshkov, M.V. (2006). Liquid Chromatography at Critical Conditions: Comprehensive Approach to Sequence-Dependent Retention Time Prediction. Anal. Chem. *78*, 7770–7777.

Granholm, V., Navarro, J.C.F., Noble, W.S., and Käll, L. (2013). Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. J Proteomics *0*, 123–131.

Hebert, A.S., Richards, A.L., Bailey, D.J., Ulbrich, A., Coughlin, E.E., Westphall, M.S., and Coon, J.J. (2014). The One Hour Yeast Proteome. Mol Cell Proteomics *13*, 339–347.

Hoopmann, M.R., MacCoss, M.J., and Moritz, R.L. (2012). Identification of peptide features in precursor spectra using Hardklör and Krönik. Curr Protoc Bioinformatics *0 13*, Unit13.18.

Houel, S., Abernathy, R., Renganathan, K., Meyer-Arendt, K., Ahn, N.G., and Old, W.M. (2010). Quantifying the Impact of Chimera MS/MS Spectra on Peptide Identification in Large-Scale Proteomics Studies. J. Proteome Res. *9*, 4152–4160.

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

Jaffe, J.D., Berg, H.C., and Church, G.M. (2004). Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics *4*, 59–77.

Janicki, J.S., Brower, G.L., and Levick, S.P. (2015). The Emerging Prominence of the Cardiac Mast Cell as a Potent Mediator of Adverse Myocardial Remodeling. Methods Mol Biol *1220*, 121–139.

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

Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002). Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. Anal. Chem. *74*, 5383–5392.

Kim, S., and Pevzner, P.A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun *5*.

Koenig, T., Menze, B.H., Kirchner, M., Monigatti, F., Parker, K.C., Patterson, T., Steen, J.J., Hamprecht, F.A., and Steen, H. (2008). Robust Prediction of the MASCOT Score for an Improved Quality Assessment in Mass Spectrometric Proteomics. J. Proteome Res. *7*, 3708–3717.

Krishnaswamy, G., Ajitawi, O., and Chi, D.S. (2006). The human mast cell: an overview. Methods Mol. Biol. *315*, 13–34.

Krokhin, O.V., Craig, R., Spicer, V., Ens, W., Standing, K.G., Beavis, R.C., and Wilkins, J.A. (2004). An Improved Model for Prediction of Retention Times of Tryptic Peptides in Ion Pair Reversed-phase HPLC Its Application to Protein Peptide Mapping by Off-Line HPLC-MALDI MS. Mol Cell Proteomics *3*, 908–919.

Lakatta, E.G., and Levy, D. (2003). Arterial and Cardiac Aging: Major Shareholders in Cardiovascular Disease Enterprises. Circulation *107*, 139–146.

Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., Stein, S.E., and Aebersold, R. (2008). Building consensus spectral libraries for peptide identification in proteomics. Nat Meth *5*, 873–875.

Levick, S.P., Meléndez, G.C., Plante, E., McLarty, J.L., Brower, G.L., and Janicki, J.S. (2011). Cardiac mast cells: the centrepiece in adverse myocardial remodelling. Cardiovascular Research *89*, 12–19.

Li, G.-Z., Vissers, J.P.C., Silva, J.C., Golick, D., Gorenstein, M.V., and Geromanos, S.J. (2009). Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. Proteomics *9*, 1696–1719.

Li, Y., Zhong, C.-Q., Xu, X., Cai, S., Wu, X., Zhang, Y., Chen, J., Shi, J., Lin, S., and Han, J. (2015). Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. Nat Meth *advance online publication*.

Li, Y.F., Arnold, R.J., Tang, H., and Radivojac, P. (2010). The Importance of Peptide Detectability for Protein Identification, Quantification, and Experiment Design in MS/MS Proteomics. J. Proteome Res. *9*, 6288–6297.

Liebler, D.C., Hansen, B.T., Davey, S.W., Tiscareno, L., and Mason, D.E. (2002). Peptide Sequence Motif Analysis of Tandem MS Data with the SALSA Algorithm. Anal. Chem. *74*, 203–210.

Luethy, R., Kessner, D.E., Katz, J.E., MacLean, B., Grothe, R., Kani, K., Faça, V., Pitteri, S., Hanash, S., Agus, D.B., et al. (2008). Precursor-Ion Mass Re-Estimation Improves Peptide Identification on Hybrid Instruments. J. Proteome Res. *7*, 4031–4039.

