# Development and Integration of Informatic Tools for Qualitative and Quantitative Characterization of Proteomic Datasets Generated by Tandem Mass Spectrometry

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Rachel Michelle Adams
August 2013

# ACKNOWLEDGEMENTS

In the realm of scientific accomplishment, few have eclipsed the seminal contributions of Sir Isaac Newton. Yet, per his letters, he concedes, “If I have seen further it is by standing on the shoulders of giants.” He no doubt was referencing the 12th century Latin phrase *nanos gigantum humeris insidentes*, which literally refers to dwarfs perched on giant shoulders. A more contemporary interpretation would be, “One who develops future intellectual pursuits by understanding and building on the research and works created by notable thinkers of the past." Dwarf stature aside, not only have I had the privilege and opportunity to stand on the shoulders of giants, but I readily acknowledge the hands that pulled me up along the way and that enabled me reach whatever academic success I have enjoyed.

In that vein, first and foremost, I would like to extend the proportionate giant-size word of gratitude to Dr. Robert Hettich for allowing me to "stand on his shoulders” throughout the past five years of my research and completion of my dissertation. I deeply appreciate Bob's confidence in my abilities as he brought me into the group, his continuous encouragement and availability in the trenches, and his management style that fostered trust, growth, and independence. Secondly, I would like to thank my dissertation committee, Dr. Arnold Saxton, Dr. Steven Wilhelm, Dr. Michael Berry, and Dr. Chongle Pan. Their support and contribution to my academic growth and knowledge are profoundly appreciated. I would also like to thank my current and former colleagues in Bob's research team, Dr. Richard Giannone, Dr. Brian Erickson, Dr. Allison Erickson, and Dr. Paul Abraham. I have truly appreciated their valuable insights, encouragement, and discussions in and outside of the lab. I would be remiss if I did not also thank Dr. Harry Richards and the SCALE-IT group for their contribution and support to my professional and academic growth and development. I’m so grateful to Dr. Cynthia Peterson for being instrumental in bringing me into the program and providing the SCALE-IT infrastructure. Additionally, I’d like to acknowledge the BESC community, including Dr. Gerald Tuskan, for their guidance and accountability in urging me to deliver meaning and value from my research.

Thirdly, I would like to thank my parents who have answered for me the ultimate question in life: “What are you going to do after you get your Ph.D.?” I’m going on a Disney cruise! They have instilled in me a love for learning, a pursuit of excellence in not only what I do but in my moral fiber and character, and a sense of humor; all of which have served me well over my life and particularly the last five years.

# ABSTRACT

Shotgun proteomic experiments provide qualitative and quantitative analytical information from biological samples ranging in complexity from simple bacterial isolates to higher eukaryotes such as plants and humans and even to communities of microbial organisms. Improvements to instrument performance, sample preparation, and informatic tools are increasing the scope and volume of data that can be analyzed by mass spectrometry (MS). To accommodate for these advances, it is becoming increasingly essential to choose and/or create tools that can not only scale well but also those that make more informed decisions using additional features within the data. Incorporating novel and existing tools into a scalable, modular workflow not only provides more accurate, contextualized perspectives of processed data, but it also generates detailed, standardized outputs that can be used for future studies dedicated to mining general analytical or biological features, anomalies, and trends.

This research developed cyber-infrastructure that would allow a user to seamlessly run multiple analyses, store the results, and share processed data with other users. The work represented in this dissertation demonstrates successful implementation of an enhanced bioinformatics workflow designed to analyze raw data directly generated from MS instruments and to create fully-annotated reports of qualitative and quantitative protein information for large-scale proteomics experiments.

Answering these questions requires several points of engagement between informatics and analytical understanding of the underlying biochemistry of the system under observation. Deriving meaningful information from analytical data can be achieved through linking together the concerted efforts of more focused, logistical questions. This study focuses on the following aspects of proteomics experiments: spectra to peptide matching, peptide to protein mapping, and protein quantification and differential expression. The interaction and usability of these analyses and other existing tools are also described. By constructing a workflow that allows high-throughput processing of massive datasets, data collected within the past decade can be standardized and updated with the most recent analyses.

# TABLE OF CONTENTS

**CHAPTER 1: The Role of Informatics in Shotgun Proteomics Experiments** **1**
*1.1 The Role of Proteomics in the Era of Systems Biology* *1*
1.1.1. Systems Biology: The Ultimate Data Integration Challenge 1
1.1.2. Complementary "Omics" Technologies 4
*1.2 Mass Spectrometry and Shotgun Proteomics* *9*
1.2.1. Shotgun Proteomics 9
1.2.2. Mass Spectrometry Instruments 17
1.2.3. Ion Activation & MS/MS Fragmentation 23
1.2.4. Spectral Interpretation 30
*1.3 Informatic Analogues of Analytical Processes in Shotgun Proteomic Studies* *31*
1.3.1. Spectrum to Peptide Matching 31
1.3.2. Peptide to Protein Mapping 34
1.3.3. Differential Protein Expression 37
1.3.4. Functional Analysis and Data Visualization 39
*1.4 Conclusions* *40*

**CHAPTER 2: Current Tools & Workflows Employed for Analysis of Large-Scale Shotgun Proteomics Experiments** **42**
*2.1 Evaluation of Existing Proteomic Informatic Tools and Approaches* *42*
2.1.1. Database Searching Algorithms 42
2.1.2. Filtering Criteria and FDR Calculations 50
2.1.3. Protein Inference Approaches 56
2.1.4. Differential Protein Expression Algorithms 59
2.1.5. Toolboxes and Software Packages 63
*2.2 Challenges of Integrating and Developing Workflows for Analyzing Large Experimental Datasets* *64*
2.2.1. Standardization of Data Formats 64
2.2.2. Impediments to Integrating Systems Biology Data 66
*2.3 Summary of Dissertation* *67*

**CHAPTER 3: Spectrum to Peptide Matching** **70**
*3.1 Matched Ion Intensities Increases Accuracy and Robustness of Peptide Identification* *70*
3.1.1. Evaluating Spectral Counts and Matched Ion Intensities 70
3.1.2. Calculating Matched Ion Intensities 78
3.1.3. Comparing PSM-level Intensities to Peptide-level Intensities 82
*3.2 Augmented and Refined Peptide Identifications from Otherwise Unassigned Spectra* *89*
3.2.1. Qualifying Peptide Assignments from Ambiguous Peptide-Spectrum Matches 89
3.2.2. Supplementing Traditional Database Searching Approaches 93
3.2.3. Evaluating Amino Acid Polymorphisms by Proximal Matched Ion Intensities (AAPProxiMIT) 99
*3.3 Conclusions* *109*

**CHAPTER 4: Peptide to Protein Mapping** **111**
*4.1 Using Cluster-Unique Sequences in Proteomes (CUSPs) to Enhance Confidence in Protein Inferences* *111*
4.1.1. Outlining Existing Solutions to the Protein Inference Problem 111
4.1.2. Choosing Appropriate Identity Thresholds 113

4.1.3. Rescuing Identifications that Would Otherwise be Lost 124
4.1.4. Spectral Balancing to Distribute Abundance Measurements 125
4.1.5. Preserving Functional Annotations 126
*4.2 Applying CUSPs to Large-Scale Proteomic Datasets 128*
4.2.1. Applying Clustering: Defining the Boundaries of Functional Genome Expression in *Populus* using Bottom-up Proteomics. 128
4.2.2. Applying Clustering: Putting the Pieces Together: High-performance LC-MS/MS Provides Network-, Pathway-, and Protein-level Perspectives in *Populus* 132
4.2.3. Applying Clustering: Metaproteomics Reveals Functional Shifts of Microbial and Human Proteins in Infant Gut Colonization 134
*4.3 Conclusions 139*

**CHAPTER 5: Protein Quantification 140**
*5.1 Using a Poisson Bootstrapping Method to Test Differential Protein Expression Based on Spectral Counts 140*
*5.2 Protease-Optimized Spectral Indexing for Relative Protein Abundances in Label-free Approaches 154*
5.2.1. Using Matched Ion Intensities for POSI 154
5.2.2. POSI: Comparing Samples Digested by Different Proteases 160
5.2.3. POSI: Comparing Samples Loaded in Different Concentration Amounts 165
*5.1. Using Reporter Ion Intensities for Relative Protein Abundances in Labeled Measurements 167*
*5.3 Conclusions 176*

**CHAPTER 6: Integrating Novel and Existing Tools into a Seamless Bioinformatic Workflow for Analyzing Shotgun Proteomic Datasets 178**
*6.1 Logistics of Developing a Bioinformatics Workflow 178*
*6.2 Improving an Existing Workflow 182*
*6.3 Conclusions 187*

**CHAPTER 7: Propelling a Dynamic, Iterative Feedback Loop between Biology and Technology: Future Outlook, Remaining Challenges, and Conclusions 188**
*7.1 Overview 188*
*7.2 Status and Remaining Challenges of Peptide-Spectrum Matching 190*
*7.3 Status and Remaining Challenges of Protein Inference 191*
*7.4 Status and Remaining Challenges of Protein Quantitation 192*
*7.5 Status and Remaining Challenges of Proteome Informatic Workflows 192*
*7.6 Concluding Perspective 193*

**LIST OF REFERENCES 195**

**VITA 209**

# LIST OF TABLES

Table 1.1. Common proteolytic enzymes used for digestion step in shotgun proteomics experiments. ..... 12
Table 5.1. Results for maximum likelihood goodness of fit test to Poisson distribution using protein SpCs from 10 replicates of R. palustris. ..... 146
Table 5.2. Definitions of protease-optimized protein lengths..... 158

# LIST OF FIGURES

Figure 1.1. Central dogma of molecular biology. ..... 3
Figure 1.2 The MudPIT strategy involves 2 chromatographic components: strong cation exchange (SCX) followed by reverse-phase (RP) separation of peptides. ..... 14
Figure 1.3. Illustration of data collected in a tandem mass spectrometry run. ..... 25
Figure 1.4. Illustration of collision induced fragmentation of a polypeptide. ..... 26
Figure 1.5. A peptide-spectrum match. ..... 28
Figure 1.6. Graphical illustration of the protein inference problem. ..... 36
Figure 2.1. Primary methods of SEQUEST and Myrimatch for generating the peptide-spectrum matches (PSMs). ..... 43
Figure 3.1. Peptides representing proteins with different SpC but similar MIT. ..... 75
Figure 3.2. Method for detecting noise. ..... 79
Figure 3.3. Validating the use of matched ion intensities instead of other simple features inherent to MS/MS scans. ..... 83
Figure 3.4. The distributions of an abundant peptide's matched ion intensity for each of the 11 salt pulses in a single run. ..... 85
Figure 3.5. Abundant peptide demonstrates different matched ion intensity distributions depending on its charge state. ..... 87
Figure 3.6. Peptide matched ion intensities are reproducible. ..... 88
Figure 3.7. Comparison of DeltCN scores between *Populus* and *E. coli*. ..... 90
Figure 3.8. Comparison of 2 possible peptide-spectrum matches for an ambiguous scan. ..... 92
Figure 3.9. SAAP-resolved peptide identification in PAL. ..... 98
Figure 3.10. Identifying the level of ambiguity between adjacent mass shift sites. ..... 101
Figure 3.11. Illustration of site-determining ions. ..... 103
Figure 3.12. Fragmentations statistics of CID and HCD spectra. ..... 108
Figure 4.1. Increased redundancy in higher eukaryotes decreases the potential of detecting unique regions within individual proteins. ..... 118
Figure 4.2. Graphs of the percent of unique peptides per protein as a function of sequence similarity thresholds applied to *E. coli, Mus musculus, Arabidopsis thaliana, Populus trichocarpa, and Zea mays* (by row). ..... 122
Figure 4.3. Possible ways to assign spectral counts when a peptide is shared among multiple proteins. ..... 126
Figure 4.4. Graph of the ratio of total human/microbial proteins with time. ..... 137
Figure 5.1. High degree of reproducibility between 10 technical runs of *R. palustris*... 144
Figure 5.2. Validation of the use of SpC for estimating relative protein abundance. ..... 149
Figure 5.3. ROC Curves for BetaBinomial (BB), Poisson Bootstrapping (PBS), and QSpec tests of differential protein expression between the standard mixture datasets. ..... 150
Figure 5.4. Comparison of abundance ratios considered significant by each significance test. ..... 152
Figure 5.5. Calculated Log2 ratios (right column) and their confidence intervals (left column) using the PBS Method. ..... 153
Figure 5.6. Normal distribution of protein-level measurements. ..... 157

Figure 5.7. Illustration of the different effective lengths suggested by POSI..... 159
Figure 6.1. Illustrations of visualization tools provided by TORPEDO..... 186

# CHAPTER 1: The Role of Informatics in Shotgun Proteomics Experiments

## 1.1 The Role of Proteomics in the Era of Systems Biology

### *1.1.1. Systems Biology: The Ultimate Data Integration Challenge*

Answers are limited by the scope of the questions asked and the technology employed for the investigation. This is particularly true of experimental conclusions generated by the specific hypotheses put forth by today's increasingly higher-resolution analytical platforms. As the questions become more specific, so do the answers. Consequently, there is a dichotomy of efforts simultaneously pushing both extremes of the spectrum: demands for highly accurate and precise data points collected within tightly controlled environments, as well as demands for highly reproducible, diagnostic, and deterministic characterizations of interactions between systems and their exceedingly complicated backgrounds. In the spirit of describing "systems biology," researchers are ambitiously pursuing the goal of creating a scalable perspective of biology. This endeavor encourages collecting, analyzing, and integrating multi-dimensional data points at all levels of observable science, from the microscopic to the cosmologic, while preserving appropriate and accurate degrees of resolution in order to propose models that explain, predict, and describe the world around us.[1-3] Although this grandiose aim graciously includes the indulgence of every scientist, allowing him to investigate his desired biological, physical, or chemical topic of interest at whatever level of depth or connectedness he desires, the grand achievement of putting the pieces together in a meaningful, lossless analysis remains a major challenge.

Significant strides have been made in characterizing groups and pieces of related data, resulting in the designation of a myriad of meta- and sub-categories of systems. In attempt to adopt a consistent nomenclature that captures the hierarchical nature of the data, biological researchers have started cataloguing all of the data related to a system under an umbrella term suffixed with *-ome*. Terms such as *biome*, describing the smallest unit of complete characterization of a plant and animal community[4], and *genome*[5],

describing the entire sequence of deoxyribonucleic acid (DNA) that encodes an organism's hereditary information, had existed since the beginning of the 20th century, but the "omics" revolution within molecular biology starting exploding in early 21st century when *genomics, proteomics, transcriptomics*, and *metabolomics* were joined by *cytomics, epigenomics, glycomics, kinomics, metallomics, secretomics,* and many more.[6] With the adoption of each new category, the existing categories had to be re-defined as to how they were different from and related to each other. For example, while genomic information for an organism may give insight as to what genes are possibly encoded in a cell and transcriptomic information reveals what genes are actually translated from ribonucleic acid (RNA), proteomic information sheds light on what proteins are encoded, translated, and ultimately expressed[7, 8] (Figure 1.1). More concisely phrased, "DNA makes RNA makes proteins," is known as the "central dogma of molecular biology." This statement's oversimplification of the transfer of information by molecular biology leads to a common misconception that measurements on each type of downstream data should agree with its biological predecessor and that any discrepancy between the profiles indicates one to be wrong. However, much like the three proverbial blind men describing an elephant, the different types of measurements have limited perspectives on the behavior of subcellular components at a given, time, location, or condition; they are only snapshots of a much larger dynamic picture. Until appropriate methods are devised to successfully integrate all of the contributing pieces of information, each individual "omics" is more than sufficiently complicated to merit focused investigation into its unique properties, biases, and limitations.

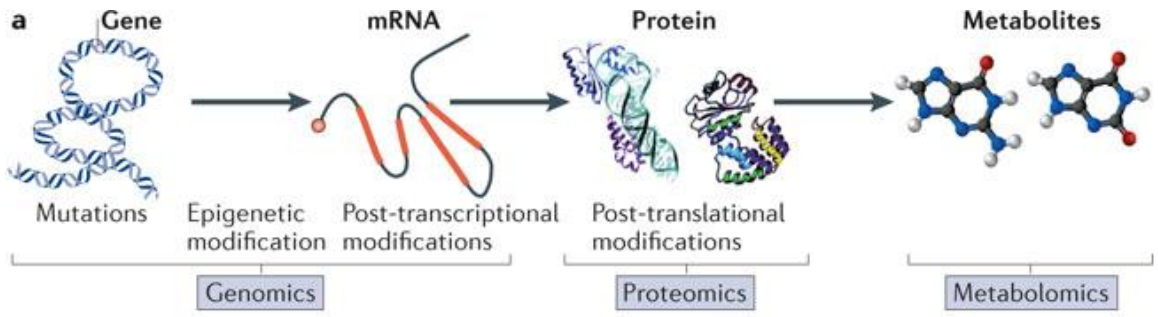



Figure 1.1. Central dogma of molecular biology.

Genomics, proteomics, and metabolomics are three of the most common "omic" fields of study. (Figure adapted from Patti, et al., *Nat Rev Mol Cell Biol*, 2012.[9])

### *1.1.2. Complementary "Omics" Technologies*

In the 1970s, molecular biology techniques heavily relied on laborious biochemical protocols to sequence the genome of small bacterial organisms, and completing a single virus genome in 1977[10] was a major accomplishment. In the next two decades, these techniques underwent a paradigm shift in their approaches, increasing the throughput of each analysis and reducing the cost by more than 6-fold.[11] By 1995, researchers had successfully implemented a "shotgun sequencing approach" that deciphered the genome of *Haemophilus influenza* using analytical and computational processes.[12] This approach involves physically shearing the DNA into smaller pieces (reads) that are more amenable to quick analytical analysis and then using computational analyses to stitch the sequences back together. The success of this strategy relies on the fragments of DNA containing overlapping regions of sufficient length to allow the reads to be assembled together into contiguous regions and ultimately, the full genome sequence.

Assembling a genome is not necessarily a straightforward computational process, especially if the process is only using information from the collected reads. *De novo* approaches are particularly problematic for organisms in which there are large regions of highly-repetitive DNA. Instead of the *overlap-layout-consensus* model,[13] a more common informatics approach is to use a genome of a closely-related organism as a scaffold or template for the new organism.[14] In order to map millions of collected reads against a reference genome, researchers must not only have sequence alignment software that they trust to correctly distinguish discrepancies between the reference and new genome, but also, more fundamentally, they need access to sequenced genomes. As the number of sequenced genomes continues to climb, online data repositories have become invaluable to creating accessible electronic catalogues of genomic information. For example, GenBank, a national genetic sequence database maintained by the National Institute of Health (NIH), collects all publicly available DNA sequences so that the scientific community has access to a single, centralized warehouse that includes the most comprehensive, up-to-date listing of genomic sequences.[15]

Although publication of the human genome sequence in 2001[16, 17] has been perhaps the most notable achievement, as of early 2013, 2,417 species have had their genomes completely sequenced, including 2,125 species of bacteria and 149 eukaryotic species.[18] Developments of additional, "next-generation" sequencing methods, including pyrosequencing[19] and massively parallel signature sequencing,[20] are continuing to add to the number of sequenced genomes.[18]

In contrast to the relatively static nature of a genome, gene expression in the form of RNA (i.e., transcripts) yields highly dynamic messages that are designed to be quick-responding indicators that anticipate and adjust to changes in environmental conditions. Determining transcript sequences is therefore much more specific to the state of a particular cell at a given time and requires a complementary method to genome sequencing. Detecting transcript sequences are important from two points of view: their presence or absence provides evidence for genes (and whether they are being turned on or off), and their relative abundance taken between two time points signifies their level of gene expression. Technologies for gene expression profiling are typically variants of DNA microarrays, in which all possible mRNA sequences for an organism are specifically arranged on a single chip such that the fluorescent intensity signal of a particular spot on the chip corresponds to the abundance of that particular mRNA within the sample.[21-23] More recently, a technology called "RNA-seq" or "Whole Transcriptome Shotgun Sequencing," exploits the advancements made by next generation genomic sequencing to read cDNA or RNA just as easily as DNA.[24, 25] The deep coverage and base-level resolution provides faster, less expensive measures of differential gene expression compared to microarray analysis, primarily because it does not rely on the manufacture of an organism-specific chip to perform measurements. Replicating the measurements is particularly helpful when the gene expression profiles are used in comparative studies that observe the change in expression levels between multiple biological conditions. These studies heavily rely on the powers of statistical tests to determine whether expression levels are significantly different between conditions. One of the major limitations to these studies is that the extraction and stabilization of mRNA

from environmental samples is not trivial and not all samples collected meet the stringent quality control criteria. Therefore, this process is still under development.

Just as the number of words in a dictionary does not dictate how many books can be written, the number of protein-coding regions in a genome does not necessarily indicate the number of proteins that can be expressed by a cell. In fact, the human genome contains 20-25,000 protein-coding genes,[26] but it has been suggested that these genes code for up to 1,000,000 distinct proteins,[27] resulting in a large discrepancy between genes and proteins (1:40 ratio, respectively). The seeming disagreement is likely due to factors such as alternative splicing and post-translational modifications of proteins. In light of this, it is not surprising that the identifications from many gene expression assays do not always align with protein expression assays.[28-30] Furthermore, transcripts may be degraded or modified after translation at a much different rate than protein turnover or modification rates. Therefore, it is also unlikely that the abundances of a transcriptomics experiment would demonstrate the same trends as a proteomics experiment. These measurements are, however, part of the same biological story, and integrating transcriptomic and proteomic measurements is becoming an increasingly popular goal.

Proteomics,[31] as the complete suite of proteins being expressed by a cell at a given time, captures information that is more of a "final product" than its biological predecessor, transcriptomics. Although recent studies have shown that transcripts can enact functions outside of carrying information about which proteins should be expressed,[32-35] proteins are responsible for most of the longer-lasting, more complex machinery. Consequently, proteins are most often the targets of biological studies investigating phenotypes, the observable characteristics of an individual that result from interactions between a genome and the environment. Other studies may seek to characterize what factors are responsible for particular molecular mechanisms, such as motility, signal transduction, transporting or secreting molecules, forming or destroying complexes. Still other proteomic experiments may seek to identify constituent members of a complex community of organisms whose species can be readily identified by expression of certain proteins.

Depending on the goal of the research, it may be reasonable to take a targeted approach (focusing on one or a few proteins at a time) or a discovery-based approach (attempting to describe as many expressed proteins as possible).

Early proteomic studies took targeted approaches because that was what the technology permitted. Previous techniques for determining what proteins were in a simple mixture primarily favored using electrophoretic gels.[36, 37] By admitting proteins into the well of a polyacrylamide gel, a voltage could be applied that would cause the proteins to move down the gel and separate by size. The smaller proteins move slower and get caught up in the fibers of the polyacrylamide, whereas the larger proteins barrel down the gel without as much difficulty. Then, the distance the proteins traveled can be compared to the distance traveled by a selection of known proteins that are injected in a nearby well. The protein mixture is generally mixed with a stain, which facilitates protein detection from the gel background, as well as a denaturant such as sodium dodecyl sulfide (SDS), which unfolds the proteins and allows them to move more predictably. If the mixture is of a few, pure proteins, then the position of the protein results in a tightly concentrated band of dye moving according to the input voltage. However, if the mixture is complex or not purely proteins, the band could be very large and more closely resemble a smear or the band could be very faint. In either case, this technique is not very amenable to high resolution (small, clean bands), high sensitivity (sufficient distances between two bands), nor high selectivity (the concentration of proteins could not be determined by the opacity of the band and a protein would not always be visible with the dye). 2D gels that separate proteins based on two-dimensions, size and hydrophobicity, can improve the fractionation of proteins and these analytical figures of merit. In fact, 2D gels are common separation techniques employed in Western blots, methods that use gels to select proteins and then nitrocellulose membranes to hybridize targeted proteins with antibodies. In general, Western blots are limited to detecting a handful of proteins in each experiment and cannot provide information about their identifications beyond observing co-alignment with a ladder or standard. If a study calls for more proteins to be identified,

or more specific quantitative comparisons, other analytical techniques should be employed.

Over the past two decades, mass spectrometry (MS) has arisen as a promising analytical technique for targeted and discovery-based proteomics approaches. Although allusions to mass spectrometry are often made without extensive detail, "mass spectrometry" actually refers to an entire family of techniques rather than a single method or type of instrument.[38-40] Essentially these techniques can provide information about the mass of entire proteins, their sequences, and/or higher-order structural detail. Variations of the three main components (ion sources, analyzers, and detectors) are generally compared based on their resolution, mass accuracy, detection specificity, speed, and cost of the analysis. The different tools also have strengths and weaknesses according to a sample's purity and available amount/concentration, but in general, compared to gel techniques, mass spectrometry has a much lower limit of detection, higher sensitivity, better resolution, and improved mass accuracy. Tandem mass spectrometry (MS/MS) exploits the dissociations of ions to give the nucleotide or amino acid sequences. In addition to determining the primary structural information, mass spectrometry can reveal secondary structure information, such as the number and location of disulfide bonds or details about alpha helices and beta sheets. Tertiary and quaternary structural information, such as how the molecule is folded and how the protein interacts with DNA or other proteins, can also be gained through coupling other techniques with mass spectrometry. All of these inferences are achieved through the interrogation of the number and behavior of ions, molecules with mass and charge. In fact, the two most important pieces of raw output generated by mass spectrometry instruments are simply the defining characteristic of the ion (its ratio of mass to charge or m/z) and how many times that ion was observed.[41, 42] Consideration of the ions' charges with respect to their mass is one of the pivotal aspects contributing to the precision of mass spectrometry instruments. Although some scientists have a tendency to deprecate the usefulness of mass spectrometry analysis as a routine analysis or mere formality in confirming the identity of their protein, many mass spectrometry methods are becoming increasingly crucial to research and development

efforts tasked with confidently analyzing the identity and quantity of a single protein amidst complicated biological background or characterizing a broad swath of proteins within complex organisms and microbial communities.

## 1.2 Mass Spectrometry and Shotgun Proteomics

### *1.2.1. Shotgun Proteomics*

There are two fundamental approaches to mass spectrometry: the measurement of intact proteins ("top-down" proteomics) or the measurement of peptides from a complex protein mixture ("bottom-up" or shotgun proteomics). Although both can be used for large-scale proteomics experiments,[43] top-down proteomics[44, 45] is far less developed than bottom-up proteomics. One of the primary challenges to top-down proteomics is figuring out a way to analyze proteins across a wide mass range. The size of an intact protein under analysis may range from 10 kDa to 300 kDa, but the mass range constraints for instruments designed for multiply charged ions (200 to 1700 m/z) may advocate the use of analyzers that can sacrifice resolution to accommodate a wider mass range. For intact proteins, which are more likely to be highly multiply charged, charge states often merge together, making it difficult to deconvolute the peaks without ultra-high resolution instrumentation, such as that afforded by FTMS.[46] In addition, intact proteins tend to be stickier on columns and unless the columns are hydrophilic enough, it is difficult to push them off. Also, since proteins maintain their higher-order structure, their retention time may be very different and not necessarily time-dependent. A last challenge faced by top-down proteomics is that the fragmentation tends to be spottier. Logistically, identification through intact proteins is more of a "hit-or-miss" approach when compared to bottom-up proteomics, in which there are numerous peptides for each protein. With intact proteins, there are fewer chances of getting the right measurements. There could also be a number of modifications, including PTMs, truncations, and metal ion adductions, which would alter the molecular mass.[47] In total, there are several unresolved challenges in the sample

preparation, instrumentation, and informatics interpretations that are currently limiting the use of top-down proteomics for high-throughput characterization studies.

In comparison, shotgun proteomics has seen tremendous strides in the past two decades. The analogous “shotgun” strategy employed in genomics and transcriptomics relies on the fact that the sample is rendered more tractable for the experimental measurement, but the subsequent success relies on informatic processes to identify the constituents of the sample based on the reconstruction of the measured fragments. In shotgun proteomics using mass spectrometry, proteins from a complex mixture are denatured, reduced, and enzymatically digested into peptides before they are analyzed. These peptides are more similar in size, composition, and the number of charges they carry compared to the more complicated array of intact proteins. However, it is common to separate the peptides into even more similar fractions to achieve more comprehensive inclusion of the various types of peptides generated from the protein mixture. In fact, there are a number of nuances within protocols to prepare samples for the best coverage achievable by mass spectrometry analysis.

Oftentimes it is helpful to first perform a series of ultracentrifugation steps in order to separate the extracellular, the cellular, and the membrane-bound proteins. Acquiring the extracellular components is easiest to do first. Sample preparation for this initial separation requires a round of washes of the intact microbial cells for simultaneous collection of the supernatant. Lysing the cells by sonication followed by a centrifugation step separates the cellular content into the supernatant. Additional centrifugation enriches the supernatant for better collection of the soluble proteins out of the whole cellular fraction. For membrane proteins, however, a slightly more rigorous protocol is in order. Maximizing protein solubility and therefore enriching the extraction of membrane proteins is best be achieved by the addition of a detergent like sodium dodecyl sulfide (SDS), but this reagent can cause significant analyte suppression in electrospray mass spectrometry. A number of variant protocols have been developed in order to improve the comprehensive solubilization of proteins, but each change or optimization in the sample

preparation needs to be taken into consideration in downstream informatics processing as these details can affect the expected and observed behavior of spectra. Proteins can then be precipitated and purified via TCA precipitation. Proteins are then denatured (typically by 8M urea) and reduced (by 10 mM DTT) so that they no longer have higher order structure or disulfide bonds that could obscure their surface area and minimize efficiency of the enzymatic digestion. Clean up with Sep-Pak and a small amount (0.1%) of formic acid in water is often needed to remove contamination as unanticipated adducts will severely lower the resolution of the spectra.

Digesting proteins into peptides is one of the key elements of shotgun proteomics, but for the approach to be the most effective, peptides should ideally have similar lengths, ionization properties, and MS-compatibility. Trypsin is the most common protease used for such digestions, primarily because it cuts at the carboxyl side of arginine and lysine residues and generates peptides approximately 10 amino acids long. These frequencies are general rules, but for membrane proteins trypsin may not be able to generate peptides between 5 and 50 amino acids long. Nevertheless, the ultimate goal in choosing the right protease is to generate peptides that provide the most sequence coverage for the most number of proteins. More specifically, the only sequence coverage that will contribute to the protein's identification will be those peptides that ionize well. This is another reason why trypsin is most commonly used: in addition to the frequency at which it cuts proteins, the basic residues yield well-defined paired with high charge states that tend to be readily identifiable with collision-induced dissociation (CID or collision-activated dissociation, CAD). Despite the general preference for trypsin, many other proteases (Table 1.1) have been chosen for specific experimental designs. Ultimately, the selectivity, pH range, optimum temperature, denaturing conditions, and digestion time all play critical roles in the number and types of peptides generated from proteins.

Although the goal of digesting proteins into peptides is to reduce the variability among the measurements, the next critical component of the shotgun proteomics experiments is to fractionate the sample to maximize the depth and breadth of peptides identified by the

Table 1.1. Common proteolytic enzymes used for digestion step in shotgun proteomics experiments.

(Table adapted from Sigma-Aldrich "Protease Profiler." www.sigmaaldrich.com/life-science/proteomics/mass-spectrometry/protease-profiler.html)

| **Enzyme** | **Specificity** | **Optimal pH** |
|---|---|---|
| Trypsin, Proteomics Grade | Carboxyl side of Arg and Lys | pH 8.0 |
| Asp-N | Amine side of Asp and Cys | pH 6.0 - 8.5 |
| Glu-C | Carboxyl side of Glu and Asp | pH 4.0 – 7.8 |
| Lys-C | Carboxyl side of Lys | pH 8.5 |
| Arg-C | Carboxyl side of Arg | pH 7.5 – 8.5 |
| Chymotrypsin | Carboxyl side of Tyr, Trp, Phe, Leu | pH 7.0 - 9.0 |

MS instrument. One of the more commonly used approaches for the next dimension of separation is an online liquid chromatography (LC) system incorporating strong cation exchange (SCX) and reverse-phase (RP) columns, also known as MudPIT (multidimensional protein identification technology, Figure 1.2).[48-50] In general, the first phase, SCX, incorporates a negatively charged stationary phase that tightly grabs the positive molecules and allows the negative molecules to elute first. Eleven consecutive pulses of increasing ammonium acetate salt concentration (0-500 mM) achieve charge-based separation. Each salt pulse is followed by a 2-hour RP gradient elution that separates peptides by their hydrophobicity, allowing the less hydrophobic peptides to elute off early before the more hydrophobic, sticky peptides requiring higher concentrations of organic solvent.

More specifically, these two orthogonal separation techniques can be easily incorporated into the middle of the three main steps in a liquid chromatography experiment: sample injection, sample elution and detection, and column re-equilibration. There are two main ways to load samples for an LC experiment: using a six-way valve and a pressure cell column. Although using the six-way valve is the most common method, the volume is limited by the column size and flow rate as the sample is loaded on to the internal or external loop. Packing and loading the sample with a pressure cell allows for a much larger range of volumes to be added. Before the elution process begins, 100% of the starting solvent should be added to the sample. Depending on the type of elution chosen, the elution composition may be constant or marked by gradual or discrete changes. If the mobile phase remains constant, the chromatographic separation is known as an isocratic elution. If solvents are added gradually over time, then the elution composition, too, will change gradually over time; this is known as a gradient elution. Step gradient elution is a modification of gradient elution that involves changing the solvents in a discrete step-wise fashion. After a certain point, however, the column reaches equilibrium and the eluents do not change regardless of the increasing solvent. Historically, detection techniques have included UV, electrochemical, and fluorescent methods,[51-54] but now

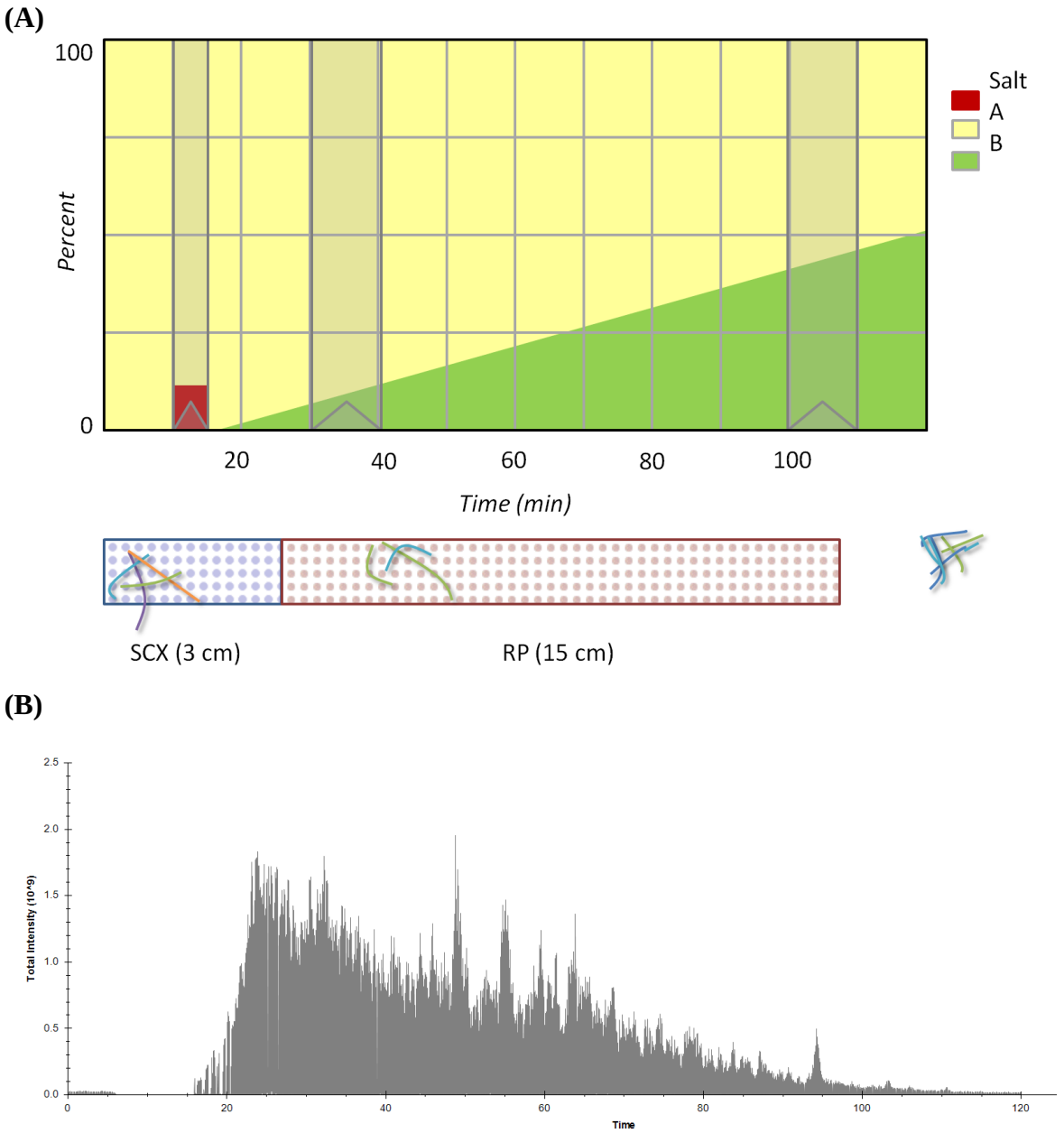



Figure 1.2 The MudPIT strategy involves 2 chromatographic components: strong cation exchange (SCX) followed by reverse-phase (RP) separation of peptides.

(A) Changing the solvents in a discrete step-wise fashion generates salt pulses that push the peptides off the columns according to their hydrophobicity. (B) Chromatograms graph the total ion currents (TIC; y-axis) collected across time (x-axis).

many HPLC experiments can be directly coupled to electrospray mass spectrometry (more details in Section 1.2.3).

Confidence in the detections relies heavily on the assumption that the column was equilibrated before the elutions began. Therefore, after every LC run it is important to re-equilibrate the column by adding incremental amounts of salts to remove any excess proteins and to "clear the slate" for the next separation. A highly organic solvent is often used to re-equilibrate SCX columns.

Both strong cation exchange and reverse phase chromatography involve a form of separation that works best with analytes that are charged or polar. SCX incorporates a negatively charged stationary phase so that it grabs those molecules that are positive. This tight binding allows the negative molecules to elute first followed by the positive ones. Strong salt solutions are added to help separate the charges and facilitate the elutions. In adsorption chromatography, the stationary or solid phase often consists of beads decorated with compounds whose properties determine which molecules are retained. Reverse phase chromatography involves a non-polar stationary phase that generally incorporates hydrophobic alkyl chains of C4, C8, or C18. Smaller alkyl chains such as C4 are more amenable to proteins because the shorter chains are less hydrophobic. C18 chains are better suited for hydrophilic peptides because they have to be pulled a little harder. The non-polar components elute more easily, but in general, bead size, column lengths, pressure, and temperature are factors that affect the absolute elution times of each peptide.

The effectiveness of these coupled separations can be measured by how well one can characterize and resolve the resulting elution peaks. In the extracted ion chromatogram, these peaks represent total ion currents (TIC; y-axis) collected across time (x-axis). Ideal chromatograms have discernible, highly resolved chromatographic peaks- not overlapping peaks that are either too close to each other to be resolved, nor more subtly, a collection of peaks that are sitting on top of each other due to co-elution. Determining an appropriate sample load can greatly affect the chromatographic performance. Insufficient

sample may be below the instrument's sensitivity limits, whereas too much sample might overload the column and dramatically impact the possible number of peptide identifications. Changing between an online and offline LC method could also impact the relationship between chromatography and initial peak observations. Offline techniques tend to clean better than online, but online chromatography touches fewer surfaces so there is less sample loss. Indications of too much salt, column bleed from a previous salt pulse not completely eluting all of the peptides, or decrease in sensitivity due to sample loss are other visual clues that may inform the researcher that the separation protocol may need to be adjusted for the particular sample under investigation.

While the 2-D separation techniques separate peptides based on charge and hydrophobicity, the success of the MudPIT strategy heavily relies on the merits of the accompanying MS instruments to provide sufficient sensitivity, resolution, high mass accuracy, and analytical dynamic range to adequately capture the range of peptides present in each fraction. Recently, sequencing speeds have dramatically improved such that more scans with high-resolution data can be collected. The new LTQ Velos (Thermo Scientific, Waltham, MA) not only has a greater speed of spectral acquisition, but it also achieves a greater analytical dynamic range and sensitivity than its predecessors. In other words, it can analyze data points from peptides that exist in the sample across a wide range of abundances, including those that may be low abundant and those that may not ionize very well (resulting in low signal-to-noise ratios).

### *1.2.2. Mass Spectrometry Instruments*

The three basic components of a generic mass spectrometer are the ion source, analyzer, and detector, whose respective main functions are to generate, sort, and identify ions.

Ion sources, such as Electron Ionization (EI), Chemical Ionization (CI), Matrix-Absorbed Laser Dissociation Ionization (MALDI) or Electrospray Ionization (ESI), generate ions from molecules in a sample that may be from solution, surfaces, or solids, so that the ions can be separated by their mass-to-charge (m/z) ratios. Among a number of additional distinctions, ionization methods can differ by how much sample is required as input, what pressure level is ideal, whether they generate ions continuously or pulsed, whether they produce singly or multiply-charged ions, and the harshness of their fragmentation. Because of the wide variety of types of samples and the strengths and weaknesses inherent to each ionization method, there is no ion source that is ideal for all applications. Similarly, there is no mass analyzer that is ideal for every possible experiment. Sorting ions in the mass analyzers falls into two types of processes, defined by their type of fields: static fields, such as those found in sector, time-of-flight, or ion cyclotron resonance mass spectrometers; or dynamic electric fields, such as those found in linear quadrupoles or quadrupole ion traps.[38-40] All of these analyzers exploit the charge of ions as a handle to steer and manipulate their positions. Whether in response to electric, magnetic, or a combination of both fields, the charged properties of ions allow them to be filtered or trapped so that detectors can distinguish the molecular components of the sample. The detectors feed their input into data systems that process the information and output the signal intensities in a human-readable fashion. Finally, one or more database searches are typically performed in order to produce a list of identifications and descriptions corresponding to the significant peaks.

Matrix-assisted laser desorption ionization (MALDI) is an ion source that is often coupled with a time of flight (TOF) analyzer to separate the singly-charged ions. MALDI, a softer ionization technique, involves forming a crystal with the sample and a

UV-absorbing matrix, pulsing a laser beam at the crystal, and lifting the ions out. The matrix helps soften the high energy of the laser, leaving craters in the crystal where desorption occurred and the analytes were lifted out into the gas phase. The ions steal protons from the matrix, typically only a single hydrogen at a time. [55]

Electrospray Ionization (ESI) is a type of ion source that captures ions by converting them from the liquid phase to the gas phase.[41, 56] The samples are first protonated in solution either by adding 0.1% acetic acid to a 50:50 mixture of water and acetonitrile. The solution is then injected into the instruments (by direct infusion or from HPLC attachment) and continuously pushed through a highly charged (3-6 kV) metal needle at a rate of approximately 1 µL/minute.

"Ideal" solvent conditions for ESI-MS often differ from "typical" solvent conditions employed in biology labs for protein studies.[57, 58] Typical biological solution phase conditions are not very amenable to clear identification of proteins of interest by ESI-MS ionization. Electrospray ionization operates on the principle that anything that is charged will be seen- including unwanted detection of solvents and buffers. Many buffers that are used in HPLC separations, such as TFA, are such good ion-pairing reagents that they create unstable sprays. Other adducts from separation techniques, such as salts with sodium or potassium ions, can give undesired peaks that contaminate the mass spectrometry profile. The presence of carrier proteins, such as bovine serum albumin (BSA), as well as denaturants like guanidine and urea, show peaks that may also overwhelm the protein of interest, or at least increase the noise level. In fact, typical MS requirements are fairly stringent for solvents, pH levels, salts, protein co-factors, and sample amounts. Ideally, MS instruments work better with pure water as a solvent, but some small amount of acetonitrile is acceptable. Because ESI-MS relies on pre-formed ions, buffers that are slightly acidic tend to produce better results. Salts and co-factors are considered contaminants and should not be present at all. Therefore, as a compromise between biological and MS preferences, solvents are generally composed of both water and acetonitrile (50:50 by volume). Their pH ranges from 3 to 7 for most proteins.

Buffers are typically low concentrations, 1-100 mM ammonium acetate. For macro-ESI, protein concentrations are uM in solution volumes of 100 to 1000 uL, but macrospray does not tolerate 100% water. Nanospray, electrospray with flow rates ~ 25 nL/min, is more tolerant to salt. Processes to ensure purity include removal of molecules that can be charged, including protease inhibitors like EDTA, cleanup by dialysis or reverse-phase C4 Zip-Tips, and desalting columns before eluting into MS.

Some samples are better-suited to ESI than others; those that are charged in solution, including inorganic anions and cations, organic acids and bases, and synthetic polymers and biopolymers, are particularly amenable because they form ions more quickly. Because ESI is a soft ionization technique that can produce multiply charged ions, special care should be taken when the mass spectra are interpreted. Mixtures can be difficult to analyze because each compound could give a number of peaks for each component with a different charge. This may result either in overlaps of signals or hiding less abundant peaks. The monoisotopic mass, which should be the peak with the smallest m/z ratio, may not be visible if there is a particularly large abundance of highly charged molecules skewing the scale of the relative intensity. However, this capability of detecting ions with multiple charges increases the range of the mass of proteins that can be detected. In fact, ESI instruments can measure proteins up to 100,000 Da. Despite the inclusiveness of large molecules, ESI can be used for samples that have very low concentrations, from 250 fmol to 10pmol, and are generally very accurate in their mass determinations. The accuracy greatly depends on the purity of the sample, as inclusions of salts, buffers, detergents, and other contaminates that carry charges cannot be discriminated from the molecules of interest.

Analyzers, the components of the mass spectrometry instruments responsible for sorting ions, generally fall into two categories that are defined by the type of fields they use to manipulate ion motion: static fields, such as those found in sector, time-of-flight, or ion cyclotron resonance mass spectrometers; or dynamic electric fields, such as those found in linear quadrupoles or quadrupole ion traps.

Selecting the most appropriate analyzer for a given analytical experiment is generally weighed by a common list of figures of merit including input limitations, quality of the measurements, and performance metrics. Mass resolution and mass accuracy are the figures of merit most often used to describe analyzer specifications, but they are also the most commonly confused. Simply put, mass resolution is the *sharpness* of the peaks, and accuracy is whether the peaks of mass/charge ratios are *in the right place*. The general formula for resolving power,

$$R_m = m / \Delta m_{resolution}$$

can be used to determine the resolving power of a mass analyzer where m is the mass/charge of the ion and $\Delta m_{resolution}$ can be estimated by measuring the peak's full width at half-maximum (FWHM). Sharper peaks, which have narrower FWHM and smaller $\Delta m_{resolution}$, result in larger numbers for the unit-less measure $R_m$.

Accuracy, which is only concerned with the lateral placement of the determined peaks from their true values, is typically described in terms of parts per million or Daltons:

$10^6 \times \Delta m_{accuracy}/\Delta m_{measured}$, where
$\Delta m_{accuracy} = \Delta m_{true} - \Delta m_{measured}$

It is important to keep in mind that the lower the number, such as 0.0001 Da, the better the mass accuracy. [59, 60]

While the higher resolution instruments also tend to follow the same trend in terms of their relative accuracies, it is possible for spectral interference to affect the measurements and cause larger deviations from the true mass/charge calculations. Using normal scan rates, the mass resolution of analyzers typically follows the general trend: linear time-of-flight (500-1.000), quadrupoles (1,000-2,000), ion traps (1,000-2,000), reflection time-of-flight (2,000-10,000), sectors (5,000-100,000), and Fourier transformations (5,000-

1,000,000). These last three analyzers offer high sensitivity and therefore more accurate mass determinations as well.

Due to the softer ionization of electrospray, ESI is most often employed in tandem mass spectrometry approaches with linear quadrupoles and ion traps to draw more information out of the fragmented the ions. Linear quadrupoles derive their name from applying alternating radio frequency (rf) and direct current (dc) voltages to four linear rods arranged symmetrically to direct ions from the source to the detector. These instruments, therefore, lend themselves to both targeted approaches in which a very narrow m/z window (down to a single m/z) is permitted or an extremely large m/z range is allowed to pass.[39 40]

Quadrupole ion traps (LTQs) are essentially 3-dimensional versions of linear quadrupoles in which the z direction is exploited as yet another motion that can be used for further separation. The electrodes supplying rf and dc voltages take the form of a ring and 2 end-caps. Although the principles operating in the mass filtering process between quadrupoles and traps are somewhat similar, the ion traps have many more favorable characteristics that extend their functionality and performance. Ion traps are high performance devices that achieve higher sensitivity, resolution, and m/z values.[39, 61]

Many of these features were preserved and improved in the recent adaptation of the quadrupole ion trap to produce the orbitrap. An orbitrap separates ions of different m/zs by monitoring their axial oscillations and rotations around a central electrode, but the most notable difference between the quadrupole ion trap is that the orbitrap assumes the ions travel along a spindle-shaped central electrode in a constrained z direction. Measured frequencies of ions moving in the angular, orbiting motion around the central electrode and the oscillations along the z-axis can then be selective while detecting the composition of m/z ratios collected. One of the most valuable attributes of both linear ion traps and orbitraps are their ability to easily facilitate multi-stage mass spectrometry ($MS^n$) for selected ions.

Multi-stage mass spectrometry, which is used to interrogate higher order protein structures as well as provide sequences of proteins and peptides, is most commonly termed as tandem mass spectrometry (MS/MS). MS/MS employs two sequential stages of mass spectrometry to achieve greater fragmentation detail: the first scan provides a general idea of what is in the sample, and the second selectively filters and fragments the precursor or parent ions of interest based on their m/z ratios. One might assume that it is always preferable to use the highest resolution and mass accuracy possible for analyzing every scan (both the precursor MS scans and following MS/MS scans), or at least to use the same type of analyzer for both, but it is becoming increasingly popular to use hybrid instruments that have quadrupole and TOF analyzers, or two different types of ion traps. In fact, using high mass accuracy for the precursor ion and low mass accuracy for the fragment ions from linear ion trap-Orbitrap instruments (LTQ-Orbitrap) is a widely applied instrumental configuration. Because low-resolution scans are much faster than high-resolution, the "high-low" strategy afforded by these instruments can increase the overall identification and quantification of peptides over "high-high" strategies. Aside from the sacrifice in resolution using LTQ for fragmentation, small fragment ions (~ 30% of the precursor ion mass) are typically difficult to detect. However, in the newer LTQ-Orbitrap instruments, precursor ions can be dissociated in different compartments: within the linear ion trap, or in a "high energy" octopole collision cell. Whereas ion activation is usually performed in the linear ion trap, the new addition of an octopole collision cell allows an additional activation technique of higher energy to fragment ions in the far side of the instrument and then transfer them back to the C-trap for analysis.

Even these new high-performance instruments cannot measure all peptides presented to the mass spectrometer. Data-dependent settings are needed to handle the hundreds of co-eluting peptides so that the MS can make "smarter," more-informed decisions about which analytes should be targeted for MS-sequencing. For example, it would not be a good use of MS time to repeat analyses of the same ions over and over again. Therefore, once a precursor ion is selected for sequencing, its m/z value is put on a dynamic exclusion list. If an ion is on this list, it will not be selected for fragmentation, at least for a specified period of time during the LC run. This enables the MS to measure ions across

a wider abundance range rather than constantly resampling the most abundant ions. Typical settings require an ion to stay on this list for 1 minute before it is automatically taken off and allowed to be considered for selection. Alternatively, a maximum list size can be specified in which case, if a sufficient number of ions have been added to the exclusion list, ions are popped off the list in first-in, last-out order until the list size meets the size requirements.

Another variable in setting the dynamic exclusion list is to also parameterize the isolation window to determine whether two ions are essentially the same. Depending on the analyzer, one may be able to resolve a difference of 0.001 Da or 0.01 Da. Therefore, setting an isolation window involves a tradeoff between specificity and sensitivity. It is advantageous to have a wide isolation window to go through as many species as possible (maximizing sensitivity), but if the window dimensions are small enough to allow two peptides, the fragmentation spectra will be very messy and difficult to assign.

### *1.2.3. Ion Activation & MS/MS Fragmentation*

One of the more popular forms of fragmentation, collision-induced dissociation (CID, or collision-activated dissociation, CAD), uses energetic collisions to cause an immediate, single fragmentation. Briefly, the precursor ions selected in the MS1 phase are accelerated by electrostatic pulses and forced to collide with a large neutral target gas, such as helium or argon. When the precursor ions hit these curtain gasses, the weakest bond in the peptide breaks and creates smaller, fragmented ions. The amount of energy transferred is a function of the energy of the ion ($E_{ion}$), the mass of the collision gas ($m_{gas}$), and the mass of the ion ($m_{ion}$). For a single collision, the center of mass collision energy ($E_{com}$) is represented by the following equation:

$$E_{com} = E_{ion} \times (m_{gas} / (m_{gas} + m_{ion}))$$

As $E_{com}$ increases, the number of fragmentations increases because the internal ion energy increases. These spectra created by peptide fragments are usually visualized by plotting

each ion's mass-to-charge ratio (m/z; x-axis) against the relative ion intensity (y-axis) [62](Figure 1.3).

For protein analysis, one can expect MS/MS fragmentation to break peptides in predictable ways. In fact, there are a limited number of types of ions that one would expect to see. A common nomenclature for the types of ions from peptide fragmentation is described below (Figure 1.4).

The notation of an *a*, *b*, or *c* ion indicates that during the cleavage, the charge was on the fragment with the N-terminus. Conversely, the notation of an *x*, *y*, or *z* ion indicates that the charge was on the fragment with the C-terminus. The most common types of ions, *b* and *y*, represent a cleavage between the carboxyl and amide group. For each amino acid at position *i* within a peptide sequence of length L, the N-terminus ions can be thought of having a relationship with a paired C-terminus ion using the equation:

$$C(i) = L + 1 - N(i)$$

where N(*i*) represents the position of the N-terminus ion, defined by N(*i*) = *i*, and C(*i*) is the complementary C-terminus ion for a given amino acid position *i*.

Depending on the type of fragmentation method used, one is more likely to see different types of N- or C-terminus ions.

The mass difference between each peptide's adjacent fragment ions represents a single amino acid, so one could manually take the differences between all of the peaks in the MS/MS scan, compare the mass differences to the masses of amino acids, and stitch the peptide sequence back together. Some complications in this process may arise due to the presence of different ion types, or additional peaks that represent noise or chemical additions that are not from the peptide fragmentation. Several computational algorithms have been developed to avoid these noise peaks and determine the best peptide-spectrum match (PSM).

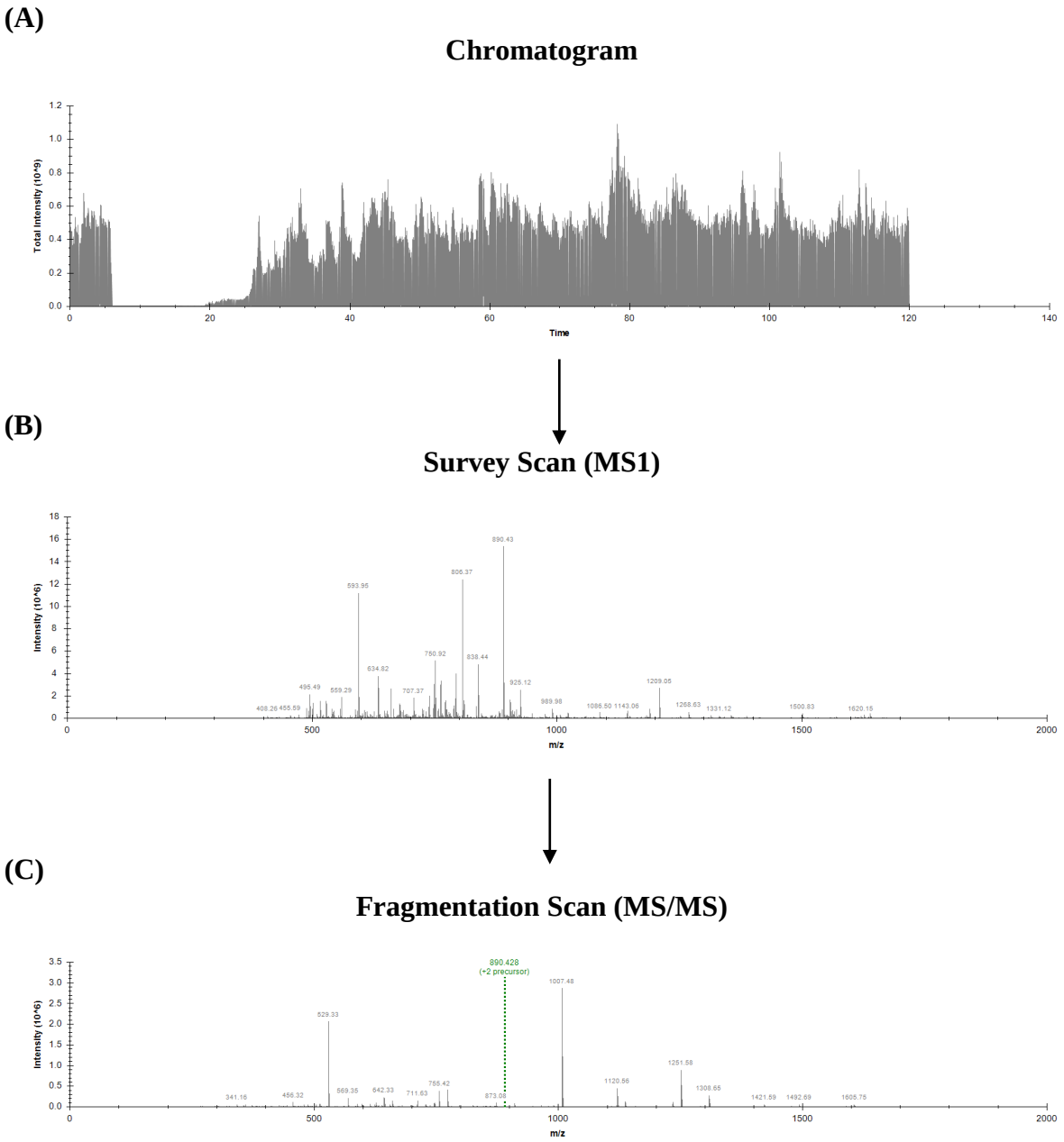



Figure 1.3. Illustration of data collected in a tandem mass spectrometry run.

(A) The chromatogram reflects the separation of peptides by liquid chromatography, graphing the total collected intensity (TIC; y-axis) by time (minutes; x-axis). (B) The survey scan details which individual ions (m/z values) are observed. Their most abundant peaks are selected for fragmentation (MS/MS). (C) The fragment ions in an MS/MS scan can be used to sequence the peptide.

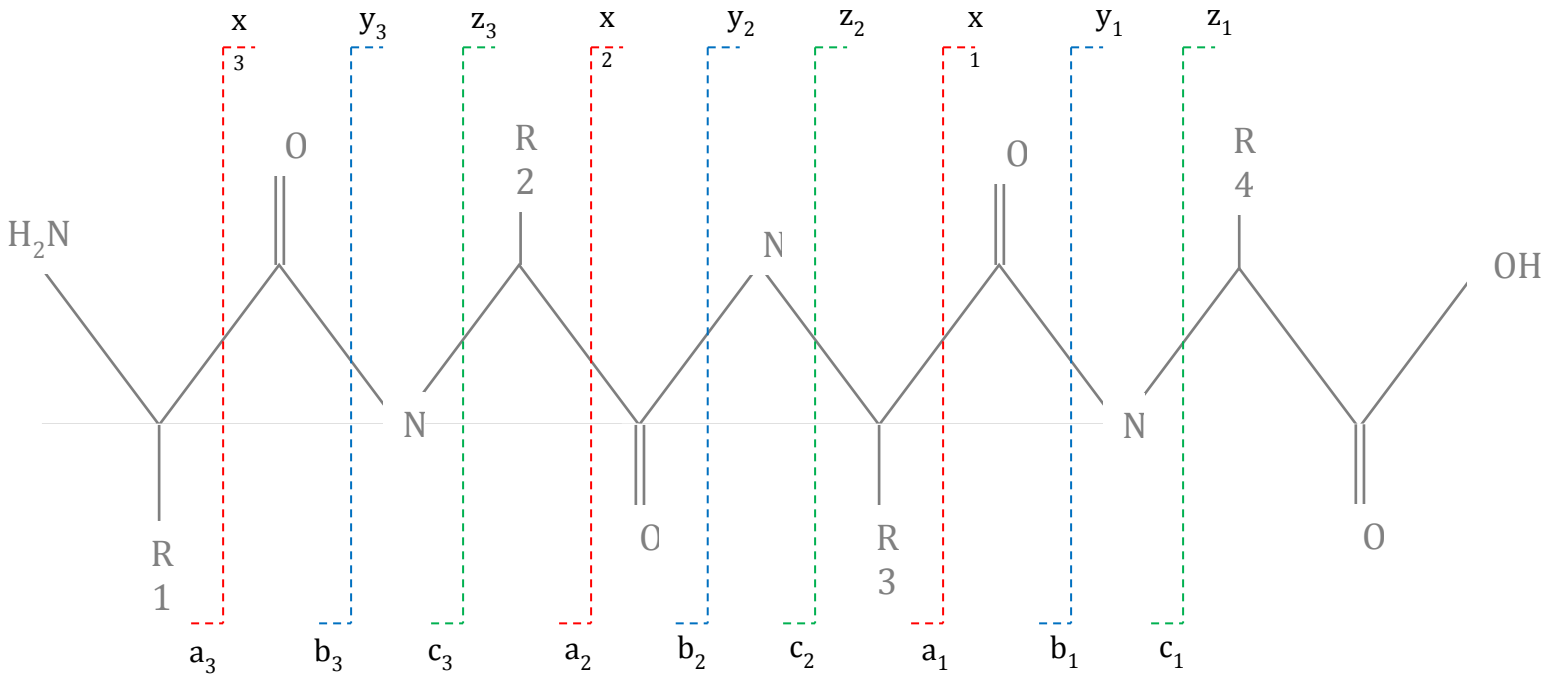
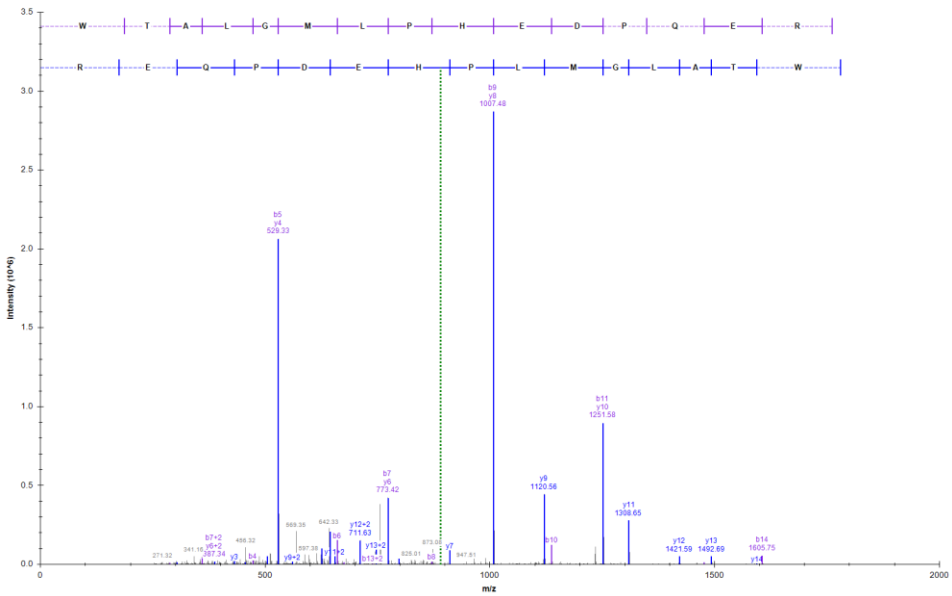



Figure 1.4. Illustration of collision induced fragmentation of a polypeptide.

A peptide backbone with four amino acid residues (R) and the types of fragment ions generated in a CID MS/MS spectrum. If the charge is retained on the N-terminal side, the fragment ion is classified as either a, b, or c. If the charge is retained on the C-terminal side, the ion type is x, y, or z.

To aid in this process, most algorithms do not work from spectral information to infer peptide sequences, but actually work in the opposite order. That is, because peptides fragment in predictable ways, *in silico* fragmentation can generate the ion series that is expected to be found in the MS/MS spectra from each peptide sequence. In CID, peptides generally fragment along the peptide bonds (as opposed to backbone bond cleavages) to generate b and y ions. Figure 1.5 illustrates the b and y ion series predicted from the peptide sequence and which peaks in the spectra correspond to the expected m/z values, thus contributing to the peptide's identification.

Energetic collisions often cleave off amino acids' post-translational modifications, so one would not typically expect to see the addition of a phosphoryl group in the masses of the ions generated from CID. This "invisibility" of modifications is not strictly true for electron transfer dissociation (ETD or electron capture dissociation, ECD). ETD is similar in concept to CID except that it typically causes more cleavages than CID and different backbone cleavages. It requires low energy electrons and long reaction times, fragmenting at the most labile bonds. Because of the low-energy added, modifications typically stay on the amino acids so one can expect to see their addition to the m/z values for the ions. For example, if a peptide had a phosphorylated serine, one would expect the masses of the ETD ions that include the modified residue to be shifted at least the mass of a phosphorylation (79.979 Da), and thus greater than the corresponding ions CID would generate for the same peptide. However, even for unmodified peptides, one might not expect to see the same types of ions present for CID and ETD fragmentations. Whereas CID favors b and y ions, ETD more commonly produces c and z ions for certain peptides. Thus, the two techniques are complementary to each other. In fact, recent studies have favored the implementation of a decision-tree instrument setting that makes a decision about which fragmentation technique to use for a particular MS scan based on the charge and m/z ratio of the ion, whether CID or ETD is most likely to yield a better distribution of ions.




| b | b(+2) | | | | y | y(+2) |
|---|---|---|---|---|---|---|
| 187.087 | 94.047 | 1 | W | 15 | 1779.848 | 890.428 |
| **288.134** | 144.571 | 2 | T | 14 | **1593.769** | **797.388** |
| **359.171** | 180.089 | 3 | A | 13 | **1492.721** | **746.864** |
| **472.255** | 236.631 | 4 | L | 12 | **1421.684** | **711.346** |
| **529.277** | 265.142 | 5 | G | 11 | **1308.6** | **654.804** |
| **660.317** | 330.662 | 6 | M | 10 | **1251.579** | **626.293** |
| **773.401** | **387.204** | 7 | L | 9 | **1120.538** | **560.773** |
| **870.454** | **435.731** | 8 | P | 8 | **1007.454** | **504.231** |
| **1007.513** | **504.26** | 9 | H | 7 | **910.401** | **455.704** |
| **1136.556** | 568.782 | 10 | E | 6 | **773.342** | **387.175** |
| **1251.583** | **626.295** | 11 | D | 5 | **644.3** | 322.654 |
| 1348.635 | **674.821** | 12 | P | 4 | **529.273** | 265.14 |
| **1476.694** | **738.851** | 13 | Q | 3 | **432.22** | 216.614 |
| **1605.737** | **803.372** | 14 | E | 2 | **304.162** | 152.584 |
| 1761.838 | 881.422 | 15 | R | 1 | 175.119 | 88.063 |

Figure 1.5. A peptide-spectrum match.

(A) Peptide WTALGMLPHEDPQER (+2) matches 24 peaks in the MS/MS spectrum. Observed peaks that match the peptide's b and y ions are highlighted in purple and blue. (B) The observed m/z values from the spectrum that matched the expected m/z values for the peptide are listed in the table in bold.

The most recently adopted fragmentation technique, higher energy collision dissociation (HCD), is most noteworthy for overcoming the limitation of CID fragmentation known as the "one-third rule," that is, the loss of mass ions less than 1/3 the parent ion mass. Whereas CID converts kinetic energy to internal, mostly vibrational energy that affect the weakest bonds, HCD uses a beam-type energy that results in fragment ions with higher levels of energy, allowing for not only more primary, but also secondary dissociation to occur. The increased energy also reduces rearrangement reactions and increases the reproducibility of HCD fragmentation spectra of the same peptide. Furthermore, HCD and CID differ in where they physically occur within the instrument. HCD activates ions in the collision cell at the far side of the instrument, requiring ions to pass through the C-trap before they are analyzed by the orbitrap. In total, then, HCD is able to achieve high resolution and high accuracy for both the precursor and fragment scans. The tradeoff for these desirable figures of merit result in diminished sensitivity compared to CID as well as slower duty cycle. However, instruments that have this extra HCD collision cell also include a new device, termed an S-lens, that is touted to improve total ion current by 10-fold, arguably minimizing the disadvantages to HCD fragmentation. In tandem mass spectrometry analyses using HCD for fragmentation, one could either send the ions to the orbitrap or LTQ for measurements. Detection of an LTQ measurement requires the ions to be accelerated towards the curved surface of a dynode cup which then directs the ions to an electron multiplier, which amplifies the signal so that the ion current leaving the detector is an amplified intensity or signal. On the other hand, detection of ions in an orbitrap involves a broadband image current detection and fast Fourier transformation algorithm, which converts the frequency of each orbiting ion into a m/z signal. Because the orbitrap is much closer, sending ions there would take a shorter transmission time and would most likely result in a higher yield, not to mention very high resolution (~100k) and highly accurate mass measurements (<1 part per million). However, the MS/MS scans themselves would be much slower than if the ions were sent to the LTQ. Depending on the goal of the experiment, one analytical strategy may be more appropriate than the other.

### *1.2.4. Spectral Interpretation*

Data from a mass spectrometer are generally very simple in their format: measurements are reported as x, y values indicating the m/z ratio and intensity value for each peak observed, whether it was for a precursor scan or MS/MS scan. Translating these results into peptide sequences is the researcher's responsibility. For the past two decades, computer software has rescued researchers from tedious manual inspection of each collected spectrum, but much effort and debate has been put forward in order to reconcile simulating a person's logic in deconvoluting spectra with computational modeling probabilistic expectations of likely measurements.[63, 64]

One of the foremost recognitions that an interpreter must note is that calculations of an expected peptide mass must consider the isotopic abundances of each element within each amino acid. Within the periodic table of elements, the average mass is almost always represented. This means that it takes into consideration the distribution of all of the possible isotopes' masses for each element. The average mass is somewhat limited to the calculations of "typical" relative abundances for the isotopes. Similarly, one may commonly find peptide masses are calculated as the sum of the monoisotopic masses of amino acids, using the most abundant isotopic mass for each element, but this, too, has limitations. When thousands to millions of ions are collected for a given amino acid, one must consider the distribution of isotopic abundances. In fact, the mass of a detected peptide may be up to a tenth of a Dalton off of the calculated, typical peptide mass.

Another factor that complicates interpretation of spectra is the inclusion of noise peaks. Not every peak that is recorded in a MS/MS scan belongs to a peptide. Although abundant peaks in a fragmentation scan are more likely to be "true" ions compared to less abundant peaks, the sensitivity and specificity of the mass analyzer needs to be taken into account when defining the noise level.

## 1.3 Informatic Analogues of Analytical Processes in Shotgun Proteomic Studies

### *1.3.1. Spectrum to Peptide Matching*

Each dissociation method fragments peptide sequences in a predictable way. If one knows the peptide sequence and charge state, all of the analyte's possible fragment ions (m/z values) can be calculated. This property of fragment scans allows interpretation of a peptide from a spectrum by scoring how well a sequence's expected list of m/z values align with the observed m/z peaks within the scan. The interpretation of a peptide from a spectrum is called a peptide-spectrum match (PSM). Although all PSM algorithms count how many observed m/z positions are within certain tolerances of the expected m/z's, many algorithms also take into consideration one or more other spectral features, such as the intensities of the matching peaks, to score a PSM. Scoring the PSM has traditionally been assessed through static measurements, such as a score of 3 is always "good" on a scale where lower numbers indicate better matches. However, it is becoming increasingly popular to give dynamic scores that are accompanied by probabilistic likelihoods. Using these newer algorithms, the same score of a 3 may be "good" in one search and "bad" in another search because their respective probabilities indicate the scores are 5% and 25% likely due to error. How different computational tools calculate these scores is discussed later in Section 2.1.1, but commonalities between these algorithms are most fundamentally based on whether the algorithms implement a database search or *de novo* method.

Database searching is one of the most popular methods for identifying peptide-derived MS/MS spectra, perhaps because the software provides answers that fall within a list of expected identifications. Database-searching algorithms operate on the assumption that the researcher has a complete list of all of the proteins expected to be present in the sample (e.g., a fasta file of protein sequences translated directly from the genome sequence of the organism or organisms). Each of the proteins within this list undergoes

an *in silico* digestion to generate a completely list of possible peptides and each peptide in turn undergoes an *in silico* fragmentation to generate a theoretical, "expected" spectrum against which each spectrum within the collected dataset can be compared. Since this entire process heavily relies on the comprehensiveness of the input (given as a protein FASTA file), an ideal list of proteins would be based on a complete genome with reliable gene calls, including all potential alternative splice variants or isoforms, as well as the addition of all possible types of post-translational modifications (single and multiple) that could be found on expressed proteins. Including all of these biological intricacies can cause a huge expansion in the search space for finding potential peptide-spectrum matches. This problem is only partially optimized by the implementation of a greedy algorithm that assumes the best short-term matches represent optimized long-term matches. There are a number of different scoring schemes with various parameters for thresholds, error rates, and other empirically-dependent metrics that provide additional suggestions for true identifications of peptides. Despite the number of possible ways one can arrive at the "true" identification of a peptide, database searching cannot identify all possible peptides from a tandem mass spectrum. The primary advantage of database searching also serves as its greatest disadvantage: one needs to know what he is looking for before performing a search. In other words, if the scientist does not make the searching software "aware" of all potential methylations, truncations, adducts, and other such biological contributors to shifts of masses, the software will not even consider those as possible sequences. In addition, the researcher is responsible for setting appropriate parameters for significant signal to noise ratios, standard deviations of analyzer resolutions, and other pertinent values to ensure that the scores returned by the peptide-spectrum matches are calibrated correctly and do not exceed or underestimate the capabilities of the instrument. Although one would conclude that all possible known modifications or additions should be included in the search and that the filtering should be less stringent to ensure all observed masses correspond to a theoretical mass, one important caveat is that not only are such searches computationally expensive, larger search spaces can dramatically increase the likelihood of false positives (incorrect

identifications). Thus, only pertinent modifications with reasonably small allowances for error are generally included as search parameters.

In response to the increasingly high-throughput of analytical technologies and informatics packages, a slightly different approach to database-searching has recently been proposed. Assuming that reproducible measurements are collected with very high mass accuracy and high precision, researchers have investigated the construction of spectral libraries as an optimization or alternative input to database-searching algorithms. Spectral libraries are collections of experimental data that have been confidently assigned to particular peptides within a proteome. It has been suggested that such spectra can take the place of the computationally-derived "expected" fragmentation spectra for all of the peptides observed in a proteome, ensuring a better, more realistic experimental-to-experimental comparison for scoring the newly generated datasets. Scores evaluating how well each spectrum from the spectral library matches against the experimentally observed spectrum still consider similar features (m/z fidelity, intensities, etc) as traditional peptide-spectrum matches, but additional meta-information can be easily acquired. For example, if the same database is to be queried repeatedly, the confident peptide-spectrum matches can be tracked and used both as diagnostic indicators of internal consistency in terms of instrument performance and sample quality as well as consistency of PSM assignments among the entire scientific community. These direct comparisons, of course, have a number of caveats, primarily involving the determination of whether the data is generated by the same types of experimental protocols for sample preparation, analytical strategies, and types of instruments. Another factor in evaluating identifications from spectral libraries is that one can only identify peptides that have been previously identified. Similar to deciding what proteins should be included in a database search, when considering if using a spectral library is appropriate, one must carefully consider whether the experiment could potentially provide evidence for sequences that were previously unsubstantiated or undetected.

*De novo* sequencing is a complementary method to database searching. This method requires finding a stretch of 3 or more amino acids described by the peaks in an MS/MS spectrum, and then submitting this sequence to BLAST to see whether the residue sequence matches any known proteins. The approach has merit when there is not any previous information available. Because it does not make any assumptions about the data, it can be helpful for identifying proteins that may not be in one's database, either because of post-translational modifications or single nucleotide polymorphisms. Coupling *de novo* sequencing to database searching can greatly optimize the efficiency and accuracy of searches through the database alone. One can either use *de novo* sequencing before database searching in order to limit how many sequences the algorithm has to search or after database searching, when there are only a few unidentified sequences remaining. Alternatively, new hybrid methods of integrated *de novo* and database-searching approaches have been proposed as a compromise between achieving confident identifications and allowing for flexibility in discovering novel or unexpected variants and modifications.

### *1.3.2. Peptide to Protein Mapping*

Interestingly, most shotgun proteomics analysis tools have separate peptide-spectrum matching software and protein inference programs. The peptide-spectrum matching algorithms generally output an exhaustive list of the top 5-10 candidate peptide sequences and charge states that could potentially have generated each MS2 spectrum. Then, a distinct program uses one or more scores to check whether any of the peptides pass a certain score threshold indicating a confident match, and then it determines which peptide candidate is the best and which protein identification(s) are inferred from the peptide sequence.

In the most straightforward scenario, a single peptide suggests the identification of a single protein within the expected proteome. This unique identification is considered a confident piece of evidence, especially when compared to a shared or *degenerate* peptide,

which may map to multiple proteins. As Figure 1.6 illustrates, ambiguous protein inferences can cause discrepancies in reporting the number of proteins identified in a sample- should one report all proteins that have at least one detected peptide (maximal list) or should one only include those proteins that have at least one unique peptide (minimal list)? Some researchers among the scientific community have decided that it largely depends on the biological system in question. Especially for higher eukaryotes and microbial communities, the protein inference problem causes much controversy as neither of the proposed solutions is desirable. The maximal list may overinflate protein counts as well as complicate quantitative measurements, while the minimal list may under-represent protein identifications and result in throwing away more than half of quality peptide-spectrum matches.

Nomenclature has been developed to qualify different levels of ambiguous protein inferences, primarily characterized by the number of unique or shared peptides contributing to a protein's identification. *Distinct* proteins are those that are identified by only unique peptides, while *differentiable* proteins are those that are identified by both unique and shared peptides. The most ambiguous identifications, *indistinguishable* proteins, are identified by only shared peptides. More specifically, an *equivalent* protein is one that is identified by an identical group of shared peptides with another protein. As a subtle difference, a *subset* protein is one in which it is identified by peptides common to another set of peptides corresponding to a larger protein. A more extreme example, a *subsumed* protein is one that is identified by peptides that are subsets of two or more larger proteins.

Depending on the software, a false discovery rate (FDR) is calculated at the peptide-spectrum level or protein level and may also influence whether a peptide is assigned to a spectrum and whether the peptide is included in the protein call. Some of the most debated filtering criteria for compiling a final protein list include the minimal number of unique and/or non-unique peptides to substantiate a protein call and the minimal number of spectral counts to provide evidence for a protein.

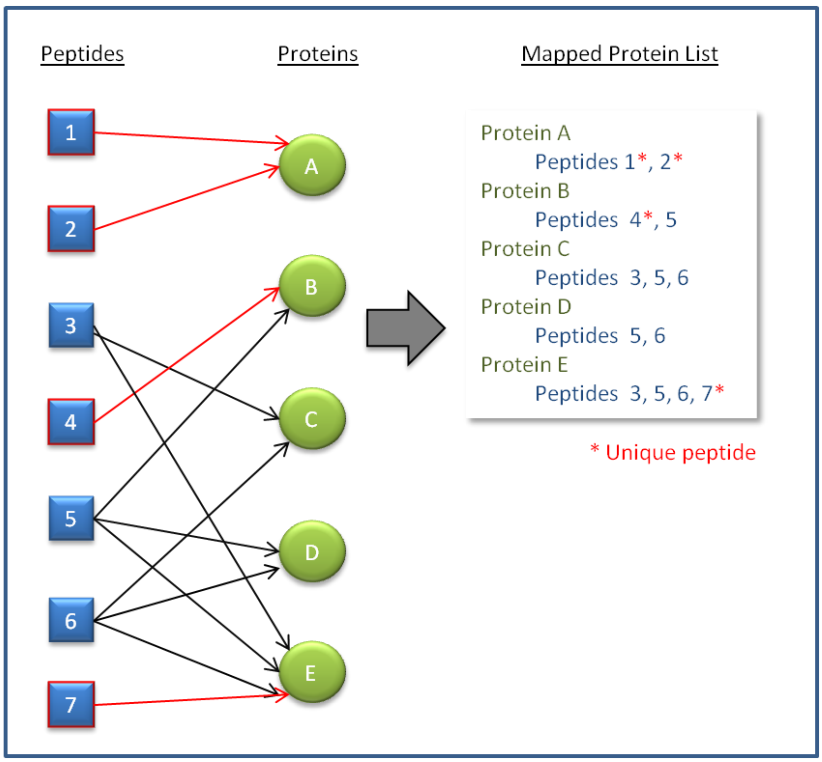



Figure 1.6. Graphical illustration of the protein inference problem.

Oftentimes a single peptide identification may map to multiple proteins, adding ambiguity to the final protein list. Proteins supported by unique peptide evidence, such as Proteins A, B, and E are considered more confident identifications. Because Protein D and E share a peptide and D does not have any unique peptides detected, we cannot be sure whether both D and E are present in the sample or just Protein E.

### *1.3.3. Differential Protein Expression*

Initially inspired by more traditional biochemical techniques, common analytical methods for comparing protein abundances involve collecting data on a known entity and systematically comparing its information to that generated by an unknown entity. The absolute quantification of each protein identified in a proteomics run can be calculated by normalizing its abundance to the measurement of a spiked internal standard, that is, a protein with known sequence and quantity added to a protein mixture. To calculate the most accurate comparisons, the selection of standard peptides should reflect the abundances as close to the sample peptides as possible (a 1:1 ratio). In the past several years, researchers have developed a number of high-throughput techniques to generate a standard peptide for every sample peptide, including methods that allow for multiplexing up to six or eight samples in a single run. Large scale chemical labeling techniques, such as AQUA,[65, 66] TMT,[67] iTRAQ,[68] and iCAT,[69] use stable isotope amino acids or add small isobaric tags to peptides so that the overall physiochemical responses of the labeled peptides mimic those of the sample peptides. In such methods, the only detectable difference between a standard peptide and a sample peptide's elution profile, electrophoretic properties, and mass is a distinguishable increase in mass in the heavy isotope's MS2 spectra. The two main disadvantages to these approaches are the unknown percentages of incomplete heavy amino acid replacements and the limited number of samples that can be compared in a single mass spectrometric run. An alternative group of isotopic labeling methods, metabolic labeling, requires the organisms to be grown in media with only heavy isotopes (such as $^{15}N$, $^{18}O$, and $^{13}C$) as well as grown in separate cultures of normal minimal media. While this process circumvents most of the shortcomings associated with chemical labeling methods, it is largely limited to use on organisms that can be cultured (i.e., bacteria and archaea). Mammals can be fed nutrients that have heavy $^{13}C$, but these SILAC methods[70] are much less widely adopted. Common to all of the labeling methods is the addition of biologically cumbersome methods and expensive synthesis and/or incorporation of isotopically-labeled amino acids. Also

ubiquitous to each of the labeling methods is an underlying dependency on one or more metrics of protein abundance.

Metrics of protein abundance are inherent characteristics associated with fragment scans or spectra measurements, including but not limited to tandem spectral counts, spectral intensities, peak area, and peak width. Spectral intensities are an intuitive measure of abundance but they may not fairly represent the peptide's abundance because the fragmentation scan may not pick the spectra at the height of its eluting peak, especially if the peak is very wide and spans several consecutive scans. It therefore might be logical to determine peaks by their width, but it is sometimes difficult to determine where one peak ends and another begins, particularly if there are co-eluting peptides or high signal-to-noise ratios. Peak widths would most likely also need to be normalized against peak height, because a peak that has low intensity but is very close to the noise-threshold may have the same area as that of a peak that is very intense but only is captured for a scan or two. Peak areas, then, are a popular combination of peak intensity and peak width, but their measurements suffer a common problem: peak identification and disambiguation. Another possible shortcoming to looking at peak shape is the effect of ion suppression. If an isotopic packet comes from a peptide that happens to ionize very well and co-elutes with precursor ions that do not ionize as well, the physiochemical properties of the dominant packet can mask or suppress the features of the secondary packets. Spectral count, the number of times a peptide is seen in a MS/MS scan, is less sensitive to these issues and has been one of the more commonly used metrics of protein abundance, primarily for its simple calculation and what it represents.

One or more of these metrics can be used with or without the labeling methods. However, if relative abundance metrics are chosen, it is important to choose effective means of normalizing the measurements. Normalization is the process of standardizing measurements so that a comparison of measurements maintains biological and statistical integrity. Most important to the normalization process is having an idea of what to expect, i.e. the general type of distribution that the numbers follow. Typical

normalization methods account for the total assigned spectral counts within a run as well as eliminate biases of protein lengths. Such methods do not mean the adjusted measurements then follow a normal distribution, but once the log10 is taken of the normalized values, it is common practice to observe a normal distribution.

Achieving a somewhat normal distribution is helpful for performing statistical tests of differential protein expression. Analysis of Variances (ANOVA) is the most powerful test that can discern whether two values are different given the context of their related measurements, but it assumes that the measurements are independent and taken from a normal distribution. ANOVA tests, therefore, are useful in comparing whether Protein A has a significantly different relative abundance value in Condition 1 compared to its relative abundance value in Condition 2. Those proteins that demonstrate evidence of significant differences are generally considered the most interesting and suggested as implicated in the cell's response to the different biological conditions. One would expect that technical replicates may show minor discrepancies within an expected variance, but biological replicates are helpful in determining whether differences observed between samples are more likely attributable to biological or external factors.

### *1.3.4. Functional Analysis and Data Visualization*

Throughout mass spectrometry experiments, visualizing the collected data is a helpful way to quickly perform quality checks, diagnose problems, and capture general behavior of the measurements. In fact, software plays a key role in calibrating the instruments, ensuring that they are running properly as scans are collected, and interpreting the results. When validating a peptide-spectrum match, researchers often find it useful to visualize the matched peaks within the scan as a bar chart of m/z ratios graphed by their relative intensities. Similarly, when looking for evidence of a protein identification based on peptide calls, overlaying the peptide sequences on top of the amino acids belonging to the protein provides a more easily comprehensible understanding of sequence coverage than a mere percentage. More importantly, after a list of proteins have been identified from a

sample, it is common to investigate the functions of the proteins and determine whether any noticeable trends or anomalies stand out. While a list of protein names might not be particularly informative to a researcher, a report or graph summarizing the presence, absence, or change in abundance of a group of proteins may be more meaningful and reinforce confidence in the significance of the findings. Differential protein expression is typically depicted as a heat map colored on a dichromatic scale to represent relatively high and low expression levels within the collected samples. Hierarchical clustering of the proteins arranges the identifications in blocks of identifications that have similar trends of abundance so that one can visually track the magnitude and agreement of the variations in measurements. Upon inspection of clustered proteins that have similar behaviors in abundances, researchers generally inquire as to the functional categories (as defined by euKarytoic Orthologous Groups, KOG or Clusters of Orthologous Groups, COG) or pathways common to a subset of proteins. This functional analysis is typically illustrated as a bar chart graphing normalized abundances per category or as a super-imposed KEGG (Kyoto Encyclopedia of Genes and Genomes) map of identifications within metabolic and regulatory pathways.

## 1.4 Conclusions

With the rising popularity of mass spectrometry for proteomic studies and the general interest in “omic” and other large-scale endeavors, there are ever-increasing needs for scientists who handle data in a way that is consistent with the current understanding of the biological system of interest, understand the biases and advantages of the analytical measurements, and the apply or develop the informatics components that transform measurements into value. There are, however, numerous tools available for each of the main informatics processes critical to mass spectrometry interpretation, each asserting that their method identifies limitations in a previous study and their new project performs comparably or overcomes these hurdles. Critical evaluation of existing tools that attempt to solve the same or similar problem is an absolutely necessary step in the development or integration of any bioinformatic workflow. Therefore, while keeping in mind that there

are computational bottlenecks in the existing informatic analyses, this dissertation attempts to survey the current state of the field and identify where there are areas for improvement or adding novel features, with particular focus on methods that match peptides to scans, infer protein identifications, and quantify relative protein abundances.

# CHAPTER 2: Current Tools & Workflows Employed for Analysis of Large-Scale Shotgun Proteomics Experiments

## 2.1 Evaluation of Existing Proteomic Informatic Tools and Approaches

### *2.1.1. Database Searching Algorithms*

Most current tools in use today can trace their ancestries to software written 10-15 years ago, but they have adopted new computational strategies, analytical philosophies, and interpretational considerations. Before using a new software package, it is important to have not only a fundamental understanding of the new features it affords, but also an appreciation of the differences with respect to its precursor. The following discussion briefly highlights similarities and differences between 2 workflows currently employed by our lab: SEQUEST[63] (v 0.27) and DTASelect[71] (v1.9) compared to Myrimatch[72] (v2.1) and IDPicker[73] (v3.463).

SEQUEST is one of the most popular choices for database searching algorithms, not only because it was one of the first tools to match raw spectra to peptide sequences, but also because it has inspired a number of adaptations and optimizations. The original SEQUEST algorithm was written by Jimmy Eng, Ashley McCormack, and John Yates in 1994. It is comprised of 4 major steps: data reduction, search method, scoring method, and cross-correlation analysis (Figure 2.1).

SEQUEST's data reduction step involves retaining only the 200 most abundant ions in each scan (ranked by intensity) and renormalizing their intensities to 100. All peaks within +/-1 Da window are equalized to the intensity of the higher value, and peaks that fall within 10 Da of the precursor ion are removed.

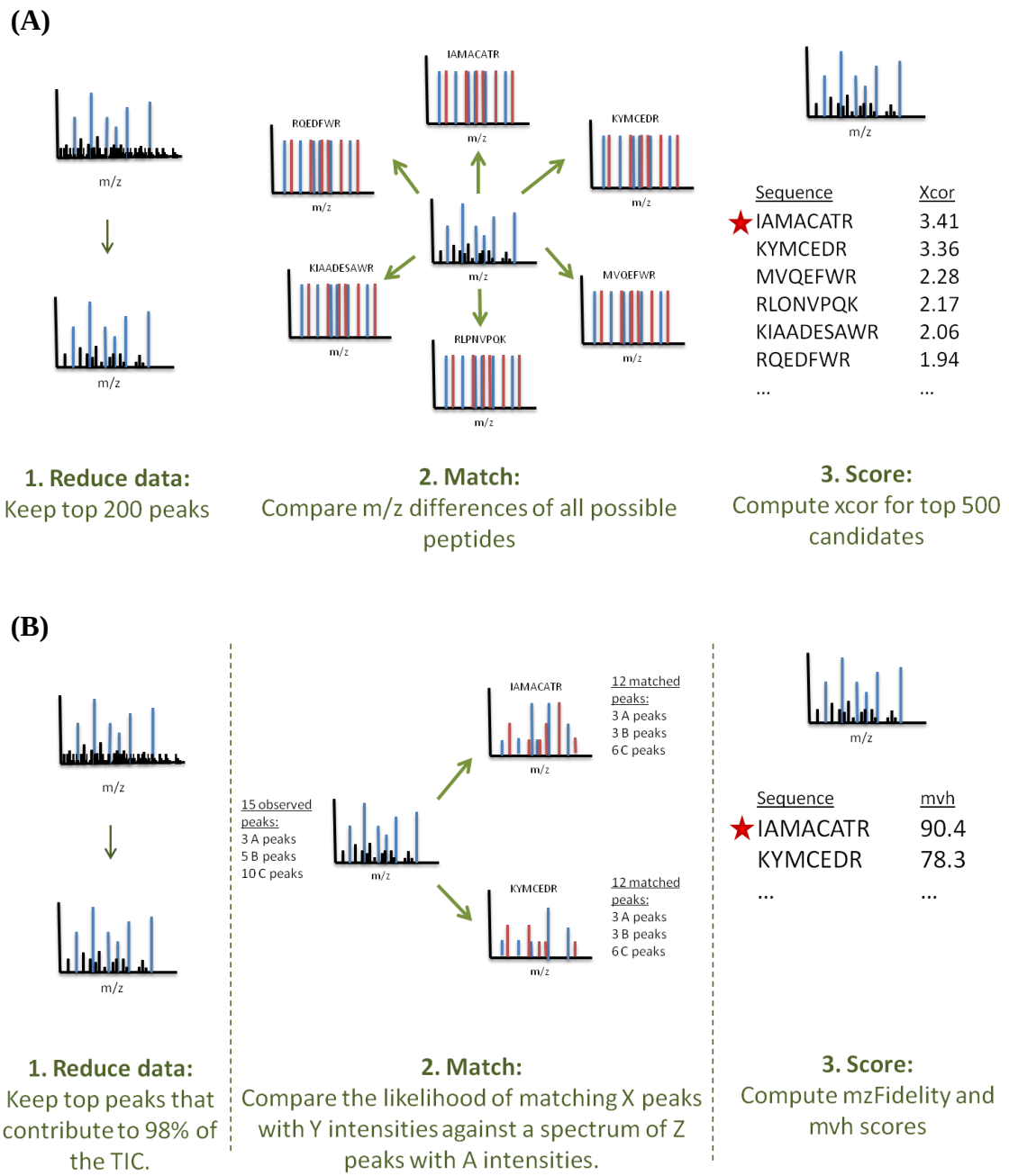



Figure 2.1. Primary methods of SEQUEST and Myrimatch for generating the peptide-spectrum matches (PSMs).

(A) SEQUEST scores PSMs by comparing the m/z ratios of theoretical spectra generated from all possible peptides in the proteome. (B) Myrimatch scores PSMs by calculating the probabilities of observed m/z ratios and their intensities matching a peptide's expected m/z ratios and their intensities.

The search method within SEQUEST involves scanning each protein sequence in the given fasta file and finding all combinations of consecutive amino acids that fall within the observed mass range of the precursor. Then, the program predicts the mass-to-charge ratio values of b and y series ions for the candidate list of peptides. Any chemical modifications to amino acids that were specified in the search parameters can be specified as "*static*," that is, they are assumed to affect every occurrence in the sequence, or "*dynamic*," indicating that the algorithm should test the fit of the amino acid with and without the modification. These mass shifts are taken into consideration for the b and y ion calculations. Additional parameters can be specified in the required *sequest.params* configuration file.

When computing the scoring method, SEQUEST first renormalizes each spectrum's observed peaks into 3 *intensity classes* based on how well they match the predicted b and y ions: "Class A": observed peaks matching b and y ions; "Class B": observed peaks matching +/- 1 Da of b and y ions; "Class C": observed peaks that match to a neutral ion loss of water and ammonia and a ions. The number of class A and B ions are summed and additional weight is given for consecutive fragment ions and immonium ions. This preliminary score (Sp) provides a quick preview of peptide-spectrum match scores, but has biases towards longer peptides.

SEQUEST then employs a cross-correlation analysis score to compare the top 500 candidate amino acid sequences with a reconstructed (theoretical) spectrum based on the m/z ratios of predicted b and y ion series, and relative intensities are assigned one of three intensity values using the classes described above. (Class A ions are predicted to be twice as intense as Class B ions, and Class B ions are predicted to be 2.5 times as intense as Class C ions.) The correlation score (*Cn*, more commonly called *xcorr*) is an average distance of differences between the observed and reconstructed spectra. To accommodate for the increased number of ions expected to be produced by sequences with higher charge states (and therefore more opportunities for deviations from theoretical spectra), there are different standard thresholds for what is considered a "good" score based on the

sequence's charge state. Typically, we accept xcorr values > 1.8 for +1 sequences, > 2.5 for +2 sequences, and >3.5 for +3 sequences. The scoring thresholds are higher as the charge state increases in order to reflect the additional numbers of fragment peaks generated by multiply charged ions, which should act as supporting evidence for the peptide identification. An additional score, the deltCN, is the difference between the 2 highest xcorr values within a spectrum, and gives an indication of how well SEQUEST could distinguish the top peptide-spectrum match compared to the second-best peptide-spectrum match.

The output from SEQUEST, sqt files, records all of the top 500 candidate sequences, their xcorrs, and the deltCN for each spectrum. Typical SEQUEST v0.27 searches take 3-4 hours for microbial isolates, 7-10 hours for complex eukaryotes like plants, and 2 weeks to a few months for metaproteome databases even on large clusters. One of the primary drawbacks to this software is the intense input and output requirements for each search. After each spectrum is compared to all of the peptides within the database, a small temporary file is created, which is only deleted once all spectra have been searched and aggregated into a single output file. Later versions of the software, including the similar algorithms that have since been developed based on the same underlying principles but with modern optimizations to improve speed and accuracy, are indeed consistently faster. Many labs today rely on the improved algorithmic and software engineering techniques found in TurboSequest, Crux, and Tide for high throughput, large-scale proteomic studies.

Over the years, SEQUEST and its accompanying protein assembly software, DTASelect, have inspired many suites of novel algorithms and software packages. New software is being constantly developed to keep pace with the evolving metrics of instrument capabilities, including improvements in their precision, accuracy, and throughput. One such set of tools, Myrimatch and IDPicker, facilitates rapid protein identification in a seamless workflow of database searching and filtering processes. Written 5 years ago by David Tabb, Christopher Fernando, and Matthew Chambers, Myrimatch is a database

searching algorithm that primarily differs from SEQUEST's analysis by including a probabilistic measurement for each peptide-spectrum match (PSM) and thus allowing a more dynamic scoring platform in which each PSM is considered independently and with respect to random chance. It is comprised of 3 major steps: data reduction, search method, and scoring method (Figure 2.1).

The data reduction step involves reading from mzML files (converted from RAW files) and retaining only the top X% of ions in each scan (ranked by intensity). This percentage can be defined by the user, but by default 98% of the TIC is kept. Additional peaks are removed that do not have the minimum number of peaks to fill each intensity class. Integral to Myrimatch's functionality, the intensity classes are characterized by the number of peaks grouped into each class compared to the class below it. For example, under the default setting of three intensity classes, "Class A" peaks have half as many peaks as "Class B" peaks, which in turn has half as many peaks as "Class C" peaks. Under such circumstances, scans must have at least 7 peaks (1 A + 2 B + 4 C peaks) to be considered for sequence matching; otherwise, they are removed from the analysis.

The search method incorporates a novel fragmentation model for defining b and y ions. While the series of b and y ions follow conventional calculations for singly and doubly charged sequences, triply charged sequences employ a different method. For each amino acid in a predicted peptide, a three-tiered weighting system is used based on the residue's basicity (its ability to hold an extra proton). More specifically, Arginine, Histidine, and Lysine are given a weight of 5, Glutamine and Asparagine are given a weight of 3, and other residues are given a weight of 1. Depending on where the fragmentation occurs, the side of the peptide that has the larger summed score is given 2 protons and the other side is given a single charge. If the "duplicate spectra" option is selected, Myrimatch will attempt to assign a sequence with a +2 and +3 charge if it cannot determine the charge state of the scan and both sequences will be reported in the output. This is a rare occurrence in high mass accuracy data, but for LTQ searches, the option is strongly recommended.

The scoring method first involves calculating whether a predicted m/z peak falls within the expected fragment tolerance and then it determines the intensity class of the matched peak. If there are multiple peaks that fall within the fragment tolerance, the closest peak is chosen. The distance from the expected m/z and the intensity class is incorporated into an initial *mzFidelity score*. Once all of the peaks are matched for a peptide, the number of missing peaks and the number of expected peaks for each intensity class are calculated for the PSM. These numbers are then used in a probability calculation to determine how likely this match could happen by chance. Because there are probabilities associated with each intensity class, Myrimatch uses a multivariate hypergeometric distribution in the calculation of this *mvh score*. Whereas SEQUEST reported the top 500 candidates for each scan, Myrimatch ranks and reports the top 5 peptide hits for each scan based on the mvh score. To accommodate users who are accustomed to SEQUEST's scoring, Myrimatch can also report its version of the *xcorr* for each PSM.