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. (2003). PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. *17*, 2337–2342.

MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L., Frewen, B., Kern, R., Tabb, D.L., Liebler, D.C., and MacCoss, M.J. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics *26*, 966–968.

Mallick, P., Schirle, M., Chen, S.S., Flory, M.R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., et al. (2007). Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotech *25*, 125–131.

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W.H., Römpp, A., Neumann, S., Pizarro, A.D., et al. (2011). mzML—a Community Standard for Mass Spectrometry Data. Mol Cell Proteomics *10*.

Marx, V. (2013). Targeted proteomics. Nat Meth *10*, 19–22.

Members, W.G., Roger, V.L., Go, A.S., Lloyd-Jones, D.M., Benjamin, E.J., Berry, J.D., Borden, W.B., Bravata, D.M., Dai, S., Ford, E.S., et al. (2012). Heart Disease and Stroke Statistics—2012 Update. Circulation *125*, e2–e220.

Michalski, A., Cox, J., and Mann, M. (2011). 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.

Mozaffarian, D., Benjamin, E.J., Go, A.S., Arnett, D.K., Blaha, M.J., Cushman, M., Das, S.R., Ferranti, S. de, Després, J.-P., Fullerton, H.J., et al. (2015). Heart Disease and Stroke Statistics—2016 Update. Circulation CIR.0000000000000350.

Murdoch, C.E., Zhang, M., Cave, A.C., and Shah, A.M. (2006). NADPH oxidase-dependent redox signalling in cardiac hypertrophy, remodelling and failure. Cardiovascular Research *71*, 208–215.

Murray, K.K., Boyd, R.K., Eberlin, M.N., Langley, G.J., Li, L., and Naito, Y. (2013). Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). Pure and Applied Chemistry *85*.

Na, S., Payne, S.H., and Bandeira, N. (2016). Multi-species identification of polymorphic peptide variants via propagation in spectral networks. Mol Cell Proteomics mcp.O116.060913.

Nesvizhskii, A.I. (2014). Proteogenomics: concepts, applications and computational strategies. Nat Meth *11*, 1114–1125.

Nesvizhskii, A.I., and Aebersold, R. (2005). Interpretation of Shotgun Proteomic Data The Protein Inference Problem. Mol Cell Proteomics *4*, 1419–1440.

Noble, W.S. (2015). Mass spectrometrists should search only for peptides they care about. Nat Meth *12*, 605–608.

North, B.J., and Sinclair, D.A. (2012). The Intersection Between Aging and Cardiovascular Disease. Circulation Research *110*, 1097–1108.

Olsen, J.V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007). Higher-energy C-trap dissociation for peptide modification analysis. Nat Meth *4*, 709–712.

Panchaud, A., Scherl, A., Shaffer, S.A., von Haller, P.D., Kulasekara, H.D., Miller, S.I., and Goodlett, D.R. (2009). Precursor Acquisition Independent From Ion Count: How to Dive Deeper into the Proteomics Ocean. Anal. Chem. *81*, 6481–6488.

Payne, S.H., Monroe, M.E., Overall, C.C., Kiebel, G.R., Degan, M., Gibbons, B.C., Fujimoto, G.M., Purvine, S.O., Adkins, J.N., Lipton, M.S., et al. (2015). The Pacific Northwest National Laboratory library of bacterial and archaeal proteomic biodiversity. Scientific Data *2*, 150041.

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

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

Plumb, R.S., Johnson, K.A., Rainville, P., Smith, B.W., Wilson, I.D., Castro-Perez, J.M., and Nicholson, J.K. (2006). UPLC/MSE; a new approach for generating molecular fragment information for biomarker structure elucidation. Rapid Commun. Mass Spectrom. *20*, 1989–1994.

Prakash, A., Tomazela, D.M., Frewen, B., MacLean, B., Merrihew, G., Peterman, S., and MacCoss, M.J. (2009). Expediting the Development of Targeted SRM Assays: Using Data from Shotgun Proteomics to Automate Method Development. J. Proteome Res. *8*, 2733–2739.

Prokisch, H., Andreoli, C., Ahting, U., Heiss, K., Ruepp, A., Scharfe, C., and Meitinger, T. (2006). MitoP2: the mitochondrial proteome database—now including mouse data. Nucleic Acids Res *34*, D705–D711.