One of the most attractive features of Myrimatch is its inclusion of many parameters that the user can define or choose to leave as the default configuration. The range of options available include different data reduction parameters (SpectrumListFilters, TicCutoffPercentage, MaxPeakCount), precursor and fragment scan filtering criteria (AvgPrecursorMzTolerance, MonoPrecursorMzTolerance, and FragmentMzTolerance), fragmentation settings (FragmentationRule, FragmentationAutoRule), modification considerations (MaxDynamicMods, DynamicMods, StaticMods), digestion information (CleavageRules, MaxMissedCleavages, MinPeptideLength, MaxPeptideLength), and many more. Some of the unique settings that are helpful but also confusing to those unfamiliar with the algorithm are the "PrecursorMzToleranceRule" and the "UseSmartPlusThreeModel" option. The PrecursorMzToleranceRule option is an acknowledgement that sometimes the mass spectrometer will pick the wrong isotope as the monoisotope of an eluting peptide (exacerbated in our new data-dependent settings). When using narrow tolerances for monoisotopic precursors, this can cause identifiable spectra to be missed. Myrimatch can be instructed to adjust the observed precursor m/z to

the expected monoisotopic precursor m/z. Additionally, Myrimatch can customize the number of fragment ions compared to each candidate peptide within the database using the Smart Plus Three Model. In most search algorithms, each peptide sequence of the same length generate the same number of fragment ions: a sequence of 10 amino acids will always have theoretical +1 y5, +2 y5, +1 b5, and +2 b5 ions. However, Myrimatch gives the user the option of interrogating the amino acid composition of the sequence and by retrieving the number and position of basic residues (those likely to carry a charge), the expected fragment ions can be generated in a sequence-specific manner. Parameters like this highlight how many algorithms can be developed on the same basic principles but additional nuances afforded by each program can provide valuable contributions to advance specific projects.

A myriad of other protein identification software exists, but most of the more recently developed tools are incorporating probabilities at one or more of the PSM, peptide, or protein levels. Many of these, like PeptideProphet[74] and ProteinProphet,[75] build on SEQUEST scores to compute probabilities and error rates for each identification. PeptideProphet, software developed by Keller in 2002, also uses a SEQUEST xcorr score and deltCN score, peptide length, the logarithm of rank of Sp score and mass accuracy, to calculate a Bayesian posterior probability for each peptide identification, which is combined and re-evaluated at the protein level within ProteinProphet. Others, like Mascot,[76] OMSSA,[77] and X!Tandem,[78] use their own PSM scores to compute probabilities and e-values. Some emerging algorithms are even specifically tailored to analyze labeling methods, such as Sipros' identification of peptide sequences and estimation of 15N atom% from stable isotope probing methods.[79] Studies have shown that direct comparisons of different algorithms chosen for peptide and protein identification consistently demonstrate ~ 70-80% overlap in the results for a given dataset.

Although database-searching algorithms are extremely useful in identifying expected peptide sequences from a dataset, many of the more recent studies are challenging the

primary prerequisite (namely, that one has a comprehensive list of expected protein sequences). Whether the problem is due to a static reference proteome used to analyze a dataset collected from a slightly different genus, or the proteome does not include all possible single amino acid polymorphisms or isoforms, or the list of expected post-translational modifications is ill-defined, there is an increasing need for sequencing proteins whose exact sequence and mass may not be known. Only recently have instrument resolutions and computational resources been able to achieve exhaustive database searching for amino acid mutations.[80] With a peptide-sequence tagging approach, such as that afforded by the combination of DirectTag[81] and TagRecon,[82] one can exploit many of the advantages of database searching algorithms while also incorporating a degree of flexibility in order to identify unanticipated sequence variants.

DirectTag is a tool that infers partial sequences ("tags" that are typically 3 amino acids long) from 4 peaks within a spectrum that match a peptide's expected consecutive ion series. For each collected spectrum, these tags are scored based on intensity, m/z fidelity, and complementarity, and each subscore is converted to a p-value. The intensity subscore is a sum of the ranks of the 4 matching ions among all the spectrum's observed ions sorted by intensity, and its associated p-value denotes whether their summed rank is representative of the entire spectrum. On the other hand, the m/z fidelity score is a summed square error (SSE) of the 4 estimates of m/z values for the first amino acid in the tag, inferred from the position of each of the other matching peaks. The p-value associated with this score conveys the probability of the SSE occurring by chance. The last subscore, complementarity, has two components that factor into its p-value: one to account for the number of peaks in the spectrum that are complementary (paired b and y ions), and the second to account for the agreement of the peptide mass estimated by the complementary peaks. For each tag in each spectrum, Fisher's Method is used to combine the p-values from the 3 subscores. Multiplying each joint p-value by the number of tags matched with each spectrum yields an "expectation value" metric that can be used to filter high-ranking, quality tag matches. These tags can then be input into another software tool, TagRecon, which uses the tags as a scaffold to infer longer peptide

sequences that may contain up to one or two amino acid mutations or mass shift modifications.

TagRecon, software designed to enhance peptide-sequence matches by allowing the comparison of specific or best-fit mass shifts to account for mutations or post-translational modifications of proteins within the expected proteome. Using the tags created by DirecTag as input, TagRecon compiles a list of extended "flanking" sequences around each tag and compares whether the calculated masses of the flanking sequence matches can be explained within the spectrum. The comparison between the flanking sequence and the rest of the spectrum can be allowed up to one or two mass mismatches. If the researcher anticipates oxidations on methionine residues, for example, a flanking peptide sequence that contains a methionine may try to match a spectrum using a mass shift (+16 Da) on the methionine and its downstream amino acids. If the researcher is interested in mutation analysis, the BLOSUM62 matrix is used to determine which amino acid substitutions are permitted in reconciling the mass mismatches. Parameters similar to Myrimatch's scoring criteria are used to determine the best peptide-spectrum match.

Compared to other mutation and modification identification software, TagRecon provides a fast, flexible platform for confidently identifying peptides that have mass shifts altering their expected mass based on a static proteome. Despite the advantages afforded by TagRecon, localizing mass shifts, constraining the number and type of expected mass shifts, and evaluating the biological interpretation of its results are still challenges that have yet to be conclusively resolved. Filtering these results and finding an appropriate false discovery rate and/or false positive rate for each of the identifications is an especially difficult but important hurdle to overcome before confidently identifying modified or unmodified peptide sequences.

### *2.1.2. Filtering Criteria and FDR Calculations*

The standardization of datasets using False Positive Rates (FPRs) and False Discovery Rates (FDRs) are a hot topic of debate among mass spectrometrists. False Positive Rate

(FPR) is a property of an *individual* spectrum as opposed to the False Discovery Rate (FDR), the property of *multiple* spectra. More specifically, FDR is the proportion of incorrect identifications among all identifications judged correct.

FDRs have many advantages and are the most pervasive form of reporting the quality of a filtered dataset. Typically, an FDR is calculated at the protein level, although it may also be reported for peptides as well. When a database-searching algorithm is adopted, it is common practice to append a decoy database to the list of possible identifications. The FDR is calculated by multiplying the number of identifications from the decoy database times 2 and dividing by the total number of identifications. Doubling the counts of decoy identifications assumes that the numbers of unidentified false positives (those that were identified as target identifications, but are not real) are as likely as the number of known false positives.

Ideally, this decoy database should resemble the target protein identifications as closely as possible so as to measure the selectivity of an algorithm as it makes peptide-spectrum matches (PSM). This approach assumes that each spectrum should be given an equal opportunity to match a target peptide and a decoy peptide so that a direct competition of their scores reflects not only how well the peptide sequence matches the spectrum, but also allows for the possibility of the PSM being a false positive. Previous studies have systematically explored whether it is most profitable for the decoy database to be comprised of shuffled sequences from the target database, reversed sequences of the target database, or sequences from the proteome of a completely different organism than the one under investigation. Proponents of a shuffled decoy database preserve the amino acid composition of the target database while simulating a random sequence by simply rearranging the order of the residues. More widely accepted, the reverse decoy databases not only maintain the composition of the target database but also the projected size of the contributing proteins. Oftentimes both of these decoy database strategies include additional distracting proteins or common contaminants, such as human keratin, BSA, or

trypsin, but including an entire separate proteome is not always deemed appropriate. While the additional proteome serves as a comparison of a true biological distribution of peptides, it is considered more likely that the different size and composition of the distracter proteome would introduce biases rather than providing truly equal chances of detecting target and decoy peptides. [83-85]

Most broadly, FDRs consider the context of the entire dataset; the entire collection of PSMs is taken into account when estimating an FDR. Not only do FDRs estimate observed false positives, but they make an effort to estimate hidden false positives as well. Practically, FDRs measure false positives according to the scoring algorithm, not a theoretical model, so the false positives are as specific to the dataset and method of identification as possible. If the target-decoy approach is executed correctly, overlap between target and decoy databases should be an exceedingly rare event. Therefore, a hit to the decoy database represents the error rate of the scoring algorithm regardless of the quality of the match. Whether the match was at the tail end of the likely distribution or near the average, all false positives are weighted the same. However, robust PSM scores can easily differentiate between decoy false positives and target peptides, even if they have equal likelihoods. Elias and Gygi demonstrated that the distributions of considered peptides were practically the same between target and decoy peptides, regardless of mass tolerance. Their study also provided evidence that top-ranked peptides showed a strong bias towards target database hits, unlike lower-ranked matches.[85] Relatively high-scoring decoy searches rarely outscore correct identifications in composite databases so researchers will not be misled to set inappropriately high scoring criteria. Therefore, if a PSM passes an FDR threshold, attributes can be investigated to distinguish features of true identifications.

Despite their widespread use, FDRs have many disadvantages as well. Antagonists to FDRs posit that assigning the same “blanket” FPR to a set of identifications with identical scores is a dangerous oversimplification since the scoring functions of existing MS/MS tools are not based on rigorous probabilistic models and are often inaccurate. In

addition, the target-decoy approach gains criticism for only looking at scores assigned by the search tool but not at the false positive rate of individual identifications. As such, it becomes possible for bogus identifications to be included in the results as long as the overall FDR among all identifications is acceptable. From a cynical perspective, scoring algorithms that identify few or no decoy hits (which seems like an analytically ideal situation) cause problems for FDR calculations, which assume equal likelihood of matching target and decoy peptides. For example, for a given spectrum, the number of matches between the spectrum and a typical decoy database of size n (where n is the same as the size of the target database) is usually zero. To obtain a reliable FPR for an individual spectrum, n could be increased to make a giant decoy database that is much larger than the target database, but this is impractical. Furthermore, FDRs are considered poor approximations of the non-exact solution. It is argued that a decoy database is simply a time-consuming way to evaluate the sum of all probabilities of a spectrum matching a random database over all spectra in the dataset, but not a good way to estimate individual probabilities. Given a database of all possible peptides with a certain length, it is possible to compute the precise number of the identified peptides and thus evaluate the error rate; however, the time required to search this entire database is unfeasible. One of the greatest proposed strengths of FDRs can also be considered a weakness: FDRs assume a virtual coin flip in the likelihood of a target or decoy identification. Ideally, the target-decoy approach assumes that the distribution of scores of incorrect identifications in the target database is the same as the distribution of scores in the decoy database, but in practice, a bogus peptide may get a higher score than a true peptide. Another principle that some consider advantageous and others find disconcerting is that FDRs are based on relative scores that are database-dependent. Searches against a composite target-decoy database yield relative scores that can differ from searches against separate target or decoy databases. Yet another prong that upsets dissenters of the target-decoy approach is that FDRs integrate raw and combined peptide scores. In other words, the algorithms make decisions using both raw PSM scores (like Xcorr) with other data-dependent information that takes into consideration the distribution of scores of all peptides in the database. To summarize, the aggregate measure of a blanket FDR is the

source of much controversy that some see as a beneficial, encompassing description and other see as a diluted approximation.[83, 84]

Those who dismiss using a False Discovery Rate (FDR) propose that False Positive Rates (FPRs) based on peptide-spectrum match (PSM) probabilities have many preferable advantages. The most notable difference in this approach is that FPRs do not require decoy databases, so the entire debate about how to simulate random distracters is completely eliminated. Instead, raw scores generated by PSM-based algorithms typically rely on generator-like functions that assign probabilities of individual PSM matches. This principle of computing FPR probabilities for individual spectra is the second point of contention. FPR proponents assert that by accurately computing probabilities for the individual spectra, the validity of MS/MS "one-hit-wonders" (a single spectrum provides the sole evidence for a protein's identification) is not a debate.[86] The lack of support from other identifications is irrelevant in the FPR philosophy so if probabilities deem the PSMs to be unlikely by chance, these otherwise questionable pieces of information can be retained. Researchers advocating FPRs tout that an exhaustive set of potential matches can be calculated with their approach. By calculating the difference between the best *de novo* spectral interpretation and the best database spectral interpretation, a spectral energy score represents the quality of the match between the chosen peptide sequence against any other possible peptide. In an effort to assure FDR supporters that the FPR approach is not missing or misrepresenting information, researchers still estimate overall numbers of false positive hits within their search even without reporting actual false positive identifications. The spectral probability metric represents the total probability of all peptides with scores exceeding a given threshold. In a recent study comparing the target-decoy and PSM probability approaches, the number of spectra that were matched in a search with a decoy database was very close to the expected number of matches computed by the generating function's spectral probability. In closing remarks in his systematic comparison of the two approaches, Pevzner et al. demonstrated that their algorithm had better sensitivity-specificity than other algorithms' combined scores, especially at very small error rates. [83]

While FPRs contest the FDR approach, using FPRs also has its disadvantages. Some degree of ambiguity is usually associated with each peptide identification, but inspection of which PSMs are correct or incorrect usually does not consider the context of the PSM within the dataset. Thus, the dependence or independence of PSM identifications based on the other identifications in the dataset is the most fundamental point of debate between the two approaches. FDR advocates often find fault in FPR studies over-valuing their statistics because the touted accuracy of the calculated probabilities may not be preserved all the way to the extreme tails of the expected distribution of scores. As one researcher snidely observed against FPRs, "Theory is needed because simulations rarely cover the extreme tails of a distribution." Additionally, while the FPR focuses on PSM-level probabilities, there is not provision for protein-level confidence values. On a final note of discrepancies between FPR and FDRs, some of the models used in estimating likelihoods for PSMs disagree or use different parameters according to the platform or expected distribution of false-positive matches, so the probabilities are not entirely absolute measurements.

Also, measurements from mass spectrometry can only report what was detected, and it is dangerous to extrapolate that the absence of a peptide or protein identification means that it is not present in the sample. However, given an identification, metrics can be applied to describe the likelihood that it is the correct identification. For example, the detection limit is the smallest signal to noise ratio that can be differentiated at a given concentration. Identifications with high signal to noise ratios are more confident than those that are closer to random observations that could be due to chemical or instrumental noise. However, if a protein is only identified by a single peptide, regardless of its spectra's signal-to-noise ratios, this protein identification is typically considered untrustworthy and removed from the analysis. Moreover, if a protein is not identified by any unique peptides, it is also considered ambiguous and typically filtered out of the analysis. One assumes that proteins with more spectra and more peptides are more likely to be the abundant proteins. The low abundant proteins, identified only by spectra that are close to

or under the signal-to-noise threshold, are commonly filtered out. When multiple runs are compared, there is increased variation caused by the lack of consistency in detection of low-abundant proteins. To improve the reproducibility of the reported results, the proteins with few spectra counts may be eliminated from the overall analyses. Pearson correlations can be used to measure the overall reproducibility of identified proteins between replicates and sample types.

### *2.1.3. Protein Inference Approaches*

While SEQUEST is an effective tool for automated identification and scoring of peptide-spectrum matches (PSMs), a complementary program is required to filter the results so that only "quality" PSMs are retained and mapped into the context of protein identifications. DTASelect,[71] written by David Tabb, Jimmy Eng, and John Yates in 2002, performs this function for SEQUEST outputs. DTASelect has 3 primary components: summarization, evaluation, and reporting.

DTASelect's summarization step first extracts the xcorr, deltCN, Sp rank, sequence, precursor m/z, protein name, intensity, and percentage of matched fragment ions for each PSM and sorts them by locus. If a peptide belongs to more than one protein, the PSM information is reported for each shared protein. This information is stored as a DTASelect.txt file so that it doesn't need to be re-extracted each time a new filter is applied. The evaluation step in DTASelect applies the PSM-level and protein-level filtering criteria. The PSM-level filtering criteria keep PSMs that are above the charge-specific thresholds (as described above). The protein-level criteria only retains proteins that have a sufficient number of different peptides ("-p 2" for 2-peptide/protein minimum or "-p 1" for 1-peptide/protein minimum). After filtering, the program stacks protein identifications that have the exact same sequence coverage for a group of peptides. The reporting step formats the filtered DTASelect file into an HTML file and allows for more interactive exploration. Typically we do not use the more advanced features of this part of the program except the plain text and html files.

After DTASelect, there are a number of additional manipulations that must be performed before a dataset is publication-ready. Some of these processes include ppm filtering, enforcement of protein evidence from a minimum of 2 peptides (at least 1 of which must be unique to the database), FDR calculations, and removal of redundant PSMs for more accurate spectral counts.

Just as SEQUEST required DTASelect to refine its search results, Myrimatch's complementary program is IDPicker. IDPicker,[87] initially written in 2007, by Bing Zhang, Matthew Chambers, and David Tabb, is a GUI wrapper program that has undergone a number of changes since its first release. The latest (IDPicker v3.0) is a user-friendly program that essentially incorporates 3 modules: FDR calculation (idpQonvert), protein assembly and filtering (IdpAssemble), and reporting (idpReportFDR).

IDPicker's PSM-level FDR calculation first extracts peptide, sequence, scan, and scoring information from Myrimatch output files (in .pepXML format). By comparing how many forward and reverse hits are allowed at each ranked PSM in the scan, the FDR is calculated for each PSM. PSM-level filters remove any scans that match below the user-defined FDR level. Typically, 5% is the recommended maximum PSM-level FDR, although situations may warrant using 1% or 2% for additional confidence. The next step in IDPicker, protein assembly**,** involves first creating peptide groups and protein groups through a "minimal list" approach to parsimony. Peptide groups provide evidence for the exact same set of proteins and protein groups share the exact same set of observed peptides. User-defined PSM-level and protein-level filters are then applied. This includes minimum spectra per peptide, minimum spectra per match, maximum protein groups, maximum distinct peptides, minimum additional peptides, and minimum spectra per protein. In the context of IDPicker, a distinct peptide is a peptide that is not only unique to the database, but that also has a unique mass. In other words, charge states and modifications to unique sequences increase the number of distinct peptides. Additional peptides provide evidence for proteins that would not be identified if only distinct

peptides were required. For most cases, a protein level between 5 and 10% is acceptable. Also key to this protein assembly step is the formation of peptide groups, protein groups, and clusters. Peptide groups gather peptides that identify a common set of proteins, and protein groups include proteins whose peptide evidence all overlap. These groups therefore represent levels of ambiguity in both directions of protein inference. One property of these peptide and protein groups is that adding evidence from additional MS runs may change the constituents of the groups. Other researchers have suggested caution in relying on this type of clustering, citing that its volatility and data-dependence are not sufficiently biologically relevant for unaided interpretation.

Finally, IDPicker's reporting step basically transforms the filtered information into user-friendly viewing panes for detailed exploration as well as facilitates exporting the information into Excel. IDPicker's initial release had a number of limitations that have since been mitigated to a certain extent, but a few properties that are inherent to its design still remain contentious drawbacks. IDPicker was primarily designed to be a browsing tool for exploring individual information contributing to a spectral identification as well as a means of clustering identifications at the peptide and protein level, based on the which pieces of supporting evidence were shared between two peptides or proteins. In the early versions of the software, an HTML page was created for each protein so that one could open the main page of the project, choose filtering settings, and upon completion, click on a hyperlink on the protein's name to see which proteins had similar information. Each of these HTML pages were re-created each time the user wanted to change the filtering settings, which took a fair amount of time. The layout also required a considerable number of mouse clicks in order to navigate around the dataset, with very little meta- information and limited spectrum-specific information. Newer versions of the software implement a sqlite database file that stores all of the identification information collected during the run, so that when the user is changing filtering settings, he is simply querying the database and generating a different report. The software has an option to export peptide, protein, or spectral level information to Excel, which helps users easily sort, transform, and summarize information. One of the remaining challenges of using

IDPicker for large-scale proteomic studies, however, is its dependence on a Windows-based platform that demands user interaction. Although the GUI is user-friendly, it does not allow the researcher to filter or query a number of runs at a time. Therefore, for complex experimental designs, one is limited to running a single instance at a time and clicking through the settings to arrive at a filtered, summarized dataset. Another debatable feature of IDPicker is not a logistical problem, but an analytical restraint: IDPicker's filtering and FDR calculations require the assignment of protein and peptide groups. IDPicker assigns proteins to a group if they share at least one peptide, which is beneficial for a researcher to have a perspective on the analytical overlap between identifications, but it does not necessarily reflect biological association. More detail about the implications of this feature is discussed in Chapter 4.

Other software packages handle protein inference by reporting all possible proteins that could be identified, choosing one representative protein, or assigning a rank or probability to each protein identification. For example, ProteinProphet[75] ranks proteins according to probabilities computed from the number of peptides, confidence in the peptide sequence, and degree to which proteins are shared between multiple proteins. DBParser,[88] on the other hand, simply ranks proteins according to those with the most peptides. Yet another software, Phenyx,[89] ranks proteins by the number of peptides identified and the protein sequence coverages, but chooses only one representative protein for shared peptides. Despite the abundance of these naming conventions and software solutions, there is yet consensus among researchers about a biologically-meaningful compromise to unambiguously infer protein identifications from shared peptides.

### *2.1.4. Differential Protein Expression Algorithms*

Among the advocates of label-free relative quantification, most prefer to use spectral counts for evaluating protein abundances. Most of the efforts in analyzing differential protein expression, then, have focused on significance tests for determining whether two

spectral count measurements for a protein are statistically different from each other. One of the primary premises of all of these statistical tests demands knowledge or estimation of an expected distribution of spectral counts within the dataset. Despite the widespread adoption of spectral counts as the metric of choice for these tests, there is a substantial disparity within the community about what kind of distribution spectral counts do and should follow, or whether Bayesian probabilities should estimate likelihoods rather than cumulative distribution functions. Consequently, the number of normalization methods attempting to adjust spectral count distributions is as numerous as the suggestions for statistically-robust measures for evaluating differences among values. Of the more noteworthy methods to date, the beta-binomial method, generalized linear mixed effects models, quasi-Poisson method, and normal distributions have been most commonly compared.

In 2008, Choi et al from Nesiviskhi's lab proposed the implementation of QSpec, [90]quantitation software using spectral counting to measure protein expression differences between two datasets. From their perspective, one of the biggest disadvantages in previous quantitative efforts was that the statistics relied too heavily on signal-to-noise ratios to adjust spectral count distributions within a run, causing biases that favored large differences in highly abundant proteins. They assert that signal-to-noise methods lose power because they are performed on a per protein basis, rather than taking into account all of the proteins within a replicate. While they admit that most other algorithms focus on the highly abundant proteins because they are the most reproducibly present across replicates or across samples, a primary aim of QSpec is to include a model that is robust enough to handle the absence of replicate samples. In short, QSpec uses hierarchical Bayes estimation of generalized linear mixed effects model (GLMM) where the spectral counts are considered random numbers from a Poisson distribution, described by a large population of proteins (those identified within a replicate). Therefore, regression parameters are modeled for each protein as random effects, and if replicate information is available for the protein, the coefficients are "shared" by each instance of the protein so that intrasubject variation is preserved and consistent across the dataset. Random effects

are also contextualized by every sample and for every treatment or condition. Model parameters are estimated using a Markov chain Monte Carlo method, and the number of iterations can be specified by the researcher. In particular, the treatment term, which is described as a random variable from a Gaussian distribution with inverse gamma-distributed variance parameters, is tested for significance, and if it is found not to be contributing to the description of the data, the model is "reduced." For each protein, a significance test is performed to determine whether there is more evidence for the "full" or "reduced" model. Proteins that have more evidence for the full model are considered statistically differentially expressed.[90] One of the primary disadvantages to this approach is that it requires pooling statistical information across all identified proteins. Another contentious decision is how they handle "missing" data: QSpec randomly generates a count from a Poisson distribution using its replicate's mean. While this ensures that the protein will not be considered significantly different, its meaning is slightly different from a true-negative.

Much like the debate between single aggregate descriptive metrics versus individual, specific scores as discussed in the context of False Discovery Rates and False Positive Rates (Section 2.1.4), a similar debate exists in the context of protein quantitation. In 2009, Pham et al. proposed a beta-binomial method to describe spectral count data collected from label-free tandem mass spectrometry-based proteomics, citing their primary contribution as distinguishing between within- and between-sample variation.[91] Therefore, instead of pooling statistical information for each protein like QSpec, this software attempts to identify the variation resulting from the random sampling process of each biological sample and the variation of random biological samples in a sample group. The two types of variation are modeled by the beta-binomial distribution, in which the parameters to estimate within-sample variation (binomial distribution) and between-sample variation (beta distribution) are based on a likelihood ratio test (G-test). Using the beta-binomial distribution, one can achieve comparable true detection rates of the differential expression of proteins when compared to the LPE test, t-test (with log-transformed data), and the G-test,[91, 92] as well as estimate a false positive rate, which is

not possible with the LPE or t-test. An important consideration, however, is the performance of the test with multiple replicates. The beta-binomial distribution can be used if there are one-replicate comparisons, but it outperforms QSpec and performs comparably with one-way ANOVA with multi-replicate experiments.[91]

As a default option, most researchers prefer to use ANOVA as a test of significance in differential protein expression. Although many computational groups have suggested various other methods of testing label-free data (spectral counts in particular), ANOVA is a very straight-forward test that is not only well-understood, but it can be easily implemented through a variety of pre-existing software packages, including Excel, R, and Matlab. ANOVA tests whether the between-group variation of collected data overlaps with the expected variation within a group. ANOVA is therefore more powerful when replicates are available and more powerful when the collected data is comprised of independent measurements from a normal distribution. A log transformation of spectral counts can approximate a normal distribution, especially if the filtering criteria is high enough to retain only the most abundant (and therefore more reproducible) protein measurements. Setting the minimum spectral count and reproducibility too high may result in undesirably significant data loss. Currently there is no "gold standard" for determining which proteins pass an appropriate cutoff for identification purposes and whether an additional stringent filter needs to be used before quantifying proteins. Even if ANOVA is performed on a well-filtered dataset, the test considers a repeated measurement of 0 (no spectral counts) to be highly consistent as well as a repeated measurement of 3 to be highly consistent. A comparison of a protein that is consistently not detected in sample 1 and consistently detected as 100 spectral counts in sample 2 may not pass through filtering criteria that require a protein to be detected in each sample, eliminating this otherwise striking change in protein expression from the final report. Additional considerations are needed to ensure that the measure of protein abundance is in accord with the chosen normalization method as well as the filtering criteria employed to generate final datasets.

### *2.1.5. Toolboxes and Software Packages*

Concomitant with the rise in popularity with mass spectrometry for shotgun proteomic analysis and the continuous advances of instrumentation resolution, speed, and throughput, it is not surprising that the last two decades have witnessed an explosion of informatics tools to facilitate the researcher's ability to aggregate, visualize, compare, and analyze the generated data. Many specific software tools have been generated to help model or predict information, such as outputting all possible peptides that could be generated from a protein subjected to a certain proteolytic enzyme, or calculating entire fragment ion series theoretically from a peptide sequence at a given charge state, or estimating the relative intensities of fragment ions generated by a peptide sequence. A common feature of most of such scripts is that they are generally designed for handling a single input at a time. Especially since most of these tools are web applications, it is sometimes difficult to find the tools as well as apply their calculations to an entire proteome, which is becoming an increasingly necessary step. Some websites, such as ExPasy[93] and PROWL,[94] contain links to many of these small tools developed exclusively for descriptive mass spectrometry questions. Notably, these tools do not attempt the computationally-intensive tasks of identifying and quantifying analytes measured within the experiments. Such processes are generally designed for personal desktops or, more recently, in cloud computing environments. In the past decade, there has been a major push to make this type of software user-friendly with graphics, selection boxes and drop-down menus, and configurable settings so that the peptide sequencing algorithms, protein assembly processing, and quantitative comparisons not only generate quality results, but also are easily understood and navigable. Many commercial software companies have spent a considerable amount of time and effort in creating a user experience designed to help the user feel more comfortable using the software as well as instill confidence in the results. A major deterrent to using such tools is the cost associated with purchasing a license as well as the proprietary algorithms that are not easily nor sometimes legally accessible. Other software that come from universities and research facilities tend to be more transparent, but other than typically having a less jazzy look and feel compared to the commercial software, they suffer from a major common

issue: inconsistent file formats. These tools primarily arise out of meeting an immediate need within a mass spectrometry lab and are designed to work with their specific bioinformatics workflow. Oftentimes this means that they require obscure file formats or extracted information from more standardized data and the additional work to pipe one tool's input to another's output is not generally within the scope of time or skillset available to a typical researcher wanting to analyze his mass spectrometry results. There has therefore been a push to build large software packages, groups of tools that seamlessly communicate with each other and internally handle the integration of multiple informatic processes. Some desktop packages such as Bumbershoot offered by Vanderbilt University, include a number of options to perform the same task (peptide sequencing) using a variety of different approaches. Bumbershoot allows the user to run Myrimatch, DirectTag, TagRecon, and Pepitome in preparation for analysis by a complementary software package, IdPicker, which filters and assembles protein identifications. Other software packages, such as the TransProteomicPipeline (TPP),[95] have the option of downloading a desktop component that communicates with their web server or using their software in a cloud account. TPP contains a number of tools that perform various tasks along the informatics pipeline transforming raw data into biologically meaningful reports and displays. However, installing the desktop software is quite complicated and using cloud services requires registration and payment for storage space and compute time. Many other software bundles have been proposed but few offer transparent algorithms that can handle multiple large datasets and report information in an easily understandable, portable format.

## 2.2 Challenges of Integrating and Developing Workflows for Analyzing Large Experimental Datasets

### *2.2.1. Standardization of Data Formats*

The goal of proteomics is to identify and quantify proteins, but as data complexity increases, accurately determining high- and low-confidence identifications and high- and

low-abundant abundances becomes non-trivial. For the past decade, MS-analysis has been centrally-processed in a workflow with primary emphasis on converting raw instrumental data (*.RAW files) to summarized reports (DTASelect files[1]) with basic spectral, peptide, and protein information. As new instruments collect approximately 28,000 spectra per fraction and a typical run consists of 12 fractions, over 308,000 scans are collected for every MS run- which translates into approximately 10 GB. For a single experiment on the newer instruments, between 120,000 and 180,000 spectra are assigned. Some experiments have been known to incorporate up to 60 MS runs, resulting in a deluge of highly dimensional high mass accuracy raw data that requires matching to peptide sequences, mapping to protein sequences, filtering, quantification, and normalization before biological interpretation can begin.

The Human Proteomics Organization's (HUPO) Proteomics Standard Initiative (PSI) is a group of researchers dedicated to standardizing data formats to improve cross-platform analyses, set standards of high quality data, and foster collaborations between institutions.[96] Their organization is divided into three primary working groups: molecular interactions, mass spectrometry and proteomics informatics, and protein separations. Each year each working group hosts a meeting to discuss the emerging needs of the current technologies and experimental designs as well as evaluate how the existing data formats, controlled vocabularies, and responsibilities are faring. Currently, most of the recommended data formats are XML-compliant, ensuring a relational structure that can not only be easily enforced with strictly defined schema but also effectively compressed into manageable file sizes. Until 2008, PSI suggested that mzData and mzXML formats should be used to capture raw data generated by the instruments. Whereas mzData was intended to be more of an index file to aggregate and point to numerous types of raw file formats from instrument vendors and not intended to replace the original files, mzXML files were created to be used as open-source substitutes for the information stored in vendor files, which were locked in proprietary formats. While mzData format is now deprecated, mzXML is still in common use. However, PSI currently recommends mzML or TraML formats instead. mzML files are expected to be ubiquitous and applicable to all

mass spectrometry instrument configurations and experimental designs, although no vendor has yet released software supporting it. TraML is a more specifically-designed format, targeting selected reaction monitoring (SRM) experiments. Both of these data formats contain scanning information to be used as inputs to search algorithms, which are, in turn, recommended to output mzIdentML files. mzIdentML files are expected to report MS scans, their MS/MS scans, the peptide sequences matching the MS/MS scans, and peptide-spectrum match (PSM) scores. These files do not contain all of the original peak data, but they have the structure in place to allow for matching fragment peaks to supplement each PSM score. For software that does not generate mzIdentML files by default, there is free software available to convert the other formats, such as dtaselect (from SEQUEST) or pepXML (from Myrimatch or MASCOT). Notably, the mzIdentML file format is not the last stage of the post-processing analysis. There is yet another step of filtering PSMs, assembling peptides into proteins, and quantifying abundances. For such processed information, PSI is currently working on finishing mzQuantML. However, mzQuantML has yet to be widely adopted, most likely due to the numerous differences on the standard procedure for filtering, assembling, and quantifying proteins.

### *2.2.2. Impediments to Integrating Systems Biology Data*

As technologies continue to improve in quality, throughput, and specialization, propelling the pursuit of scientific knowledge into new frontiers, it is not surprising that the newly acquired information does not immediately suggest clear-cut mathematical models, completely agree with all existing theories, or self-organize in a way that can easily be documented, stored, and accessed. In fact, there are three main impediments to integrating systems biology data. First, with each new discovery and accumulation of information, assimilation of theories, and unexpected breakthrough, the meaning and context of scientific ideas keep changing. Some of these revelations are truly revolutionary while others take a while to refine, become accepted, and incorporated into our understanding of how things work. Secondly, the progression of science requires work to provide context and details. A single discovery cannot stand on its own, but once

it has been recognized and adopted, the scientific community has to work together to understand the meaning and implications behind the discovery. Even from a purely technical point of view, the effort in integrating multiple analyses, services, storing information in a way that not only makes sense with the current architecture, but is amenable to extraction and adaptation in anticipation of future expansion. The instrumentation, file formats, and data recorded keeps changing faster than people can accommodate. Thirdly, people are motivated to further scientific research by a number of different drivers, including funding opportunities reward systems, popular science topics, or personal interests. Most of this work requires extensive collaborations across disciplines, institutions, and cultures, and figuring out how to seamlessly coordinate with others can be a challenge in itself. With the rise of the internet and cloud infrastructures improving data accessibility, computing resources, and methods of communication, some of these hurdles are easier to overcome than others. As we continue to move forward, it is becoming ever more important for computational biologists to keep analyzing data in its proper, albeit dynamic context and implement modular, scalable software solutions.

## 2.3 Summary of Dissertation

The objective of this dissertation is to develop and integrate tools that enhance the entire spectrum of proteomics analysis by mass spectrometry- from detection of raw data to interpretation of biological story. These tools alleviate computational bottlenecks at each step of analysis by providing statistically-sound software in order to deliver biologically-relevant and meaningful output without distorting information, losing data, or adding artifacts. The intentionally modular design of this toolbox provides an environment for each tool to perform its individual function in response to specific bioinformatic queries, as well as sets the framework for all of the tools to interact in a seamless, holistic manner for a more comprehensive understanding of the biological questions under investigation. By looking through a computational biologists' lens at a vast array of biological studies implementing mass spectrometry for proteomic analysis, a number of inter-related but functionally-distinct informatics processes present themselves. The development of a tool

for each process provides a mechanism of answering biological inquiries ranging from focused, hypothesis-driven questions, such as the increased ratio of a structural cellulase protein CipA in condition 1 compared to condition 2, to more global, discovery-based investigations, such as the identification of a core group of proteins expressed across a collection of plant tissues.

Answering these questions requires several points of engagement between informatics and analytical understanding of the underlying biochemistry of the system under observation. Deriving meaningful information from analytical data can be achieved through linking together the concerted efforts of more focused, logistical questions. This study focuses on the following aspects of proteomics experiments: spectra to peptide matching (Chapter 3), peptide to protein mapping (Chapter 4), and protein quantification and differential expression (Chapter 5). The interaction and usability of these analyses are also described (Chapter 6).

While it is important for informatic tools to be able to handle large datasets, it is becoming increasingly crucial for tools to also handle the biological complexity associated with more intricate experimental designs. Although some existing tools can scale computationally and maintain biological relevance, most of the time new tools need to be developed to appropriately address these concerns. The overwhelming volume and complexity of these experiments requires that the new and existing tools are not only optimized for speed and interpretation, but they also necessitate seamless communication with each other in an integrated workflow. By constructing a workflow that allows high-throughput processing of massive datasets, data collected within the past decade can be standardized and updated with the most recent analyses. Once these analyses are complete, meta-analyses can identify global analytical and biological trends.

Technological and informatic improvements are continuously accelerating the scope and complexity of biological investigations. As such, defining how a question is answered is becoming just as important as determining what both the question and answer should

look like. In fact, clearly identifying appropriate analytical and informatics methods is half of the work in solving these biological problems. Method optimization, versatility, and specialization become ends worthy of research in themselves. Although collections of measurements are motivated by biological enquiries and ultimately exist to reveal biological significance, the data points that act as intermediary, empirical evidence of interactions between genotypic and phenotypic information could arguably be considered more “real” and reproducible than their initial biological drivers and final interpreted conclusions. However, data cannot be useful until it is contextualized as information and interpreted as knowledge. In these processes, the truth or value of the data may be altered due to misinterpretations of newly annotated data, such as causal instead of correlative conclusions, or tendencies to over-fit, normalize, or filter results in an effort to arrive at pre-conceived outcomes. Therefore, in order to continue the iterative feedback loop of inspiring and answering biological questions, it is becoming ever more important to also ensure that the informatics validating, analyzing, and interpreting collected data preserve and reflect the integrity of the analytical measurements.

# CHAPTER 3: Spectrum to Peptide Matching

*Data presented in Section 3.1 has been adapted from the following journal article ready for submission to the* Journal of Proteome Research:

Rachel M. Adams, Richard J. Giannone, Paul Abraham, Robert L. Hettich. "Protease-Optimized Spectral Indexing Enhances Protein Identification and Quantification in Shotgun Proteomics Datasets." Sample preparation and experiments were performed by Richard J. Giannone. Data analysis was performed by Rachel M. Adams.

*Data presented in Section 3.2 has been adapted from the following journal article*:

Paul Abraham*, Rachel M. Adams*, Richard J. Giannone, Robert L. Hettich. "Defining the Boundaries and Characterizing the Landscape of Genome Expression in Vascular Tissues of *Populus* using Shotgun Proteomics." * Authors contributed equally to this work. Sample preparation and mass spectrometry experiments were performed by Paul Abraham. The bioinformatic workflow for evaluating sequence redundancy was developed by Paul Abraham, Rachel Adams, Richard Giannone and implemented by Rachel Adams. The supplemental database for single nucleotide polymorphism detection was created by Rachel Adams. Quality of spectra was evaluated using software written by Brian Erickson. Biological data analysis was performed by Paul Abraham.

*Data presented in Section 3.3 has been adapted from the following journal article*:

Paul Abraham, Rachel Adams, Gerald Tuskan, Robert Hettich. "Moving Away from the Reference Genome: Evaluating Single Amino Acid Polymorphism Identifications from a Peptide Sequencing Tagging Approach for the Genus *Populus*". *Journal of Proteome Research (**In review**)*. Sample preparation, mass spectrometry experiments, and manuscript preparation were lead by Paul Abraham. In-house scripts for matching ion intensity information and evaluating the site-determining ions of modified amino acids were developed by Rachel Adams.

## 3.1 Matched Ion Intensities Increases Accuracy and Robustness of Peptide Identification

### *3.1.1. Evaluating Spectral Counts and Matched Ion Intensities*

Proteomics, the characterization of the complete suite of proteins expressed in a cell, is commonly used as a discovery-based component to identify and quantify protein expression in an organism under a given condition. Considering that differing cell types produce different levels of complexity, the biological dynamic range of proteins analyzed

in a proteomics experiment can span up to 7 orders of magnitude and cover an equally large mass range. To reduce the complexity of this wide range of protein types and abundances, researchers commonly adopt the shotgun proteomics strategy, i.e., digesting proteins into peptides that have a smaller range of sizes, masses, and physiochemical properties. By sequencing the resulting peptides via mass spectrometry (MS), each experiment's detected peptides can be computationally mapped back to their proteins. In fact, this dependence on informatics to reflect the analytical measurements and inform the biological interpretation is not a single requirement. Each of the experimental steps that simplify the set of analytes being measured at a given time requires an analogous informatics process that deconvolutes the measurements and reconstructs the identifications within their local and global context. Thus, *in silico* analyses must be performed in reverse order of their experimental counterparts and scored/evaluated/assessed according to the scope of the measurement before proceeding to a larger interpretation: 1) spectra are matched to peptides, 2) peptides are mapped to proteins, 3) proteins are quantified in the context of the run, 4) proteins are compared between technical replicates, and finally, 5) proteins are compared between experimental conditions. For this bottom-up strategy to accurately identify and quantify proteins, all variables and optimizations within the data collection processes must dictate which and how the data analysis methods should be applied.

However, one does not want to over-fit or transform raw data in such a way that it loses analytical accuracy and biological relevance. As Occam's Razor suggests, the most simple, straightforward metrics are generally considered the more confident methods for inferring protein identification and quantification.[75] Label-free protein quantification, in fact, assumes that one can accurately compare relative protein abundances using inherent features of the collected data without introducing any additional analytes to the sample. The most common measure of label-free protein abundance, spectral count (SpC), is the number of MS/MS scans that match peptides belonging to a protein. While the simplicity of SpCs has merit, this single-dimensional measure relies on a number of assumptions and shifts the more complicated calculations to downstream analysis. For example, the

discrete property of spectral counts coupled with the stochastic nature of low-abundance protein identifications, diminishes the likelihood of SpC following a normal distribution and suggests a quasi-Poisson, modified binomial, or some other mixed model distribution. These more complicated distributions can limit the statistical powers of significance tests attempting to discern which proteins are differentially expressed in two samples. However, recent studies have suggested that instead of counting the MS/MS events that identify a protein, one can achieve greater analytical accuracy by summing the intensities of the individual fragment ions contributing to a peptide-spectrum match. The protein's spectral index (SpIn) therefore inherently includes SpC information but also captures the chromatographic contexts of the identifications. This multidimensional measure embodies analytical nuances provided by the TIC, the number of total MS/MS peaks, and number of matching peaks but simplifies them as an aggregate data point. In other words, matched ion intensities preserve quantitative features that articulately reflect how well the peptides were separated, the degree of competition for charge, and each peptide's specific ion contribution within ambiguous peptide-spectrum matches due to co-fragmentation or indiscernible charge states. In addition, the number and range of individual matched ion intensities contributing to protein identifications far exceeds that of spectral counts. Like all analytical measures relying on a sampling process rather than a fully comprehensive collection, there are inherent biases and advantages to both spectral counts and matched ion intensities.

Although both spectral counts (SpC) and matched ion intensities (MIT) reflect a successful peptide identification, whether the peptide-spectrum match should be a digital or weighted measurement is the primary point of contention between SpC and MIT. It is highly intuitive that the more times the instrument detects a peptide within a set period of time, the more abundant that peptide must be among the group of peptides under analysis. However, the number of times a peptide is observed (SpC) is not just a function of abundance within the sample- it also dependent on instrument settings, the physiochemical properties of the peptide sequence, the number and types of peptides that are eluting around the same time, and the peptide-spectrum matching algorithm.

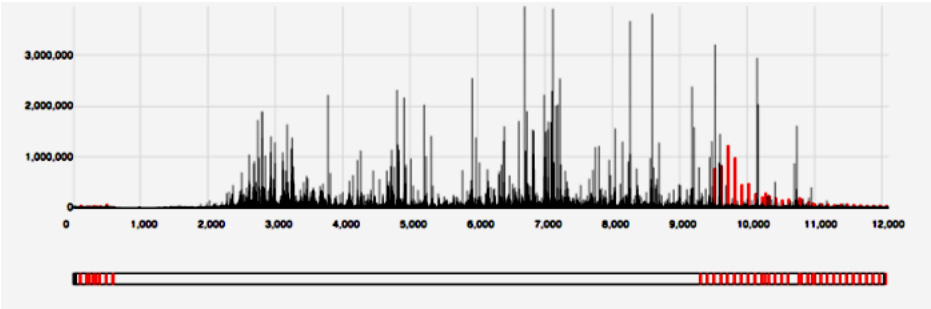
If the goal of an experiment is to identify the primary components of a sample, then the researcher may choose data-dependent settings on the instrument in that focus on acquiring deeper measurements of the more abundant analytes (more SpC per peptide) rather than achieving a broader survey of all possible constituents in the sample, including capturing those peptides that may be low-abundant. That is, in a discovery-based experiment seeking to acquire comprehensive identifications, instruments may be instructed to skip over analytes that had previously been measured so that less-abundant peptides may get an opportunity to be detected. In such a scenario, all detected peptides would receive a spectral count of 1 and no quantitative information could be inferred at all (other than the presence of one or more peptides supporting each protein identification). For those types of experiments, however, the instruments more often implement a dynamic exclusion list, allowing an analyte to be temporarily ignored from additional measurements if it had been previously detected within a certain window of time or after it has been measured a certain number of times within the run. With these types of rules enforced, it is important to keep in mind that the peptides are following an elution peak (where the x axis is time and y axis is typically the precursor intensity) in response to a changing salt gradient and that at the time a peptide is put on an exclusion list, it may not be measured at its highest point. Under these circumstances, SpCs are robust measurements because they are not concerned with the intensities at which the peptides were analyzed- just whether or not the measurement was acquired. In general, a peptide that has a wide elution peak is more likely to be sampled (and therefore have a higher SpC) than one that has a very narrow elution peak. For experiments in which the biological dynamic range is relatively small or when deep, repetitive measurements are more important than the breadth of identifications, SpC are consistent metrics that can be reliably used for relative abundance measurements.

In LC-MS experimental designs, however, the elution profile of a peptide is characterized not only by its width, but also its height. The intensity or height of a peak has been argued to be just as indicative of a peptide's abundance compared to the peak's width

(SpC). Whereas SpC is applauded for decoupling the individual measurements from the context in which their measurements were taken, proponents of intensity measurements boast of empirically incorporating the quality of the measurement with respect to the other analytes in the background. Across analytical platforms there is general agreement that observations may be useless, or at least statistically insignificant, if they are indistinguishable from noise measurements. Following this trend, the signal to noise ratio has often been used to filter SpC measurements by allowing only those spectra that were collected above a certain threshold to be used for quantitative measurements. From there, it is not difficult to imagine how one would extend these digital comparisons of "quality" or "poor" spectra to a more continuous scale by summing the precursor intensities of each peptide identification.

To illustrate how interpretation based solely on raw SpC can be misleading, two examples are shown below. Figure 3.1A compares the MS/MS assignments of two proteins from each salt pulse collected from a single MS run analyzing the *C. thermocellum* proteome. These two proteins, Clo1313_0465 and Clo1313_0296 have very different SpC (20 and 113), ranking them in the 25$^{th}$ and 75$^{th}$ percentiles, respectively. When the TICs of their MS/MS scans are compared (6.07E7 and 6.12E7), the proteins both fall within the 40$^{th}$ percentile. While it is not advisable to directly compare protein A against protein B within the same sample or between conditions, the relative percentage of SpC assigned to a protein is used in the most common form of protein abundance normalization (NSAF). It is assumed that two proteins of similar size and similar SpC have the same relative abundances within a sample. Figure 3.1B illustrates a scenario when this would not appear to be a fair assumption. Proteins Clo1313_0837 and Clo1313_0467, both 256 amino acids long, generated about the same number of peptides (6 and 7 peptides, respectively) and approximately the same SpC (26 and 27). When their intensities were compared, however, Clo1313_0467 was almost half as intense as Clo1313_0837. As the picture illustrates, the less intense protein was sampled at very different parts in the chromatogram compared to the more intense protein. Many such proteins receive additional spectral counts from the low-complexity

A. YAFMGGSNLVIFNSSK
SpC: 45, MIT: 1.501 E7




B. AHTIANLAGFEVPETTK
SpC: 11, MIT: 1.393 E7

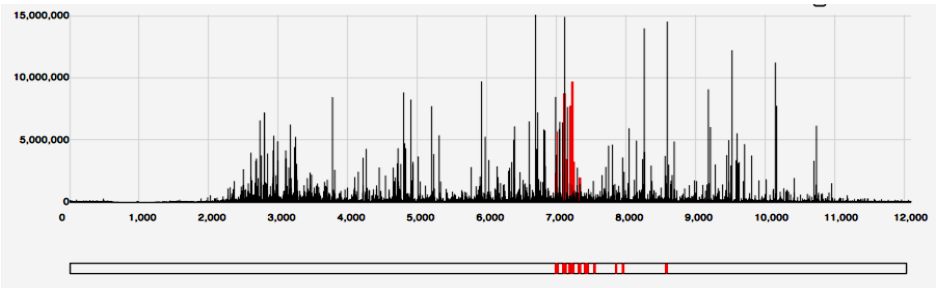



Figure 3.1. Peptides representing proteins with different SpC but similar MIT.

(A) High spectral counts may be over-inflated due to over-sampling in low-complexity regions. (B) Low spectral counts may be underrepresented due to competition in high-complexity chromatographic regions.

regions of the chromatogram, typically at the very end and beginning of each salt pulse. These spectra are generally low-intense peaks that are continuously re-sampled for lack of competition, but it is evident from visual comparison that the additional SpC collected during these times are not at all comparable to the SpC collected in regions of higher chromatographic complexity. Therefore, intensity measures can be highly indicative of the chromatographic background in which the identification was assigned and give a more sensitive abundance measurement.

When using intensity measurements as a proxy for relative protein abundances, a weighty assumption is that the observed intensity truly affects the quality of the measurement and not just the ionizability of the peptide. However, depending on the physiochemical properties of a peptide sequence, a peptide may not be very compatible to ESI and therefore may not have the same opportunities to be detected as another peptide that is just as abundant (or less abundant) but is very amenable to enzymatic digestion, SCX-RP separation, and salt gradient elution. In fact, studies have shown that 2 peptides presented to the mass spectrometer at the same concentration can behave up to X-fold different in their intensities due to differences in their sequence's hydrophobicity, length, and number of immonium ions. For these reasons, simply picking the most intense point of a peptide's elution peak is not sufficient; taking multiple samples along the peak more accurately describes the behavior of a peptide compared to the other peptides it is competing against. For quantitative purposes, some researchers suggest that taking the sum of each intensity measurement sufficiently captures the shape of the peak, whereas others insist that taking the area under the curve of the elution peak is more accurate. Taking the area under the curve, a method commonly employed in labeling approaches even for absolute quantitative measurements, is indeed more accurate when one is comparing the two observations of two analytes that have the exact same chemical composition. However, deviations in peptide length, charge states, post-translational modifications of residues all significantly affect the ionizability of the analyte and therefore its elution peaks.

More recently, researchers have proposed that not only are precursor intensity measurements more descriptive than SpC, but that even more information can be gleaned by looking at the individual intensities of peaks within an MS/MS scan. More specifically, instead of using the intensity of the MS precursor or taking the sum of all of the peaks within an MS/MS scan, also called the total ion current (TIC), one could sum the intensities of each peak within a scan that matches a peptide's expected fragment ion. These matched ion intensities (MIT) for each MS/MS scan can then be summed for each peptide and in turn, summed for each protein, to give an overall spectral index (SpIn) value. (See Chapter 5 for more details about the downstream calculations).

One of the most practical advantages about this MIT approach to quantitation is that it is not an entirely new concept or measurement: a form of the method has already been essentially implemented into every automated peptide-spectrum matching algorithm over the past decade. That is, every search algorithm compares a list of observed and expected m/z's and performs some type of scoring assessment, usually taking into consideration the corresponding intensity measurements of the matching fragment ion peaks. The choice of the best peptide identification for a given spectrum is not always clear and depending on the database-searching algorithm and its criteria for scoring peptide-sequence matches, the list of which peaks contribute to the peptide identification may change. Although each algorithm constructs internal lists of matched observed and expected m/z values for each peptide and spectrum compared, none of the approaches provide explicit reports of which of those peaks are the matched ions. We were also interested in looking at the contribution of secondary fragment ions (in a case study with HCD), which most search algorithms currently do not take into account for identifications. Therefore, for this study, it was necessary to devise our own matching analysis that compared expected ion series and observed peaks within the spectrum.

The goal of POSI is not to make any new peptide-spectrum matches, but rather work within the scope of the supplied information to make qualitative and quantitative calls about the identifications. In other words, this study's implementation of a peak-matching

algorithm did not compare all possible peptides to all possible spectra. Instead, it was limited to the user-defined list of peptide-spectrum pairs and only compared the collected peaks of selected spectrum against the theoretical ion series of its associated peptide and charge state.

One of the most relevant characteristics of matched ion intensities is their ability to span multiple orders of magnitude, compared to the limited number of spectral counts acquired during MS. The analytical dynamic range of matched ion intensities more accurately aligns with the biological dynamic ranges captured in an MS/MS sample, and even approaches the increased accuracy generally associated with absolute quantitative methods. However, the relative nature of these measurements necessitate normalization methods that accurately attribute precise abundance information to confident protein identifications, ensuring that each inference is interpreted within the analytical background in which the data was collected and the informatic context in which assignments were made.

### *3.1.2. Calculating Matched Ion Intensities*

*Removing Noise from Analysis*

Each spectrum has different distributions of peak intensities and therefore different noise peak cutoffs, so we looked for the inflexion points where the bottom-ranked peak intensities indicated strong linearity ($R^2 > 0.8$). Additional $R^2$ values were explored as well as comparison against the TIC, the number of total peaks in the spectrum, the number of matched peaks, keeping the most intense peaks that accounted for X% of the TIC, and retaining only the top X peaks per scan (see Figure 3.2). Also, for comparison to other existing methods, rather than simply removing all of the peaks determined to be in the noise (the “remove” method), the intensity of the highest peak determined to be in the noise was subtracted from all of the non-noise peaks (the baseline “subtract” method). Quality scores were calculated based on the sum of the matched ion intensities compared to the sum of the unmatched ion intensities.

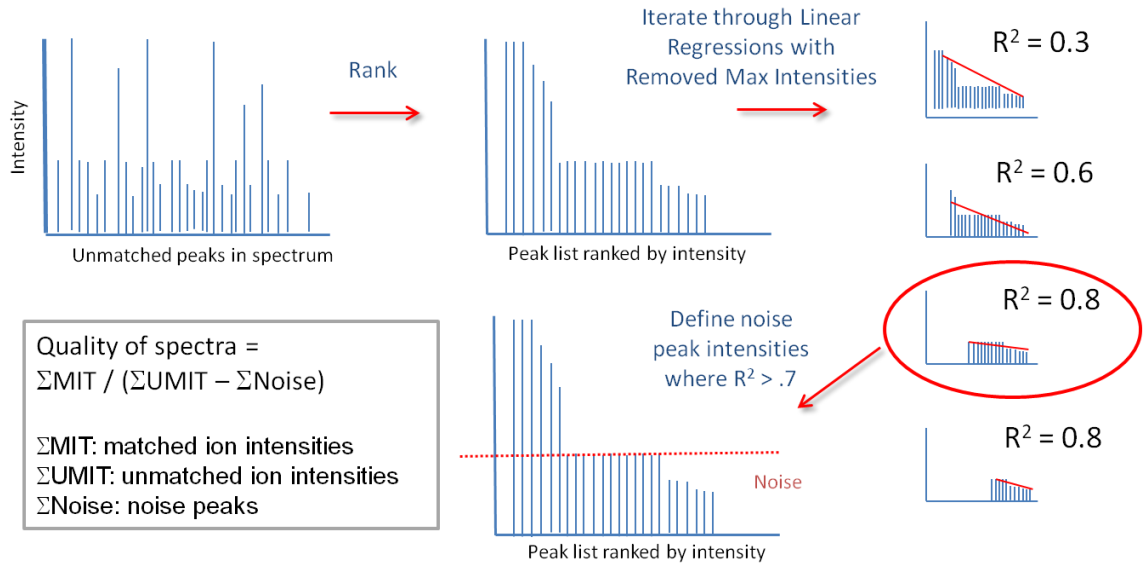



Figure 3.2. Method for detecting noise.

This approach goes through each spectrum, ranks all of its peaks by decreasing intensity, and looks for inflexion points where the intensities become linear, indicating a level of randomness in the data. Any peaks below the inflexion point are considered noise and removed from the analysis.

*Matched Ion Intensity Calculations*

Only peaks that have intensities higher than the spectrum's calculated noise level should be considered for matching ions. Matching ions are peaks within the spectrum that contribute to the peptide's identification. Because peptides fragment in predictable patterns, each peptide's theoretical ion series can be calculated *in silico* and then compared and scored against the observed peaks in approaches described above (see PSM scoring). For this study, it was necessary to devise our own matching analysis that compared expected ion series and observed peaks within the spectrum. It is important to note that the input to the overall POSI algorithm requires a list of filtered peptide-spectrum matches that the researcher has confidently generated from a previous database-searching algorithm.

*Matching Ions for PSMs*

The list of candidate expected ions for a given sequence was developed to be highly parameterizable. The researcher could define which ion series ("a", "b", "c", "x", "y", or "z"), which losses ("-H20","-NH3"), as well as which static and dynamic modifications ("C+57", "M+16") should be considered. Parameters can also be set for secondary fragmentation ions (neutral fragment losses, see Section 3.2.3 for details). Each of the analyses performed for this study only used the b and y ion series, static cysteine carbidomethylation modification (+57), dynamic N-terminal modification (+43), and dynamic methionine oxidation modification (+16). For each sequence observed at a given charge state, each ion series was calculated with a charge of +1 to precursor – 1. Within each scan that identified a peptide, lists of matched fragment ions were generated by sorting the observed m/z's by intensity and then assigning the sequence's closest expected fragment ion within a user-defined tolerance (0.5 Da by default). The newly matched fragment ions were then removed from the candidate list for the rest of the scan.

After matched ions were identified for each peptide-spectrum match, it was necessary to sum each peptide's matched ion intensities for each of its scans. For 97% of the scans, summing the intensities of the matched ions for that peptide and scan was

straightforward. The remaining scans had ambiguous peptide-spectrum matches primarily because the searching algorithm could not determine charge state and instead assigned 2 peptide sequences to the scan, one for +2 and one for +3. To avoid double-counting the intensities of matched ions that were strong candidates for both peptides associated with the same scan, matched ion intensities that were assigned to two peptides were proportionally distributed among the peptides according to the number of matched ions assigned to each peptide. This careful summing of matched ion intensities was carried through for each peptide, and a summed matched ion intensity was calculated for each scan as well. We also explored additional aggregate functions including taking the mean, median, and selecting the top 3 scans' intensities for each peptide.

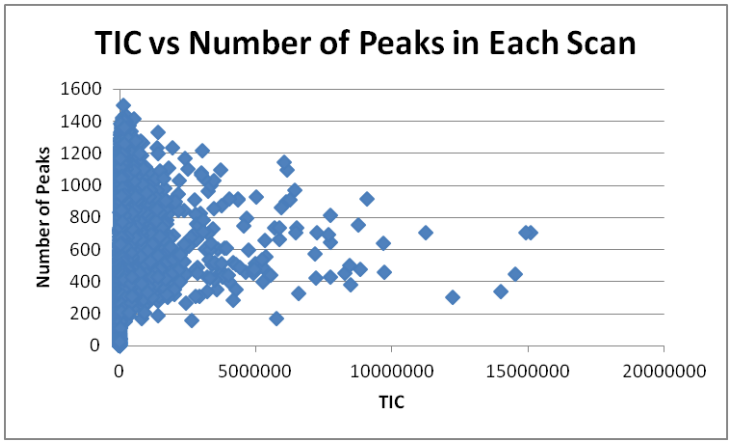
*Generating Report with Protein Spectral Indexes*

Peptide matched ion intensities can be summed to generate protein spectral indexes. To account for peptides that are shared between multiple proteins, the matched ion intensities for the redundant peptides are apportioned among the shared proteins according to each protein's number of unique peptides identified. The number of unique peptides that provide evidence for protein identification is proportional to the confidence we have in that particular protein call. Therefore, a weighted fraction of matched ion intensities are directed to each protein that has at least one unique peptide. After the matched ion intensities are balanced and summed into protein spectral indexes, the spectrum, peptide, and protein information is reported in a format similar to the common DTASelect –t0 output (an unfiltered tab-delimited format). In other words, the report includes details about every spectrum contributing to a peptide identification and a protein call. Most notably, the generated output from POSI also contains additional information about the number of matched ions, matched ion intensities for the peptide sequence, matched ion intensities for the scan, and how many times each peptide sequence appears in the protein. For portability purposes, this output can also be converted into an mzIdentML file format, a standardized report for database-searching algorithms.

### *3.1.3. Comparing PSM-level Intensities to Peptide-level Intensities*

*Data reduction step*

When using intensities to estimate relative protein abundances, it is important to make sure that we are not inflating measurements with data from instrument or chemical noise. Just as the TIC varies dramatically from scan to scan, even within a single salt pulse, the intensity and number of noise peaks in each MS/MS scan also change. Each spectrum has different distributions of peak intensities and therefore different noise peak cutoffs, so we looked for the inflexion points where the bottom-ranked peak intensities indicated strong linearity. Due to the high variability of individual scans, these noise levels ranged widely in their absolute values (between 5-300 peaks and 1000-5000 summed intensities due to noise), On average, this noise-reducing method resulted in removing about 80% of the peaks and roughly 20% of the TIC, suggesting superior performance that neither intensity nor peak counts could individually achieve (



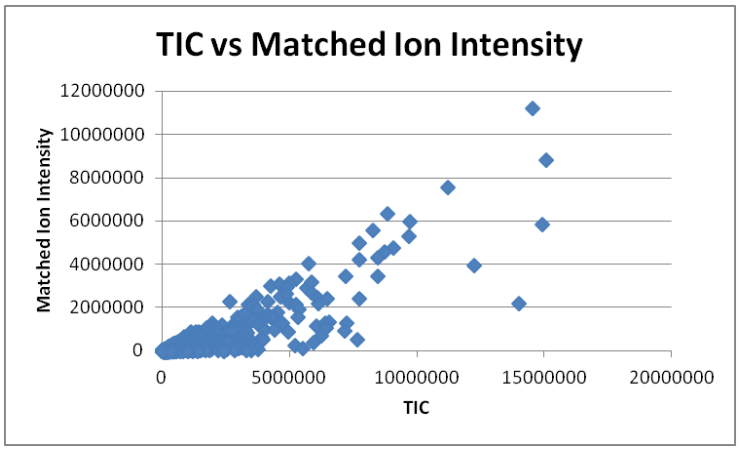
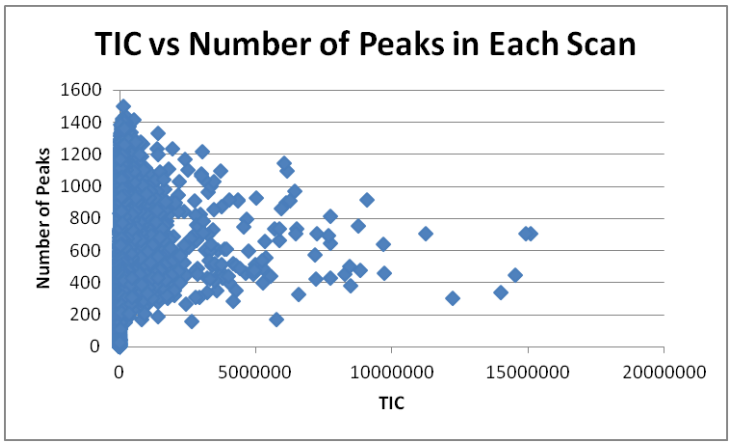



Figure 3.3). However, further examination revealed that the data reduction step was not noticeably improving the matched ion intensities for a given peptide, so this method was not used in the final quantification analyses.

*Scan-based metrics*

Each scan averaged an MIT of 2.87e5 with a standard deviation of 1.34e6, accounting for 22% (+/- 12%) of the TIC. Whereas the TICs ranged from 1.29e4 to 7.07e8, the MITs ranged from 1.95e3 to 2.64e8. MITs were not correlated with the TICs, so the matching process was demonstrated to be an informative calculation step. Similarly, the MITs were not correlated with the number of peaks (or quality peaks) in a scan, so they could not have been substituted by those quality metrics. Each salt pulse contributed a slightly different distribution of scans- as noted by the distribution of TICs and the average MITs. Figure 3.4 below also illustrates how the number and range of matched ion intensities of each salt pulse differed for a single run. Almost a 50:50 split between the number of b and y ions were observed. For each scan, an average of 11.78 and 11.99 b and y ions were matched, contributing an average of 4.8e4 and 1.25e5 to a scan's MIT, respectively. On average, 24 peaks (+/- 9) matched within a scan, predominately reflecting the number

**(A)**




**(B)**

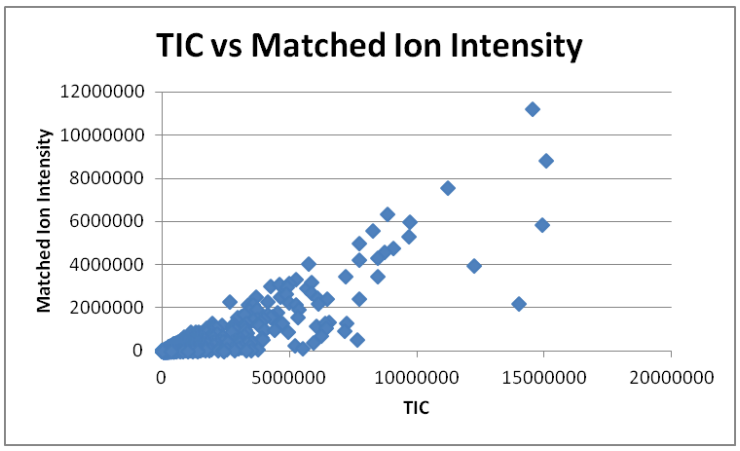



Figure 3.3. Validating the use of matched ion intensities instead of other simple features inherent to MS/MS scans.

(A) Each MS/MS scan's TIC was compared to the number of fragment peaks within the scan to see if there was a correlation.
(B) For each peptide-spectrum match that passed the typical filtering criteria, the MS/MS scan's TIC was compared to the matched ion intensity to see whether the matched ion intensity was a consistent fraction of the TIC.

of matched peaks from the most abundant charge states (+2 and +3), which matched 22.5 and 23.45 peaks per scan. The +1 scans averaged 15 matching peaks and the +4 scans averaged 53.8 peaks, differing from the other scans' metrics primarily due to their relatively increased and decreased number of possible peaks matched.

*Peptide-based metrics*

In keeping with the NSAF assumption that more opportunities to sample an analyte would increase its abundance, MIT measurements were compared to peptide length in order to assess whether there was a correlation. Similarly, peptides with a higher charge state have more opportunities to generate fragment ions, so the correlation between MITs and charge state (and number of possible fragment ions). PSM-level MITs grouped into peptide MITs were not biased for more opportunities based on any of these metrics. As Figure 3.5 suggests, peptide MITs were, however, different between charge states and salt pulses. The distribution of MITs for a highly abundant peptide, TVIEVLVENGNVSK (700 total SpC and 2.45e8 MIT) is illustrated in Figure 3.4. The average of the MITs collected for each salt pulse are slightly different for the exact same analyte. The shift downwards (smaller intensities) with each consecutive salt pulse reveals that the peptide is continuing to be measured even amidst growing competition for identification. Looking at any of these salt pulses individually would be a misrepresentation of the peptide's behavior across the entire run. Even looking at the cumulative distribution of MITs observed for this analyte does not completely capture the behavior of this peptide sequence. As the graph in Figure 3.5 suggests, this peptide behaves like quite different analytes depending on its charge state- perhaps just as differently as two peptide sequences altogether. If one is trying to validate the distribution of the peptide's MITs as a component of the protein's abundance within a run, it is more accurate to compare the peptide's distribution across technical replicates than it is to compare two peptides from the same protein within a single run.

Figure 3.6A highlights the similarities in MIT distributions between the same peptide identified in two technical replicates. To determine whether the peptide distribution

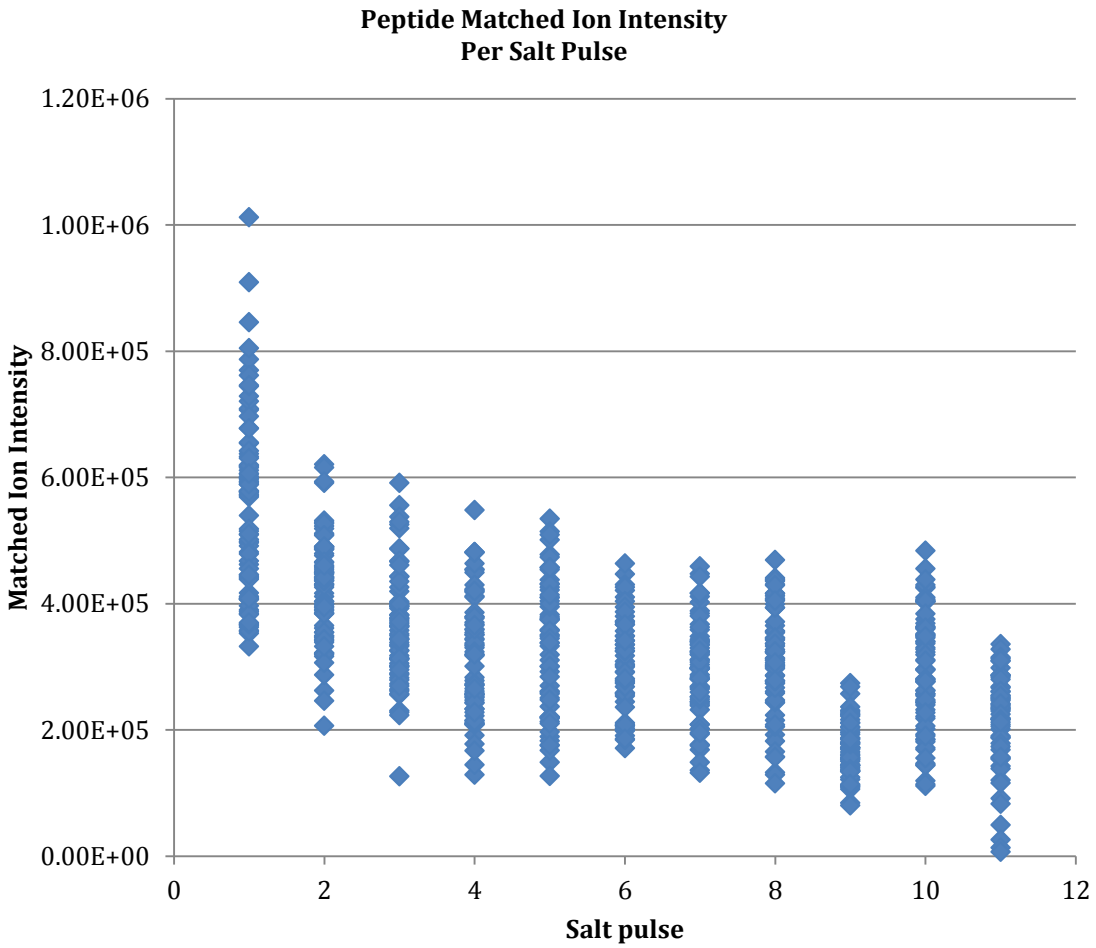
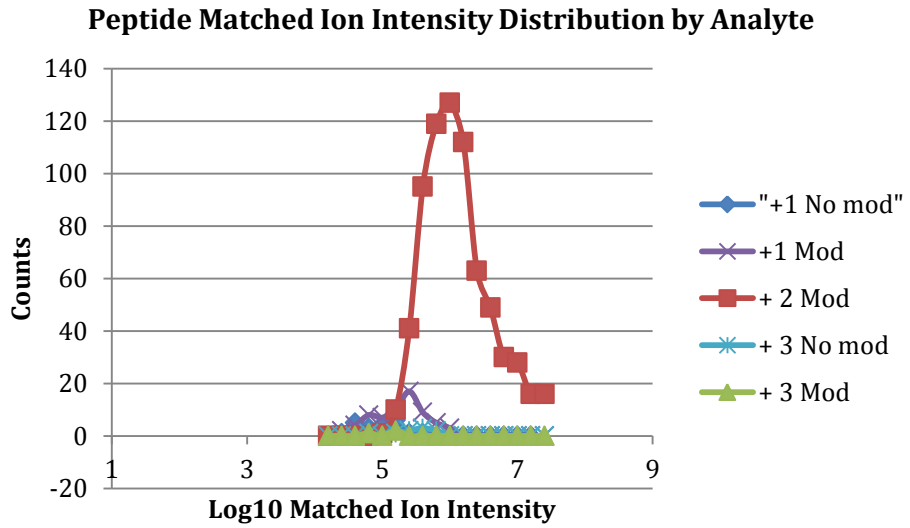



Figure 3.4. The distributions of an abundant peptide's matched ion intensity for each of the 11 salt pulses in a single run.

would behave the same across different loading amounts, the same sample was loaded on to a column in 2 different concentrations (25 μg and 67 μg). Figure 3.6B graphs how the peptide MIT distribution follows the same shape and general trend between the two concentrations and systematically reflects the expected shift in intensities between the two runs. Therefore, peptide MITs are considered reproducible across replicates and across loading amounts. However, not all peptides were identified in all replicate measurements. When the peptides are assembled into protein measurements, these inconsistencies in identification warrant careful consideration to either filter or normalize for the disparities.

(A)




(B)

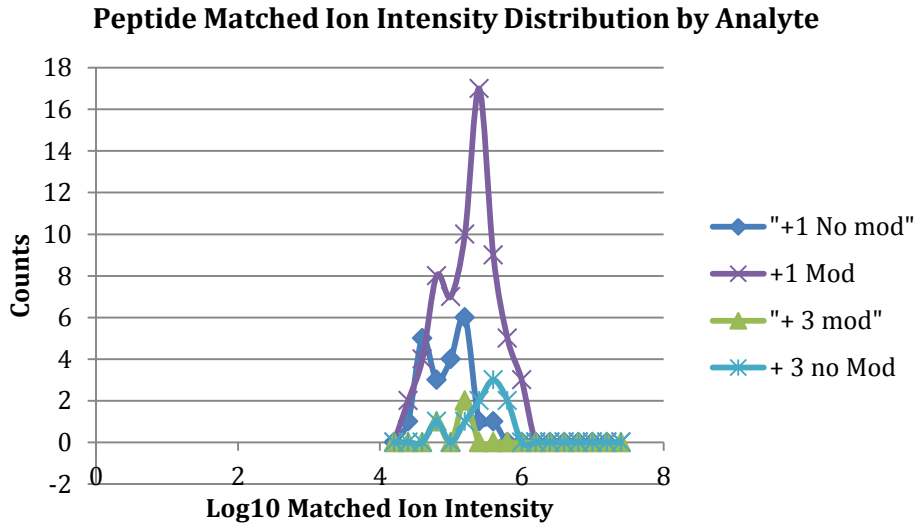
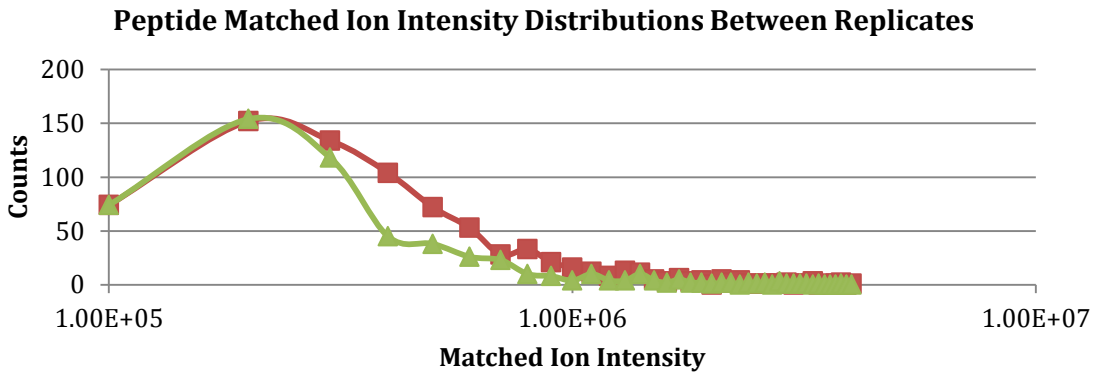



Figure 3.5. Abundant peptide demonstrates different matched ion intensity distributions depending on its charge state.

(A) The same peptide sequence was captured by vastly different SpC throughout a single run. The +2 species was observed over 700 times, compared to the 3-60 SpC detected by the other species. (B) An inset of the graph above to illustrate that the carbamylated (N-terminus + 43) species of +1 and its non-modified form followed the same general trend, as did the +3 modified and non-modified species.

**(A)**




**(B)**

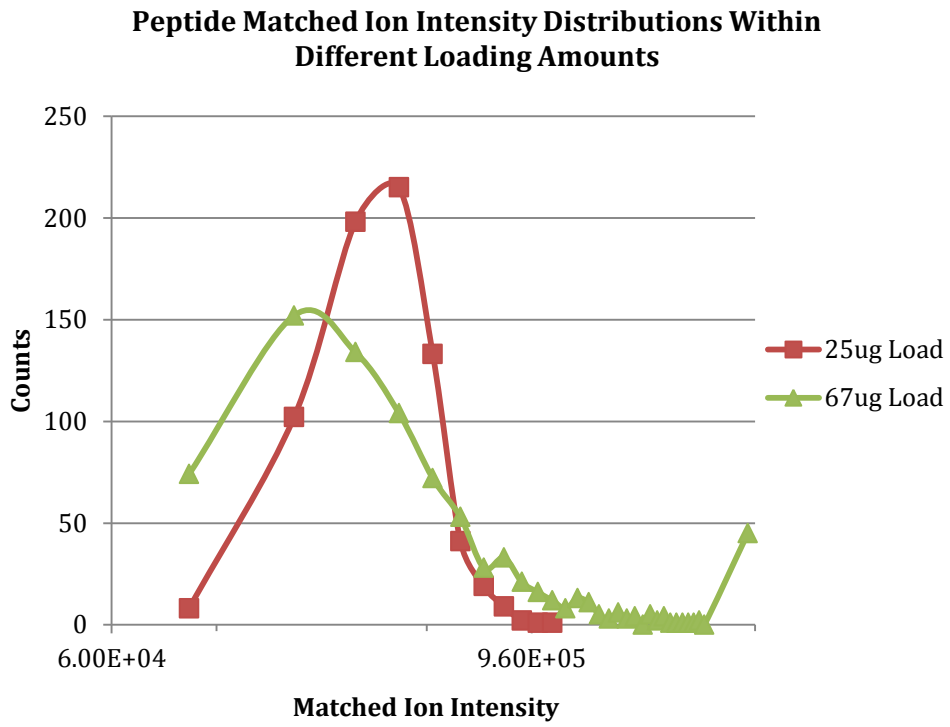



Figure 3.6. Peptide matched ion intensities are reproducible.

(A) Peptide matched ion intensities are consistent across technical replicates. (B) Peptide matched ion intensities may reflect the relative differences in the amount of sample analyzed by MS.

## 3.2 Augmented and Refined Peptide Identifications from Otherwise Unassigned Spectra

### *3.2.1. Qualifying Peptide Assignments from Ambiguous Peptide-Spectrum Matches*

Intensity information not only helps identify which proteins are more abundantly detected within a sample, but it can also highlight which peptide-sequence matches (PSMs) are more confident than others. As database size and redundancy increases, there is an increased likelihood of several peptide sequences receiving sufficiently high XCorr values for the same observed MS/MS spectrum. To account for these occurrences, SEQUEST uses 'DeltCn' to measure how the lower ranked peptide scores differ from the XCorr of the best-matched peptide sequence. Therefore, a higher DeltCn score of the second ranked match means that the best-matched peptide sequence is most likely correct. Since prokaryotic and eukaryotic genomes can vastly differ in size and genetic redundancy, the abundance of peptide sequences, which are similar in sequence identity, dramatically increases in plants. The *Populus* genome is highly convoluted by genetic redundancy that has resulted in two-thirds of the genes that express proteins to have a high sequence similarity (~90% or higher sequence identity). To address how this phenomenon could impact spectral processing, we compared the DeltCn distribution of a 24-hour MudPIT for *E. coli* and *Populus*. By plotting DeltCn values for the best-matched peptide sequence for every scan, the distributions for *E. coli* and *Populus* differ in both location and shape (Figure 3.7).In comparison to *E. coli*, the DeltCn distribution for *Populus* clearly shifts towards zero, indicating an increase in quality MS/MS spectra matching to more than one peptide sequence. This characteristic shift reflects the increased amount of genetic redundancy within the plant genome, which contains more indistinguishable peptide sequences. Previous studies show that a DeltCn threshold of 0.08-0.1 provides the necessary FDR with an appropriate balance between false-positives and false-negatives. This study suggested that perhaps high mass accuracy should be incorporated in future studies to decrease ambiguity and allow the DeltCn filters to become more liberal.

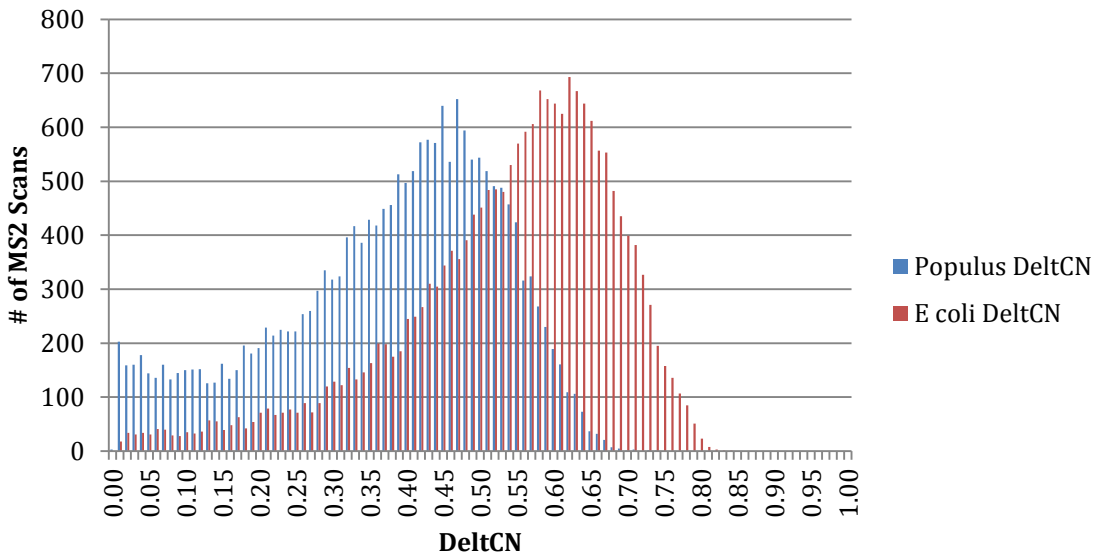



Figure 3.7. Comparison of DeltCN scores between *Populus* and *E. coli*.

SEQUEST's DeltCN score is the percent difference between the top and second-best xcorr values assigned to an MS2 scan. If these two xcorr scores are very similar to each other, (indicated by a DeltCN value close to 0), the algorithm was not able to clearly choose one candidate peptide sequence over another. Overall, the DeltCN scores in the *Populus* dataset were systematically shifted towards 0 when compared to the *E. coli* dataset, suggesting that *Populus* has more ambiguous peptide-spectrum matches than *E. coli*.

An alternative approach, however, is to determine which peptide has the highest matched ion intensity or which matched ion intensity best falls in line with the other unambiguous intensity information assigned to the candidate peptides. In most scenarios, one would most likely attribute the scan to either the protein with the overall highest abundance, or the peptide with the most matching peaks. However, with matched ion intensity (MIT) information, we can distinguish which peaks matched which peptide and possibly give non-redundant intensity support to both proteins. Figure 3.8 below illustrates an example in which the scan 05.29208 matches both WEIEFFK (+2) and VVDLIVHMASVDAK (+3). The peptide WEIEFFK, which belongs to protein Clo1313_1808 (932 SpC, total MIT of 1.17e5) matched 12 peaks in the scan and received a Myrimatch MVH score of 32.3, while peptide VVDLIVHMASVDAK, which belongs to protein Clo1313_2095 (5911 SpC, total MIT of 9e5), matched 23 peaks and received an MVH score of 42.8. Based on the number of matched peaks, the mvh score, and the overall protein intensity, one would most likely attribute this scan's intensity measurements to the second peptide, but it in fact matched 2% of the TIC, whereas the smaller peptide matched 23% of the TIC. Peptide WEIEFFK (+2) had an MIT of 6.50e5, which was an order of magnitude greater than the other peptide's MIT (6.31e4). Upon examination of each peptide's MIT distribution within its respective proteins, it is interesting to note that WEIEFFK had 2 ambiguous scan assignments within the run and VVDLIVHMASVDAK had 33 ambiguous scan assignments within the run. In total, Clo1313_2095 had 49 ambiguous scan assignments (whose MITs sum to 8.72e6, 0.2% of the protein's total MIT). We suspect that these ambiguous scan assignments may be over-inflating protein Clo1313_2095's SpC, but the impact of these incorrect or co-fragmentation identifications are more appropriately handled by the MIT measurements.

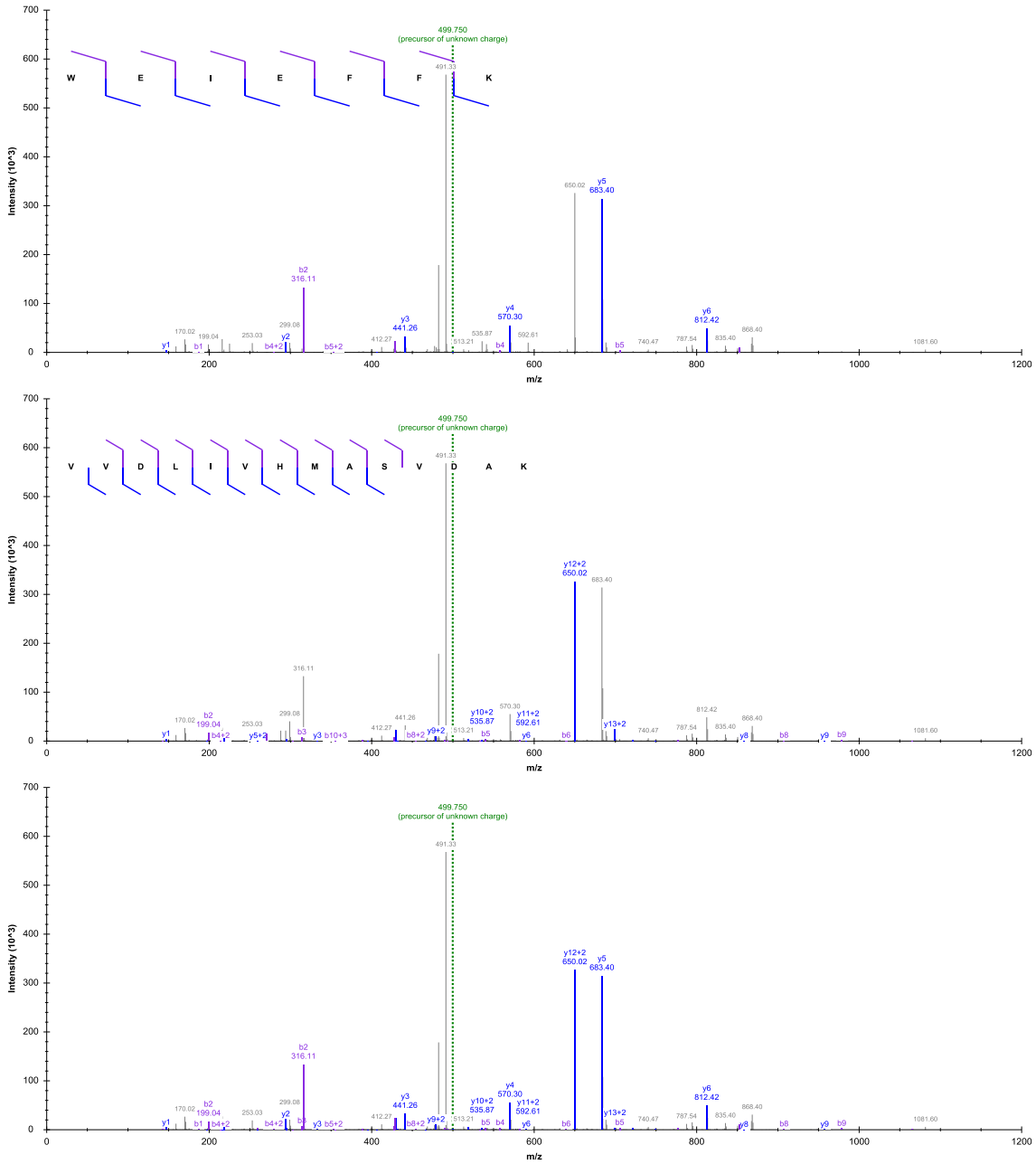



Figure 3.8. Comparison of 2 possible peptide-spectrum matches for an ambiguous scan.

Because the precursor charge state could not be determined for this MS2 scan, the searching algorithm tried to assign a peptide-spectrum match as if the charge state was +2 and +3. The candidate peptides for +2 (WEIEFFK, Fig A.) and +3 (VVGLIVRMASFGAK, Fig B.) scored similarly and their matched ions contained no overlapping peaks (Fig C.).

### *3.2.2. Supplementing Traditional Database Searching Approaches*

One of the greatest heuristics that contributes to the success of database-searching approaches also has a complementary limitation: regardless of the quality of peptide-derived spectra, algorithms will only match spectra to peptides that exist within user-defined sequence variations. Peptide sequencing by mass spectrometry is most commonly performed via collisional-induced dissociation (CID), in which peptide ions fragment in a predictable manner to produce dissociation products that yield sequence information. Though widely used for its simplicity and effectiveness, more than 50% of MS/MS spectra collected in a typical shotgun proteomic experiment do not result in high-confidence peptide identifications when using automated search algorithms such as SEQUEST or MASCOT. Even though these low identification rates can be partially explained by the presence of spectra arising from concurrent fragmentation of multiple precursor ions, incomplete fragmentation of peptides, and chemical noise, a large fraction of peptide-derived spectra remain unassigned because of the quality and completeness of the proteome database[97, 98]. Neither prokaryotic nor eukaryotic protein databases typically include protein isoforms or alterations/modifications, and furthermore their omission has a more dramatic effect on higher eukaryotes in which sequence variations and unexpected splice variants are more prevalent. Thus, by not anticipating the presence of these peptides, database search algorithms are more likely to interpret fewer peptide-derived MS/MS spectra when analyzing proteomes of higher eukaryotes. Reanalysis of unassigned tandem mass spectra was performed to determine the magnitude of peptide-derived spectra that remained unmatched to a sequence, thereby providing the proportion of "missing" peptide identifications in a run.

To compare the rates of peptide-spectrum matching (PSM) between eukaryotes and prokaryotes, we contrasted MS/MS data from *Populus* with a simpler bacterium, *Escherichia coli*.[99] In both cases, proteolytic peptides were measured on the same instrument using identical methods to minimize experimental biases. The instrumental acquisition and chromatographic distribution of all MS/MS spectra collected were similar for both organisms. However, the ability to successfully match experimental MS/MS

spectra to theoretical database sequences was superior in *E. coli*. A greater percentage (86%) of *Populus* MS/MS spectra remained unassigned, as compared to only 63% of the MS/MS spectra collected for *E. coli*. A closer look at the proportion of unassigned peptide-derived spectra was used to determine if the observed discrepancies in peptide identifications could be attributed to the incompleteness of the reference database. Spectral quality assessment was used to identify the number of unassigned high-quality spectra, i.e., a population of spectra that likely represents mutated, modified or novel peptides. A conservative set of criteria, based on previous implementations of spectral analysis was utilized in the assessment of MS/MS spectral quality.[100, 101] A spectrum was considered high quality if the parent charge state was calculated to be greater than +1 and if the spectrum contained three or more peaks within 20% of the base peak intensity with a minimum intensity of 2,500 counts. Using this approach, we performed an assessment of MS/MS spectra quality to distinguish high-quality unassigned spectra from low-quality unassigned spectra in the representative MS runs from *Populus* and *E. coli*. Spectra analysis revealed that, of the total MS/MS spectra collected for *Populus* and *E. coli*, the percentage of high-quality MS/MS spectra (45%) within the representative MS run for *Populus* contained almost twice the percentage (24%) in the *E. coli* run. Nonetheless, the ability to successfully match the high-quality experimental MS/MS spectra to database sequences remained more common in *E. coli*. A greater percentage of *Populus* high-quality MS/MS spectra (77%) remained unassigned, as compared to only 45% of the high-quality MS/MS spectra collected for *E. coli*. This suggests a critical need to evaluate bioinformatic approaches to rescue the lost, high-quality spectra.

To explore the prevalence of single amino acid polymorphisms (SAAPs), a single MS run from within the 60 described above was searched against an expanded *Populus* database that included a list of tryptic peptides generated from predicted SAAP variants in the database. In brief, high-throughput single nucleotide polymorphism (SNP) discovery through deep (30X depth per genotype) resequencing of 19 trees yielded 16 million SNPs in the *Populus* genome (485 Mb) (unpublished results). For this analysis, a subset of these SNPs present in 2 *P. trichocarpa* and 2 *P. deltoides* genotypes were considered. Of

the 17 million amino acid positions found in *P. trichocarpa's* 45,778 protein-coding gene models, ~400,000 amino acid positions due to non-synonymous SNPs (SAAP) were investigated. All possible combinations of SNP-influenced peptides (SAAP peptides) were predicted and subjected to *in silico* tryptic cleavage using PeptideSieve[102] software with the following parameters: maximum mass criterion of 5000, minimum sequence length of 6, maximum sequence length of 50 and allowing for 4 missed cleavages. Some of the non-synonymous amino acid changes resulted in new tryptic cleavage sites or resulted in disappearance of these sites. These were taken into consideration while predicting the peptides. To detect the expression of a SAAP peptide, experimental MS/MS spectra from one MS run were compared to theoretical tryptic peptide sequences generated from a target database consisting of the protein database of *P. trichocarpa* (v2.0) and all predicted SAAP peptides. Each SAAP peptide was concatenated to the target database as a new protein entry, in which ten tryptophan residues flanked both sides of the peptide sequences. For SAAP peptides that originated from the N-terminus of a protein, the tryptophan residues were excluded from the beginning of the SAAP peptide. Similarly, for each SAAP peptide that originated from the C-terminus of a protein, the tryptophan residues were excluded at the end of the SAAP peptide. With the high frequency of SAAPs in *Populus*, over 700,000 distinct SAAP positions and 7,200,000 new peptides were included in our database. All MS/MS were searched with SEQUEST and filtered by DTASelect as described previously.[103] Once peptide-spectrum matches were identified, filtering criteria were controlled to yield peptide FDRs less than 1%. We found that *Populus* proteins on average contained 17 SAAPs. When identifying SAAPs from MS/MS spectra, it is important to differentiate these from post-translational modifications (PTMs) or peptide modifications generated during sample processing that result in mass shifts which are isobaric to several amino acid substitutions. For example, the covalent addition of a methyl group to a K, R, E, or Q produces a mass shift that is similar to the following amino acid changes: D to E, S to T, V to I/L, and G to A. Therefore, all spectra interpreted as both a PTM and a SAAP were discarded to lower the identification of false positives. To identify a targeted common set of PTMs, MS/MS spectra were analyzed by an automated software tool, InSpecT,[104] at a peptide FDR of

2%. In total, 271 spectra that matched to both a PTM and a SAAP peptide were removed from the analysis. Using conservative search criteria, we were able to identify a total of 1,354 peptides containing a SAAP and 201 peptides that become tryptic due to a K or R substitution. Although the new SAAP peptides account for 2% of high-quality unassigned spectra, these newly identified peptides correspond to 502 proteins. Among these, we identified 97 proteins that had not been previously identified. Interestingly, for those proteins containing a SAAP peptide, their overall peptide coverage increased by an average 25%.

Due to the widespread distribution of SAAP peptides in the database, it seems probable that the detected SAAP peptides would map randomly across the proteome. However, our data suggests that the detected population of proteins containing a SAAP peptide map to specific and functionally similar groups. Grouping the SAAP proteins into KOGs, the vast majority of SAAP proteins belonged to the four specific functional categories: unknown function, signal transduction, post-translational modification, and carbohydrate transport and metabolism. Although these functional categories are among the most abundant categories in phloem and xylem, we note that other abundant functional categories, such as general function and translation, do not contain a large number of proteins containing SAAPs. Therefore, it appears that the overrepresentation of non-synonymous substitutions for the aforementioned functional categories is not a result of their expression levels, but rather that these proteins are under low selective pressure. Although it is unclear how many of these proteins represent evolutionary novelties, future comparative proteomics studies may identify expression patterns that reveal the outcomes of such mutations. In some instances, the location of these mutations could compromise or benefit an enzyme: replacing catalytic, binding, or substrate determining residues with amino acids differing in size, polarity, or hydrophobicity can either disrupt or modulate the activity of an enzyme.

For example, when looking at the monolignol biosynthesis pathway, we identified a SAAP within phenylalanine ammonia lyase (PAL), the entry enzyme into the

phenylproponoid pathway. As shown in Figure 3.9, a mass shift of +1 Da and the experimental b- and y- ion fragmentation pattern coincides with the predicted SAAP substitution of an asparagine (N) with an aspartic acid (D) at position 138. While the effect of the observed polymorphism is unknown, the localization of the substitution within a few amino acids of the substrate-binding site may impact the binding of coumarate to the substrate specificity residues.[105] Because studies have shown that PAL serves as a regulatory control point for the entire pathway,[106] any mutations compromising or altering the activity of the enzyme will, in fact, impact the overall lignin content.

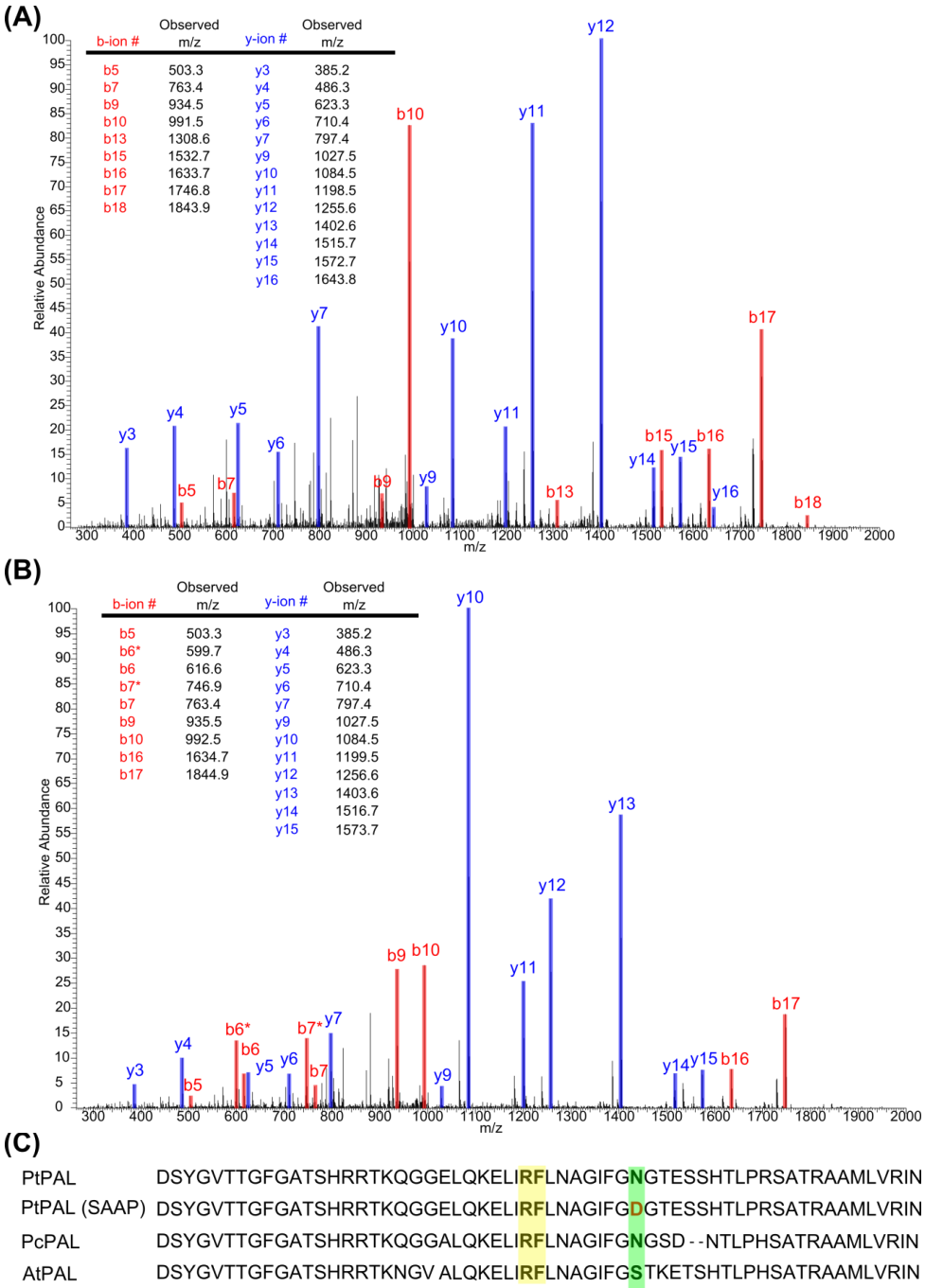



Figure 3.9. SAAP-resolved peptide identification in PAL.