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

Rardin, M.J., Schilling, B., Cheng, L.-Y., MacLean, B.X., Sorenson, D.J., Sahu, A.K., MacCoss, M.J., Vitek, O., and Gibson, B.W. (2015). MS1 Peptide Ion Intensity Chromatograms in MS2 (SWATH) Data Independent Acquisitions. Improving Post Acquisition Analysis of Proteomic Experiments. Mol Cell Proteomics mcp.O115.048181.

Reiter, L., Rinner, O., Picotti, P., Hüttenhain, R., Beck, M., Brusniak, M.-Y., Hengartner, M.O., and Aebersold, R. (2011). mProphet: automated data processing and statistical validation for large-scale SRM experiments. Nat Meth *8*, 430–435.

Rosca, M.G., and Hoppel, C.L. (2013). Mitochondrial dysfunction in heart failure. Heart Fail Rev *18*.

Rose, C.M., Merrill, A.E., Bailey, D.J., Hebert, A.S., Westphall, M.S., and Coon, J.J. (2013). Neutron Encoded Labeling for Peptide Identification. Anal. Chem. *85*, 5129–5137.

Röst, H.L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S.M., Schubert, O.T., Wolski, W., Collins, B.C., Malmström, J., Malmström, L., et al. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotech *32*, 219–223.

Sawyer, D.B., Siwik, D.A., Xiao, L., Pimentel, D.R., Singh, K., and Colucci, W.S. (2002). Role of Oxidative Stress in Myocardial Hypertrophy and Failure. Journal of Molecular and Cellular Cardiology *34*, 379–388.

Schriner, S.E., Linford, N.J., Martin, G.M., Treuting, P., Ogburn, C.E., Emond, M., Coskun, P.E., Ladiges, W., Wolf, N., Remmen, H.V., et al. (2005). Extension of Murine Life Span by Overexpression of Catalase Targeted to Mitochondria. Science *308*, 1909–1911.

Serang, O., MacCoss, M.J., and Noble, W.S. (2010). Efficient Marginalization to Compute Protein Posterior Probabilities from Shotgun Mass Spectrometry Data. J. Proteome Res. *9*, 5346–5357.

Shi, G.-P., Bot, I., and Kovanen, P.T. (2015). Mast cells in human and experimental cardiometabolic diseases. Nat Rev Cardiol *12*, 643–658.

Shteynberg, D., Deutsch, E.W., Lam, H., Eng, J.K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R.L., Aebersold, R., and Nesvizhskii, A.I. (2011). iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates. Mol Cell Proteomics *10*, M111.007690.

Siegel, M.P., Kruse, S.E., Percival, J.M., Goh, J., White, C.C., Hopkins, H.C., Kavanagh, T.J., Szeto, H.H., Rabinovitch, P.S., and Marcinek, D.J. (2013). Mitochondrial targeted peptide rapidly improves mitochondrial energetics and skeletal muscle performance in aged mice. Aging Cell *12*, 763–771.

Silva, J.C., Denny, R., Dorschel, C., Gorenstein, M.V., Li, G.-Z., Richardson, K., Wall, D., and Geromanos, S.J. (2006). Simultaneous Qualitative and Quantitative Analysis of theEscherichia coli Proteome A Sweet Tale. Mol Cell Proteomics *5*, 589–607.

Smith, L.M., Kelleher, N.L., and Proteomics, T.C. for T.D. (2013a). Proteoform: a single term describing protein complexity. Nature Methods *10*, 186–187.

Smith, L.M., Kelleher, N.L., and Proteomics, T.C. for T.D. (2013b). Proteoform: a single term describing protein complexity. Nat Meth *10*, 186–187.

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

Swaney, D.L., Wenger, C.D., Thomson, J.A., and Coon, J.J. (2009). Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. PNAS *106*, 995–1000.

Szeto, H.H. (2006). Cell-permeable, mitochondrial-targeted, peptide antioxidants. AAPS J *8*, E277–E283.

Szeto, H.H. (2014). First-in-class cardiolipin-protective compound as a therapeutic agent to restore mitochondrial bioenergetics. Br J Pharmacol *171*, 2029–2050.

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

Takai, S., and Jin, D. (2016). Improvement of cardiovascular remodelling by chymase inhibitor. Clin Exp Pharmacol Physiol *43*, 387–393.

Taylor, J.A., and Johnson, R.S. (1997). Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. *11*, 1067–1075.