(A) MS/MS spectra of the genomic peptide (FLNAGIFGNGTESSHTLPR) and the (B) SAAP peptide (FLNAGIFGDGTESSHTLPR). (C) A partial sequence alignment of *P. trichocarpa* (PtPAL) with other members of the phenylalanine ammonia-lyase family (PcPAL, *P. Crisum* and AtPAL, *A. thaliana*). The yellow box highlights the substrate specificity residues and the green box highlights the SAAP position.

### *3.2.3. Evaluating Amino Acid Polymorphisms by Proximal Matched Ion Intensities (AAPProxiMIT)*

In a later study, we sought an alternative approach to appending predicted protein sequence variations to the original database in order to detect novel protein forms. The main disadvantage of that approach is the requirement of *a priori* knowledge of SNPs. Moreover, it was preconditioned on both the coverage and quality of the predictions when they are available. Therefore, in a follow-up study, we argue that a more attractive approach considers unexpected single amino acid polymorphisms, relying on matched ion intensity information to help discriminate false positive identifications.

The high-throughput discovery of protein sequence variants (truncations, post-translational modifications, or mutations), has seen tremendous advancements in recent years, with the identification of unexpected variants particularly emerging as investigations of interest.[107] Many database-searching algorithms have been recently designed to effectively identify unanticipated (blind) sequence variants at a global level. One class of such algorithms uses *de novo* sequencing in order to infer full-length peptide sequences from tandem mass spectra without requiring a sequence reference database.[104, 108, 109] A strength of this approach is that the concept of variant peptides is not relevant; each spectrum is given an equal opportunity to match any combination of amino acids, regardless of whether the researcher anticipated detecting the sequence or not. This technique, however, greatly increases the number of candidate peptides compared to each spectrum, consequently incurring not only significant costs to processing time but also unacceptable false discovery rates (FDR).[110] In addition, mass spectrometrists have developed and routinely used a hybrid approach between traditional database searching and *de novo* approaches: here peptide sequence tagging (PST) algorithms can detect unexpected sequence variants as extensions of partial sequences identified from a database.[111-114] In particular, the proteome informatics group led by David Tabb recently released a two-step methodology involving the DirecTag algorithm[81] for highly accurate PST tag generation, followed by the TagRecon software[82] for the detection of peptide sequence variants through tag reconciliation. In brief, short sequence “tags” are directly

inferred from a tandem mass spectrum and then tags are automatically reconciled against representative peptides from a protein database while making allowances for unexpected mass shifts (i.e., mutations and post-translational modifications). PSTs serve as a filter to effectively reduce the number peptide-spectrum matches being scored, which in turn improves costs in processing time, sensitivity, and specificity.[64]

To evaluate a peptide sequence tagging approach for *Populus* with the ultimate goal of globally identifying unknown SAAPs, we employed DirecTag and TagRecon software. Using the state-of-the-art LTQ-Orbitrap-Pro platform, we profiled and compared two genotypes of *P. trichocarpa* and revealed a large number of unexpected SAAPs that would have otherwise been missed by a traditional database search. The sequence variants leveraged from TagRecon demonstrates the value of using peptide sequence tagging algorithms to interrogate proteomics data sets, provided that a SAAP location could be confidently identified. Therefore, while our initial aim was to comprehensively identify SAAPs, we focused on our most abundant sequence variant to show that confident site localization remains an important yet challenging task. Since others have shown that HCD fragmentation improves the coverage of peptide sequences overall, in particular for tryptic peptides up to 15 amino acids in length, we exploited HCD fragmentation to further refine a subset of the dataset.

The procedure described above identified a total of 76 types of sequence variants (each type denoting an amino acid with a mass shift corresponding to a mutation). Noticeably, the occurrence of variants in both genotypes is similar (Pearson correlation = 0.99). Peptides and fragment ions containing an oxidation mass shift (+15.99 Da) were the most prevalent variant type, representing ~38% of the total assigned spectra for variant peptides. While this observation may suggest the two most prominent SAAPs are Ala→Ser and Phe→Tyr, we critically evaluated the results by validating each variant through manual verification of the MS/MS spectra. In the course of this inspection, we observed that the site of +16 Da mass shifts were often in close proximity to a methionine residue (see Figure 3.10), which is frequently oxidized during sample processing.

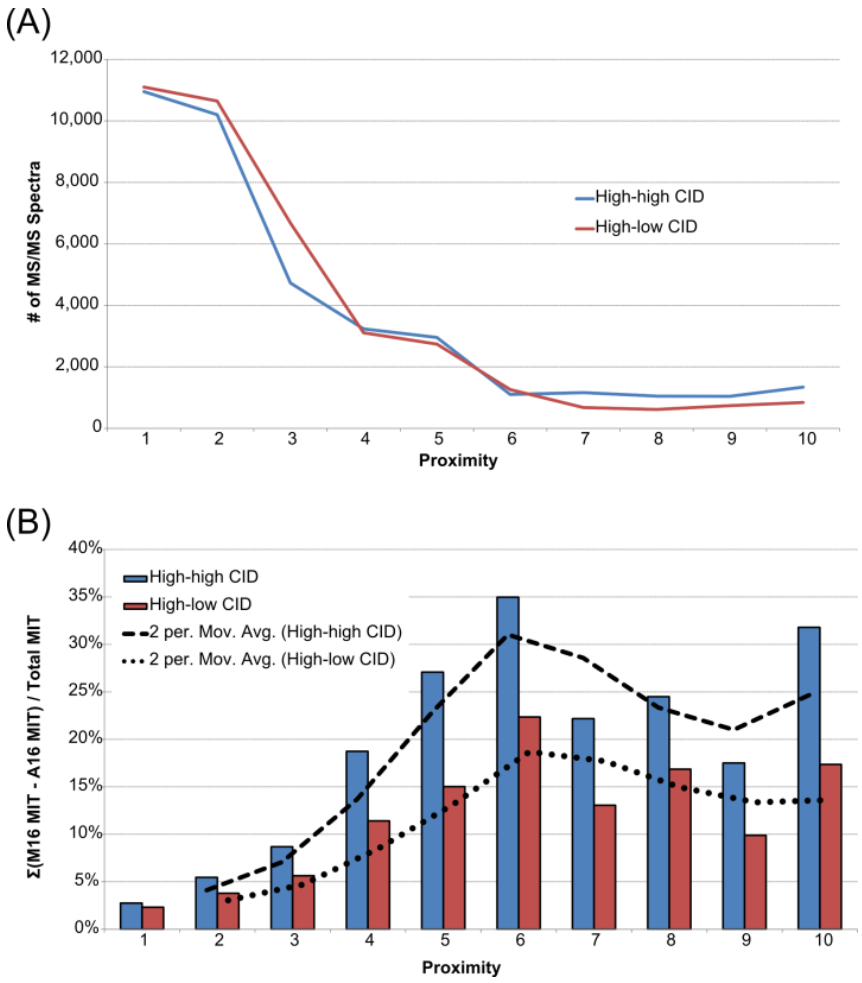



Figure 3.10. Identifying the level of ambiguity between adjacent mass shift sites.

When MS/MS spectra were collected using a high-high (blue) and high-low (red) strategy, some spectra that matched to the same peptide sequence but differed in the placement of the modification (i.e., at alanine or methionine). (A) The frequency distribution of CS illustrates that level of ambiguity is strongly dependent on the distances between two potential modifications sites. (B) A matched ion intensity (MIT) was calculated for the two site positions and the difference between the matched ion intensity values was calculated for each CS as a function of the proximity. A moving average trendline was provided for both the high-high (dashed-line) and high-low (dotted-line) strategy to highlight the earliest maximal difference in the matched ion intensities.

Correspondingly, the site of a ΔA=32 Da mass shift, which can correspond to double-oxidation event or two singly-oxidized alanine residues, was also often found near methionine residues. Therefore, the source of the most frequent and abundant SAAPs could perhaps be explained away as a "shadow" of the most common sampling processing artifact.

Though the presence of a mass shift changes the ion fragmentation pattern of the corresponding ions, the fragmentation process is often incomplete. Some mass shifts will lead to unique fragmentation patterns, enabling a site to be unambiguously located. On the other hand, a mass shift that can occur at adjacent residue sites can introduce ambiguity and lead to incorrect localization; the candidate peptide variants will have similar theoretical fragmentation patterns and thus similar statistical scores. As the distance between the two sites increases, complementary site-determining b- and y-type ions together should increase a scoring algorithm's ability to mitigate the ambiguity (Figure 3.11). Therefore, we objectively evaluated how this ambiguity diminishes as the adjacency decreases.

The analysis was constrained to 'DENA' leaf samples, which contained the highest frequency and abundance of ΔA=16 Da mass shifts. Since high mass accuracy of fragment ions can help unambiguously annotate fragment ion peaks, a MS run using a 'high-high' strategy, which means full scans (MS) and tandem mass spectra (MS/MS) are detected in the Orbitrap analyzer at high resolution and high mass accuracy, was simultaneously evaluated with the a MS run that acquired MS/MS scans in the ion trap ('high-low'). The collected spectra were searched by MyriMatch using a directed method; only a user-defined mass shift was considered. For both MS runs, two directed searches were performed: either a methionine (+16 Da) or an alanine (+16 Da) was allowed as a dynamic modification. By searching for the modifications independently, the search algorithm interpreted each spectrum, identified the mismatch region containing a permissible modification and determined the most probable position of the mass shift on either the methionine or alanine. This approach enabled the identification of spectra

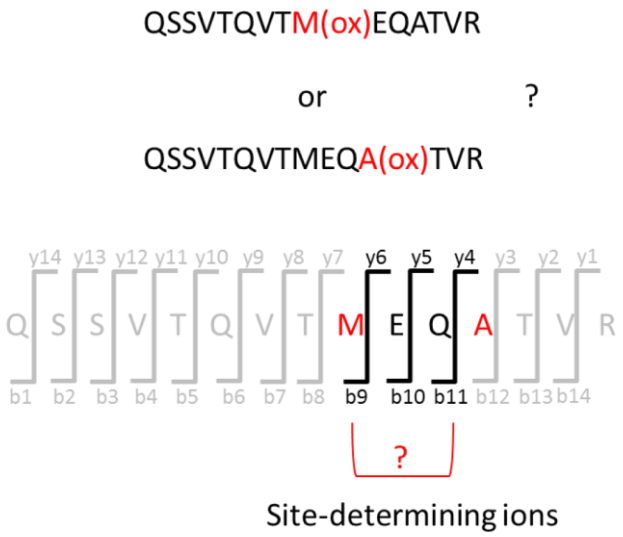



Figure 3.11. Illustration of site-determining ions.

Peptide sequence-tagging approaches can readily identify whether a peptide sequence has a modification (mass shift), but confident localization of the modified residue still remains a challenge. If there are two potential residues that could have modifications, the ions between the two sites are the only pieces of information that could provide evidence for the modification of one residue over another. Residues that are further away from each other within the peptide sequence have more site-determining ions and therefore more opportunities for localizing the modification.

that were annotated similarly, having the same underlying peptide sequence but differing by the location of the mass shift, either on a methionine or a neighboring alanine. For discussion purposes, these spectra will be referred as 'contentious spectra' (CS). In total, the MS searches identified nearly the same number of CS for each analysis strategy – 37,776 and 38,399 for high-high and high-low, respectively.

As anticipated, the number of CS declined as the distance between the methionine and alanine sites increased (Figure 3.10). This observation is the result of an overall increase in the number of discriminatory b- and y-ions, which provides a more definitive spectral fingerprint. Also shown in this figure, the frequency of CS decreased at a similar rate for the two MS strategies. This was expected as both strategies perform collision-induced dissociation (CID); the MS/MS spectra will contain the same percentage of backbone fragmentation. Interestingly, both MS strategies show a clear inflection point when the proximity was ~6 amino acid residues. We suspect that this point represents the distance that provides the most discrimination between the two types of mass shifts, 1) those belonging to a methionine sulfoxide and 2) those more likely due to a SAAP. For distances greater than 6, the mass shift locations likely approach the terminal ends of the peptide sequence. In general, mass shifts located near the ends of a peptide sequence tend to be assigned less reliably than those near the center, which explains why a level of ambiguity remains. These observations are further corroborated by comparing the total matched ion intensity (MITs) of the b- and y-ion series for each peptide sequences that differed only by the location of a +16 Da mass shift. That is, for each ambiguous spectrum, we calculated the difference between the total MIT of the methionine (+16 Da) sequence and the total MIT of the alanine (+16 Da) sequence. Figure 3.10 shows the distribution of the percent difference between two potential sites for each distance. As shown, the maximum difference between the two theoretical mass shift sites occurred when the site locations were 6 amino acids apart. Although we suspected a high level of uncertainty for proximal sites, we demonstrated the likelihood of precise site localization is severely diminished when the number of site-determining b and y ions fall below 12. Notably, the vast majority of the CS (68% high-high and 70% high-low) belong to

peptides containing two potential possibilities that are less than four amino acids apart. Cleary, these spectra have little or no site-determining information for proper site placement, which would be necessary for confident SAAP identification.

As others have shown, these observations highlight how precise site localization can be challenging for search algorithms when there are few site-determining fragment ions261. Presently, additional software is available to calculate the probability of correct localization for each site.[115-118] Though calculating a probability-based score provides a measure of certainty, spectra with insufficient site-determining ions (i.e., peptides with proximal residue sites and spectra featuring incomplete fragmentation) remain logistical problems. In other words, precise site localization in CID fragmentation spectra can be difficult when the distance between the two likely sites is less than 6 amino acids apart. Nevertheless, an immediate alternative approach is available to provide additional information for discriminating between SAAPs and what we suspect is the most common chemical modification mistaken for SAAPs: methionine oxidations.

For peptide-sequence tagging, we employed collision induced dissociation (CID), which is by far the most frequently used technique in proteomics for peptide sequencing. When CID fragmentation techniques are applied, the widely accepted model that describes the dissociation process designates b- and y-ion series as the most prevalent types.[119, 120] The primary fragment ions and their contribution to the overall intensity coverage for a single CID run are illustrated in Figure 3.12A. In principle, complete coverage of the entire b- and y-ion series ions allows full annotation of the amino acid sequence of a peptide. As detailed in the section above, this information may be insufficient for definitively localizing mass shifts. However, there are alternative fragmentation processes that could benefit this task.

Introduced in 2007, higher energy collisional dissociation (HCD) fragmentation became available on the Orbitrap platforms.[121] In a dedicated collision cell, peptide ions are subjected to a beam-type fragmentation process, where primary fragment ions retain

kinetic energy and are therefore more likely to fragment again. In general, HCD ion types are expected to follow the fragmentation rules modeled from CID. Therefore, regular ions (b- and y-type ions) derived from backbone fragmentation are expected to be among the most abundant types observed. Besides a slightly lower contribution of the b- and y-ion series to the total TIC collected in each scan, the observed primary fragment ions and their overall intensities in a HCD run are comparable to CID (Figure 3.12B). A prominent difference, however, is larger contribution of the a-type ion series, which are derived from b-ions by losing CO. Moreover, as a direct consequence of the beam-type fragmentation process, the primary fragment ions are subjected to additional fragmentation pathways and consequently give rise to various ion types beyond those typically observed in CID.[122] A large portion of such ions are those involving neutral losses; the loss of water and ammonia are by far the most frequently observed. Another frequently observed class is the neutral loss of an amino acid side chain. In fact, the side chain of methionine sulfoxide is prone to cleavage,[123] producing ions with a specific neutral fragment loss (NFL). Since search algorithms only consider backbone fragmentation (i.e., a-, b-, and y- ions) and some of their neutral losses ($NH_3$ and $H_2O$), a large percentage of the content in HCD spectra remain unassigned. Though many of these peaks belong to internal fragment ions and immonium ions, there are peaks which can be unambiguously assigned as neutral losses from methionine sulfoxide, based on the knowledge of how they fragment and the calculation of their fragment masses. Therefore, we exploited HCD fragmentation to identify the presence and precise location of methionine oxidations.

Again, the analysis was constrained to ‘DENA’ leaf samples and measurements were collected by the LTQ Orbitrap Pro mass spectrometer, which features improved sensitivity and HCD capability compared to its predecessors. HCD fragmentation was performed in the dedicated octopole collision cell and fragment ions were detected in the Orbitrap. To test the suitability of this approach, the collected spectra were searched by MyriMatch using a directed method: alanine (+16 Da) was considered as the only dynamic modification. With this approach, the search algorithm considers the location of

the mass shift irrespective of a neighboring methionine sites. Methionine was intentionally neglected during the peptide-spectrum matching process to eliminate the MyriMatch scoring system from the discrimination process. HCD spectra that matched a peptide sequence containing a modified alanine (+16 Da) and at least one methionine were further interpreted. This step restricted the analysis to 4,943 spectra, which matched to 1,175 peptides. When annotating HCD peptide-spectrum matches, we looked for the presence of the characteristic neutral loss ions from the primary fragment ions (a, b, and y) of a peptide containing methionine sulfoxide (Figure 3.12A). As mentioned previously, the loss of water and ammonia from primary fragment ions are frequently observed. Therefore, these additional small molecule losses were taken into consideration when applicable.

For each spectrum, we calculated the percent gain in matched ion intensity when considering peaks attributable to the cleavage of a methionine sulfoxide side chain. Figure 3.12 depicts their overall contribution for each ion series: 96% of the spectra and 81% of the peptides exhibit at least one neutral loss from a methionine sulfoxide residue. With only a slight increase in the relative abundance of b-ions, the trends observed for each ion series (Figure 3.12C) agree with their expected contribution in a typical HCD run (Figure 3.12B). The most prominent fragmentation process observed was the neutral loss of methane sulfenic acid (CH4SO). This chemical species exhibited a higher percentage of side change cleavage relative to the frequencies of the other fragment ions and could be observed in 83% of all MS/MS spectra exhibiting side chain loss. Despite only occurring when a fragment ion contains a methionine sulfoxide residue, i.e., $CH_4SO$, $C_3H_6SO$ and $C_3H_8SO$, the three species could be found relatively abundant in the spectra, 3%, 1%, and 1% respectively. While their mean contribution to the overall intensity coverage was 5%, the maximum additional coverage achievable was 31% (Figure 3.12D). The gain in spectral information is promising: if searching algorithms could consider these characteristic permutations during the identification process, the false localization rate of oxidation events would be minimized. It should be noted, that the HCD fragmentation process is not only beneficial for the localization of methionine

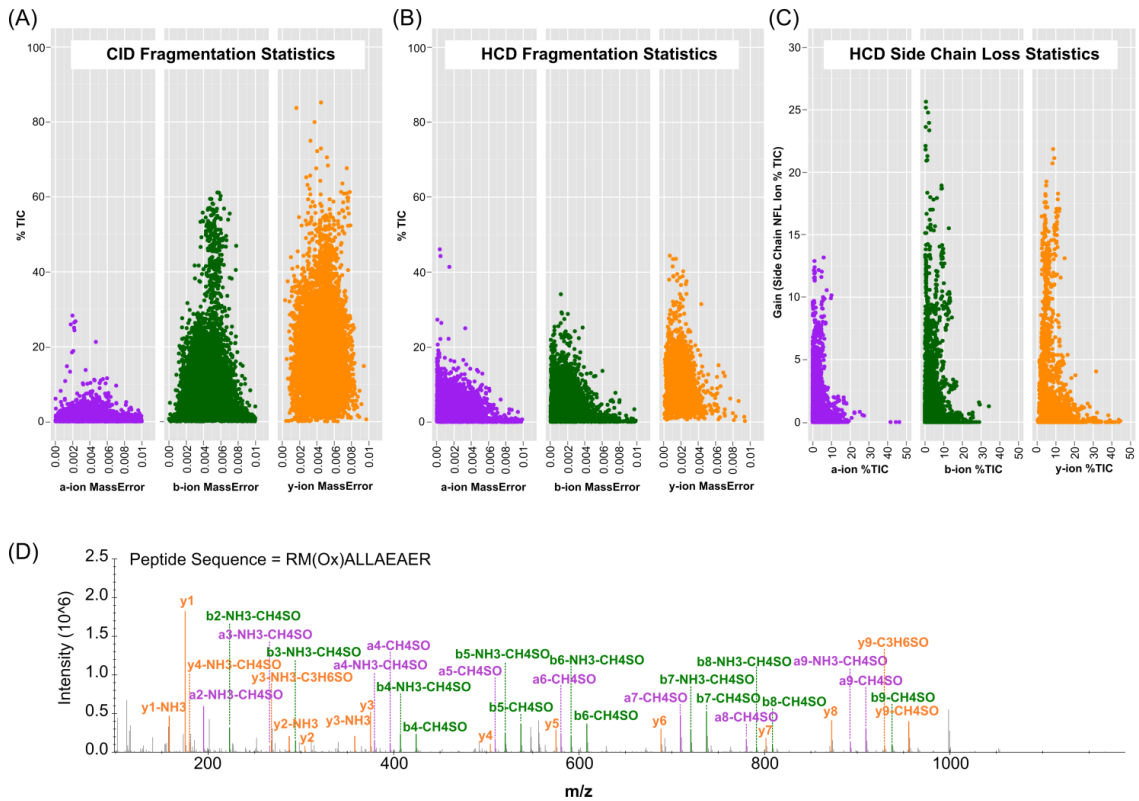



Figure 3.12. Fragmentations statistics of CID and HCD spectra.

a- (purple), b- (green), and y- (yellow) series were plotted. For each CID spectrum (A) and HCD spectrum (B), the percentage of the total ion current (TIC) attributable to a particular fragment ion series was plotted. (C) If a spectrum contained peaks that could unambiguously assigned as neutral losses from methionine sulfoxide, the additional intensity coverage for ambiguous spectra was calculated. (D) As an example, the HCD spectrum with the maximum additional coverage achievable (31%) was provided. Here, only the top 20 most abundant fragment ions were highlighted.

oxidations, but also for other modification events that have characteristic neutral losses, such as phosphorylations.

In summary, the identifications of amino acid polymorphisms remains a huge challenge in shotgun proteomic experiments. New methods have been developed that allow researchers to identify unexpected variants, which improves the speed and throughput of these analyses, but the false discovery rate for these identifications is not quite at an acceptable level. For many of these analyses, one would more likely prefer fewer false positives in order to be very conservative about claiming sequence variants as an amino acid change. Using matched ion intensities, however, one can focus on the site determining ions surrounding the modified residue and compare quantitative values for the evidence supporting that mass shift's position. An advantage that this project's method affords over other similar algorithms is the inclusion of parameterizable ion values, such as the neutral fragment losses from a methionine oxidation measured in HCD. By including the intensities of these additional ions, we can more confidently support the identification of a common chemical modification (methionine oxidation) over the unlikely co-occurrence of a nearby mass shift identifying a mutation from the reference genome.

## 3.3 Conclusions

These studies directly compared the behavior and advantages of spectral counts and matched ion intensities in peptide identification, concluding that using matched ion intensities for qualitative and quantitative purposes avoids many of the biases exhibited by spectral counts. Most notably, matched ion intensities are more sensitive to their chromatographic neighborhood, weighting identifications made in low-complexity regions of the chromatogram less than those identifications made in the rich regions of the chromatogram. Spectral counts, on the other hand, may be biased in over-representing or under-counting identifications in the varying contexts. Matched ion intensities paired with peptide sequence tagging affords a finer level of detail in discriminating the correct localization of mass shifts along a peptide sequence. In fact, matched ion intensities

collected from HCD analysis of peptide-sequence-tagging results refined false positive identifications and suggested that the most abundant mutation identified may in fact be a known chemical modification.

# CHAPTER 4: Peptide to Protein Mapping

*All of the data presented in Section 4.1.1 and 4.2.1 has been adapted from the following journal article:*

Paul Abraham,* Rachel Adams,* Richard Giannone, Udaya Kalluri, Priya Ranjan, Brian Erickson, Manesh Shah, Gerald Tuskan, Robert Hettich. "Defining the Boundaries and Characterizing the Landscape of Functional Genome Expression in Vascular Tissues of *Populus* using Shotgun Proteomics." *Journal of Proteome Research* ***2012*** 11(1): 449-460. *Authors contributed equally to this work. Sample preparation and mass spectrometry experiments were performed by Paul Abraham. The bioinformatic workflow for protein grouping was developed by Paul Abraham, Rachel Adams, and Richard Giannone. The in-house scripts for protein grouping were created by Rachel Adams. Biological data analysis was performed by Paul Abraham.

*All of the data presented in Section 4.1.2 has been adapted from the following journal article in preparation for submission*:

Rachel M. Adams, Richard J. Giannone, Paul Abraham, Robert L. Hettich. "Using Cluster-Unique Sequences in Proteomes (CUSPs) to Enhance Confidence in Protein Inferences from Shotgun Proteomics Studies of Organisms with Complex Genomes." The bioinformatic workflow for protein grouping was developed by Rachel Adams, Paul Abraham, and Richard Giannone. Data analysis was performed by Rachel M. Adams.

*All of the data presented in Section 4.1.3 has been adapted from the following journal article in review*:

Jacque C. Young, Chongle Pan, Rachel Adams, Brandon Brooks, Jillian F. Banfield, Michael J. Morowitz, Robert L. Hettich. "Metaproteomics Reveals Functional Shifts of Microbial and Human Proteins in Infant Gut Colonization." Submitted to *Molecular Systems Biology*. Sample preparation, mass spectrometry experiments were performed by Jacque Young. Database searching and reanotation of protein identifications performed by Rachel Adams. Biological data analysis was performed by Chongle Pan.

## 4.1 Using Cluster-Unique Sequences in Proteomes (CUSPs) to Enhance Confidence in Protein Inferences

### *4.1.1. Outlining Existing Solutions to the Protein Inference Problem*

As a property of evolution, genetic redundancy is rampant across the eukaryotic kingdom. In fact, many eukaryotic organism genomes have been duplicated more than

once in their evolutionary past. As a result, the majority of genetic redundancy observed is between gene homologues. Immediately after gene duplication, these genes (i.e., proteins) are believed to be functionally redundant. It is generally assumed that one of the redundant genes is initially free of all selective pressure, allowing the gene to acquire advantageous (i.e., neofunctionalization) that may lead to a new function. Consequently, the function of some duplicated genes may only be partially redundant.

Redundancy in the genomes, and therefore proteomes, of the samples under investigation often results in a single peptide mapping to multiple proteins. The protein inference problem acknowledges the ambiguity in asserting the presence of a protein whose peptides are shared among other proteins[74, 97, 124]. There are two traditional approaches to counting protein identifications in light of the protein inference problem: 1.) *maximal lists* ignore the ambiguity and count all proteins that are implicated by at least one detected peptide, or 2.) *minimal lists* avoid the ambiguity and count only proteins that are implicated by at least one database-unique peptide.[125] Depending on the proportion of unique peptides detected within an MS sample, a single dataset can be interpreted to have two wildly different protein identification counts. For higher eukaryotes that have a high degree of proteomic redundancy due to large protein families, multiple gene duplications, and different gene models, there is a need for an intermediate protein identification list that is flexible enough to accommodate these ambiguities without leading to over-inflation.

A recent nomenclature developed by Yang et al in 2004 has been used to classify proteins according to their number of unique peptides.[88] This parsimonious naming system consists of three main classes of proteins: *indistinguishable* proteins are those identified without any unique peptides detected, *differentiable* proteins are those identified by detection of some unique and some non-unique peptides, and *distinct* proteins are those identified by detection of only unique peptides.[88] While classifying proteins based on their degree of uniqueness highlights which proteins have more definitive or ambiguous peptide evidence than others, the naming convention does not provide means to handle

differentiable proteins. Although the most common approach to addressing the protein inference problem is to apply Occam's Razor (minimal lists) using this parsimony nomenclature, the number of confident protein identifications (i.e., proteins with at least one distinct peptide) is severely reduced in eukaryotic organisms because of the prevalence of shared peptides.[75] Therefore, objectives of this study encompass developing a bioinformatic workflow that reflects our ability to analytically and biologically distinguish identifications of closely-related proteins.

Since genes (i.e., proteins) with extensive sequence similarity have a high likelihood of performing similar biological roles in a cell, they can be collapsed together by sequence homology algorithms. By grouping homologous proteins together, this consolidates indistinguishable proteins into a meaningful report, while preserving biological information. The research presented in this dissertation has demonstrated that this provides a means to alleviate the majority of ambiguity associated with shared peptides. Similar to a peptide being unique to a protein within the database, many shared peptides are found to be unique to a particular protein group. Although in some cases it may not be clear as to which member of a protein group is actually present in a given sample, the identification of peptides belonging to a particular protein group likely indicates the presence of a shared functional process. Despite sacrificing some level of protein resolution, this approach accurately resolves protein ambiguity as a result genetic redundancy.

### *4.1.2. Choosing Appropriate Identity Thresholds*

In order to provide a means of identifying and grouping proteins that share a large portion of their constituent peptides, proteins that shared a high degree of sequence similarity were clustered into protein groups. The software UClust v4.0 sorted the protein database by descending length and performed a pairwise comparison of each protein's sequence to the running list of "seed" proteins. If a protein shared a high degree of sequence similarity with a seed, the protein was considered a "hit" and joined the protein group. If

a protein did not pass the minimal sequence similarity threshold with any of the existing seeds, it became a new seed against which subsequent proteins could be compared. The final seeds provided a simplified representative list of proteins that were referenced for further downstream analysis including quantitative and functional annotation.

The shared peptide problem is essentially when a single peptide can map to multiple proteins. This problem is exacerbated if multiple peptides are identified which each may infer the presence of several proteins or worse yet, if the majority of the proteins can only be described by shared peptides. Before any measurements are collected, one can predict the number of unique peptides within a database and calculate how many of the proteins are likely to be detected by a unique peptide. It may be the case that only a few peptides are shared by many proteins or a few proteins are shared by many peptides, in which case, the relative small number of ambiguous assignments may minimally impact the collected data. Graphs in the right half of Figure 4.1 illustrate the number of unique peptides per protein across the proteomes of *E. coli, Mus musculus, Arabidopsis thaliana, and Populus trichocarpa*. *Zea mays* and *Oriza sativa* were also analyzed, but are not featured in this figure. Red dashed lines along the diagonal represent where the data points would fall if all of the peptides within a protein were unique (100% distinct peptides per protein), while black dotted lines indicate linear regressions of the observed data. Of the proteomes considered, the *Zea mays* proteome contains the most redundancy. The data's large deviation from the diagonal ($R^2$ = 0.02) indicates that many of its proteins had dissimilar proportions of unique peptides, but more notably, over 75% of its proteins had >95% shared peptides. In other words, only a quarter of the proteome was theoretically not in danger of suffering an ambiguous assignment due to shared peptides. Among the plant proteomes, *Arabidopsis* seemed to be the least affected by shared peptides, although 25% of its proteins had more than 95% shared peptides. This may be due to the fact that *Arabidopsis* is very well-characterized and/or because it is a diploid that underwent a whole genome duplication event much further ago than the other plants considered in this analysis. Interestingly, the length of the protein has very little

correlation with the number of shared peptides, so proteins with few or many peptides are equally likely to be affected by the shared peptide problem.

*E. coli*
4,286 total proteins

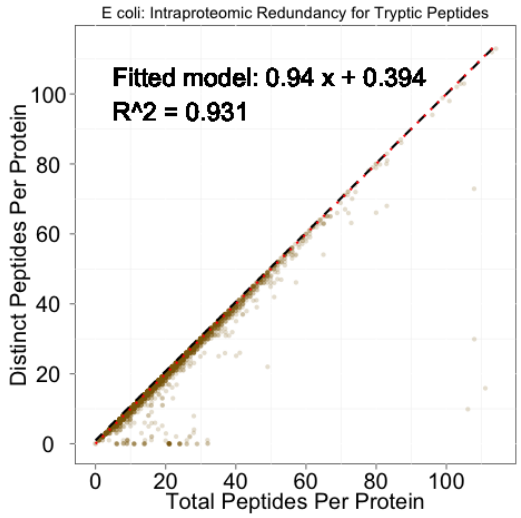



***Figure A:***
*Highly redundant:* 78 (1.82%)

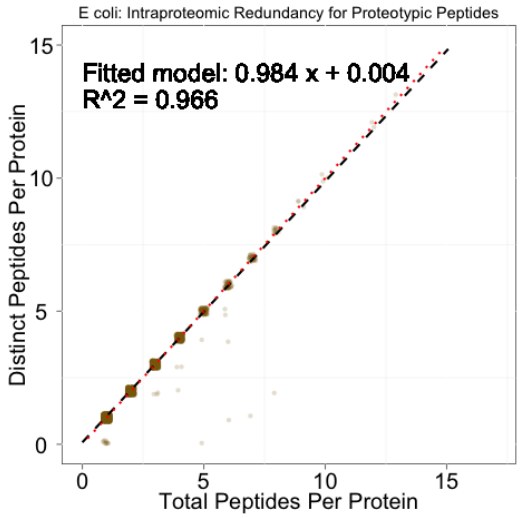



***Figure B:***
*Expect to detect:* 2,522 (58.84%)
*Expect to detect & highly redundant*: 38 (0.89%)

*Mus musculus*
43,895 total proteins

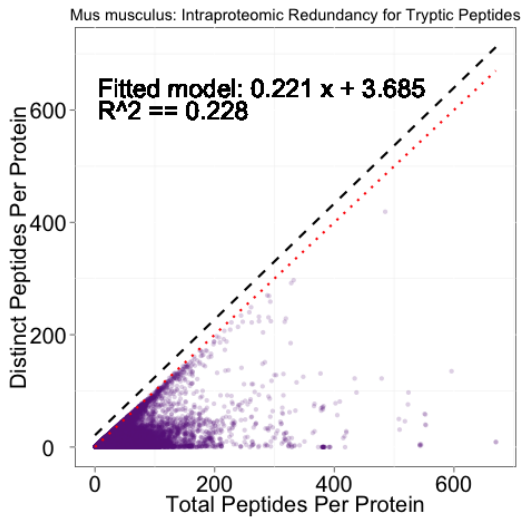



***Figure C:***
*Highly redundant:* 6,724 (15.33%)

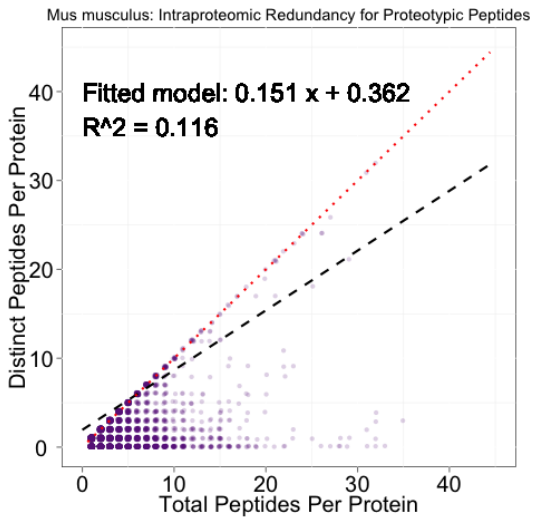
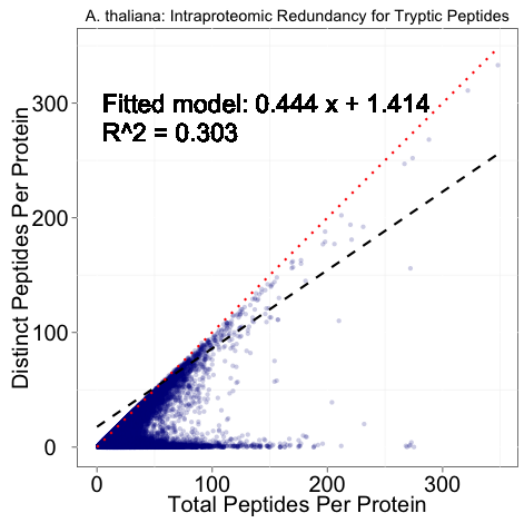



***Figure D:***
*Expect to detect:* 25,098 (54.97%)
*Expect to detect & highly redundant*: 14,470 (33.01%)

(continued on next page…)

*Arabidopsis thaliana:*
35,360 total proteins



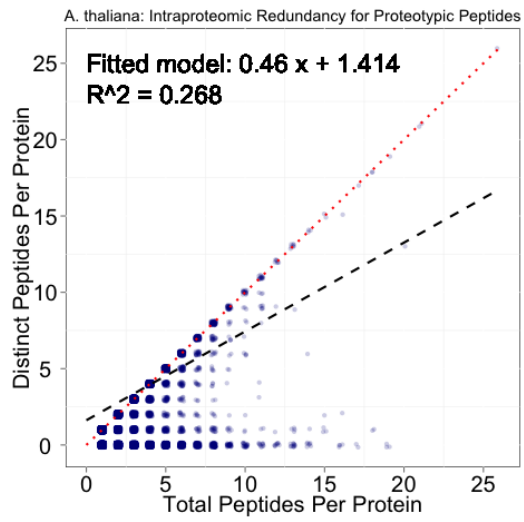
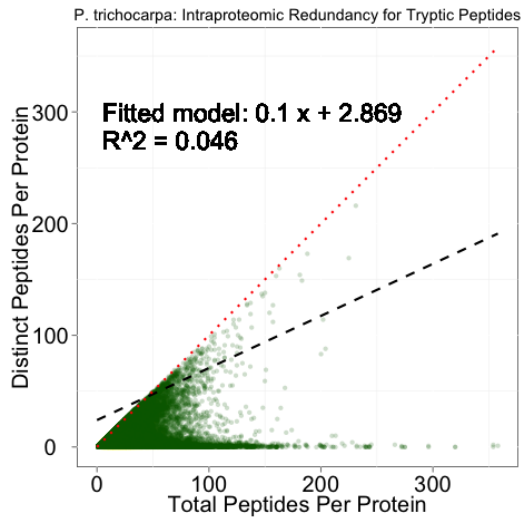




***Figure E:***
*Highly redundant:* 8.860 (25.06%)

***Figure F:***
*Expect to detect:* 22,154 (62.65%)
*Expect to detect & highly redundant*: 9,197 (26.01%)

*Populus trichocarpa:*
73,014 total proteins



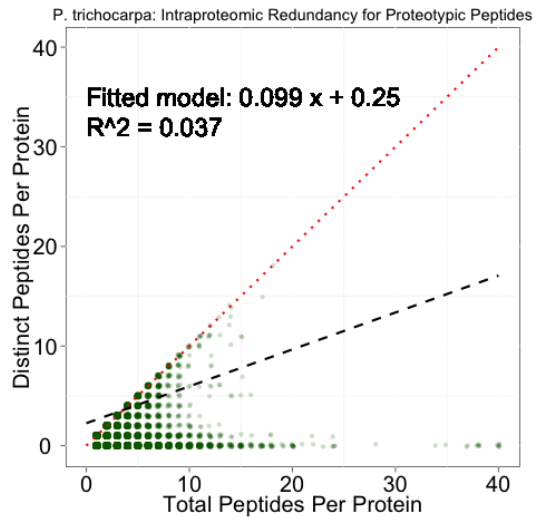




***Figure G:***
*Highly redundant:* 32,185 (44.10%)

***Figure H:***
*Expect to detect:* 46,992 (64.39%)
*Expect to detect & highly redundant*: 33,736 (46.22%)

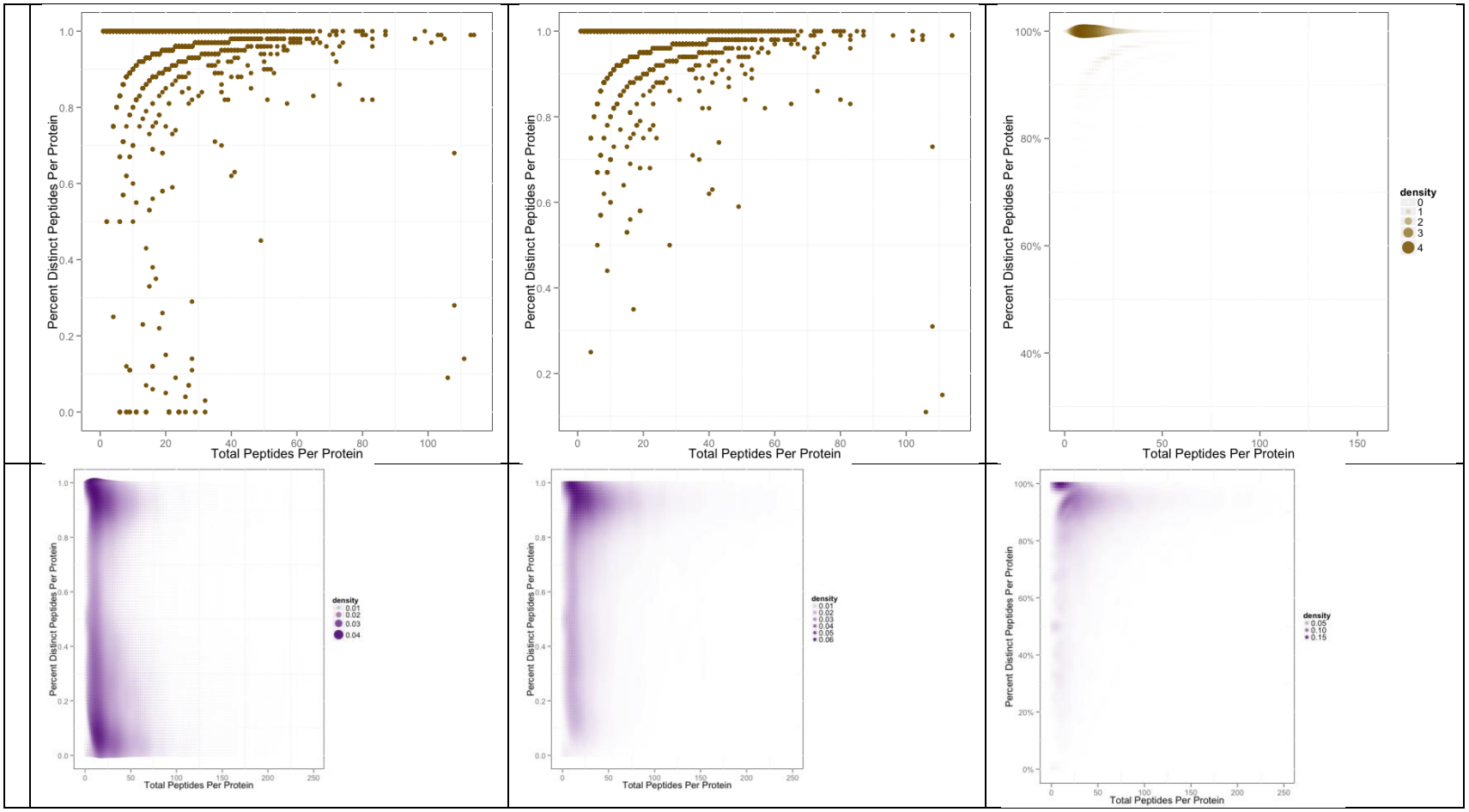
(continued on next page…)

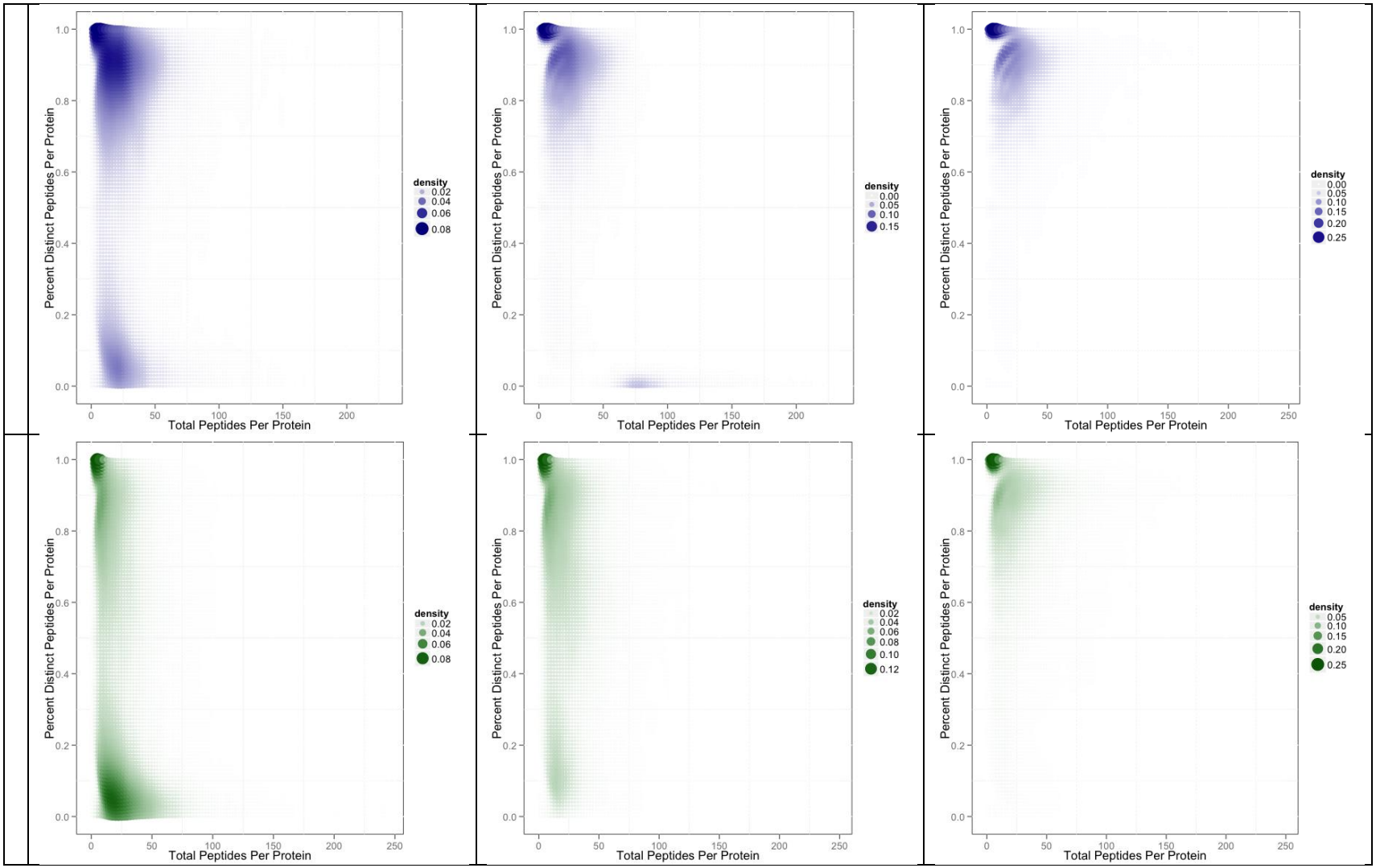
Figure 4.1. Increased redundancy in higher eukaryotes decreases the potential of detecting unique regions within individual proteins.

Graphs in the right half of this figure (ACEG) illustrate the number of distinct peptides per protein across the proteomes of *E. coli*, *Mus musculus*, *Arabidopsi thaliana*, and *Populus trichocarpa*. Black dotted lines indicate linear regressions. Protein redundancy is evaluated by the percentage of distinct peptides per protein ("highly redundant" proteins have less than 5% of their peptides distinct within the proteome). Graphs in the left half of Figure 4.1 (B, D, F, H) illustrate the number of distinct peptides per protein that are likely to be detected by an ESI-MUDPIT experiment ($p > 0.9$), according to PeptideSieve.

Optimistically, we hypothesized that many of the shared peptides may not be MS-friendly, and due to their minimal probability of detection, might alleviate the severity of the shared peptide problem among these proteomes. After running the proteomes through PeptideSieve,[102] software that calculates the likelihood of peptide detection by ESI-MudPIT analysis, graphs of unique peptides per protein were constructed in the same format as before, but this time, with only those peptides that were likely to be detected (p-value > 0.9). From this analysis we were looking for 2 metrics: how many of the proteins did we expect to detect (at least one peptide with p-value > 0.9), and of those proteins, how many were highly-redundant (>95% shared peptides). In general, the trends across proteomes were about the same as the previous analysis. Interestingly, the proteome with the highest percentage of detectable proteins was *Populus* (64%), which had 46% of its detectable proteins designated as highly redundant. In comparison, *Arabidopsis* had 63% detectable proteins, but only 26% of these proteins were highly-redundant. *Zea*, on the other hand, had very few proteins expected to detect, but of those detectable proteins, 80% were highly redundant. Therefore, while *Populus* and *Zea* show different manifestations of a shared peptide problem, they will both have difficulty with protein inference.

The goal of clustering a database by the protein sequence identity is to anticipate the prevalence of shared peptides within the proteome and to alleviate ambiguous protein inferences by grouping together proteins that we are unlikely to analytically distinguish under standard conditions. However, in the process of collapsing multiple protein identifications into one representative, it is just as important to ensure that whatever distinguishable evidence was initially available can still be used to confidently identify as many individual proteins as possible. Lowering the threshold of sequence similarity for clustering proteins together results in larger, fewer protein groups in the proteome and may potentially result in loss of noteworthy distinctions among protein identifications and abundances. Therefore, it's important to analyze the tradeoff between lowering a threshold of sequence similarity and the reducing the number of representative proteins.


Percent Distinct Peptides Per Protein
Total Peptides Per Protein
density
0
1
2
3
4
0.01
0.02
0.03
0.04
0.05
0.06
0.05
0.10
0.15


Percent Distinct Peptides Per Protein
Total Peptides Per Protein
density
0.02
0.04
0.06
0.08
density
0.00
0.05
0.10
0.15
density
0.00
0.05
0.10
0.15
0.20
0.25
density
0.02
0.04
0.06
0.08
density
0.02
0.04
0.06
0.08
0.10
0.12
density
0.05
0.10
0.15
0.20
0.25

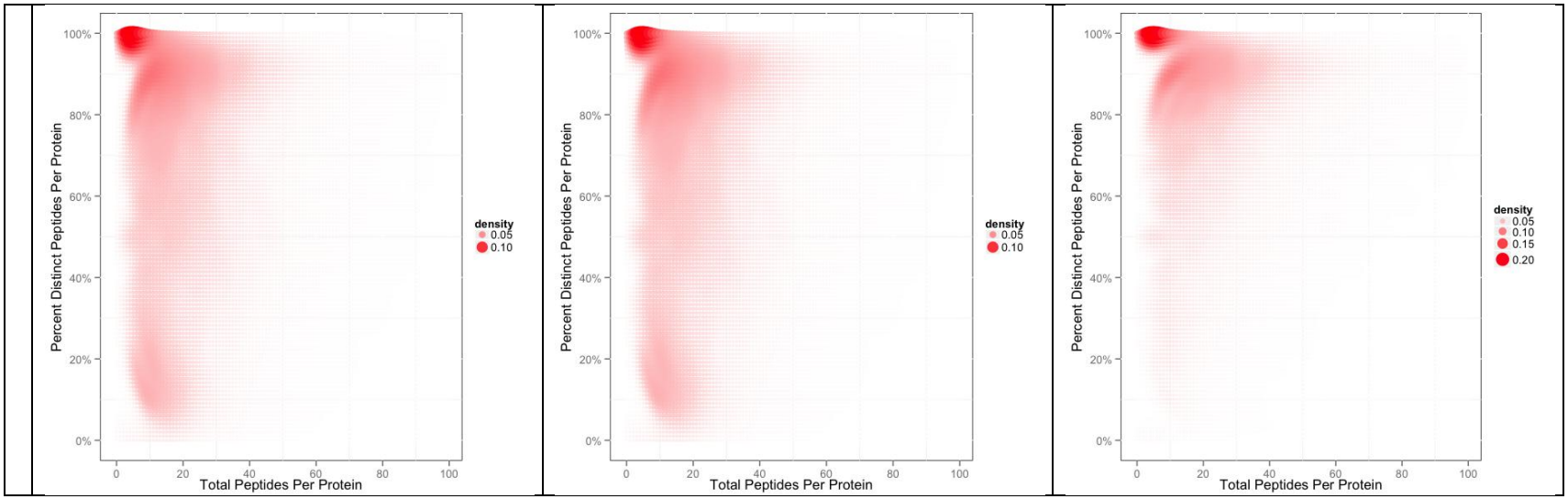



Figure 4.2. Graphs of the percent of unique peptides per protein as a function of sequence similarity thresholds applied to *E. coli*, *Mus musculus*, *Arabidopsis thaliana*, *Populus trichocarpa*, *and Zea mays* (by row).

The columns represent the proteomes clustered at 100%, 95%, and 85%.

To determine the impact of clustering on the number of unique identifications within a proteome, the percent of unique identifications compared to the total number of identifications was graphed across a range of identity thresholds (Figure 4.2). Four proteomes were clustered at conservative thresholds (100%, 95%, 90%, 85%, and 80%) and the resulting percent of unique peptide (y-axis) were graphed according to each protein group's number of total peptides (x-axis). For the purposes of the remainder of this discussion, we will refer to the identity thresholds as their fractional values (id = 1, 0.95, 0.9, 0.85, and 0.8). These graphs help visually compare how the unique peptides are distributed among protein groups at each identity threshold, but the goal of this exercise was to pinpoint at which threshold the majority of the protein groups were characterized by 80% unique peptides. For E. coli, id = 1 clustering was sufficient to achieve this characteristic, but for the eukaryotic organisms surveyed, the thresholds varied slightly. Clustering *Arabidopsis* at id = .95 resulted in 89% of the proteins >80% unique peptides. In fact there was very little difference in the number of proteins >80% unique peptides when the proteome was clustered at 90% (74%) and 85% (78%). Clustered at id = 1, the *Populus* proteome almost looked like an inversion of the *Arabidopsis* graph at id = 1. That is, in the *Populus* proteome, there were two main distributions of proteins: those that had 0-20% unique peptides (representing the majority of the protein groups) and those that had 80-100% unique peptides. Based on the number of protein groups with >80% unique peptides, the *Populus* proteome could be clustered at either id = 0.85 or id = 0.9. The differences between these annotations (76% of protein groups clustered at id = 0.9 or 80% of protein groups clustered at id = 0.85) are quite small, suggesting that one should probably choose the more conservative threshold (id = 0.9). *Oriza* also demonstrated little gains, but its contentious thresholds were id = 0.8 and id = 0.75. At these clustering levels, the *Oriza* proteome generated 74% and 78% protein groups with mostly unique peptides. Again, to be as conservative as possible, the id = 0.8 was suggested to be the identity threshold that would be most appropriate. *Zea*, however, behaved unlike all of the other proteomes considered. Even at id = 0.85, it had only 60% of its protein groups having >80% unique peptides. Once the threshold was lowered to id = 0.75, 70% of its protein groups had predominately unique peptides. In summary, each of the genomes

considered suggested different identity thresholds in their assessment of the number of unique identifications. These recommended values are not meant to be exact rules, but merely guidelines based on what is most likely analytically distinguishable under theoretical conditions. Had we encountered a proteome that was still exceedingly redundant and lacked unique identifiers even after clustering at id = 0.5, we would suggest that clustering based on such low values may substantially lose biological meaning. Under such circumstances, in addition to the assumed critical evaluation of a clustering threshold selected during routine analyses, it is more important that a researcher's discretion should be relied upon to inspect and validate the chosen clustering threshold.

### *4.1.3. Rescuing Identifications that Would Otherwise be Lost*

One method of evaluating the application of a certain threshold is to consider the number of proteins identified in a mass spectrometry run that would have been thrown away due to their lack of database-unique peptides, but that are rescued by evidence of cluster-unique peptides. Whereas the previous section focused on the theoretical distribution of peptides that could be generated from a proteome and their unique status, assessing the impact of clustering on a real dataset of *Populus trichocarpa* proved to be particularly helpful in validating the guidelines we had chosen. Figure 3 graphs proteins that were thrown out of a search using the 1 unique peptide rule, but were rescued after the results were reannotated into protein groups. The x-axis marks the range of identity thresholds that we tested, while the y-axis indicates how many proteins were rescued. The figure suggests that our recommendation of id = 0.85 or 0.9 match the inflexion points where the number of proteins rescued quickly diminishes as the threshold becomes more conservative.

Numbers of peptides, proteins, and protein groups observed in one MS run were compared using the two traditional ways to count proteins and the suggested clustering method. In one MS run from the *Populus* dataset, 10,154 non-redundant peptides were

identified. After clustering the database by 90% sequence similarity, the number of protein groups identified (2,312) was between the counts tallied by the traditional approaches: maximal (3,968 proteins) and minimal (1,880 proteins). Of these 2,312 protein groups identified, there were 2,055 that had at least one cluster-unique peptide.

The degree of confidence in protein identification is directly related to the number of unique peptides assigned to a protein, but the overall unique status of a protein can be more specifically classified by a widely-used nomenclature, as mentioned previously. After clustering the database by 90% sequence similarity, the percentage of non-unique protein groups identified (11%) was far greater than the percentages tallied by the traditional approaches (52%). Overall, the percentage of protein groups identified as cluster-unique was approximately the same as the percentage of all protein groups identified, demonstrating increased confidence in the identifications assigned by the clustering method. The percentage of distinct proteins for clustered data (66%) was higher than the combined percentages of differentiable (37%) and distinct (26%) proteins for unclustered data. The shift in the percentage of distinct protein groups increases confidence in protein group identifications. Therefore, we feel confident in the use of id = 0.9 for the *Populus* proteome after assessment of its behavior theoretically and in a real dataset.

### *4.1.4. Spectral Balancing to Distribute Abundance Measurements*

Typically, a protein's spectral count is the sum of its peptides' spectral counts but clustering proteins into protein groups necessitated methods of assigning spectral counts to the representatives of the protein groups. After an MS dataset was searched against an unclustered database, the protein identifications were reannotated to their representative seed protein names. For peptides that were shared by multiple protein groups, there are several ways we could have assigned spectral counts, but they generally fell into one of two categories: adding the peptide's full spectral count to each of its proteins, or adding a partial spectral count to each of the peptide's proteins (Figure 4.3 below).

For each spectral count $s$ associated with peptide $i$, when $i$ is found in proteins $\mathbf{I_1}$ to $\mathbf{I_n}$,

1.) each protein I in $n_I$ could get $s_i$ spectral counts from $i$
2.) each protein I in $n_I$ could get $s_i \times {}^{1}/_{n}$ spectral counts from $i$
3.) each protein I in $n_I$ could get $s_i \times {}^{S_I}/_{tot_{S,n}}$ spectral counts from $i$
4.) each protein I in $n_I$ could get $s_i \times {}^{U_I}/_{tot_{U,n}}$ spectral counts from $i$

where $\boldsymbol{tot_{S,n}} = \sum_{n}^{I=1} S_I$ , $\mathbf{U_I}$ is the count of unique peptides found in protein *I*, and $\boldsymbol{tot_{U,n}} = \sum_{n}^{I=1} U_I$

Figure 4.3. Possible ways to assign spectral counts when a peptide is shared among multiple proteins.

Spectral balancing, one of the more effective methods to calculate what portion of a spectral count is added to a particular protein, involves a weighting system determined by each protein's number of unique peptides (Figure 4.3, Eq. 2). The algorithm collects all of the proteins that share the peptide and sums the total number of unique peptides found within those proteins. Then, it calculates the proportion of unique peptides contributed by each protein and divides the shared peptide's spectral count accordingly. Ultimately, the proteins with the greater proportion of unique peptides get a greater share of the shared peptide's spectral counts. In order to apply this method into a clustered dataset, all of the peptides that once belonged to the individual proteins were added to the seed proteins and their uniqueness was reassessed. The seed proteins' spectral counts were re-calculated using a spectral balancing method based on cluster-unique peptides.

### *4.1.5. Preserving Functional Annotations*

Clustering proteins based on shared sequence similarity was primarily motivated to improve confidence in reporting protein identifications as well as protein abundance measurements, but the take home message from most comparative proteomic analyses primarily revolve around tethering identifications and abundances to functions. What is

this protein or group of proteins doing differently in this condition compared to that condition? Ultimately, then, clustering has a huge impact in terms of what and how many functional components are observed within a sample. Therefore, we were wanted to determine whether applying the clustering method changed the overall distribution of functional categories in the *Populus* proteome. The distribution of KOG categories within the *Populus* proteome were compared based on the percentages of identifications and spectral counts of unclustered and clustered data. First, a "theoretical" comparison was made using the entire list of identifications in the unclustered proteome and the clustered proteome (id = 0.9). Interestingly, the theoretical KOG category distribution of proteins within a proteome does not change significantly when the identifications come from the unclustered database (including all proteins from the proteome) versus the clustered database (using only representatives from each protein group). This suggests that each functional category benefits equally from the clustering threshold set at 90% and that each representative protein in fact captures the same general information of its members. From this observation, we feel confident that this clustering method preserves qualitative information. The second graph in Figure 4 compares the quantitative observations from a single representative MS analysis of *Populus* leaf. The detected KOG category distributions of proteins from two different samples maintain integrity whether the spectral count comparison is generated from an unclustered or clustered database. The ratios of spectral counts between samples are kept intact for each KOG category, suggesting clustering preserves quantitative information as well.

## 4.2 Applying CUSPs to Large-Scale Proteomic Datasets

### *4.2.1. Applying Clustering: Defining the Boundaries of Functional Genome Expression in Populus using Bottom-up Proteomics.*

In this study, current experimental and computational approaches were employed to obtain a broad proteome profile of *Populus* vascular tissue. The experimental context includes 1) a large *Populus* sample set consisting of two genotypes grown under normal and tension stress conditions182, 2) bioinformatics clustering to effectively handle gene duplication, and 3) an informatics approach to track and identify single amino acid polymorphisms. Together, the integration of deep proteome measurement on an extensive sample set with protein clustering and characterization of peptide sequence variants has provided a level of proteome characterization for *Populus* that has not yet been observed.

To generate a high-coverage proteome profile, we performed bottom-up proteomics on a large sample set consisting of subcellular fractions (soluble, pellet) of two tissue types (xylem, phloem) from two *Populus* species: *P. deltoides and P. tremula x alba*. Using the most recent *Populus* genome draft (v2.0, http://www.phytozome.net/cgi-bin/gbrowse/poplar/), tandem mass spectra from 60 *Populus* proteome measurements collectively identified 7,505 total proteins and 33,233 tryptic peptide sequences with an overall false discovery rate of <1% at the protein level. Combining the proteome measurements together provided a global view of protein expression involved in vascular tissue development, resulting in protein assignments for ~17% of the predicted *Populus* proteome. Approximately 40% of all detected proteins belonged to three specific functional categories based on 24 EuKaryotic Orthologous Groups (KOGs): 1) unknown function, 2) post-translational modification and turnover, and 3) signal transduction. The remaining identified proteins are scattered across the other functional categories.

Shotgun proteomics employs a peptide-centric approach that relies on the ability to accurately assemble and assign thousands of measured peptides to reference proteins in biological samples. Although this is the conventional method for identifying proteins in

large-scale studies, this approach presents several challenges when assigning peptides to proteins in higher eukaryotes. The most common issue deals with inferring a protein's existence through the identification of peptides that constitute its primary structure. Protein inference becomes problematic when two or more proteins share peptides.[75, 97, 124] Shared or degenerate peptides are natural occurrences that originate from protein homology, conserved protein domains among various proteins, splice variants, and redundant entries due to gene duplication events, all of which are common in plants184-185. Compared to *A. thaliana*, the *Populus* genome is highly genetically redundant, such that two-thirds of protein-coding genes share sequence similarity greater than 90% (Figure 3.2A-B). After performing an *in silico* digest of the *A. thaliana* protein reference database, there were ~4.3 million fully tryptic peptides in the database. Out of those, ~320,000 peptides are shared between two or more proteins. After completing an *in silico* digest of the *P. trichocarpa* reference protein database, ~6.3 million fully tryptic peptides were present and, of those, ~2 million are shared between two or more proteins. Clearly, the level of genetic sequence redundancy is extensive in the *Populus* proteome. Therefore, within these large data sets emphasis must be placed on accurate identification and validation of proteins, accounting for highly conserved, shared peptides.

In previous studies, the categorical nomenclature of Yang et al. (2005) has been adapted to rationally organize the peptide data from each LC-MS/MS experiment.[88] Several research groups have shown that this nomenclature can be coupled with Occam's razor constraints to provide a minimal list of proteins to explain all observed peptides.[75] Using this classification method, we consolidated protein assignments by their level of uniqueness. Proteins that consist of only uniquely identified peptides were classified as distinct proteins. Proteins were classified as differentiable when they contain at least one peptide that is unique to that locus, as well as one or more peptides that map elsewhere in the proteome. The indistinguishable proteins consisted of only measured non-unique peptides that map elsewhere in the data set. Within our entire data set, only 50% of the tryptic peptides identified were classified as unique to the database. Therefore, out of the 7,505 total protein identifications in the present study, 3,510 proteins were uniquely

identified (classified as distinct or differentiable) and 3,995 proteins were categorized as non-unique or indistinguishable (Figure: 3.2C-D).

Using the nomenclature above, we generated a minimal list of proteins that were conclusively determined to be present within the data set. However, due to the inherent ambiguity of the *Populus* proteome, less than 50% of the proteins categorized by the above-mentioned criteria could be used for biological interpretation. In addition, due to the extensive homology within the database, a vast majority of the proteins were classified as indistinguishable. As most of the proteins in this category contain no unique peptides, it was difficult to determine which specific proteins were present in the sample using an MS-based approach. As shown in other studies, one approach for proteins that cannot be distinguished on the basis of identified peptides is to collapse these into protein groups to provide a more accurate and informative data set.[126, 127] In an attempt to reconcile this problem, a bioinformatics workflow was incorporated to better handle proteins sharing high sequence homology (90%) to increase qualitative accuracy by avoiding the over- and under-identification of homologous proteins.

Briefly, proteins sharing 90% or more sequence identity were clustered into groups by UCLUST, a clustering algorithm functionally equivalent to BLASTP.[128] Each protein group was defined by a representative protein sequence called a seed, where each seed shares >90% sequence identity to each protein in that cluster. By applying the clustering algorithm to the *Populus* database, the number of protein entries decreased from 64,689 proteins to a total of 43,069 protein groups. Implementation of clustering to the data set reduced the 7,505 observed proteins to a total of 4,226 protein groups (see Methods), in which 2,016 were singletons (i.e., a one-member group). This reduction implies that ~50% of the observed proteins were clustered into groups that shared extensive sequence homology. Therefore, this approach effectively consolidates indistinguishable proteins into a meaningful report. Although grouping proteins by high sequence similarity undoubtedly sacrifices some level of protein resolution, it is reasonable to assume that proteins with this level of sequence homology share similar biological functions.

Furthermore, integrating the clustering approach with the initial SEQUEST analysis provided a means to categorize which members of a protein group were unique.

Due to the peptide-centric nature of shotgun proteomics, it was imperative to report peptides in the context of proteins groups. As expected, clustering proteins into groups alleviated some of the ambiguity associated with shared peptides. Similar to a peptide being unique to a protein within the database, we found many peptides were unique to a particular protein group within the clustered database. In fact, 68% of previously shared peptides that were classified as non-unique to the *Populus* database were reclassified as unique to the clustered database. Moreover, the bioinformatics workflow generated a data set where 84% of the detected peptides were classified as unique. Therefore, rather than disregarding these peptides from the analysis, they were rescued and used for biological insight. While it may not be clear as to which member of a protein group is actually present in a given sample, the identification of peptides belonging to a particular protein group likely indicates the presence of a shared functional process, especially considering the relatively stringent similarity cut-off (90%) applied to the protein database.[129]

Here, we investigate the growth and development of the tree vascular network, which involves a complex system that integrates both molecular signaling components and regulation of protein expression. In higher plants, this elaborate network exists in two vascular tissues, phloem and xylem. Spanning the entire length of plants, these extensive vascular networks are responsible for the distribution of water and essential nutrients across long distances to vital locations. A recent study used bottom-up proteomics to examine proteins expressed during xylem development.[130] This approach demonstrated an ability to robustly characterize xylem tissue in *Populus* by vastly increasing the number of proteins identified and characterized relative to previous *Populus* proteome studies.[131] In the current study, a similar experimental approach was applied to identify and contrast the relationship and dissimilarities between the xylem and phloem proteomes. A “core” proteome was extracted from the entire data set, consisting of 2,627 protein groups that were confidently identified in both xylem and phloem. The core

proteome, encompassing 59% of the total proteins identified in the *Populus* data set, includes proteins representing each KOG category. The core metabolic signature is consistent with other studies that show an overrepresentation of proteins that are involved in energy production and translation.[132] Moreover, a similar quantitative distribution profile was also observed during xylem development.[130, 133] In addition, these functionally and spatially separate vascular networks contain tissue-specific proteins: 606 unique xylem proteins-groups and 461 unique phloem protein groups, each having a distinct metabolic profile.

### 4.2.2. Applying Clustering: Putting the Pieces Together: High-performance LC-MS/MS Provides Network-, Pathway-, and Protein-level Perspectives in *Populus*

In an effort to generate a high-density proteomic atlas that accurately captures the predicted *Populus* proteome, individual proteome maps of the four major organ-types were integrated. In total, we performed multiple (5-6 each) LTQ Velos ion-trap mass spectrometry measurements on proteome extracts from root, stem and both mature (fully expanded, leaf plastichronic index (LPI) 10-12) and young leaf (LPI 4-6) samples. The resulting tandem mass spectra (MS/MS) were searched (SEQUEST) against the most recent protein database of *P. trichocarpa*, containing 45,778 predicted proteins and supplemented with the chloroplast and mitochondrial proteomes.

In plants, the task of assigning identified peptides to their respective proteins is not trivial. Due to the peptide-centric nature of shotgun proteomics, peptides that map to multiple proteins in a reference database can lead to ambiguous identifications. Within higher eukaryotes, this imposes a considerable challenge because shared or degenerate peptides, which result from segmental duplications, homologous proteins or splicing variants and comprise a large fraction of total extracted peptide library.[134-136] To date, there are different methods for aggregating MS evidence for protein assembly.[97] As discussed in Chapter 3, the most advantageous framework to classify and validate protein

identifications in higher eukaryotes should include the following: 1) a means to report the minimum of proteins implicated by at least one unique peptide and 2) the ability to account for database redundancies by clustering similar proteins into groups by sequence homology.

Using the principle of parsimony with Occam's razor constraints, 7,720 *Populus* proteins were confidently identified (classified as distinct or differentiable), and 4,520 proteins were categorized as indistinguishable. Although widely used, the guidelines in the suggested nomenclature make data interpretation more complicated and less accurate, especially in highly redundant proteome databases like *Populus*.

For this reason, we proposed a strategy that incorporates additional supporting information (i.e., sequence homology) to better infer the existence of proteins. While this approach can be applied to bottom-up proteomic studies of plants in general, it confers demonstrable advantages for *Populus* specifically. Proteins sharing 90% or more sequence identity within the *Populus* database were clustered into groups. Each protein group was defined by a single representative protein sequence called a seed, where each seed shares ≥90% sequence identity with all other members of that group. Observed peptides from the originally searched protein entries were then directly referenced back to the clustered database. For the current data set that included 63,056 tryptic peptides, ~25% were previously shared within the original *Populus* database (non-unique/degenerate) but were reclassified as unique to a particular protein group in the newly constructed database. This illustrates the advantage of implementing a "protein group-centric" approach, such that including information about sequence homology allows the interpreter to readily assess the relatedness between shared peptides of indistinguishable proteins derived from gene duplication and splice variants. Moreover, as clustered proteins are ≥90% similar to one another, members of a particular group likely exhibit similar functional roles which, when applied to semi-quantitative proteomics, allows for a more robust analysis of functional signatures across conditions,

time points or organ types. In other words, this strategy effectively reduces the complexity of the functional analysis and biological interpretation of plant data.

Based on this approach, a total of 11,692 protein assignments across all organ- types were reduced into 7,538 protein groups at an average false-discovery rates of <1% at the peptide level. Protein groups were populated by as many as 21 members, with one-membered groups (i.e., singletons) representing only 36% of the total. In total, we were able to measure 25% of the predicted proteins for *Populus*. Generating complete proteome maps of higher organisms is a difficult task as it is unlikely the entire ensemble of polypeptide species encoded by a genome will be expressed at any given time. Nevertheless, this integrated data set provides an "information backbone" that captures baseline protein expression across spatially and functionally distinct pathways. This holistic view of plant-wide protein expression will provide a better understanding of the detected components (i.e., proteins, pathways, etc.) in the context of relationships between organs.

### *4.2.3. Applying Clustering: Metaproteomics Reveals Functional Shifts of Microbial and Human Proteins in Infant Gut Colonization*

Microbial colonization of the human gastrointestinal tract plays an important role in establishing health and homeostasis. However, the time-dependent functional signatures of microbial and human proteins during early colonization of the gut have yet to be determined. Thus, we employed shotgun proteomics to simultaneously monitor microbial and human proteins in fecal samples from a healthy preterm infant during the first month of life. Microbial community complexity and functions increased over time, with compositional changes that were consistent with previous metagenomic and rRNA gene data indicating three distinct colonization phases. Overall microbial community functions were established relatively early in development and remained stable. Detected human proteins included those responsible for epithelial barrier function and antimicrobial activity. Some neutrophil-derived proteins increased in abundance early in the study

period, suggesting activation of the innate immune system. Likewise, abundances of cytoskeletal and mucin proteins increased later in the time course, suggestive of subsequent adjustment to the increased microbial load. This study provides the first snapshot of coordinated human and microbial protein expression in the infant gut during early development.

A search database was generated from the predicted protein sequences of dominant members reconstructed from metagenomic sequences collected on days 10, 16, 18, and 21 from matched samples. These included a *Serratia* species *UC1SER*, two closely related *Citrobacter* strains, *UC1i* and *UC1ii*, an *Enterococcus* species *UC1ENC*, and associated virus and plasmids *UC1ENCp*, *UC1ENCv*, and *UC1CITp*. Since samples from early time points were not represented in the metagenomic sequences, the following additional isolate sequences, selected based on 16S rRNA data, were also included in the database: *Arcobacter butzleri* RM4018, *Acinetobacter junii* SH205, *Bacteroides fragilis* NCTC 9343, *Bifidobacterium adolescentis* ATCC 15703, *Bifidobacterium longum infantis* ATCC 15697, *Campylobacter concisus* 13826, *Clostridium sporogenes* ATCC 15579, *Enterobacter cancerogenus* ATCC 35316, *Escherichia coli* K12 DH10B, *Eubacterium rectale* ATCC 33656, *Fusobacterium* sp. 1_1_41FAA, *Klebsiella* sp. 1_1_55, *Lactococcus lactis* subsp. *lactis* KF147, *Lactobacillus reuteri* 100-23, *Leuconostoc mesenteroides cremoris* ATCC 19254, *Pseudomonas aeruginosa* PAO1, *Staphylococcus aureus* 04-02981, *Streptococcus* sp. 2_1_36FAA, *Weissella paramesenteroides* ATCC 33313 (acquired from JGI: http://www.hmpdacc-resources.org/cgi- bin/img_hmp/main.cgi in January of 2011 ).

Since mass spectrometry based proteomics identifies proteins by their corresponding peptide sequences, data analysis must take into consideration the high levels of protein redundancy within and between species to avoid inflating the total number of proteins identified or misinterpretation of the biological conclusions by over- representing proteins with the same function. Therefore, we applied a bioinformatic clustering algorithm to the database in order improve confidence in protein identification and

quantification. Different similarity thresholds were chosen to reflect the higher level of redundancy in the human genome due to gene duplications, splice variants, and multiple protein isoforms. Microbial proteins were clustered using more stringent criteria in order to preserve species information and distinguish functional contributions of different community members. Specifically, using the publically-available software, USEARCH v.5.0,[128] microbial proteins were clustered into a protein group if they shared 100% amino acid identity, and human proteins were clustered into a protein group if they contained ≥90% amino acid similarity. These differing similarity thresholds were chosen based on the higher numbers of paralogous proteins present within the human genome, and were supported by plotting similarity thresholds ranging from 0.5-1 against the percent proteome reduction via clustering. In fact, the clustered microbial metaproteome had 0.5% of its protein groups with more than one member and the clustered human proteome had 36% of its protein groups characterized by multiple members. Spectral counts were assigned, balanced, normalized, and adjusted according to methods previously described, yielding adjusted NSAF values.[103, 137, 138] In total, 4,413 microbial and 3,062 human protein groups were detected across the dataset. Protein groups range from singletons to groups that contain multiple protein isoforms.

By measuring both microbial and human proteins simultaneously in each run, we observed an increased complexity of the microbial composition and a decrease in the ratio of total human/microbial proteins with time (Figure 4.4). At the earliest time point, when the initial microbial communities were being established, human proteins comprised ~96% of all proteins identified (day 7). The low microbial load may be a consequence of antibiotic administration during the first week of life for this particular infant. Human proteins comprised ~72% of the identified protein dataset on day 13, and by day 15 the percent of human proteins decreased to ~30%, with a concomitant increase in the number of microbial proteins detected. The ratio of human to microbial proteins remained at this level for the remainder of the times measured, with the exception of day 20, when an unexpected rise in human proteins was detected. Microbial proteins detected in this time course study are consistent with metagenomic inference of three distinct

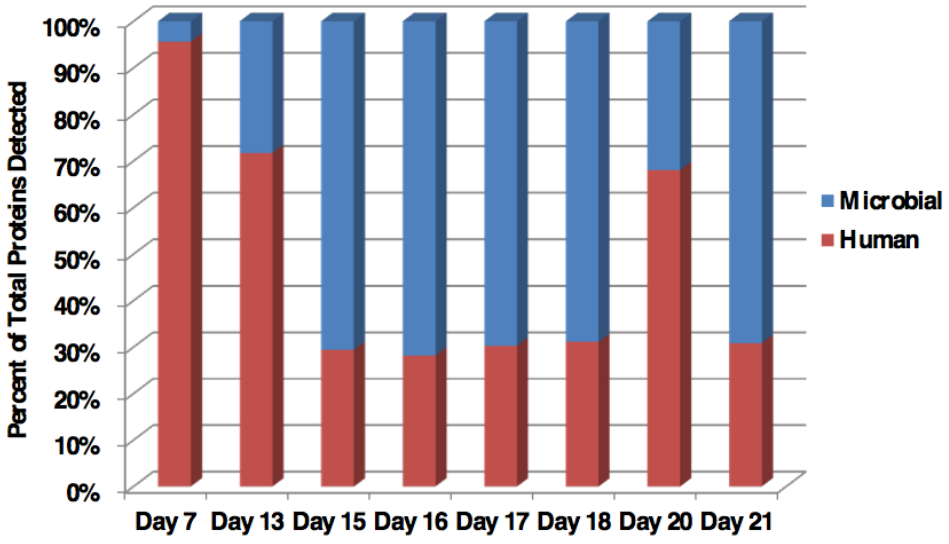



Figure 4.4. Graph of the ratio of total human/microbial proteins with time.

colonization phases with vastly different species composition. Despite temporal changes in microbial community composition, the overall functions of the community increase in complexity with time, stabilize relatively early, and remain remarkably conserved thereafter. Thus, this study provides detailed information about the microbial and human proteins in fecal samples from a newborn premature infant during the first month of life, and reveals the complex-but-synergistic interplay of host adaptation to microbiome establishment.

## 4.3 Conclusions

In this project, we developed a potential solution to the protein inference problem: clustering protein databases by sequence similarity groups together proteins that we are unlikely to analytically distinguish while also taking into consideration shared biological functions. While other existing approaches group proteins based on the observed peptides detected within a run or experiment, our approach, Clustering Unique Sequences in Proteomes (CUSPs), provides more stable grouping that only changes with the database-not with observed data. In addition, by comparing entire protein sequences rather than partial sequences, we are more confident that the proteins are grouped based on similar biological function (i.e., multiple domains and motifs). We suggest using two approaches to identify an appropriate clustering threshold: the reduction of proteome size as a result of clustering and the number of distinguishable identifications from the clustered database. While lowering the threshold for grouping proteins will create more groups, it is also possible to lose unique information that could be helpful in confidently pinpointing which proteins are identified within the sample. We considered these tradeoffs for a number of complex proteomes, including *Mus muculus*, *Populus trichocarpa*, *Oriza sativa*, and *Zea mays*. In total, we suggest that each proteome has different properties that would recommend different identity thresholds, so future studies would need to adopt this methodology to find the most appropriate identity for grouping the proteome of interest. Case studies of *Populus trichocarpa* and the infant gut microbiome demonstrated successful implementation of CUSPs to gain crucial insight into the identification and quantification of proteins that would have otherwise been excluded from their analyses. Therefore, CUSPs not only removes ambiguity from protein reports, but also rescues and strengthens the confidence in the protein identifications and abundances measured in complex proteomic studies.

# CHAPTER 5: Protein Quantification

*Data presented in Section 5.2 has been adapted from the following journal article ready for submission to the* Journal of Proteome Research:

Rachel M. Adams,* Richard J. Giannone,* Paul Abraham, Robert L. Hettich. "Protease-Optimized Spectral Indexing Enhances Protein Identification and Quantification in Shotgun Proteomics Datasets." * Authors contributed equally to this work. Sample preparation and mass spectrometry experiments were performed by Richard J. Giannone. Data analysis was performed by Rachel M. Adams.

*Data presented in Section 5.3 has been adapted from the following journal articles*:

Zhou Li,* Rachel M. Adams,* Karuna Chourey, Gregory B. Hurst, Robert L. Hettich, and Chongle Pan. "Systematic Comparison of Label-Free, Metabolic Labeling, and Isobaric Chemical Labeling for Quantitative Proteomics on LTQ Orbitrap Velos". *Journal of Proteome Research 2012* 11(3):1582-90. * Authors contributed equally to this work. Sample preparation, mass spectrometry experiments, and manuscript preparation were lead by Zhou Li and Chongle Pan. In-house scripts for comparison of HCD with dual HCD/CID identifications and iTRAQ and TMT intensity summarization were written by Rachel Adams.

## 5.1 Using a Poisson Bootstrapping Method to Test Differential Protein Expression Based on Spectral Counts

Recently the traditional quantitative MS methods using isotopic labeling have received criticism for being biologically and computationally cumbersome. In light of these drawbacks, our endeavors have been focused on pursuing label-free quantitative analysis of MS data. Strong cases have been made for running each sample separately and considering inherent characteristics of the measured data, primarily spectral count (SpC), in order to compare protein abundances.[139-141] By counting the number of times each protein is detected with respect to the other proteins in a sample, a relative abundance can be calculated. Likewise, a relative abundance can be calculated by comparing the spectral count of a protein between samples. Appropriate normalization and consideration of statistical guidelines for determining differences between groups of values, spectral

counts in this case, are left to the discretion of biostatisticians with the implicit understanding that the methods implemented should be biologically and analytically defendable.

The normalized spectral abundance factor (NSAF) is a widely-used normalization method for spectral counts that takes into consideration protein length and the run's overall total spectra collected.[138] This normalization method assumes that longer proteins are more likely to have more peptides and therefore higher spectral counts than shorter proteins. While protein length is an important consideration when comparing two proteins within a sample, it becomes irrelevant when comparing the same protein between two conditions. Normalizing total spectral counts between runs, however, is exceedingly important; depending on any number of instrumental or experimental factors, two runs may have quite different total spectral counts. Due to their prevalence in label-free shotgun proteomic studies, NSAFs are benchmark measures against which any newly-proposed label-free measures or normalization methods are compared. In fact, NSAF is the prominent label-free measure that has pushed spectral counts as the forerunner among label-free features, including spectral intensities, peak area or peak height, or a combination of these features. It is therefore beneficial to compare the descriptive behavior of any normalized label-free measure to NSAFs. An objective of this research will be to properly employ a normalized label-free measure to determine and compare relative protein abundance between samples.

Alongside the debates over the most effective label-free features to be used for measuring protein abundance, many studies have suggested that certain statistical methods are more appropriate for detecting and assessing changes in protein abundances. Most proposed algorithms, such as the t-test and G-test, assume normally-distributed data and can only handle pairwise comparisons,[142] which can become computationally cumbersome for experiments involving multiple conditions. Pooling replicates' data into a single representative value is an additional limitation characteristic of several algorithms previously implemented in label-free quantification. An additional aim of this study is to

optimize a statistical test that best determines whether two spectral counts are significantly different from each other.

One of the most straightforward approaches to normalization is to transform all of the run's total assigned spectra counts to the same amounts, much like a percentage but with the sum being a number that is more representative of the spectral counts assigned. To achieve this method, one first divides each protein's raw SpC by the total SpC for that run. Then the normalized SpC is divided by the sum of the protein's normalized SpC and multiplied by the sum of the protein's raw SpC. For validation, the sum of each protein's normalized SpC's should equal the sum of the protein's raw SpC.

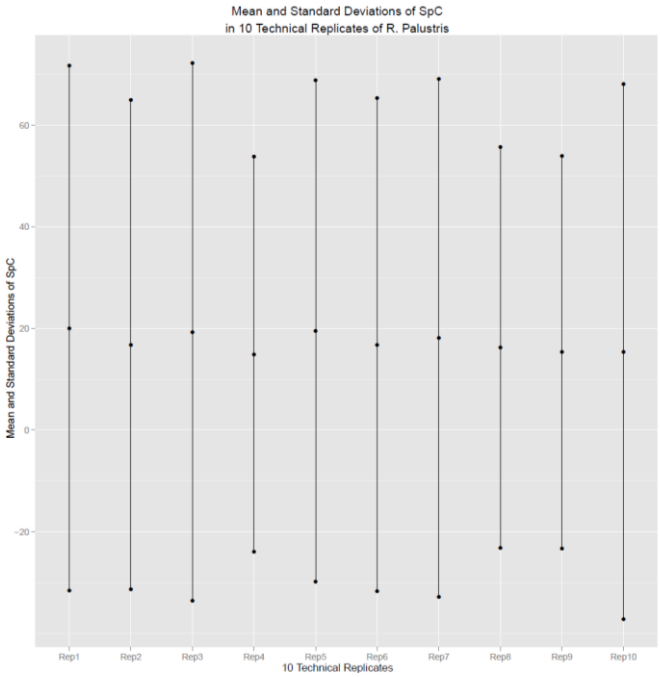
A second approach, normalization by means, ensures that each run's average SpC is the same. For this method, the average spectral count of each run needs to be identified as well as the average spectral count of the entire dataset. Next, each run's total spectral count is multiplied by an adjustment factor (the run's spectral count divided by the average spectral count for the entire data set). This method may be preferable over the normalization by totals only if one is confident that the variation between runs is approximately the same.

Before developing an improved method for determining differential expression in spectral counts, we wanted to get a sense of the overall behavior and characteristics of raw spectral count (SpC) measurements. Looking at non-normalized SpCs allowed us to inspect the distribution of unaltered measurements without any presuppositions about how the data should look. To ensure that our sampling space was large and reproducible enough to make unbiased observations, 10 samples of *R. palustris* were prepared with identical protocols and measured using an LTQ. A total of 3,114 proteins were identified in these 10 runs. Our first, most general inquiry sought to ascertain basic descriptive statistics about the entire dataset's proteins and spectral counts. Each run identified 2,005-2,329 proteins, and each run assigned 47,985-64,623 SpC. 44% of the proteins were identified in all 10 runs compared to the 12% of proteins that were unique to a

single run, highlighting the high-degree of reproducibility between these technical replicates. The plots in Figure 5.1A illustrate that the 1500 proteins found in all 10 runs have a similar SpC distribution as the entire dataset, so those highly reproducible proteins can be used to represent the dataset for later analyses. Additional validation of the reproducibility of the dataset was measured by calculating the standard deviations and variances for each protein's spectral counts across the 10 runs compared to that protein's overall mean spectral count.

From these basic measurements of means, standard deviations, and variances, we were able to start exploring the overall distribution of spectral counts. Figure 5.1B graphs the relationship between the means and variances for each of the 3,114 proteins. The strikingly linear relationship between the means and variances ($R^2 = 0.89$) with a slope close to 1 (1.09) suggests that for the majority of the proteins in this study, the mean spectral count equals the variance. In other words, the more abundant proteins (those with larger spectral counts) have a larger standard deviation than the less abundant proteins (proteins with smaller spectral counts). While this observation seems intuitive, it has very significant implications in supporting the data's Poisson-like distribution. Poisson distributions are characterized by a single variable: the mean (lambda, $\lambda$), which is equal to the variance and therefore the only degree of freedom. Normal distributions, on the other hand, are characterized by 2 variables: the mean and standard deviation. If the data had followed a normal distribution, the standard deviation (and therefore the variance) would have been approximately constant for all proteins, regardless of their means. Figure 5.1's demonstration of the mean dictating the variance suggests raw spectral count data follows a Poisson-like distribution.

**(A)**




**(B)**

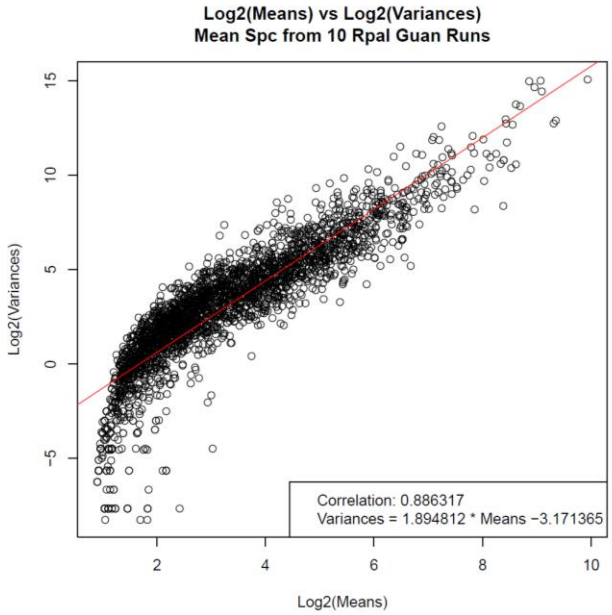



Figure 5.1. High degree of reproducibility between 10 technical runs of *R. palustris*.

(A) Means and standard deviations of 1500 proteins found in all 10 technical replicates.

(B) Means versus Log2(Variances) of all 3,114 proteins found in the 10 runs.

While the linear correlation between mean and variance was very strong for the raw spectral counts, it was not known whether the data behaved the same way after normalization. The same dataset of 10 runs was transformed by NSAF, an in-house normalization, and additional modifications, including log and square-root transformations, to see whether the correlation between average and standard deviations improved (i.e, whether the data became more Poisson-like after normalizations were applied). There is a highly linear correlation ($R^2$ = 0.94) between the averages and standard deviations for NSAF values, but the in-house normalization method produces a stronger correlation coefficient ($R^2$ = 0.95) between spectral counts' averages and variances. There is also a strong linear correlation ($R^2$ = 0.95) between the average and the relative standard deviation of normalized spectral counts. The correlation coefficient for the square-root transformation ($R^2$ = 0.92) is less than that of the logarithmic transformation ($R^2$ = 0.9435). In total, all of these strong linear correlations between the average and standard deviations of spectral counts suggest the data follows a Poisson distribution. However, the additional transformations and normalizations do not greatly improve the raw correlation between means and variances ($R^2$ = 0.95) and therefore they do not greatly impact the fit of a Poisson-like distribution.

In fact, goodness of fit (GOF) tests showed the majority of the proteins in the dataset were consistently a Poisson-like distribution when the datasets were raw and normalized by means. Since comparisons of multiple datasets increases the sparseness of the data (by generating more data points produced by proteins found in only one or a few experimental runs), additional GOF tests were performed to determine if the number of runs in which a protein was identified affects whether that protein follows a Poisson-like distribution. As Table 5.1 displays, although the percentage of proteins that followed a Poisson-like distribution did not vary much according to the number of replicates identifying a protein, the highest number of proteins following a Poisson distribution were those found in all 10 runs. Of the 1368 proteins found in all 10 runs, 85% of these proteins passed a Poisson GOF test ($p < 0.05$). Similar results were found for normalizing the proteins by the means. These proteins found in all 10 runs are most likely our more

Table 5.1. Results for maximum likelihood goodness of fit test to Poisson distribution using protein SpCs from 10 replicates of R. palustris.

| # Exact Reps | # Proteins | % Proteins | $p < 0.05$ # Proteins | $p < 0.05$ % Proteins |
|---|---|---|---|---|
| 3 | 173 | 5.56% | 128 | 73.99% |
| 4 | 124 | 3.98% | 109 | 87.90% |
| 5 | 138 | 4.43% | 108 | 78.26% |
| 6 | 135 | 4.34% | 94 | 69.63% |
| 7 | 165 | 5.30% | 110 | 66.67% |
| 8 | 160 | 5.14% | 104 | 65.00% |
| 9 | 221 | 7.10% | 161 | 72.85% |
| 10 | 1368 | 43.93% | 1207 | **88.23%** |

confident identifications; therefore, their GOF tests further enhance our assertion of the Poisson distribution.

To demonstrate whether ratios of spectral counts can approximate known relative protein abundances, a dataset was analyzed from a standard mixture of metabolically-labeled proteins in which *R. pal* was grown in normal ($^{14}N$) minimal media and heavy ($^{15}N$) minimal media and then mixed in ratios of 1:1, 1:5, 1:10, 5:1, and 10:1. The graphs in Figure 5.2 demonstrate that ratios of $^{14}N$:$^{15}N$ spectral counts can approximate differences in known relative protein abundances whether those differences span an order of magnitude (1:10 or 10:1) or whether the differences are negligible (1:1). Some of the proteins were found in only one of the conditions, so their ratios could not be calculated and therefore are displayed separately. The histograms of 3,115 proteins distribute around the expected $^{14}N$:$^{15}N$ spectral count ratios, indicated by the thick red lines.

Confident that not only are spectral counts valid measurements of protein abundances and that they tend to follow a Poisson distribution, we propose a novel parametric bootstrapping method of normalization. For the explanation of this method, we assume that there are at least two technical replicates for two biological conditions. For each condition, a single random measurement (SpC) selected from either replicate is chosen. For each measurement, 1000 bootstrapped values are then randomly generated from a Poisson distribution described by $\lambda$ = the value of the random sample. A random pair of bootstrapped values are chosen to represent each condition. The ratio of the two values is calculated, followed by calculating the p-value.

To benchmark the performance of other existing significance analysis tests, we compared QSpec[90] and BetaBinomial[91] tests on our *R. pal* standard mixture datasets in addition to the PBS method. In Figure 5.3, ROC curves of the false positive rate (FPR) against the true positive rate (TPR) provide a visual representation of the tradeoffs between *significance* and *power* for each algorithm. Area-under-the-curve (AUC) calculations confirm that the PBS method outperforms the other methods on the *R. palustris* labeled

standard mixture datasets. However, the power of the algorithms at 95% significance was a more practical measure of assessing the algorithm's performance at a typical p-value cutoff ($p < 0.05$).

Figure 5.3B illustrates that the power of our Poisson Bootstapping (PBS) method (63%) far surpasses that of QSpec (>1%) and BetaBinomial (33%) at 95% significance. In fact, even with exceedingly stringent cutoff criteria ($p < 0.001$), PBS can discriminate numerous differences in protein expression where the other methods cannot detect any changes in abundance at all.

Of the 1647 proteins compared, 845 were considered differentially expressed by at least one method at $p < 0.05$. PBS identified 670 (80%) of these proteins, and 104 proteins were found to be identified by PBS only. To determine how our method was improving over the other methods, we sought to classify the types of abundances being compared (Figure 5.4). Each protein abundance was categorized as "high" (SpC > 50), "medium" (5 < SpC < 50), "low" (2 < SpC < 5), or "zero" (0 SpC) so the comparison of a protein in condition A and condition B would be considered a "high-high" abundance comparison or "high-low" or "medium-low" or "medium-zero", etc… For each of these classifications, PBS identified 78-97% of all significant proteins except for "low-zero" proteins, which are the least confident identifications. In comparison, QSpec identified at least 92% of the "high-high," "high-medium," "medium-medium," and "medium-low" proteins but missed a majority of the "medium-zero" and "low-zero" proteins. BetaBinomial, on the other hand, identified 93% of the "low-zero" proteins and 29-69% of the other proteins. While QSpec emphasized significance in high abundant-proteins and BB emphasized significance in low-abundance proteins, PBS identified the largest overlap and unique set of proteins across the entire dynamic range. Therefore, the significance analysis achieved by PBS excels in correctness, lack of bias, and comprehensiveness compared to other existing approaches.

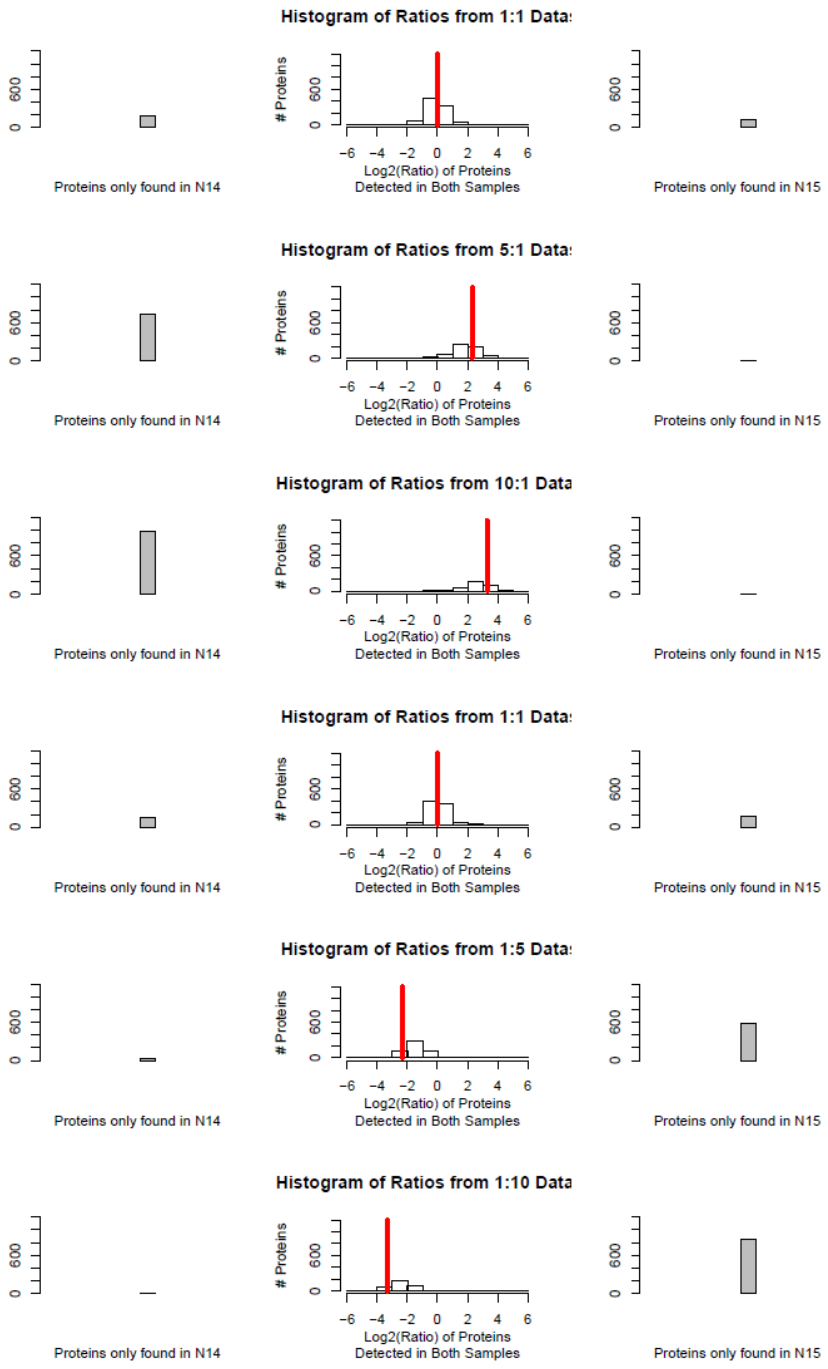
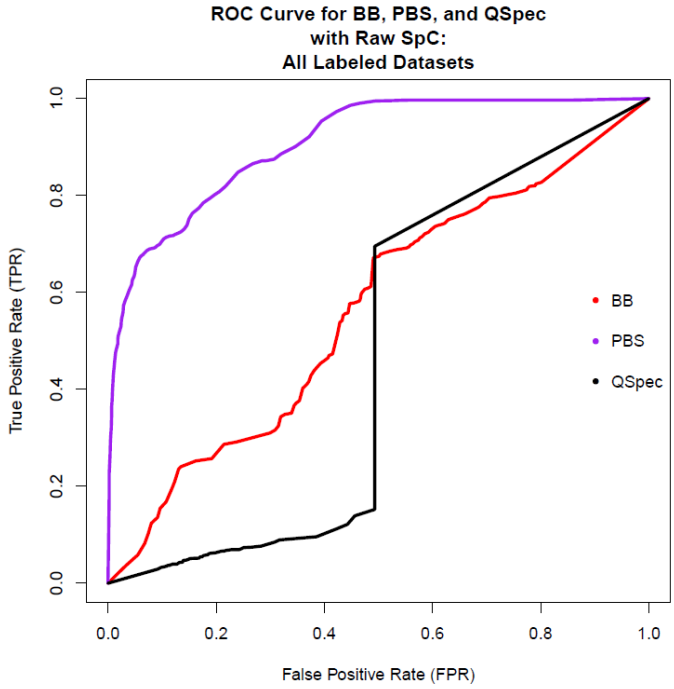



Figure 5.2. Validation of the use of SpC for estimating relative protein abundance.