Tecce, M.A., Pennington, J.A., Segal, B.L., and Jessup, M.L. (1999). Heart failure: clinical implications of systolic and diastolic dysfunction. Geriatrics *54*, 24–28, 31–33.

Ting, Y.S., Egertson, J.D., Payne, S.H., Kim, S., MacLean, B., Käll, L., Aebersold, R., Smith, R.D., Noble, W.S., and MacCoss, M.J. (2015). Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. Mol Cell Proteomics *14*, 2301–2307.

Tsou, C.-C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A.-C., and Nesvizhskii, A.I.

(2015). DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat Meth *12*, 258–264.

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

Wang, J., Bourne, P.E., and Bandeira, N. (2011). Peptide identification by database search of mixture tandem mass spectra. Mol Cell Proteomics mcp.M111.010017.

Wang, J., Bourne, P.E., and Bandeira, N. (2014). MixGF: Spectral Probabilities for Mixture Spectra from more than One Peptide. Mol Cell Proteomics *13*, 3688–3697.

Wang, J., Tucholska, M., Knight, J.D.R., Lambert, J.-P., Tate, S., Larsen, B., Gingras, A.-C., and Bandeira, N. (2015). MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. Nat Meth *advance online publication*.

Webb-Robertson, B.-J.M., Wiberg, H.K., Matzke, M.M., Brown, J.N., Wang, J., McDermott, J.E., Smith, R.D., Rodland, K.D., Metz, T.O., Pounds, J.G., et al. (2015). Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. J. Proteome Res. *14*, 1993–2001.

Weisbrod, C.R., Eng, J.K., Hoopmann, M.R., Baker, T., and Bruce, J.E. (2012). Accurate Peptide Fragment Mass Analysis: Multiplexed Peptide Identification and Quantification. J. Proteome Res. *11*, 1621–1632.

Wu, C.C., and MacCoss, M.J. (2002). Shotgun proteomics: tools for the analysis of complex biological systems. Curr. Opin. Mol. Ther. *4*, 242–250.

Yen, C.-Y., Houel, S., Ahn, N.G., and Old, W.M. (2011). Spectrum-to-Spectrum Searching Using a Proteome-wide Spectral Library. Mol Cell Proteomics *10*, M111.007666.

Yi, E.C., Marelli, M., Lee, H., Purvine, S.O., Aebersold, R., Aitchison, J.D., and Goodlett, D.R. (2002). Approaching complete peroxisome characterization by gas-phase fractionation. ELECTROPHORESIS *23*, 3205–3216.

Zhang, B., Pirmoradian, M., Chernobrovkin, A., and Zubarev, R.A. (2014). DeMix Workflow for Efficient Identification of Cofragmented Peptides in High Resolution Data-dependent Tandem Mass Spectrometry. Mol Cell Proteomics *13*, 3211–3223.

Zhang, N., Li, X., Ye, M., Pan, S., Schwikowski, B., and Aebersold, R. (2005). ProbIDtree: An automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. Proteomics *5*, 4096–4106.

Zhao, K., Zhao, G.-M., Wu, D., Soong, Y., Birk, A.V., Schiller, P.W., and Szeto, H.H. (2004). Cell-permeable Peptide Antioxidants Targeted to Inner Mitochondrial Membrane inhibit Mitochondrial Swelling, Oxidative Cell Death, and Reperfusion Injury. J. Biol. Chem. *279*, 34682–34690.