Standard mixtures of *R. palustris* grown in labeled ($^{15}N$) and unlabeled ($^{14}N$) media were mixed in 1:1,1:5, 5:1, 1:10, and 10:1 ratios before MS analysis. The ratios of SpC between the $^{14}N$ and $^{15}N$ proteins can approximate the known relative protein abundances.

**(A)**




**(B)**

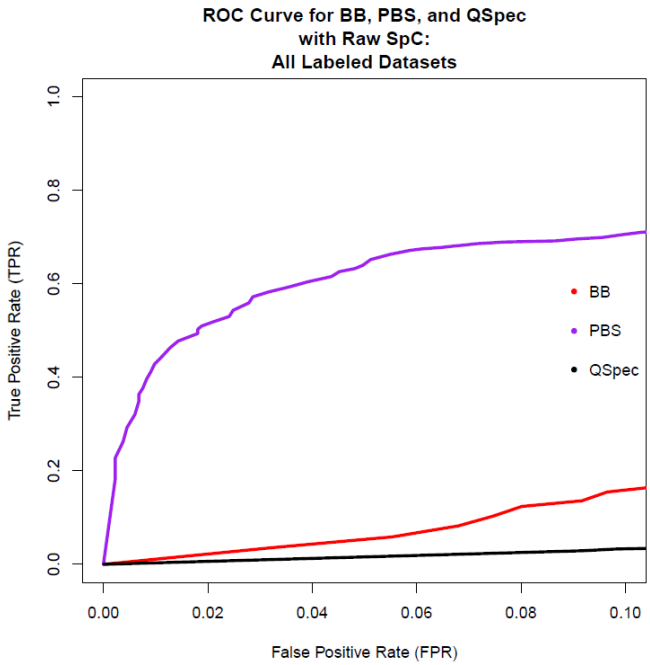



Figure 5.3. ROC Curves for BetaBinomial (BB), Poisson Bootstrapping (PBS), and QSpec tests of differential protein expression between the standard mixture datasets.

(A) Full ROC Curve analysis (up to 100% FPR) for these three methods illustrate quite different powers of significance testing. (B) An inset analysis (up to 10% FPR) of the same results highlights the discrimination of PBS over the other two methods.

An additional feature of the PBS method is that we are able to calculate confidence intervals for the ratios between the two measurements. As a complementary figure to Figure 5.2, we generated the 95% confidence interval for each protein in the standard mixture dataset (Figure 5.5). We expected to see the confidence intervals of bootstrapped ratios coincide with the pre-determined 1:1, 5:1, and 10:1 ratios. As evidenced in this figure, the proteins compared in the 1:1 dataset are in fact centered on 1. Furthermore, the proteins in the 5:1 dataset and 10:1 dataset have fairly tight distributions around 5 and 10, respectively, without elongating the range of bootstrapped values and therefore changing the kurtosis of expected data point distributions.

| | Proteins Significant in Only One Test | | | Proteins Significant in Two Tests | | | Proteins Significant in At Least One Test | | | Proteins Significant in All Tests |
|---|---|---|---|---|---|---|---|---|---|---|
| **Significance Tests** | BB | PBS | Q-Spec | BB & PBS | BB & Q-Spec | PBS & Q-Spec | > BB | > PBS | > Q-Spec | BB, PBS, & Q-Spec |
| **Comparison Types** | | | | | | | | | | |
| high-high | 0.00% | 0.00% | 0.00% | 0.00% | 7.69% | 46.15% | 53.85% | 92.31% | 100.00% | 46.15% |
| high-med | 0.00% | 0.00% | 0.00% | 0.00% | 2.99% | 61.19% | 38.81% | 97.01% | 100.00% | 35.82% |
| med-med | 1.20% | 0.00% | 10.24% | 0.00% | 4.82% | 60.24% | 29.52% | 83.73% | 98.80% | 23.49% |
| med-low | 2.40% | 5.29% | 12.50% | 0.00% | 5.77% | 39.90% | 41.83% | 78.85% | 92.31% | 33.65% |
| med-zero | 6.35% | 28.57% | 1.27% | 43.17% | 0.32% | 0.63% | 69.52% | 92.06% | 21.90% | 19.68% |
| low-zero | 82.22% | 4.44% | 2.22% | 0.00% | 8.89% | 0.00% | 93.33% | 6.67% | 13.33% | 2.22% |

Figure 5.4. Comparison of abundance ratios considered significant by each significance test.

The BetaBinomial (BB), Poisson Bootstrapping (PBS), and Q-Spec tests often disagreed on whether a protein was significantly different or not. BB uniquely identified many low abundance proteins to be significantly different, while the PBS test agreed with the medium-low abundance differences detected by BB and the high-high, high-medium, medium-medium, and medium-low differences detected by QSpec.

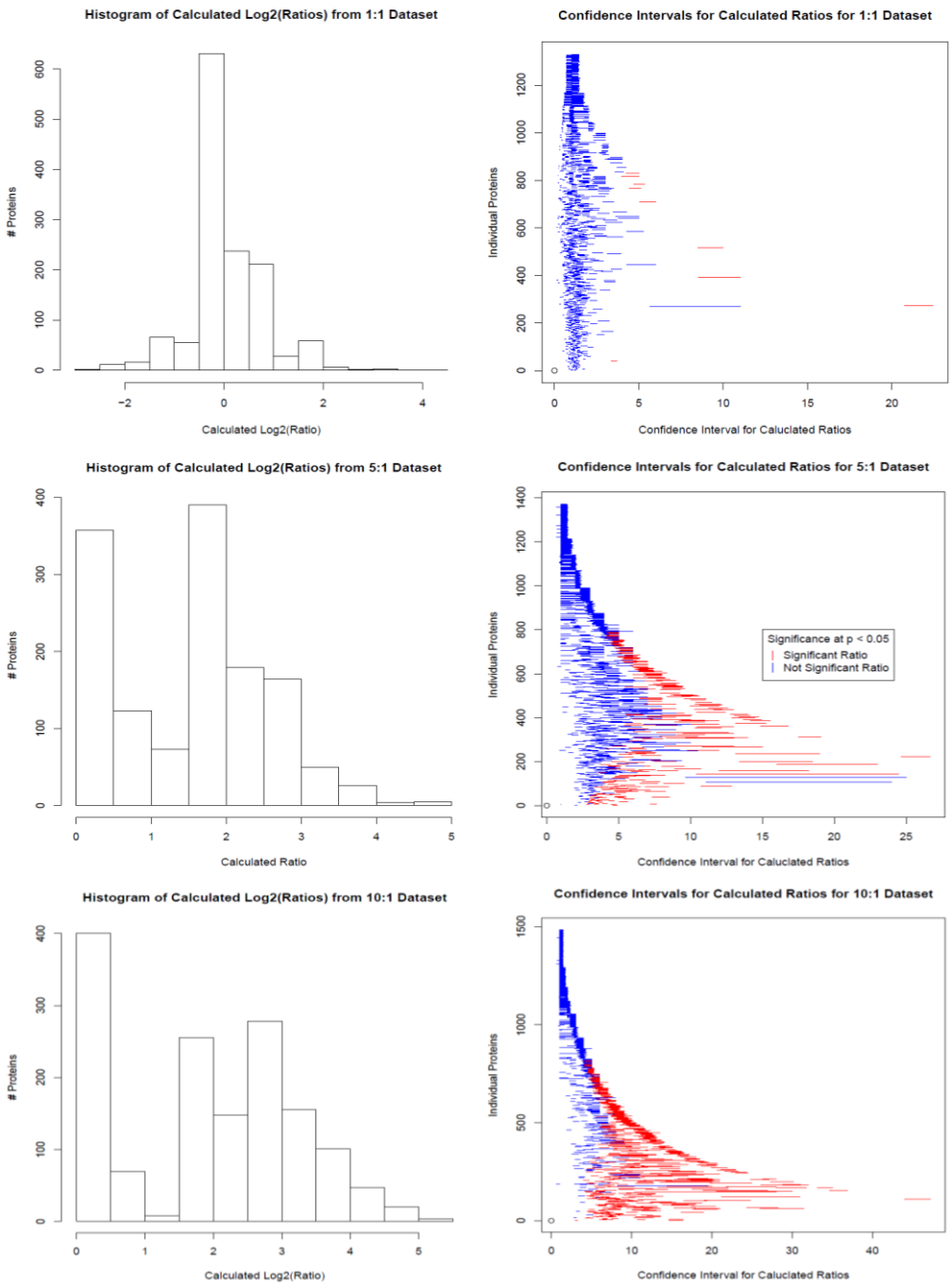



Figure 5.5. Calculated Log2 ratios (right column) and their confidence intervals (left column) using the PBS Method.

Proteins that were determined to be significantly different ($p < 0.05$ are colored in red) and those without enough evidence of difference are in blue. Each row represents a different dataset (1:1, 5:1, and 10:1, respectively).

## 5.2 Protease-Optimized Spectral Indexing for Relative Protein Abundances in Label-free Approaches

### *5.2.1. Using Matched Ion Intensities for POSI*

As described in Chapter 3, matching all fragment ion intensities within an MS/MS scan provides reproducible results for peptides measured across technical replicates. In addition, comparing the sum of matched fragment ion intensities for an entire run can serve as an indicator of the relative difference in two loading amounts. Whereas the previous discussion focused on how the matched ion intensities correspond to the peptide-level measurements, it is equally important to give thoughtful consideration to the method of aggregating PSM-level matched ion intensities into protein-level measurements. Therefore, many of the same metrics that were discussed previously (distribution of measurements, reproducibility, and comparison across loading amounts) will be revisited from a protein perspective.

One of the touted strengths of matched ion intensities is that they capture an extra dimension of information for each spectrum collected. Each fragment m/z value is quantified by the number of electrons measured by the detector as a measure of quantify the analyte's abundance observed at that given time.[140, 143] Therefore, by summing the intensity value of each m/z from each scan for each peptide, one is actually accumulating a distribution of intensity data with thousands of data points for each peptide and millions to billions of data points for a given protein. One would suspect that if the random sampling of mass spectrometry instruments is truly random, and if all of the spectra are representatives of the protein's relative abundance within a run, then the distribution of these data points would most likely follow a normal distribution. Indeed, previous studies have demonstrated that spectral counts tend to follow a log-normal distribution,[138] but it has not been previously asserted whether the additional information provided by matched ion intensities supports or rejects this hypothesis.

Using a single run as an example, we tested the null hypothesis that each protein'sPSM matched ion intensities came from a normally distributed population (after the log 10 of each intensity measurement was taken). Using the Shapiro-Wilk's test, in which smaller p-values indicate less probability of a normal distribution, 89% of the 1190 *C. thermocellum* proteins were not normally-distributed ($p < 0.05$). We suspected that closer examination would demonstrate that the remaining 11% of the proteins would be the most abundant proteins (by SpC and MIT), because their increased measurements would be more representative of the "truer" (more normal) distribution. However, we found that the proteins that initially passed the Shapiro-Wilks test were proteins of "medium" abundance (50-100 SpC). As depicted in Figure 5.6, the test was actually more sensitive to samples that had more data points, so the test asserted that there was sufficient evidence to say matched ion intensities from proteins that had more than 100 SpC were highly unlikely to come from a normal distribution.

While only 25% of the dataset had proteins > 100 SpC, the majority of the dataset was plagued by a different problem. Scans that were assigned to multiple peptide sequences, either due to an isobaric sequence (isoleucine or leucine ambiguity) or lack of evidence for confident charge state identification, were strongly shaping the intensity distributions of the low-abundant proteins. In a second normality test, these ambiguous scans were removed from each protein and their distributions re-evaluated. This time, the number of proteins that were not normally-distributed diminished considerably (30% of the 1190 proteins, $p <0.05$). A third normality test was performed in attempt to also address the problem of high-abundant proteins. For each protein > 120 SpC, we selected 120 measurements from the protein and ran the normality test. This selection of 120 measurements was repeated 1000 times for each protein in order to achieve representative sampling of the protein. Their average p-value after 1000 of these "bootstrapped" samples was considered for their assessment of normality. After this method was applied, 21% of all proteins had sufficient evidence to say they did not follow a normal distribution ($p < 0.05$). When applied to other datasets, averages between 80-90% of the proteins failed to reject the null hypothesis that their measurements came from a log-

normal distribution. These findings were instrumental in the path of considering various methods of normalizing protein abundances, as discussed in 5.2.2. Unique, unambiguous peptide intensities have long since been the metrics of choice for representing protein abundances in labeling methods, such as TMT and iTRAQ, which depend on the intensity of a specific tag for relative or even absolute quantitative comparisons. If similar performance could be achieved without the labeling process, that method may be a more attractive and economical approach for routine MS analyses.

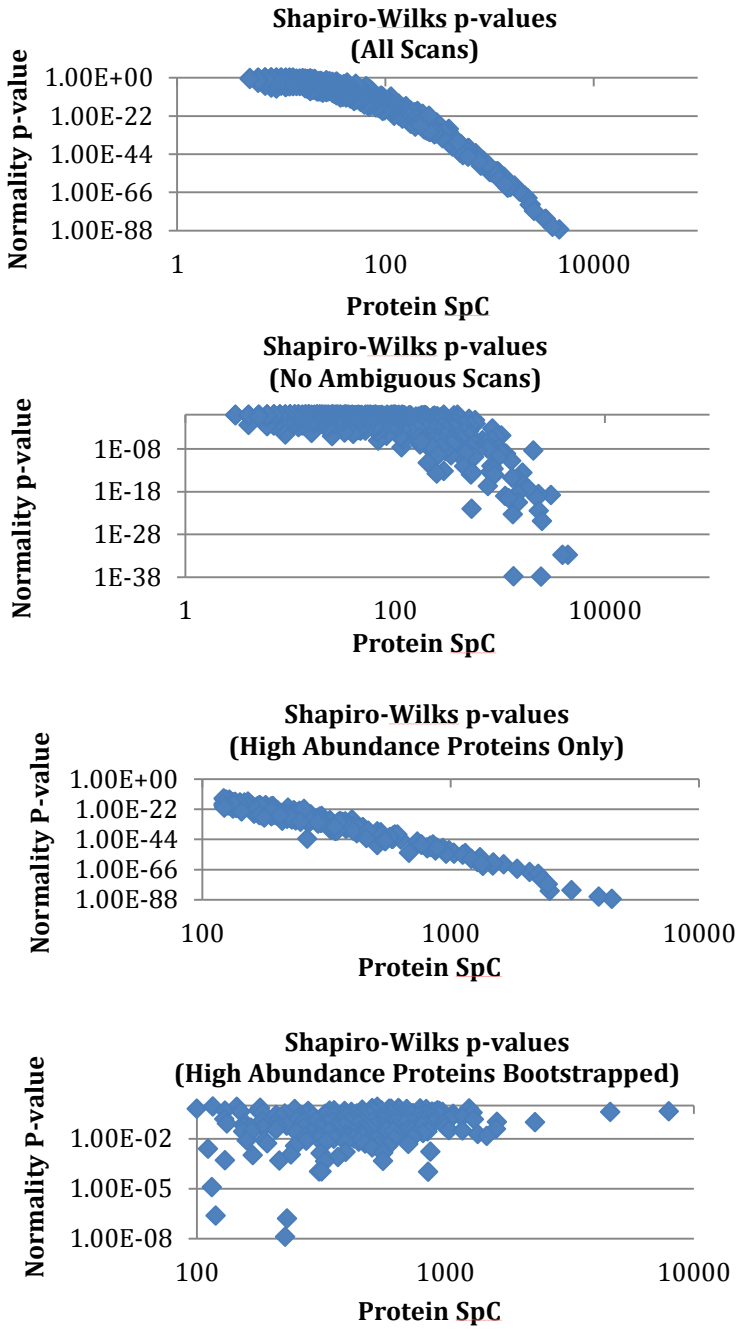



Figure 5.6. Normal distribution of protein-level measurements.

(A) Original normality test of all scans, graphed by SpC and their p-value. (B) Ambiguous scans were removed and the normality tests were rerun. (C) Insert of p-values from first normality test, limited to proteins with >100 SpC. (D) Normality test results for high abundance proteins after bootstrapping method was applied.

Table 5.2. Definitions of protease-optimized protein lengths.

| Effective Protein Length | Description |
|---|---|
| Coverage | Number of amino acids “covered” by a peptide |
| PeptideSum | Sum of (non-redundant) peptide lengths |
| NumPeptide | Number of non-redundant peptides<br>**Note: Valid peptides must be between 5 and 50 amino acids long and have a mass between 400 and 6000 Da* |

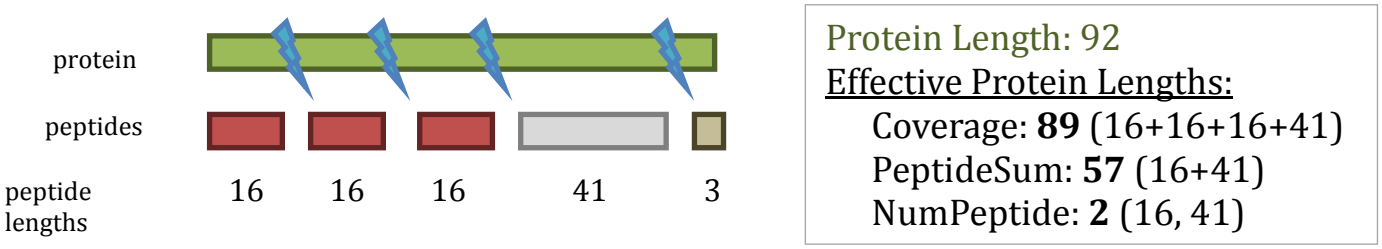



Figure 5.7. Illustration of the different effective lengths suggested by POSI.

### *5.2.2. POSI: Comparing Samples Digested by Different Proteases*

Instead of normalizing proteins by their lengths, we propose that proteins should be normalized by protease-optimized effective protein lengths (Table 5.2, Figure 5.7). Protease-optimized effective lengths take into consideration the number, length, and redundancy of peptides generated by specific proteases on each protein.

The first effective protein length is the number of amino acids covered by a proteotypic peptide ("coverage length"), or the sum of the redundant peptide lengths. This measurement will most likely look very similar to the original protein length, but it could be slightly shorter due to long stretches of hydrophilic amino acids without a cut site. In other words, if a region of a protein does not have a cut site and therefore doesn't contribute to peptide identification, this first effective protein length will not consider those amino acids as part of the protein length. For the purposes of this calculation, no missed cleavages were allowed and a peptide had to fall within 5 to 50 amino acids in length. In the *C. thermocellum* proteome, 12% of the proteins have at least one 40-amino acid stretch that is not MS-compatible under a tryptic digestion and 5% of the proteins are "invisible" to the MS instruments for more than half of their amino acids. Figure 3A shows an example of a protein that has a reduced coverage length when it is digested with each protease. This protein, Clo1313_2479, has 2300 amino acids but under a tryptic digestion, it will only generate peptides that account for 1629 amino acids (70% of the original size). Even more staggering, when this protein is digested with chymotrypsin, it will generate an effective coverage length of 47 amino acids (just 2% of the original protein size). This protein under a digestion with chymotrypsin under a digestion will produce 137 amino acids that will be amenable to MS analysis, accounting for 6% of the original size. Figure 3B illustrates, however, that these differences are not typical for any given protein. Their "coverage" lengths do not significantly affect most proteins, but those that respond generally have extreme reductions in size. Overall, the average protein length reduction by coverage is 93% with trypsin, 98% with chymotrypsin, and 75% with

glu-C. Based on a comparison of the traditional protein length and the calculated coverage length, 266 proteins in the database are expected to be affected by their trypsin differences, 59 proteins could be affected by their chymotrypsin differences, and 961 proteins could be affected by their glu-C length differences (based on those with >25% length differences).

The second effective protein length is the number of non-redundant peptides generated by the specific protease. This measurement is particularly aimed at addressing questions raised by researchers interested in proteins that have large regions of repeated domains within a protein. For example, in a recent dataset analyzing the *C. thermocellum* proteome, many of the peptides in Clo1313_2479 occur multiple times throughout the protein. In fact, one of its peptide has 7 separate locations. Thus, the protein's full length could be said to be "inflated" by internal redundancy. If these regions are MS-compatible and ionize well, these repeated peptides could make the protein look far more abundant, simply because it has a surplus of advantageous peptides. On the other hand, if redundant proteins do not have many cut sites or are not MS-friendly, then the protein will be unduly penalized by its longer length and reduced opportunities for detection. This attempt to resolve whether a protein "putting all of its eggs in one basket" is more likely to be rewarded or penalized for its high degree of internal redundancy may give conflicting results for different proteins, but we are interested in whether there are trends that suggest this is a valid alternative for true protein length. Within a tryptic digestion, 2784 proteins are predicted to have more than 25% of their cut sites lead to redundant or otherwise inadmissible peptides. The calculated number of peptides after a gluC digestion of the *C. thermocellum* proteome suggests that 1814 proteins will have >25% of their cut sites fail to generate a novel peptide. Interestingly, only 59 proteins are predicted to be affected by this effective protein length after a chymotrypsin digestion.

The third effective protein length is the sum of the non-redundant peptide lengths. This measure is a hybrid of the first and second effective protein lengths. Unlike the first effective protein length that summed the length of each instance of a peptide but like the

second effective length that only considered non-redundant peptides, this method of calculating a protein's length reduces the plurality of identical peptides. This method rewards those peptides that have longer sequences and gives advantage to proteins that have more distinct peptides. The more distinct peptides a protein generates, the closer this length will look like the first "coverage" length. Conversely, the more redundant peptides generated by a protein, the more this length will resemble the second effective protein length. For our example protein Clo1313_2479 digested under trypsin, its coverage length was reduced by 70% due to incompatible peptides and 22 of its 90 peptides occurred more than once within the protein, resulting in a final "sum" length of 1256 (55% of the original length and 77% of the coverage length). Similar to the other protease-optimized lengths, this effective length is not predicted to affect the majority of the proteins (226 in trypsin, 60 in chymotrypsin, and 862 in gluC), but the emphasis of this study is not to affect the overall distribution of protein lengths expected in an MS analysis; rather, we anticipate that these minor deviations will be extremely helpful to a few proteins and not harm the remainder of the identifications.

To observe whether varying proteolytic enzymes within the digestion step of a shotgun proteomic experiment affects the relative protein abundances of its complex protein mixture, a sample of *C. thermocellum* was digested by three proteases (trypsin, chymotrypsin, and glu-C) and examined by three replicate measurements for each digestion. While each individual run identified between 1120 and 1388 proteins, 1543 proteins were identified collectively (representing 45% of the proteome). Moreover, 1231 proteins were identified within all 3 replicates of at least one digestion type, suggesting there is a high degree of consistency between each technical replicate. As further validation of the reproducibility within this data set, the Pearson correlation between replicates is greater than 97% and the Pearson correlation among digestion types is greater than 80%. In other words, there is a substantial overlap in the protein identifications between each type of digestion.

Confident in this high correlation of protein identifications between datasets, we then considered the reproducibility of the protein abundances assigned across protease sets. Across all of the runs, the raw totals of assigned spectra varied by nearly 3-fold differences in abundances (from 8,3496 to 213,898 SpC per run), but the assigned intensities spanned multiple orders of magnitude (from 1.5e11 to 5.34e15). After the traditional NSAF approach was applied to apportion protein abundances according to the protein length and total abundances assigned in the run, 37% of the proteins demonstrated significant variation (ANOVA, $p < 0.05$) between their SpC abundances under different protease conditions. Using MIT as the metric for protein abundance, 69% of the proteins demonstrated significant variation between the proteases. Interestingly, when we applied the POSI normalization methods, the number of proteins identified as significantly different between the datasets increased. One of the unforeseen implications of our dataset was that the tight reproducibility of technical replicates allowed each digestion series to be distinguishable. Those proteins that had larger standard deviations in their measurements were those that the ANOVA analysis calculated to be significantly different. In fact, the behavior of proteins across protease sets appeared far more similar when the relative rank of each protein was compared. For NSAF MIT measurements, the Pearson correlations between the average ranks within the different protease datasets were between $R2 = 0.82$ and $R2 = 0.84$ with a slope around 0.9. For the POSI MIT measurements, the correlation values were between 0.6 and 0.74, still suggesting slightly different protein abundances between the enzymatic digestions. From this data, we can surmise that the peptide-spectrum match abundances are differently distributed among proteins depending on the protease used in the experimental protocol. If a study wanted to compare its relative protein abundances to that of another study whose data was collected using a different enzymatic digestion method, the two datasets could not be assumed to be comparable. Rather, the individual ionizability of each peptide in the biological sample and the stochastic, random sampling of MS instruments create substantially different subpopulations of measurements for each protein and each group of proteins. For the time being, relative protein quantitation is most suitable for

comparing protein A to protein A between two samples collected under different biological conditions (but assuming their experimental/analytical protocol is the same).

When the proteins' NSAF ranks were compared to each of the POSI ranks, the majority of the proteins followed the same trend in both types of normalization methods (R2 = 0.98 for coverage and NSAF, $R^2$ = 0.89 for sum and NSAF, $R^2$ = 0.78 for num). There were a few proteins, however, that were dramatically different in their ranks between normalization methods. For example, Clo1313_2540 was ranked 698 out of 856 proteins in the tryptic digestions when normalized with NSAF, whereas it was ranked 305 when normalized with the coverage length. This protein, which has an original size of 749 amino acids, was reduced to a calculated effective coverage length of 99 in a simulated trypsin digestion, causing the protein to appear in the bottom 10% of protein lengths rather than the top 90%. Clearly this magnitude of a discrepancy would affect relative abundance measurements dependent on apportioning values based on expected protein lengths. The ranks between proteins normalized by coverage lengths and (non-redundant) sum lengths were overwhelmingly similar (only 3 of 856 proteins differed by more than 10 positions). Consistent with our previous discussion in anticipating differences among the protease-optimized normalization methods, Clo1313_2479 not only demonstrated cause for different positions in its NSAF and coverage normalized ranks (627 and 572), but it moved to position 519 once it was normalized by the sum effective length. Clo1313_1021, a large protein (7955 amino acids) that was calculated to only generate 48 non-redundant peptides out of the 711 tryptic cut sites, was by far the least consistent between its tryptic NSAF rank and rank of values normalized by the proteins' numbers of peptides was. Its jump from ranks 529 (normalized by NSAF) to 123 (normalized by number of peptides) reflected a large reward for the protease-optimized protein length, supporting the expected implication that it far fewer opportunities to be measured than its original size would suggest. In contrast, when the same ranks are compared for this protein's behavior in the chymotrypsin runs, the ranks are far more similar (only 19 away) because the predicted number of peptides under a chymotryptic digestion (687) aligns more with the large original size of the protein. Interestingly, some proteins that

were expected to be greatly affected by protease-optimized lengths did not show much difference from their NSAF values, especially for the highly abundant proteins. Clo1313_0627, for example, has an original size of 1352 amino acids but because it is a membrane protein with large, repetitive hydrophilic regions, its expected number of contributing peptides is 56 (approximately 40% fewer than other proteins of the same size). However, it was consistently identified as one of the top 3 most abundant proteins, regardless of normalization method. In total, the results generated from the comparisons of protease datasets highlight the impact of enzymes on a sample, creating truly different populations of peptides that can be detected and measured for protein identification and relative quantification. The protease-optimized normalization methods do not standardize the relative abundance measurements across proteolytic digestions, but rather more sensitively apportion measurements within the dynamic context of their analytical backgrounds.

### *5.2.3. POSI: Comparing Samples Loaded in Different Concentration Amounts*

While matched ion intensities provide more accurate information than spectral counts and the normalization methods adjust the measurements to account for biases in instrumental detection and analytical viability, protease-optimized spectral indexes are still a relative quantitative measure- not an absolute one. In other words, the normalization methods help compare the measured expression of one protein in a sample against the measured expression of another protein in the same sample relative to all proteins identified in the sample, but comparing the absolute abundance of one protein with respect to another is still beyond the scope of label-free measurements. However, given the improvements in accuracy and robustness of multidimensional measures like intensity information, one can consistently observe fold-changes between groups of proteins measured within the same sample as well as fold-changes between proteins identified in a sample loaded onto the LC column in two different amounts.

To determine the level of sensitivity of the protease-optimized spectral indexing method, a standard mixture of 48 proteins was analyzed following a tryptic digestion amidst a proteomic background and their intensities were compared against the expected 6 orders of magnitude differences. 15 of the 16 proteins within the top two tiers of the standard mixture ("tier 1": 8 proteins at initial concentrations of 50000 fmoles and "tier 2": 7 proteins at 5000 fmoles) were consistently identified across all 3 replicates of two sample sets (25 μg and 67 μg loads) and were therefore the focus of this specific investigation. For both sample loads, pairwise comparisons of the abundances in tier 1 proteins versus tier 2 proteins yielded fairly disparate results, regardless of whether the measurements were calculated (spectral counts or intensities) or normalized (raw, nsaf, coverage, sum, or num). Although the average intensity ratio of tier1 to tier 2 abundances was around 10 (sd 15) for the 25 μg load, the average was closer to 15 (stdev 15) for the 67 μg load. Similarly, the average spectral count ratio of tier1 to tier2 was 7.3 (stdev 7) in the 25 μg load and a little higher (8.75 avg and stdev 12.31) in the 67 μg load. In an effort to remove individual protein-dependent biases in the measurements, we then considered how the tiers behaved as a whole. When the ratio of the sums of intensities between tier 1 and tier 2 were calculated, the averages were between 7.65 and 13.82 with a standard deviation of 2.13 for the 25 μg load. The improvement by grouping proteins was less noticeable for the 67 μg load, moving the average to 10.35-17.34 (standard deviation of 2.56). Interestingly, the raw summed intensities had the smallest deviations and the closest ratio to 10 when compared to the normalized summed intensities in both samples (Figure 5). Visual inspection of the standard proteins separated by tiers highlights how one protein in particular (gold) does not seem to follow the trend of the other proteins in its group. UPS_P0131 had an initial concentration that places it in the tier 1 group, but it behaved more similarly to the tier 2 group in every run. This protein is extremely small (74 amino acids) compared to the other tier 1 proteins, which average 200 amino acids in length, and is the least affected by the protease-optimized lengths. The differences observed here support the claim that peptide ionizability strongly affects our ability to collect absolute quantitative measurements at the protein level and that comparisons between protein A and protein B within the same run is not very reliable.

On the other hand, comparison of protein abundances between the two runs that only differ by their loading amounts demonstrate extreme consistency in their run-level, protein-level, and even peptide-level fold-changes. The total raw matched ion intensities collected for the 3 technical replicates in which 67 μg was loaded summed to an average of 6.87e10, whereas the 3 technical replicates in which 25 μg was loaded summed to an average of 1.49e11. Ratios of the raw summed matched ion intensities were between 2.01 and 2.29, very similar to the concentration ratio (67/25 = 2.68). In fact, measurements of the flow-through (the material that was loaded but not analyzed by MS) revealed that an average of 80% of the 67 μg material was presented to the instrument and an average of 90% of the 25 μg material. Therefore, a closer value for the loaded concentration ratio is 2.38, which is even more similar to the observed intensity ratios. Pairwise comparison of each protein identified in all of the 67 μg and 25 μg runs (850 proteins) revealed that the average ratio of a protein's intensity between the 67 μg and 25 μg runs was 2.58 (with a standard deviation of 1.58). Among the standard mixture proteins in particular, their average intensity ratio was 2.42 with a standard deviation of 1. Even more interesting, one of the most abundant peptides (TLTVELGVSSLNEGTYK, +2) from one of the most abundant proteins, Clo1313_3011, had a SpC of 729 in the 67 μg rep 1 and 643 SpC in 25 μg rep 1 (a ratio of 1.13), but its matched ion intensities demonstrated the same overall fold-change as the ratio between the runs (6.32e8 from 67 μg and 2.45e8 from 25 μg resulting in a ratio of 2.58). These results support the use of matched ion intensities instead of spectral counts in order to detect relative abundance differences between runs overall, the same protein between two runs, and even abundant peptides detected in two runs.

## 5.1. Using Reporter Ion Intensities for Relative Protein Abundances in Labeled Measurements

Quantitative proteomics measures abundance changes of many proteins among multiple samples in a high-throughput manner.[144] Results from such measurements provide

information on how biological systems respond to environmental perturbations at a genomic scale. A number of methods have been developed for quantitative proteomics to obtain high proteome coverage, accurate quantification, and wide applicability to different types of samples.[145] In proteomics analysis based on 2-dimensional gel electrophoresis (2D-GE),[37] quantification is achieved by measuring staining intensities of protein spots. To eliminate gel-to-gel variability, proteomes under comparison can be labeled separately using different fluorescent cyanine dyes (Cy2, Cy3, and Cy5) and then combined for 2D-GE analysis.[146] However, both identification and quantification are difficult for gel spots containing multiple comigrating proteins.[147] Only one of those co-migrating proteins may be identified in such a gel spot, and that protein may not be the one responsible for the differential expression. In addition, the capability of 2D-GE proteomics is also limited by the number of quantifiable proteins in a gel, a bias against membrane proteins, and a low sample throughput.[144]

In the shotgun proteomics approach, proteins are typically digested using proteases into peptides, which are then analyzed using liquid chromatography coupled with tandem mass spectrometry (LC−MS/MS).[48] Without using any isotopic or chemical modification of proteins or peptides, label-free quantification can be achieved by correlating protein abundance with either mass spectrometric signal intensities of peptides[148] or the number of MS/MS spectra matched to peptides and proteins (spectral counting).[138] Label-free quantification is widely used because it allows simultaneous identification and quantification of proteins without a laborious and costly process of introducing stable isotopes into samples, and this approach is applicable to samples from any source. However, because samples to be quantified are prepared and measured separately, label-free approaches have limited quantification performance in terms of accuracy, precision, and reproducibility. To improve quantification performance, many approaches were developed on the basis of stable isotope labeling, including metabolic labeling,[149] enzymatic labeling,[150] and chemical labeling.[69] In metabolic labeling, stable heavy isotopes are incorporated into proteins by growing cells in controlled media containing an 15N-enriched nitrogen source[151] (15N labeling) or isotopically labeled essential amino

acids (stable isotope labeling by amino acids in cell culture or SILAC[70]). Metabolic labeling allows samples grown in different states to be combined at the cell level. Therefore, any bias in the downstream sample preparation and measurement would alter protein abundances from different samples to the same extent, making their ratios relatively unchanged. However, many biological systems are not amenable to efficient metabolic labeling, such as natural microbial communities.[152] To overcome this, chemical or enzymatic methods have been developed to label proteins or peptides using different isotopic tags. For example, after cell lysis, extracted proteins can be labeled using isotope-coded affinity tags (ICAT).[69] After protein digestion, peptides can be labeled enzymatically at the C-terminus using $H_2{}^{18}O$.[151] Peptides can also be labeled on the primary amine group at the N-terminus and lysine side chain using reductive dimethylation (ReDi).[153] In proteomics measurements based on these stable-isotope labeling strategies, the abundance ratios of mass-different isotopic variants of peptides are determined using their signal intensities in full parent ion scans of the LC− MS/MS analysis. Abundance ratios of peptides are then used to infer abundance ratios of their parent proteins.

Recently, two similar isobaric chemical labeling methods, isobaric tag for relative and absolute quantification (iTRAQ)[68] and tandem mass tag (TMT),[67] have become increasingly popular for quantitative proteomics. After proteolysis, samples are labeled separately with different isotopic variants of iTRAQ or TMT and are then combined for LC−MS/MS analysis. Both iTRAQ and TMT tags contain three functional parts: a reporter ion group, a mass normalization group, and an amine-reactive group. The amine-reactive group specifically reacts with N-terminal amine groups and epsilon-amine groups of lysine residues to attach the tags to peptides. The mass normalization groups balance the mass difference among the reporter ion groups such that different isotopic variants of the tag have the same mass. Peptides labeled with different variants of the tag are indistinguishable in full scans, which prevents increasing the full-scan complexity after mixing multiple samples. In MS/MS scans, reporter ions of different masses are dissociated from isolated peptide species. The mass of a reporter ion is associated with a

specific variant of the tag,[154] and the relative intensity of the reporter ions measures the relative abundance of the peptide labeled with that specific tag variant. 6-Plex TMT[155] and 8-plex iTRAQ[156] allow comparing up to 6 and 8 samples in a single LC−MS/MS analysis, respectively. Multiplexing is a unique capability of iTRAQ and TMT in comparison to the other labeling techniques.

Each of the described methods has its advantages and disadvantages for quantitative proteomics. A comparison of SILAC and spectral counting showed that spectral counting provided less precise quantification to proteins with low spectral counts.[157] A comparison of 14N/15N metabolic labeling with spectral counting showed that spectral counting was less sensitive to detecting small fold changes.[158] iTRAQ was also compared to a label-free quantification method based on normalized chromatographic peak intensity.[159] While the number of identified proteins and reproducibility were comparable between these two methods, proteome coverage was significantly higher in the label-free method. To date, no study has systematically compared label-free, metabolic labeling, and isobaric chemical labeling with iTRAQ or TMT using the same analytical platform. In this study, performances of spectral counting, 14N/15N metabolic labeling, iTRAQ, and TMT were benchmarked using standard proteome samples prepared from a model microorganism, *Pseudomonas putida* F1.[160] *P. putida* F1 is a gram-negative soil microbe, known for its diverse metabolism and ability to degrade aromatic hydrocarbons. Its unique bioremediation potential is frequently exploited for remedying contaminated soils. Measurements for all four methods were performed using the LTQ Orbitrap Velos.[161] The higher-energy collisional dissociation (HCD) capability and the improved ion extraction efficiency of LTQ Orbitrap Velos enabled excellent measurement of iTRAQ- or TMT-labeled samples.

For iTRAQ and TMT analysis, every full scan was followed by four CID-HCD dual MS2 scans, in which a selected parent ion was first fragmented by CID for peptide identification and then by HCD for quantification. HCD offers higher fragmentation efficiency and lower minimum m/z detection limit than CID, which enables measurement

of reporter ions in Orbitrap analyzer with high signal-to-noise ratio. However, because of the extra time needed for HCD analysis, the duty cycle of MS2 acquisition was significantly lower in the CID-HCD dual-scan configuration than the CID-only configuration used for the other analyses. Furthermore, previous studies have shown that the presence of fragment ions as a result of losing isobaric tags from precursor ions complicates the interpretation of spectra by database searching algorithms.[162] Therefore, fewer peptides and fewer proteins were identified in isobaric chemical labeling than in label-free and metabolic labeling. Similar protein identification results were observed between iTRAQ and TMT. 1473 unique proteins were detected from the iTRAQ-labeled sample (FDR = 2%) and 1404 in the TMT-labeled sample (FDR = 3%). 73% of proteins were identified reproducibly between duplicate runs in iTRAQ and 76% in TMT.

Because HCD spectra can be used for both peptide identification and quantification, TMT and iTRAQ samples can be analyzed using only HCD.[163] We found that less than 30% of identified spectra were from HCD fragmentation. Less than 10% of those identified HCD spectra have a paired CID spectrum that did not identify a peptide, whereas approximately 60% of identified CID spectra have a paired HCD spectrum that did not identify a peptide. This indicates the value of CID for peptide identification. The duty cycle of the CID-HCD configuration was not significantly lower than the HCD-only configuration because the acquisition time for CID coupled with ion-trap detection is only a fraction of the acquisition time for HCD coupled with Orbitrap detection in the dual scan.

Isobaric mass tags were chemically linked to N-terminus amine groups and the epsilon-amine group of lysine. In one database search, derivatization of the N-terminus was set as a static modification and dynamic modification was set at lysine residue. >98% of lysine residues in the identified peptides were labeled, indicating high labeling efficiency of lysine in sample preparation. A separate search for peptides with an unmodified N-terminus using dynamic modification at lysine identified only a few hundred peptides

with a greater than 50% FDR, which suggests a high labeling efficiency of the N-terminus by iTRAQ and TMT.

Ross et al. observed that the ratio of Lys-terminated peptides to Arg-terminated peptides (Lys/Arg peptide ratio) increased from 0.79 in an unlabeled sample to 0.98 in an iTRAQ labeled sample.[68] However, in this study, the Lys/Arg peptide ratios from TMT and iTRAQ were not significantly higher than those from label-free or metabolic labeling. An expected Lys/Arg peptide ratio of 0.50 (170,662 Lys-ending peptides and 342,497 Arg-ending peptides.) was calculated based on *in silico* digestion34 of the P. putida F1 proteome. The observed Lys/Arg peptide ratios in all runs were higher than the expected ratio.

All MS/MS spectra were searched using SEQUEST[63] against the *P. putida* F1 genome database containing in FASTA format a total of 5251 predicted proteins and 44 common contaminants (trypsin, keratin, etc.). The reversed sequences of all proteins were appended into the database for calculation of false discovery rate (FDR).[164] The SEQUEST searches for label-free samples and 14N/15N-labeled samples were performed as described previously.[160] Two SEQUEST searches were performed for each iTRAQ and TMT run. The first search used static modification at the N-terminus and dynamic modification at the lysine residue by the labeling reagents. The second search used only dynamic modification at the lysine residue. The output data files were then filtered and sorted using the DTASelect v1.2[71] algorithm as described previously.[160] Perl scripts were developed to process iTRAQ and TMT data sets for protein quantification.

In the CID/HCD dual scan configuration, peptide identification can be obtained from the CID scan, the linked HCD scan, or both. Reporter ions for all peptide identifications were extracted from small windows (±0.02 Da) around their expected m/z in the HCD scan. If multiple peaks were found within the accepted m/z window of a reporter ion, the one with the highest intensity was considered to represent the reporter ion. The total intensity at a reporter ion channel for a protein was calculated as the sum of this reporter ion's

intensities from all constituent unique peptides from this protein.[165] The abundance ratio of a protein was estimated using the ratio between the protein's total intensities in different reporter ion channels.

More specifically, all LC−MS/MS data sets from iTRAQ and TMT experiments were converted from the Xcalibur Raw file format to the MS2 flat file format using the Raxport[166] program freely available at http://code.google.com/p/raxport/. This tab-delimited data format contains 3 types of information lines: 1) the header lines contain scan id, its precursor scan, the precursor mass, 2) the peak header lines contain the parent scan number, and 3) the peak information lines contain the fragment m/z value and its intensity. For the dual HCD/CID runs, there is not any information that explicitly identifies which scans are "siblings," except by their reference to the same parent scan and mass. Due to the random nature of the scanning process, one could not quickly index all MS2 scans based on their scanned order. An additional complication was that the parent scan reported was not the survey scan id; it was an enumerated precursor scan id that restarted from 1 with every survey scan. Similarly, the parent mass reported for each fragment scan was in fact the m/z value acquired during the survey scan, and not the exact m/z selected for the subsequent 4 fragment scans. In other words, each sibling MS/MS scans was reported with a slightly different parent m/z value. Therefore, a 1 ppm window was used to find sibling HCD and CID fragment scans. A window of +/- 0.02 Da from the TMT or iTRAQ reporter ion mass was searched within each precursor scan's fragment peaks and the reporter ion with the highest intensity was selected as the representative for that precursor mass and fragmentation method. These intensities and their mass errors were then added to the DTASelect file so that peptide sequencing and protein assembly information could be more readily evaluated. Each peptide in the DTASelect file was assigned a reporter ion intensity if the tag was identified (in both CID and HCD fragmentation for the dual scan configuration), but only the HCD fragmentation was used for quantitative purposes. Since SEQUEST searches were performed both with and without consideration of a dynamic lysine terminus, there was an expectation that there would be valuable unique and overlapping information in the

comparison of the identifications from both searches. Therefore, all of the peptide-spectrum matches from both searches were merged into one file and of the scans that were assigned to sequences in both files, only the PSM with the best xcorr score was retained. Then, for each reporter tag, peptide intensities were summed into protein intensities only if the peptide sequence was unique to that protein. The abundance ratio of a protein was estimated using the ratio between the protein's total intensities in different reporter ion channels.

A total of 1980 unique proteins were identified using the label-free method (on average approximately 1600 non-redundant proteins from a run, FDR = 2%). 79% of all identified proteins in the duplicate runs of a sample were identified reproducibly in both duplicate runs. A total of 1606 unique proteins were identified using the metabolic labeling method with 77% identification reproducibility between duplicate runs (FDR = 3%). 1473 unique proteins were detected from the iTRAQ-labeled sample (FDR = 2%) and 1404 in the TMT-labeled sample (FDR = 3%). 73% of proteins were identified reproducibly between duplicate runs in iTRAQ and 76% in TMT. This shows that the label-free method had the highest number of protein identifications and provided the deepest coverage of the genome (~30%). Identification reproducibility between duplicates was similar among all four methods.

In label-free quantification, each sample of interest must be prepared and analyzed by LC−MS/MS separately. The semi-random sampling nature of the peptide identification process in a shotgun proteomics run also contributes to the variability of spectral counting for protein quantification. Therefore, relatively poor quantification results were observed with the spectral counting method. Several alternative MS/MS acquisition methods have been developed, which could overcome this limitation. Venable et al. introduced a data independent acquisition method based on sequential isolation and fragmentation of a series of predetermined precursor windows.39Carvalho et al. extended this method and developed an algorithm to identify multiplexed spectra acquired with CID and electron transfer dissociation.[167] In the MSE approach, a quadrupole time-of-

flight mass spectrometer was used to fragment all precursor ions in an elevated-energy mode.[168] These data-independent methods will probably increase the reproducibility of label-free quantification. Alternative data analysis methods have also been developed to improve labelfree quantification. For example, chromatographic peak areas of peptides, instead of spectral counts, can be used as the measure of protein abundance for quantification.[169] The normalized spectral index (SIN) method estimates protein abundance by combining spectral counts and total ion intensity of MS/MS spectra.[143] In contrast to label-free quantification in terms of sample preparation, metabolic labeling allows the mixing of samples at the very beginning of preparation. Samples representing two states are prepared and measured together, which minimizes potential bias in these processes. The relative abundance ratio of a protein between samples is maintained. Thus, accurate and reproducible quantification results can be obtained from metabolic labeling. In iTRAQ and TMT analysis, samples from different conditions are processed separately until peptides are generated and labeled with different tags. After that, these samples are pooled for subsequent LC−MS/MS measurement. HCD provides efficient ion extraction and fragmentation for generation of reporter ions, allowing detection of reporter ions with high signal-to-noise ratio in Orbitrap analyzer. In comparison to metabolic labeling, MS detection of reporter ions in an Orbitrap MS2 scan may be better for quantifying a peptide than detection of precursor ions in a series of Orbitrap MS1 scans. Thus, although TMT and iTRAQ require samples to be mixed at a later sample preparation stage than metabolic labeling, they produced better overall quantification results. The comparison results provided guidance for choosing an appropriate approach for a proteomics experiment. The label-free method has the largest dynamic range for protein identification; however, high spectral counts are required for reliable quantification. In addition, special care is necessary to minimize sample-to-sample variability during sample preparation and measurement. Both metabolic labeling and isobaric chemical labeling provide accurate, precise, and reproducible quantification for many proteins, but each has advantages and disadvantages. Metabolic labeling is ideal for samples that need to undergo extensive preparation steps at the protein level, such as fractionation and enrichment, which may introduce a significant amount of error without

pooling samples together. However, metabolic labeling is feasible only for selected microorganisms and cell cultures. The unique advantage of TRAQ and TMT is the capability to multiplex more than two samples in a measurement. This not only saves instrument time but also simplifies experimental design. However, iTRAQ and TMT require advanced MS instruments, such as Q-TOF and LTQ Orbitrap Velos.

In this study, four quantitative proteomic approaches, label-free, metabolic labeling, and isobaric chemical labeling by iTRAQ or TMT, were compared using an LTQ Orbitrap Velos mass spectrometer for protein identification and quantification. Our results indicate that the label-free method provides the deepest proteome coverage. However, the quantification is not as efficient as in the labeling-based approaches, especially for low-abundance proteins. Metabolic labeling and isobaric chemical labeling have improved quantification accuracy, precision, and reproducibility. iTRAQ and TMT have similar performance in all aspects.

## 5.3 Conclusions

This project explored the use of the most common metrics for relative protein abundance: spectral counts (SpCs) and matched ion intensities (MITs). We looked at the behaviors and distributions of both measures, concluding that SpCs most closely followed a Poisson-like distribution and that MITs followed a normal distribution at the protein level. These studies also looked at the most appropriate ways to normalize these measures within and between MS runs. Surprisingly, both metrics performed rather well in discriminating known differences in protein abundances without normalization. However, normalizations were applied to adjust for run-to-run variation in total SpC and MITs assigned. For MITs, we also considered slight deviations from the traditional NSAF normalization, which uses the protein's length as a second dimension for normalization. POSI (Protease-Optimized Spectral Indexing) assumes that each protein will generate different peptides, depending on the enzymatic digestion used, thereby changing the number of opportunities for each protein to be detected and analyzed by MS. Since

different peptides ionize differently and the observed intensity measurements are largely dependent on the competition between co-eluting peptides, it is more accurate to use the length of the protein as it is presented to the instrument after digestion. The suggested POSI normalizations confirmed that each protein appeared to have a different ranked relative abundance in the same sample digested by three different proteases. Comparison of the spectral indexes (sums of the peptide MITs) were also able to identify fold-changes in loading amounts when the same sample was loaded using 25 μg and 67 μg. The normal distribution of matched ion intensities supports the use of ANOVA for determining differences between relative abundance measurements. For SpC, however, we developed a new method for assessing whether two proteins are differentially expressed using the Poisson Bootstrapping Method (PBS). The PBS method seemed to give superior performance and lack of bias compared to other existing methods of testing significance.

As an alternative to label-free matched ion intensities, labeled ion intensities use a reporter ion (one with a very specific, known mass) to act as the representative peak for that peptide's intensity. This study found that label-free methods afford better representation of protein identifications but labeling methods achieve more accurate, reproducible quantification measurements.