Appendix A

Variant-specific peptides detected in DIA plasma library

Feature identifier	Accession	Variant	dbSNP	Peptide
VAR_025657	P00450	Asp544Glu	rs701753	MYYSAVEPTKDIFTGLIGPMK
VAR_025657	P00450	Asp544Glu	rs701753	MYYSAVEPTK
VAR_006711	P00734	Glu200Lys	rs62623459	SKGSSVNLSPPLEQCVPDR
VAR_011781	P00734	Thr165Met	rs5896	NPDSSTMGPWCYTTDPTVR
VAR_005294	P00738	Asn129Asp	rs199926732	TEGDGVYTLNDEK
VAR_006580	P00740	Asn283Asp	-	ITVVAGEHDIEETEHTEQK
VAR_006533	P00740	Glu54Gly	-	LEGFVQGNLER
VAR_017356	P00740	Ile344Leu	-	EYTNLFLK
VAR_011779	P00747	Ile46Arg	rs1049573	EECAAKCEEDEEFTCR
VAR_014336	P00748	Ala207Pro	rs17876030	LCHCPVGYTGPFCDVDTK
VAR_016277	P00751	Lys565Glu	rs4151659	EEAGIPEFYDYDVALIK
VAR_027451	P01008	Cys32Arg	-	HGSPVDICTAKPR
VAR_027452	P01008	Tyr95Cys	-	FATTFCQHLADSK
VAR_006995	P01009	Gln180Glu	-	EINDYVEK
VAR_006996	P01009	Glu228Lys	rs199422208	DTKEEDFHVDQVTTVK
VAR_007010	P01009	Glu400Asp	rs1303	FNKPFVFLMIDQNTK
VAR_026820	P01023	Asn639Asp	rs226405	DLTGFPGPLNDQDDEDCINR
VAR_063217	P01024	Asp1115Asn	rs121909585	QKPNGVFQEDAPVIHQEMIGGLR
VAR_063219	P01024	Gln1161Lys	-	DICEEKVNSLPGSITK
VAR_048853	P01042	Asp430Glu	rs5030084	RHEWGHEK
VAR_073349	P01602	Lys72Asp	-	LLIYDASSLESGVPSR
VAR_003897	P01834	Val83Leu	-	LYACEVTHQGLSSPVTK
VAR_003897	P01834	Val83Leu	-	HKLYACEVTHQGLSSPVTK
VAR_068700	P01860	Asn245Asp	-	VVSVLTVLHQDWLDGK
VAR_068700	P01860	Asn245Asp	-	VVSVLTVLHQDWLDGKEYK
VAR_003903	P01871	Gly191Ser	-	ESDWLSQSMFTCR
VAR_014602	P01876	Glu176Asp	rs1407	DASGVTFTWTPSSGKSAVQGPPDR
VAR_014602	P01876	Glu176Asp	rs1407	SAVQGPPDR
VAR_003102	P02042	Gly25Asp	rs34460332	VNVDAVDGEALGR
VAR_003103	P02042	Gly26Asp	rs34389944	VNVDAVGDEALGR
VAR_000612	P02647	Ala119Asp	-	DKVQPYLDDFQK
VAR_000617	P02647	Glu134Lys	-	WQKEMELYR
VAR_000618	P02647	Glu160Lys	rs121912718	LQEKLSPLGEEMR
VAR_000625	P02647	Glu222Lys	rs121912717	ATKHLSTLSEK
VAR_000615	P02647	Lys131Met	rs4882	MWQEEMELYR
VAR_000649	P02649	Gln99Lys	-	SELEEKLTPVAEETR
VAR_013093	P02675	Pro265Leu	rs6054	KGGETSEMYLIQPDSSVK
VAR_013093	P02675	Pro265Leu	rs6054	GGETSEMYLIQPDSSVK
VAR_014170	P02679	Gly191Arg	rs6063	LYFIKPLK

VAR_036018	P02751	Asp940Asn	rs752106647	VNVIPVNLPGEHGQR
VAR_061486	P02751	Val2170Ile	rs1250209	GATYNIIVEALK
VAR_061486	P02751	Val2170Ile	rs1250209	GATYNIIVEALKDQQR
VAR_007591	P02766	Arg124Cys	rs745834030	CYTIAALLSPYSYSTTAVVTNPKE
VAR_038967	P02766	Asp58Ala	-	KAAADTWEPFASGK
VAR_038968	P02766	Asp58Val	-	AAVDTWEPFASGK
VAR_007585	P02766	Glu109Gln	rs121918082	ALGISPFHQHAEVVFTANDSGPR
VAR_010659	P02766	Glu109Lys	-	ALGISPFHKHAEVVFTANDSGPR
VAR_038976	P02766	Glu74Lys	-	TSESGKLHGLTTEEEFVEGIYK
VAR_007583	P02766	Ile104Asn	-	ALGNSPFHEHAEVVFTANDSGPR
VAR_038985	P02766	Ile127Met	-	YTMAALLSPYSYSTTAVVTNPK
VAR_007576	P02766	Ile88Leu	rs121918085	TSESGELHGLTTEEEFVEGLYK
VAR_007594	P02766	Leu131Met	rs121918073	YTIAALMSPYSYSTTAVVTNPK
VAR_007570	P02766	Leu78His	rs121918069	TSESGELHGHTTEEEFVEGIYK
VAR_038961	P02766	Ser43Asn	-	VLDAVRGNPAINVAVHVFR
VAR_007595	P02766	Tyr134Cys	rs121918075	YTIAALLSPCSYSTTAVVTNPKE
VAR_000527	P02768	Asp389His	rs77187142	CCAAAHPHECYAK
VAR_000530	P02768	Asp399Asn	rs77514449	VFNEFKPLVEEPQNLIK
VAR_000542	P02768	Asp587Asn	rs76587671	ADNKETCFAEEGK
VAR_000508	P02768	Asp87Asn	rs78574148	TCVADESAENCNK
VAR_000509	P02768	Glu106Lys	rs80296402	KTYGEMADCCAK
VAR_000509	P02768	Glu106Lys	rs80296402	TYGEMADCCAK
VAR_000511	P02768	Glu143Lys	rs75522063	LVRPKVDVMCTAFHDNEETFLK
VAR_000511	P02768	Glu143Lys	rs75522063	VDVMCTAFHDNEETFLK
VAR_000511	P02768	Glu143Lys	rs75522063	VDVMCTAFHDNEETFLKK
VAR_000526	P02768	Glu382Lys	rs75791663	KCCAAADPHECYAK
VAR_000532	P02768	Glu400Gln	rs79047363	VFDQFKPLVEEPQNLIK
VAR_000531	P02768	Glu400Lys	rs79047363	VFDKFKPLVEEPQNLIK
VAR_000531	P02768	Glu400Lys	rs79047363	FKPLVEEPQNLIK
VAR_000533	P02768	Glu406Lys	rs76483862	EPQNLIK
VAR_014294	P02768	Glu420Lys	-	QNCELFKQLGEYK
VAR_000536	P02768	Glu525Lys	rs75523493	KFNAETFTFHADICTLSEK
VAR_000536	P02768	Glu525Lys	rs75523493	FNAETFTFHADICTLSEK
VAR_000537	P02768	Glu529Lys	rs74826639	EFNAKTFTFHADICTLSEK
VAR_000537	P02768	Glu529Lys	rs74826639	TFTFHADICTLSEK
VAR_000543	P02768	Glu589Lys	rs75709682	KTCFAEEGK
VAR_000512	P02768	His152Arg	rs80095457	LVRPEVDVMCTAFRDNEETFLK
VAR_000515	P02768	Lys249Gln	rs79804069	FPQAEFAEVSK
VAR_013016	P02768	Lys383Asn	rs75069738	LAKTYETTLENCCAAADPHECYAK
VAR_013012	P02768	Val146Glu	rs77752336	LVRPEVDEMCTAFHDNEETFLK
VAR_013012	P02768	Val146Glu	rs77752336	LVRPEVDEMCTAFHDNEETFLKK
VAR_058199	P02787	Ile448Val	rs2692696	SDNCEDTPEAGYFAVAVVK
VAR_058199	P02787	Ile448Val	rs2692696	SDNCEDTPEAGYFAVAVVKK

VAR_012000	P02787	Pro589Ser	rs1049296	SVEEYANCHLAR
VAR_012000	P02787	Pro589Ser	rs1049296	DYELLCLDGTRK
VAR_012000	P02787	Pro589Ser	rs1049296	KSVEEYANCHLAR
VAR_016286	P03952	Arg560Gln	rs4253325	ITQQMVCAGYK
VAR_059582	P04114	Ile2313Val	rs584542	INDVLEHVK
VAR_029342	P04114	Pro877Leu	rs12714097	LEVANMQAELVAK
VAR_061558	P04114	Tyr1422Cys	rs568413	NTFTLSCDGSLR
VAR_024429	P04196	Asn493Ile	rs1042464	HPLKPDIQPFPQSVSESCPGK
VAR_018369	P04217	His52Arg	rs893184	LETPDFQLFK
VAR_038628	P04264	Ala454Ser	rs17678945	LNDLEDALQQSK
VAR_000627	P06727	Glu44Lys	-	KAVEHLQK
VAR_046821	P07225	Cys121Tyr	-	SCVNAIPDQYSPLPCNEDGYMSCK
VAR_033800	P07357	Asp458Asn	rs17114555	YNPVVINFEMQPIHEVLR
VAR_011889	P07357	Gln93Lys	rs652785	KAQCGQDFQCK
VAR_011892	P07357	Glu561Gln	rs1342440	QCDNPAPQNGGASCPGR
VAR_019406	P08603	Cys959Tyr	-	YFEGFGIDGPAIAK
VAR_025093	P08603	Ser890Ile	rs515299	SSQEIYAHGTKLSYTCEGGFR
VAR_023836	P08603	Val62Ile	rs800292	SLGNIIMVCR
VAR_072438	P08779	Asn125Asp	rs58608173	VTMQNLDDR
VAR_069154	P0C0L4	Leu141Val	rs9296005	GHVFLQTDQPIYNPGQR
VAR_069154	P0C0L4	Leu141Val	rs9296005	RGHVFLQTDQPIYNPGQR
VAR_069160	P0C0L5	Pro478Leu	-	LTVAAPPSGGPGFLSIER
VAR_033799	P10643	Thr587Pro	rs13157656	DGFVQDEGPMFPVGK
VAR_001214	P12259	Lys858Arg	rs4524	LLSLGAGEFR
VAR_069914	P12814	Glu225Lys	rs387907350	MLDAKDIVGTARPDEK
VAR_017475	P14136	Glu362Asp	rs28932768	LALDIDIATYR
VAR_050173	P19652	Gly141Arg	rs12685968	NWRLSFYADKPETTK
VAR_004020	P19827	Gln595Arg	rs1042779	MSLDYGFVTPLTSMSIR
VAR_044226	P35579	Lys910Gln	rs554332083	QQELEEICHDLEAR
VAR_007639	P49747	Asp518Asn	-	INVCPENAEVTLTDFR
VAR_012857	P68032	Glu101Lys	rs193922680	VAPKEHPTLLTEAPLNPK
VAR_062436	P68133	Ile77Leu	-	YPIEHGLITNWDDMEK
VAR_062427	P68133	Pro40Leu	-	AVFPSIVGR
VAR_003031	P68871	Asn109Lys	rs34933751	VLVCVLAHHFGK
VAR_003077	P68871	Asn140Asp	rs33910475	VVAGVADALAHK
VAR_002886	P68871	Asn20Asp	rs34866629	VDVDEVGGEALGR
VAR_002887	P68871	Asn20Lys	rs63750840	VDEVGGEALGR
VAR_002891	P68871	Asp22Asn	rs33950093	VNVNEVGGEALGR
VAR_002890	P68871	Asp22Gly	rs33977536	VNVGEVGGEALGR
VAR_003058	P68871	Gln128Glu	rs33971634	EFTPPVEAAYQK
VAR_002927	P68871	Gln40Glu	rs76728603	LLVVYPWTER
VAR_003048	P68871	Glu122Gln	rs33946267	QFTPPVQAAYQK
VAR_003049	P68871	Glu122Lys	rs33946267	KFTPPVQAAYQK

VAR_002897	P68871	Glu23Gln	rs33959855	VNVDQVGGEALGR
VAR_002793	P69905	Asp75Asn	rs281864857	VADALTNAVAHVNDMPNALSALSDLHAHK
VAR_034541	Q13748	Val75Leu	rs36215077	AVFVDLEPTVLDEVR
VAR_027870	Q14624	Gln669Leu	rs2276814	LLGLPGPPDVPDHAAYHPFR
VAR_014761	Q16610	Gly415Ser	rs13294	DILTIDISR
VAR_032337	Q6UXB8	Thr50Pro	rs1405069	AQVSPPASDMLHMR
VAR_049062	Q9UGM5	Lys360Arg	rs7999	LVVLPFPR

VITA

Ying Sonia Ting graduated with a Bachelor of Science in life sciences in 2006 from the National Taiwan Normal University, Taipei, Taiwan, where she first showed interests in algorithms and bioinformatics. In 2008, she earned her M.S. in Bioinformatics from the National Yang-Ming University, Taipei, Taiwan. Here she became fascinated with the challenges in analyzing mass spectrometry data under the mentorship of Victor Ng. During her master training, she was awarded with a scholarship to study lipid characterization with mass spectrometry in the laboratory of Dr. David R. Goodlett at the University of Washington. Upon graduation, she moved to the United States and continued her work with Dr. Goodlett for three years as a visiting scientist, during which she developed a keen interest in mass spectrometry and its applications in modern life sciences. In 2011, Sonia joined the Department of Genome Sciences at the University of Washington. She carried out her doctoral research in the laboratory of Dr. Michael J. MacCoss, focusing on developing new methodologies and technologies in mass spectrometry-based proteomics. Outside the lab, she enjoys cooking, traveling, and the company of her cats and dog friends.