# CHAPTER 6: Integrating Novel and Existing Tools into a Seamless Bioinformatic Workflow for Analyzing Shotgun Proteomic Datasets

## 6.1 Logistics of Developing a Bioinformatics Workflow

Shotgun proteomic experiments provide qualitative and quantitative analytical information from biological samples ranging in complexity from simple bacterial isolates to higher eukaryotes such as plants and humans and even to communities of microbial organisms. Improvements to instrument performance, sample preparation, and informatic tools are increasing the scope and volume of data that can be analyzed by mass spectrometry (MS). To accommodate for these advances, it is becoming increasingly essential to choose and/or create tools that can not only scale well but also those that make more informed decisions using additional features within the data. Incorporating novel and existing tools into a scalable, modular workflow not only provides more accurate, contextualized perspectives of processed data, but it also generates detailed, standardized outputs that can be used for future studies dedicated to mining general analytical or biological trends.

Tightly coupled to the advancements in sample preparation and instrument technology, there is an increasing demand for software improvements to make sense of and report the collected data in a meaningful way. Each new parameter that can be adjusted in the experimental protocol or tuning of the instrument adds to the opportunity for an optimized combination of settings that provide the best-case scenario for deep, accurate measurements. Therefore, data is collected on instrument statistics (voltage, % salt, DE window, etc) as well as spectral-level statistics (elution time, precursor ion selection, spectral counts, measured MS1 peak intensity, calculated peak area, etc). Once the data has been collected, interpreting the data requires algorithms to perform peptide to spectra matching (PSM scores/likelihoods) and protein to peptide matching (FDRs). The existing algorithms that provide scores and suggest assignments have a mixture of competing and

complementary benefits and disadvantages, so it is conceivable one may want to compare the results of multiple software algorithms in order to come up with the most comprehensive understanding of the components collected in the biological sample. Several software programs use index-based information retrieval so that one isn't always moving every piece of data with each analysis. For example, protein assembly software generally does not maintain information about individual ion series distributions for each PSM; the data is usually linked in some way so that the user can explore to his or her desired level of detail, or the data's represented by an aggregate measure. However, with a centralized repository of raw input data as well as processed results, it is a much more straightforward task to provide means of easily extracting cross-referenced information, transforming or filtering it, and sharing it with others. Some of the most beneficial steps taken by other research groups along the way include standardizing their input and output formats for informatics software. Making data results portable not only increases the speed and efficiency at which a new tool can be evaluated and adopted into an existing workflow, but it helps standardize vocabularies, establish quality control, and move the community closer to diagnostic and deterministic assessments of datasets' behaviors. Therefore, in our implementation of a bioinformatic workflow, TORPEDO (Tools and Omnibus of Resources for Proteomic Experimental Datasets Online), we have endeavored to receive and generate the common standardized outputs.

With so many tasks to accomplish, and for occupation by multiple users at a time, such a workflow requires support by adequate hardware infrastructure. Successfully integrating this workflow within the existing computing architectures was not possible. The distinct computational resources currently available require multiple data transfers between users, adaptation of analysis scripts to accommodate different operating systems, numerous transformations of the data into various input and output file formats, and non-linear documentation of analyses performed on the data. Therefore, we proposed developing cyber-infrastructure that would allow a user to seamlessly run multiple analyses, store the results, and share processed data with other users.

Specifically, we built a web-based front-end to facilitate data exploration. Users have to sign in with an account to run analyses, but they can choose to make their results available to the public or private. In addition, users can opt to upload data for one-time analyses, or users can create a persistent project with longer-term data storage and invite other users to view their results. In short, this project offers an easy-to-use interface for running multiple proteomics analyses tools backed by sizable computing resources and a platform for sharing data with other researchers for enriched collaborations.

The alternative hardware solutions we explored involved complicated communication between two existing resources: a host computer and a compute cluster. For this setup to work, the host computer handled user interactions and constantly pinged the remote compute cluster for notifications of completed jobs and retrieving the results. Checks had to be made both at the user- and processing-end to ensure all parameters were in place, even though the software only resided on the compute cluster. Security requirements also provided a significant hurdle to protect the computing cluster from attacks originating through our website.

Simplifying this architecture into one machine minimized redundant validation steps, mitigated communication errors, allowed real-time job status updates, and simplified the overall design concept to create truly modular program development. By building a computer system that can handle both responsibilities of hosting and computing processes, computing tasks can later be distributed onto other machines as infrastructure expands. Thus, in the future the single computer will not become obsolete, but instead can easily be repurposed as a login node facilitating load balancing and job execution on compute nodes.

This proposed solution is not just for large labs on the cutting-edge of large-scale data-centric experiments; it also provides a gateway for small labs doing one-off experiments that do not necessitate a dedicated informatics solution. This solution fits well with the NSF-recognized need for national cyber-infrastructure for research and provides a

starting framework for future projects to expand capabilities. Specifically, we will leverage this project to apply for the Annual Research Cluster Grant from Silicon Mechanics. By providing centralized pre-built options for analysis, we engage an audience that otherwise may not participate in data-centric biological experiments and provide a functional education for best-practices in experimental design and data analyses.

To ensure the usability, performance, and integrity of the data in these analyses, it is critical to have efficient ways to store, access, and interpret information. These needs translate into tangible computational specifications. For example, current state-of-the-art mass spectrometry instruments are generating twice as many spectra as their predecessors, which means algorithms that are optimized for multi-threading and MPI communication are becoming increasingly essential to efficiently deconvoluting spectra into protein identifications. In addition, filtering true protein identifications is far more effective when a user can dynamically score matches, but re-evaluating the large amount of multidimensional data is a highly memory-intensive, user-interactive process. Once the data is properly filtered, it is common to normalize datasets against technical and biological replicates and compare the results between biological conditions or experimental methods. Since each experiment can easily scale to 10-20 GB, having adequate data storage is especially important to obtaining proper perspectives on the analytical quality and biological significance of these proteomics experiments.

While it is important for informatic tools to be able to handle large datasets, it is becoming increasingly crucial for tools to also handle the biological complexity associated with more intricate experimental designs. The overwhelming volume and complexity of these experiments requires that the new and existing tools are not only optimized for speed and interpretation, but they also necessitate seamless communication with each other in an integrated workflow. By constructing a workflow that allows high-throughput processing of massive datasets, data collected within the past decade can be

standardized and updated with the most recent analyses. Once these analyses are complete, meta-analyses can identify global analytical and biological trends.

## 6.2 Improving an Existing Workflow

When attempting to overhaul an existing workflow, it is important to consider what the major functionality of the workflow needs to be, as well as who is going to be using it. For researchers within our lab, we identified 3 major tasks that this architecture would provide: 1) search raw data against a protein database, 2) apply normalization methods in preparation for differential expression tests, and 3) store and access raw and processed data. With these three chief aims in mind, we set out to evaluate existing software and decide what would be most appropriate for the scale and quality of the proteomic analyses currently underway as well as anticipated computational bottlenecks of future projects.

One of the most immediate informatic needs that aligned with the objectives of the new workflow was the adoption of a new search algorithm. Whereas local projects were analyzed by SEQUEST for the last decade, the version that was most commonly used was out of date and no longer sufficiently fast or reliable for large-scale experiments that include highly complex organisms and metaproteomes of bacterial communities. Myrimatch, a database searching platform released in 2007, not only has multi-threaded MPI compatibility for optimized speed performance but also highly customizable search parameters that allow the researcher to specifically define what he is looking for. Another appealing feature about Myrimatch is that it can run on a personal desktop or it can be run on a more powerful Linux computer for batch submissions and high-throughput analyses.

Myrimatch takes standardized input files that merely need to be converted from the direct output from the instrument (.RAW files) and run through a freely-available conversion software, MSConvert, into .mzML or .mzXML files. These XML files compress well and

can be used for other searching software, should one want to compare their identifications. The structure of these files accommodates quite a bit of data and lends itself to quick parsing by PERL or Python scripts. Pure data collection information can easily be retrieved and graphically displayed, such as the total ion current (TIC) compared to the number of peaks collected in each MS2 scan (Figure 3.3). This ability, while a feature, reinforced the decision to adopt Myrimatch as the default search engine.

While it is becoming more common to move away from the reference genome and look for dynamic modifications and sequence variants that may be mutations, it was also necessary to include searching algorithms that could handle flexible search parameters for unexpected or multiple modifications. Complementary programs, DirectTag and TagRecon, which were written by the same research group, were natural choices for their shared vocabularies and options with Myrimatch, as well as their demonstrated performance in large-scale studies. Therefore, Myrimatch, DirecTag, and TagRecon were implemented into the TORPEDO workflow.

While these database searching algorithms are very comprehensive, their outputs are pepXML files containing all possible peptide-spectrum matches between the collected scans and the specified protein database. Another program is required to filter the peptide-spectrum matches and retain only the high-scoring peptide sequences. Generally this program also assembles the peptide sequences into proteins as well. IDPicker is the recommended protein assembly and filtering tool for pepXML files, but it is a Windows-only GUI program that does not lend itself to batch submissions or Linux web hosts. Therefore, a current limitation in this workflow is that it requires offline filtering by the user and submission of a filtered list of identifications. Until a comparable program has been identified or developed, temporary scripts were written to extract the information from the (filtered and assembled) IDPicker output files and recast them into tab-delimited files that associated peptide-spectrum matches with protein identifications.

In order to integrate the CUSPs method we developed, which clusters protein databases based on a degree of sequence similarity, we added an option for the users to select that will reannotate the IDPicker results in terms of protein groups. In fact, there is a program that generates pictures of the reduction in proteome size as a result of clustering at a range of identity thresholds (id = 0.5 to id = 1) so that the user can make an informed decision about how to group the protein sequences. Peptide uniqueness and redistribution of protein abundance measurements are automatically calculated upon selection of a clustering threshold.

From these filtered files, whether they are clustered or not, we aimed to implement the second goal: normalize and quantify protein abundances. TORPEDO houses and runs scripts that apply NSAF and/or Protease-Optimized Spectral Indexing (POSI) normalizations on either spectral counts (SpC) or matched ion intensities (MITs). Users can compare the results of any of these methods and a preliminary ANOVA p-value is performed to highlight which proteins are ranked differently for each normalization method. The POSI method requires the mzML file to be uploaded as it contains the original scan information with the collected m/z values and their intensities. One feature of choosing the POSI method is that it generates a mzIdentML file with the matching ions and their intensity information so that individuals can easily lookup and/or parse the information that was used to generate the protein abundances. Especially since most peptide-spectrum matching algorithms do not explicitly reveal which fragment ions were identified, this listing is particularly useful for those who desire detailed inspection of the identifications. In general, the output can be exported as a tab-delimited file that can be easily introduced to JMP Genomics or another software program that performs significance tests.

The third goal of this project, a resource to store and access data, is perhaps one of the most appealing aspects of this infrastructure. The web-based front-end of this project allows users to easily upload raw or processed files into a project or user-specific directory. This feature allows the researchers to have control over what data is shared

with collaborators or kept private. Due to the increasing number of projects a researcher may be working on, it is becoming increasingly impractical to keep all raw, processed, and analyzed data for all projects on a person's personal computer. Therefore, terabytes of storage were purchased and added to the storage on TORPEDO in order to free users from worrying about crowding their computers with data files. Each file kept on TORPEDO has an export or download option (if the user has permission), so that data can be easily retrieved as necessary. In addition, the system was RAIDed in order to backup the data kept on the system to help protect the computer from unintentionally losing user data.

Some of the more novel aspects of this architecture are the availability of streamlined analyses that have been developed in-house for individual projects or intended for widespread adoption within the group. This architecture allows researchers to pick and choose at which level he would like to analyze his data: *in silico* digestions or fragmentations, peptide-spectrum matching, peptide to protein assembly, or quantitative comparisons. Specifically, tools such as a viewer for sequence coverage viewer, a summarizer for matched ion intensity, a modified residue locator that uses matched ion intensities to support site localization, and an extracted ion chromatogram of MS2 intensities among many others. Many of these tools can be used in tandem for predictive studies, such as the *in silico* digestion tool followed by the sequence coverage viewer, in order to have a preliminary snapshot of what peptides are expected to be generated for a given protein and visualization for how these peptides could affect their POSI effective lengths (Figure 6.1). More commonly, the tools can be used in concert to identify peptides and proteins within or across technical replicates, such as viewing the gains afforded by additional runs or the discrepancy between proteins abundances measured under different biological conditions.

Figure 6.1. Illustrations of visualization tools provided by TORPEDO.

The sequence coverage tool in TORPEDO can highlight regions of proteins that were identified by multiple runs, regions that are predicted to be identified by an *in silico* digest, regions that vary by observed abundances, and regions that occur multiple times within a protein. Collected data can also be interactively explored using a dynamic MS1/MS2 viewer.

## 6.3 Conclusions

In total, the TORPEDO platform has been designed to be a user-friendly bioinformatic workflow that meets the immediate computational needs of researchers performing large-scale proteomic analyses, as well as includes a few unique features that set it apart from other comparable tools. Users can upload their data straight from the instrument, sequence peptides, quantify the identifications, and query their results with a number of visualization tools designed to further investigations and evoke new questions to be explored. Housing all of the informatic components in a single repository also makes the tools easier to disseminate, update, and access. The intentionally modular design also allows users to pick and choose their analyses for a customized experience that facilitates quality scientific research.

# CHAPTER 7: Propelling a Dynamic, Iterative Feedback Loop between Biology and Technology: Future Outlook, Remaining Challenges, and Conclusions

## 7.1 Overview

As the adoption of mass spectrometry analyses becomes more routine analyses for large-scale proteomic analyses, the potential opportunities that lay ahead appear infinite. A large part of its current success is due to the field's general receptiveness and outright eagerness in collaborating across disciplines to pair biochemical sample preparation with exquisitely purposeful manipulation of molecular physics in order to develop high-throughput, high-resolution analytical platforms that continue to outperform their previous depths, breadths, and accuracies of complex protein mixtures. As the instrumentation and analyses continue to progress and there are even more questions to ask and investigate, one also discovers new challenges that have not been encountered before. The informatics components to shotgun proteomic experiments, in particular, are experiencing a roller coaster ride of accomplishments and new hurdles that makes for an exhilarating non-stop journey of high-velocity loops, unexpected turns, and generally obscured views of the path ahead. That being said, there are a number of key mechanics along the bioinformatic workflow that are crucial to delivering biologically meaningful pieces of information from raw instrumental values. This dissertation particularly focused on preparing a virtual vehicle (TORPEDO, Chapter 6) to customize and optimize the informatics processes of peptide-spectrum matching (Chapter 3), peptide to protein assembly (Chapter 4), and protein quantitation (Chapter 5).

First, the principles underlying peptide-sequence matching were examined. Due to their sensitivity to their chromatographic neighborhood, matching fragment ion intensities demonstrated to be useful in discriminating which MS2 scans were over-inflating spectral count identifications and suggesting which peptides may have been under-represented based on their spectral counts. Matched ion intensities were also helpful in refining ambiguous scan identifications as well as improving site localization for modified

peptides. In fact, matched ion intensities from HCD data discriminated false positive identifications of single amino acid polymorphisms and provided evidence for a new attestation rule for calling sequence variants: modified sequences that are within close proximity of a known chemical modification should make use of additional information to confidently localize the modification or else be discounted as an ambiguous identification. For the confident peptide identifications, a notoriously contentious challenge was addressed: the protein inference problem. Peptides that map to multiple proteins add layers of ambiguity to the affected protein identifications and quantifications. We proposed clustering the protein database by sequence similarity in order to form groups of identifications whose members we would most likely not be able to analytically distinguish. Using the CUSPs (Clustering Unique Sequences in Proteomes) framework, sequence identity thresholds are recommended by analyzing the tradeoffs of the clustered proteome's reduction in size from the original database as well as the clustered proteome's number of unique identifications that can still be teased apart. This approach not only allows researchers to rescue up to 50% of their data that would have otherwise been lost, but also improves confidence in the identifications in total. The clustering method also has provisions for accurately distributing measurements among protein groups so that quantitative analyses can follow. While this study considered the use of spectral counts and matched ion intensities for quantitative purposes, matched ion intensities demonstrated a number of advantages, particularly with respect to their ability to reflect expected fold-changes at the run, protein, and peptide level when different amounts of sample was loaded. Additionally, a number of POSI (Protease-Optimized Spectral Indexing) normalization methods were explored on the same sample digested by 3 protease sets. From this study, we propose that the peptide distributions generated by each protease are very specific to the enzyme and that it may not be reasonable to expect protein A's relative abundance in a run digested with trypsin to be the same as protein A's relative abundance in a run digested with gluC.

## 7.2 Status and Remaining Challenges of Peptide-Spectrum Matching

One of the most controversial topics in proteome informatics is the derivitization and application of a false discovery rate on a dataset. Some firmly believe that the FDR should be calculated with respect to all other candidates within the database, whereas others seek to capture the likelihood of an identification with respect to all possible sequences. Where the FDR should be applied (PSM, peptide, or protein level) as well as at what amount of error (1%, 2%, 5%, or even 10%) are also contested. Unfortunately, regardless of what approach one chooses, there are biases in letting only one FDR dictate a dataset's filtering criteria. PSM-level and peptide-level FDRs have instrument biases and protein-level FDRs have proteome biases. Changing one slight setting on the instrument could have quite large ramifications that propagate through to the final analysis. There are also a number of arguments against the protein FDRs (how do protein lengths, proteome redundancy and size affect FDR calculations?), but in summary, collectively resolving how protein data sets should be filtered is a major need in the proteomics field. There may not be one standard approach that becomes the only rule to adopt, but FDRs are exceedingly crucial to keep in mind as one draws biological conclusions from analytical data that has a certain probability of being due to error.

In addition, an area of improvement that has become more evident by the analyses addressed in Chapter 3 is the confident identification of modified residues. Although genomes can be sequenced at much higher rates and with better accuracies than even 5 years ago, the proteomes of all organisms cannot wait for their sequencing. Despite the use of reference genomes, the variability from one genus to the next can be just as different as one species to the next, depending on the complexity of the organism. It is therefore imperative that database-searching software include flexibility in determining mutations or unexpected sequence variants from the given protein database. Such inclusion would ideally also extend to the conclusive identification of post-translational modifications, which considerably impact amino acid masses, and therefore impact

peptide-spectrum matching. The key to many biological questions being explored with proteomic experiments may not be found in the primary sequence; post-translational modifications such as glycosylations, acetylations, and phosphorylations may be the most interesting nuances present in the sample. Of the modifications that are specifically searched for within a dataset, there is a high level of ambiguity in site localization, as suggested in Chapter 3. Instrumentation or algorithms in the near future will surely be able to confidently select which residue contains the mass shift, but at the present, it is in large part, a guessing game. Despite the probabilistic calculations, these identifications also factor into the murky controversies over false discovery rates and complicate the standards even further.

## 7.3 Status and Remaining Challenges of Protein Inference

There are a number of solutions available that tout that they have the best way to handle the protein inference problem. However, most of them rely on data-dependent observations and are therefore volatile in which proteins may be grouped together in one run compared to another. The proposed clustering method, CUSPs, is able to group analytically ambiguous identifications that also most likely have shared biological function. However, this method is more beneficial for single, complex organisms and must be carefully applied to metaproteomes. Metaproteomes, which may contain several closely-related organisms or several distinct organisms that are most likely present in the same sample, could cause a similar issue to the volatility observed in the data-dependent clustering methods. Depending on which databases are combined into a metaproteome, it may make more sense for some proteins to be collapsed into a single representative, or it may be more useful to keep them as separate identifications. The methods recommended for identifying an appropriate clustering threshold take the entire database into consideration; it does not make any delineations for species or genus. For example, the infant gut microbiome discussed in Chapter 4 contained both human and bacterial proteomes. The human proteomes were clustered at id=0.9, while the microbial proteomes were clustered as a group at id=1.0. The applicability of this method on other

metaproteomic datasets has yet to be explored, but there is currently not a standard approach that meets this informatic need and maintains biological integrity and consistency.

## 7.4 Status and Remaining Challenges of Protein Quantitation

Perhaps the most natural next step after the discussion in Chapter 5 is the investigation and improvement of peak-picking algorithms for determining the area-under-the-curve for label-free relative protein quantitation. While spectral counts (SpC) provide "width" information about how many times the peptide was sampled across a run, and matched ion intensity (MIT) provides a "height" metric indicating the abundance of the analyte measured at the time of fragmentation, the true behavior of the peptide is an elution peak and is best described by the integration of the SpC and MIT. However, all peptides are not constantly sampled at every time point, so there needs to be some method of inferring where the peptide would have been measured if its precursor had been selected for fragmentation. Currently, the ionizability of a peptide cannot be directly computed among the complicated background of all peptides generated in a proteomic digestion. Changing the instrument's data acquisition settings can greatly affect whether each peptide is sampled once (optimized for proteome breadth) or sampled many times (optimized for proteome depth). Whereas label-free methods generally allow for the identification of more peptides, labeling methods are more accurate and consistent for quantification purposes. Therefore, for large-scale proteomic studies, one must choose upfront qualitative or quantitative data, understanding the limitations and biases of the selected approach.

## 7.5 Status and Remaining Challenges of Proteome Informatic Workflows

Despite the abundance of proteome informatics workflows available, there is one common issue that cannot be escaped: with each new algorithm or approach that is

suggested, a substantial amount of work is required to implement and integrate the tool into an existing series of analyses. Software packages that try to be too comprehensive lack the flexibility and versatility of many smaller tools that have been individually developed. Finding a program that performs a desired function and has an easy to use input and output format can sometimes be quite difficult, especially for specialized tools that were designed for a specific purpose. The turnover for such tools is quite high, many of which are the culmination of a graduate student's thesis, implemented on a specific platform and not easily maintained (if at all) by the original research group. Many labs are pushing towards open source development, allowing researchers to peer inside the inner-workings of algorithms and prompting "transparency" in each process, which helps computational biologists apply existing tools to meet his or her specific needs. However, for those researchers that do not have the time or skill set to make modifications, it can be somewhat overwhelming to study the nitty-gritty details of each algorithm. Integrating computer scientists in biological workspaces, allowing each expert to become more conversational in each other's language, would greatly facilitate the communication and adoption of informatics workflows.

## 7.6 Concluding Perspective

An increasingly popular mentality among the scientific community is an adaptation of the Central Limit Theorem: as one collects more data points, you are more likely to be looking at the data's "true" distribution, from which you can make more accurate interpretations and estimations about its behavior. If we were trying to describe the characteristics of a single variable in a 2 dimensional space, simply collecting more data points might be sufficient. But the growing trend of pursuing "systems biology"- integrating complementary bioscience information across the entire scale of biological interrogations- redefines the process, rationale, and personnel involved in collecting "more data points." In fact, it leads to a growing dichotomy of holism versus reductionism, and an ever-increasing gap between scientific integration and specialization. With each new discovery, there is an overwhelming amount of associated vocabulary, hidden implications, and revolutionizing truths that may not be immediately

apparent. Computational scientists are tasked with not only trying to mine significance out of yesterday's answers in hopes of evoking new questions, but also trying to implement architectures that can use and anticipate today's technology with tomorrow's questions. Proteomics experiments using mass spectrometry are continuing to push the envelope on what it means to produce more accurate, precise, and comprehensive protein identifications and quantifications. Improvements to sample preparation, instrument capabilities, and informatic reconstruction and interpretation are creating new, higher standards with each study. But it is not the protocols, the instruments, nor the algorithms that are moving the field forward- it is the people. The greatest successes arise out of collaborative efforts by analytical chemists, molecular physicists, bioinformaticians, and biologists. It is becoming increasingly impossible for every student of life sciences to be an expert in all fields of research, and therefore ever more necessary for young researchers to be conversant with other advanced and specialized scientists. The ability to immerse oneself in the current vocabularies, struggles, and discoveries of other relevant fields is not only a way to enrich one's academic repertoire, but also provides a cultural appreciation for the differences and commonalities of the contributions made in pursuit of scientific truths. If one is provided the tools that allow one to collaborate with others while maintaining personal excellence in his study, then investigating biological drivers with analytical technologies that are processed by computational resources and assessed by statistical tests will assuredly propel the scientific process into an exciting new era of successful integration.

# LIST OF REFERENCES

1. Kitano H. Systems biology: a brief overview. Science. 2002 Mar 1;295(5560):1662-4. PubMed PMID: 11872829.
2. Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. Annual review of genomics and human genetics. 2001;2:343-72. PubMed PMID: 11701654.
3. Westerhoff HV, Palsson BO. The evolution of molecular biology into systems biology. Nat Biotechnol. 2004 Oct;22(10):1249-52. PubMed PMID: 15470464.
4. Dictionary OE. "biome, n.": Oxford University Press.
5. Dictionary OE. "genome, n.": Oxford University Press.
6. Lander ES, Weinberg RA. Genomics: journey to the center of biology. Science. 2000 Mar 10;287(5459):1777-82. PubMed PMID: 10755930. Epub 2001/02/07. eng.
7. Crick F. Central dogma of molecular biology. Nature. 1970 Aug 8;227(5258):561-3. PubMed PMID: 4913914. Epub 1970/08/08. eng.
8. Crick FH. On protein synthesis. Symposia of the Society for Experimental Biology. 1958;12:138-63. PubMed PMID: 13580867.
9. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol. 2012 Apr;13(4):263-9. PubMed PMID: 22436749.
10. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature. 1977 Feb 24;265(5596):687-95. PubMed PMID: 870828.
11. Nowak R. Bacterial genome sequence bagged. Science. 1995 Jul 28;269(5223):468-70. PubMed PMID: 7624767.
12. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995 Jul 28;269(5223):496-512. PubMed PMID: 7542800. Epub 1995/07/28. eng.
13. Pop M, Phillippy A, Delcher AL, Salzberg SL. Comparative genome assembly. Briefings in bioinformatics. 2004 Sep;5(3):237-48. PubMed PMID: 15383210.
14. Gnerre S, Lander ES, Lindblad-Toh K, Jaffe DB. Assisted assembly: how to improve a de novo genome assembly by using related species. Genome Biol. 2009;10(8):R88. PubMed PMID: 19712469. Pubmed Central PMCID: 2745769.
15. Benson D, Lipman DJ, Ostell J. GenBank. Nucleic Acids Res. 1993 Jul 1;21(13):2963-5. PubMed PMID: 8332518. Pubmed Central PMCID: 309721.
16. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001 Feb 16;291(5507):1304-51. PubMed PMID: 11181995.
17. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921. PubMed PMID: 11237011.
18. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, et al. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 2012 Jan;40(Database issue):D571-9. PubMed PMID: 22135293. Pubmed Central PMCID: 3245063.
19. Ronaghi M. Pyrosequencing sheds light on DNA sequencing. Genome Res. 2001 Jan;11(1):3-11. PubMed PMID: 11156611.

20. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007 Oct 19;318(5849):420-6. PubMed PMID: 17901297. Pubmed Central PMCID: 2674581.
21. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, et al. Empirical analysis of transcriptional activity in the Arabidopsis genome. Science. 2003 Oct 31;302(5646):842-6. PubMed PMID: 14593172. Epub 2003/11/01. eng.
22. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, et al. Global identification of human transcribed sequences with genome tiling arrays. Science. 2004 Dec 24;306(5705):2242-6. PubMed PMID: 15539566. Epub 2004/11/13. eng.
23. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, et al. A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci U S A. 2006 Apr 4;103(14):5320-5. PubMed PMID: 16569694. Pubmed Central PMCID: 1414796. Epub 2006/03/30. eng.
24. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57-63. PubMed PMID: 19015660. Pubmed Central PMCID: 2949280. Epub 2008/11/19. eng.
25. Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlen M, et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Res. 2012 Nov 1;40(20):10084-97. PubMed PMID: 22965124. Pubmed Central PMCID: 3488244. Epub 2012/09/12. eng.
26. International Human Genome Sequencing C. Finishing the euchromatic sequence of the human genome. Nature. 2004 Oct 21;431(7011):931-45. PubMed PMID: 15496913.
27. Mirza SP, Olivier M. Methods and approaches for the comprehensive characterization and quantification of cellular proteomes using mass spectrometry. Physiol Genomics. 2008 Mar 14;33(1):3-11. PubMed PMID: 18162499. Pubmed Central PMCID: PMC2771641. Epub 2007/12/29. eng.
28. Rogers S, Girolami M, Kolch W, Waters KM, Liu T, Thrall B, et al. Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. Bioinformatics. 2008 Dec 15;24(24):2894-900. PubMed PMID: 18974169. Epub 2008/11/01. eng.
29. de Groot MJ, Daran-Lapujade P, van Breukelen B, Knijnenburg TA, de Hulster EA, Reinders MJ, et al. Quantitative proteomics and transcriptomics of anaerobic and aerobic yeast cultures reveals post-transcriptional regulation of key cellular processes. Microbiology. 2007 Nov;153(Pt 11):3864-78. PubMed PMID: 17975095.
30. Dhingra V, Gupta M, Andacht T, Fu ZF. New frontiers in proteomics research: a perspective. Int J Pharm. 2005 Aug 11;299(1-2):1-18. PubMed PMID: 15979831.
31. Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, et al. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Biotechnology (N Y). 1996 Jan;14(1):61-5. PubMed PMID: 9636313. Epub 1996/01/01. eng.

32. Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. 2009 Jul 1;23(13):1494-504. PubMed PMID: 19571179. Pubmed Central PMCID: 3152381.
33. Haasch D, Chen YW, Reilly RM, Chiou XG, Koterski S, Smith ML, et al. T cell activation induces a noncoding RNA transcript sensitive to inhibition by immunosuppressant drugs and encoded by the proto-oncogene, BIC. Cellular immunology. 2002 May-Jun;217(1-2):78-86. PubMed PMID: 12426003.
34. Verheije MH, Olsthoorn RC, Kroese MV, Rottier PJ, Meulenberg JJ. Kissing interaction between 3' noncoding and coding sequences is essential for porcine arterivirus RNA replication. J Virol. 2002 Feb;76(3):1521-6. PubMed PMID: 11773426. Pubmed Central PMCID: 135790.
35. Braidotti G, Baubec T, Pauler F, Seidl C, Smrzka O, Stricker S, et al. The Air noncoding RNA: an imprinted cis-silencing transcript. Cold Spring Harb Symp Quant Biol. 2004;69:55-66. PubMed PMID: 16117633. Pubmed Central PMCID: 2847179.
36. Klose J. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. Humangenetik. 1975;26(3):231-43. PubMed PMID: 1093965. Epub 1975/01/01. eng.
37. O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. J Biol Chem. 1975 May 25;250(10):4007-21. PubMed PMID: 236308. Pubmed Central PMCID: 2874754. Epub 1975/05/25. eng.
38. Wollnik H. Time-of-Flight Mass Analyzers. Mass Spectrometry Reviews. 1993 Mar;12(2):89-114. PubMed PMID: ISI:A1993MH87000001. English.
39. March RE. Quadrupole ion trap mass spectrometry: a view at the turn of the century. International Journal of Mass Spectrometry. 2000 Dec 25;200(1-3):285-312. PubMed PMID: ISI:000166178700019. English.
40. Schwartz JC, Jardine I. Quadrupole ion trap mass spectrometry. Methods Enzymol. 1996;270:552-86. PubMed PMID: 8803984. Epub 1996/01/01. eng.
41. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science. 1989 Oct 6;246(4926):64-71. PubMed PMID: 2675315. Epub 1989/10/06. eng.
42. Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem. 1988 Oct 15;60(20):2299-301. PubMed PMID: 3239801. Epub 1988/10/15. eng.
43. Yates JR. Mass spectral analysis in proteomics. Annu Rev Bioph Biom. 2004;33:297-316. PubMed PMID: ISI:000222339700016. English.
44. Tipton JD, Tran JC, Catherman AD, Ahlf DR, Durbin KR, Kelleher NL. Analysis of intact protein isoforms by mass spectrometry. J Biol Chem. 2011 Jul 22;286(29):25451-8. PubMed PMID: 21632550. Pubmed Central PMCID: 3138281.
45. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, et al. Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature. 2011 Dec 8;480(7376):254-8. PubMed PMID: 22037311. Pubmed Central PMCID: 3237778.
46. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003 Mar 13;422(6928):198-207. PubMed PMID: ISI:000181488900054. English.

47. Kelleher NL, Lin HY, Valaskovic GA, Aaserud DJ, Fridriksson EK, McLafferty FW. Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry. Journal of the American Chemical Society. 1999 Feb 3;121(4):806-12. PubMed PMID: ISI:000078471800026. English.
48. Washburn MP, Wolters D, Yates JR, 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol. 2001 Mar;19(3):242-7. PubMed PMID: 11231557.
49. Lambert JP, Ethier M, Smith JC, Figeys D. Proteomics: from gel based to gel free. Anal Chem. 2005 Jun 15;77(12):3771-87. PubMed PMID: 15952756. Epub 2005/06/15. eng.
50. Koller A, Washburn MP, Lange BM, Andon NL, Deciu C, Haynes PA, et al. Proteomic survey of metabolic pathways in rice. P Natl Acad Sci USA. 2002 Sep 3;99(18):11969-74. PubMed PMID: ISI:000177843100078. English.
51. Rabilloud T. Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. Proteomics. 2002 Jan;2(1):3-10. PubMed PMID: 11788986.
52. Renart J, Reiser J, Stark GR. Transfer of proteins from gels to diazobenzyloxymethyl-paper and detection with antisera: a method for studying antibody specificity and antigen structure. Proc Natl Acad Sci U S A. 1979 Jul;76(7):3116-20. PubMed PMID: 91164. Pubmed Central PMCID: 383774.
53. Towbin H, Staehelin T, Gordon J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. Proc Natl Acad Sci U S A. 1979 Sep;76(9):4350-4. PubMed PMID: 388439. Pubmed Central PMCID: 411572.
54. Issaq HJ, Chan KC, Blonder J, Ye X, Veenstra TD. Separation, detection and quantitation of peptides by liquid chromatography and capillary electrochromatography. J Chromatogr A. 2009 Mar 6;1216(10):1825-37. PubMed PMID: 19131068.
55. Spengler B, Kirsch D, Kaufmann R, Lemoine J. Structure-Analysis of Branched Oligosaccharides Using Post-Source Decay in Matrix-Assisted Laser-Desorption Ionization Mass-Spectrometry. Org Mass Spectrom. 1994 Dec;29(12):782-7. PubMed PMID: ISI:A1994QF69200009. English.
56. Hunt DF, Henderson RA, Shabanowitz J, Sakaguchi K, Michel H, Sevilir N, et al. Characterization of Peptides Bound to the Class-I Mhc Molecule Hla-A2.1 by Mass-Spectrometry. Science. 1992 Mar 6;255(5049):1261-3. PubMed PMID: ISI:A1992HG67700045. English.
57. Samalikova M, Matecko I, Muller N, Grandori R. Interpreting conformational effects in protein nano-ESI-MS spectra. Anal Bioanal Chem. 2004 Feb;378(4):1112-23. PubMed PMID: 14663547.
58. Pabst M, Altmann F. Influence of electrosorption, solvent, temperature, and ion polarity on the performance of LC-ESI-MS using graphitic carbon for acidic oligosaccharides. Anal Chem. 2008 Oct 1;80(19):7534-42. PubMed PMID: 18778038.
59. McLuckey SA, Wells JM. Mass analysis at the advent of the 21st century. Chem Rev. 2001 Feb;101(2):571-606. PubMed PMID: ISI:000167137400013. English.

60. Mann M, Kelleher NL. Precision proteomics: the case for high resolution and high mass accuracy. Proc Natl Acad Sci U S A. 2008 Nov 25;105(47):18132-8. PubMed PMID: 18818311. Pubmed Central PMCID: 2587563.
61. Jonscher KR, Yates JR, 3rd. The quadrupole ion trap mass spectrometer--a small solution to a big challenge. Anal Biochem. 1997 Jan 1;244(1):1-15. PubMed PMID: 9025900.
62. Shukla AK, Futrell JH. Tandem mass spectrometry: dissociation of ions by collisional activation. J Mass Spectrom. 2000 Sep;35(9):1069-90. PubMed PMID: 11006601.
63. Eng JK, Mccormack AL, Yates JR. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. J Am Soc Mass Spectr. 1994 Nov;5(11):976-89. PubMed PMID: ISI:A1994PP71300004. English.
64. Mann M, Wilm M. Error Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. Analytical Chemistry. 1994 Dec 15;66(24):4390-9. PubMed PMID: ISI:A1994PW74600008. English.
65. Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci U S A. 2003 Jun 10;100(12):6940-5. PubMed PMID: 12771378. Pubmed Central PMCID: 165809. Epub 2003/05/29. eng.
66. Rivers J, Simpson DM, Robertson DH, Gaskell SJ, Beynon RJ. Absolute multiplexed quantitative analysis of protein expression during muscle development using QconCAT. Mol Cell Proteomics. 2007 Aug;6(8):1416-27. PubMed PMID: 17510050. Epub 2007/05/19. eng.
67. Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem. 2003 Apr 15;75(8):1895-904. PubMed PMID: 12713048. Epub 2003/04/26. eng.
68. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics. 2004 Dec;3(12):1154-69. PubMed PMID: 15385600.
69. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol. 1999 Oct;17(10):994-9. PubMed PMID: 10504701. Epub 1999/10/03. eng.
70. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics. 2002 May;1(5):376-86. PubMed PMID: 12118079.
71. Tabb DL, McDonald WH, Yates JR, 3rd. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J Proteome Res. 2002 Jan-Feb;1(1):21-6. PubMed PMID: 12643522. Pubmed Central PMCID: 2811961.
72. Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome

Res. 2007 Feb;6(2):654-61. PubMed PMID: 17269722. Pubmed Central PMCID: 2525619. Epub 2007/02/03. eng.
73. Holman JD, Ma ZQ, Tabb DL. Identifying proteomic LC-MS/MS data sets with Bumbershoot and IDPicker. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2012 Mar;Chapter 13:Unit13 7. PubMed PMID: 22389012. Epub 2012/03/06. eng.
74. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 2002 Oct 15;74(20):5383-92. PubMed PMID: 12403597.
75. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 2003 Sep 1;75(17):4646-58. PubMed PMID: 14632076.
76. Weatherly DB, Atwood JA, 3rd, Minning TA, Cavola C, Tarleton RL, Orlando R. A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. Mol Cell Proteomics. 2005 Jun;4(6):762-72. PubMed PMID: 15703444. Epub 2005/02/11. eng.
77. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, et al. Open mass spectrometry search algorithm. J Proteome Res. 2004 Sep-Oct;3(5):958-64. PubMed PMID: 15473683. Epub 2004/10/12. eng.
78. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004 Jun 12;20(9):1466-7. PubMed PMID: 14976030. Epub 2004/02/21. eng.
79. Pan J, Stephenson AL, Kazamia E, Huck WT, Dennis JS, Smith AG, et al. Quantitative tracking of the growth of individual algal cells in microdroplet compartments. Integrative biology : quantitative biosciences from nano to macro. 2011 Oct;3(10):1043-51. PubMed PMID: 21863189.
80. Hyatt D, Pan C. Exhaustive database searching for amino acid mutations in proteomes. Bioinformatics. 2012 Jul 15;28(14):1895-901. PubMed PMID: 22581177.
81. Tabb DL, Ma ZQ, Martin DB, Ham AJ, Chambers MC. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. J Proteome Res. 2008 Sep;7(9):3838-46. PubMed PMID: 18630943. Pubmed Central PMCID: 2810657. Epub 2008/07/18. eng.
82. Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJ, Tabb DL. TagRecon: high-throughput mutation identification through sequence tagging. J Proteome Res. 2010 Apr 5;9(4):1716-26. PubMed PMID: 20131910. Pubmed Central PMCID: 2859315. Epub 2010/02/06. eng.
83. Gupta N, Bandeira N, Keich U, Pevzner PA. Target-decoy approach and false discovery rate: when things may go wrong. J Am Soc Mass Spectrom. 2011 Jul;22(7):1111-20. PubMed PMID: 21953092. Pubmed Central PMCID: 3220955.
84. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007 Mar;4(3):207-14. PubMed PMID: 17327847.
85. Elias JE, Gygi SP. Target-decoy search strategy for mass spectrometry-based proteomics. Methods Mol Biol. 2010;604:55-71. PubMed PMID: 20013364. Pubmed Central PMCID: 2922680.

86. Gupta N, Benhamida J, Bhargava V, Goodman D, Kain E, Kerman I, et al. Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. Genome Res. 2008 Jul;18(7):1133-42. PubMed PMID: 18426904. Pubmed Central PMCID: 2493402.
87. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, et al. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. J Proteome Res. 2009 Aug;8(8):3872-81. PubMed PMID: 19522537. Pubmed Central PMCID: 2810655.
88. Yang X, Dondeti V, Dezube R, Maynard DM, Geer LY, Epstein J, et al. DBParser: web-based software for shotgun proteomic data analyses. J Proteome Res. 2004 Sep-Oct;3(5):1002-8. PubMed PMID: 15473689.
89. Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics. 2003 Aug;3(8):1454-63. PubMed PMID: 12923771. Epub 2003/08/19. eng.
90. Choi H, Fermin D, Nesvizhskii AI. Significance analysis of spectral count data in label-free shotgun proteomics. Mol Cell Proteomics. 2008 Dec;7(12):2373-85. PubMed PMID: 18644780. Pubmed Central PMCID: 2596341.
91. Pham TV, Piersma SR, Warmoes M, Jimenez CR. On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. Bioinformatics. 2010 Feb 1;26(3):363-9. PubMed PMID: 20007255.
92. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem. 2007 Oct;389(4):1017-31. PubMed PMID: 17668192. Epub 2007/08/02. eng.
93. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, et al. ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W597-603. PubMed PMID: 22661580. Pubmed Central PMCID: 3394269.
94. Beavis R, Fenyo D. Finding protein sequences using PROWL. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2004 Oct;Chapter 13:Unit 13 2. PubMed PMID: 18428719.
95. Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, et al. A guided tour of the Trans-Proteomic Pipeline. Proteomics. 2010 Mar;10(6):1150-9. PubMed PMID: 20101611. Pubmed Central PMCID: 3017125.
96. Orchard S. Data Standardization and Sharing-The work of the HUPO-PSI. Biochim Biophys Acta. 2013 Mar 20. PubMed PMID: 23524294.
97. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data - The protein inference problem. Molecular & Cellular Proteomics. 2005 Oct;4(10):1419-40. PubMed PMID: ISI:000232207900001. English.
98. Johnson RS, Davis MT, Taylor JA, Patterson SD. Informatics for protein identification by mass spectrometry. Methods. 2005 Mar;35(3):223-36. PubMed PMID: 15722219. Epub 2005/02/22. eng.
99. Verberkmoes NC, Hervey WJ, Shah M, Land M, Hauser L, Larimer FW, et al. Evaluation of "shotgun" proteomics for identification of biological threat agents in complex environmental matrixes: experimental simulations. Anal Chem. 2005 Feb 1;77(3):923-32. PubMed PMID: 15679362. Epub 2005/02/01. eng.

100. Bern M, Goldberg D, McDonald WH, Yates JR, 3rd. Automatic quality assessment of peptide tandem mass spectra. Bioinformatics. 2004 Aug 4;20 Suppl 1:i49-54. PubMed PMID: 15262780. Epub 2004/07/21. eng.
101. Salmi J, Moulder R, Filen JJ, Nevalainen OS, Nyman TA, Lahesmaa R, et al. Quality classification of tandem mass spectrometry data. Bioinformatics. 2006 Feb 15;22(4):400-6. PubMed PMID: 16352652. Epub 2005/12/15. eng.
102. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, et al. Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol. 2007 Jan;25(1):125-31. PubMed PMID: 17195840.
103. Abraham P, Adams R, Giannone RJ, Kalluri U, Ranjan P, Erickson B, et al. Defining the boundaries and characterizing the landscape of functional genome expression in vascular tissues of Populus using shotgun proteomics. J Proteome Res. 2012 Jan 1;11(1):449-60. PubMed PMID: 22003893.
104. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. Identification of post-translational modifications via blind search of mass-spectra. Proceedings / IEEE Computational Systems Bioinformatics Conference, CSB IEEE Computational Systems Bioinformatics Conference. 2005:157-66. PubMed PMID: 16447973.
105. Louie GV, Bowman ME, Moffitt MC, Baiga TJ, Moore BS, Noel JP. Structural determinants and modulation of substrate specificity in phenylalanine-tyrosine ammonia-lyases. Chem Biol. 2006 Dec;13(12):1327-38. PubMed PMID: 17185228. Pubmed Central PMCID: 2859959. Epub 2006/12/23. eng.
106. Howles PA, Sewalt VJH, Paiva NL, Elkind Y, Bate NJ, Lamb C, et al. Overexpression of L-phenylalanine ammonia-lyase in transgenic tobacco plants reveals control points for flux into phenylpropanoid biosynthesis. Plant Physiology. 1996 Dec;112(4):1617-24. PubMed PMID: ISI:A1996VY30400026. English.
107. Lu B, Xu T, Park SK, McClatchy DB, Liao L, Yates JR, 3rd. Shotgun protein identification and quantification by mass spectrometry in neuroproteomics. Methods Mol Biol. 2009;566:229-59. PubMed PMID: 20058176. Epub 2010/01/09. eng.
108. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal Chem. 2005 Feb 15;77(4):964-73. PubMed PMID: 15858974. Epub 2005/04/30. eng.
109. Searle BC, Dasari S, Wilmarth PA, Turner M, Reddy AP, David LL, et al. Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm. J Proteome Res. 2005 Mar-Apr;4(2):546-54. PubMed PMID: 15822933. Epub 2005/04/13. eng.
110. Kapp E, Schutz F. Overview of tandem mass spectrometry (MS/MS) database search algorithms. Current protocols in protein science / editorial board, John E Coligan [et al]. 2007 Aug;Chapter 25:Unit25 2. PubMed PMID: 18429324. Epub 2008/04/23. eng.
111. Tabb DL, Saraf A, Yates JR, 3rd. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem. 2003 Dec 1;75(23):6415-21. PubMed PMID: 14640709. Pubmed Central PMCID: 2915448. Epub 2003/12/04. eng.

112. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, et al. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem. 2005 Jul 15;77(14):4626-39. PubMed PMID: 16013882. Epub 2005/07/15. eng.
113. Sunyaev S, Liska AJ, Golod A, Shevchenko A. MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. Anal Chem. 2003 Mar 15;75(6):1307-15. PubMed PMID: 12659190. Epub 2003/03/28. eng.
114. Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, et al. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. Mol Cell Proteomics. 2007 Sep;6(9):1638-55. PubMed PMID: 17533153. Epub 2007/05/30. eng.
115. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol. 2006 Oct;24(10):1285-92. PubMed PMID: 16964243. Epub 2006/09/12. eng.
116. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell. 2006 Nov 3;127(3):635-48. PubMed PMID: ISI:000241937000025. English.
117. Bailey CM, Sweet SMM, Cunningham DL, Zeller M, Heath JK, Cooper HJ. SLoMo: Automated Site Localization of Modifications from ETD/ECD Mass Spectra. Journal of Proteome Research. 2009 Apr;8(4):1965-71. PubMed PMID: ISI:000264928200035. English.
118. Chen Y, Chen W, Cobb MH, Zhao YM. PTMap-A sequence alignment software for unrestricted, accurate, and full-spectrum identification of post-translational modification sites. P Natl Acad Sci USA. 2009 Jan 20;106(3):761-6. PubMed PMID: ISI:000262809700018. English.
119. Wysocki VH, Tsaprailis G, Smith LL, Breci LA. Mobile and localized protons: a framework for understanding peptide dissociation. J Mass Spectrom. 2000 Dec;35(12):1399-406. PubMed PMID: 11180630. Epub 2001/02/17. eng.
120. Boyd R, Somogyi A. The mobile proton hypothesis in fragmentation of protonated peptides: a perspective. J Am Soc Mass Spectrom. 2010 Aug;21(8):1275-8. PubMed PMID: 20547071. Epub 2010/06/16. eng.
121. Olsen JV, Macek B, Lange O, Makarov A, Horning S, Mann M. Higher-energy C-trap dissociation for peptide modification analysis. Nat Methods. 2007 Sep;4(9):709-12. PubMed PMID: 17721543. Epub 2007/08/28. eng.
122. Michalski A, Neuhauser N, Cox J, Mann M. A Systematic Investigation into the Nature of Tryptic HCD Spectra. Journal of Proteome Research. 2012 Nov;11(11):5479-91. PubMed PMID: ISI:000311190600031. English.
123. Reid GE, Roberts KD, Kapp EA, Simpson RJ. Statistical and mechanistic approaches to understanding the gas-phase fragmentation behavior of methionine sulfoxide containing peptides. Journal of Proteome Research. 2004 Jul-Aug;3(4):751-9. PubMed PMID: ISI:000223319000008. English.
124. Rappsilber J, Mann M. What does it mean to identify a protein in proteomics? Trends Biochem Sci. 2002 Feb;27(2):74-8. PubMed PMID: 11852244.

125. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods. 2007 Oct;4(10):787-97. PubMed PMID: 17901868.
126. Friso G, Majeran W, Huang M, Sun Q, van Wijk KJ. Reconstruction of metabolic pathways, protein expression, and homeostasis machineries across maize bundle sheath and mesophyll chloroplasts: large-scale quantitative proteomics using the first maize genome assembly. Plant Physiol. 2010 Mar;152(3):1219-50. PubMed PMID: 20089766. Pubmed Central PMCID: 2832236. Epub 2010/01/22. eng.
127. Meyer-Arendt K, Old WM, Houel S, Renganathan K, Eichelberger B, Resing KA, et al. IsoformResolver: A Peptide-Centric Algorithm for Protein Inference. J Proteome Res. 2011 Jun 7. PubMed PMID: 21599010. Epub 2011/05/24. Eng.
128. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 Oct;26(19):2460-1. PubMed PMID: ISI:000282170000016. English.
129. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, et al. The Protein Information Resource. Nucleic Acids Res. 2003 Jan 1;31(1):345-7. PubMed PMID: 12520019. Pubmed Central PMCID: 165487. Epub 2003/01/10. eng.
130. Kalluri UC, Hurst GB, Lankford PK, Ranjan P, Pelletier DA. Shotgun proteome profile of Populus developing xylem. Proteomics. 2009 Nov;9(21):4871-80. PubMed PMID: ISI:000272124600005. English.
131. Plomion C, Lalanne C, Claverol S, Meddour H, Kohler A, Bogeat-Triboulot MB, et al. Mapping the proteome of poplar and application to the discovery of drought-stress responsive proteins. Proteomics. 2006 Dec;6(24):6509-27. PubMed PMID: ISI:000243277500015. English.
132. Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, et al. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. Science. 2008 May 16;320(5878):938-41. PubMed PMID: 18436743. Epub 2008/04/26. eng.
133. Yang XH, Tschaplinski TJ, Hurst GB, Jawdy S, Abraham PE, Lankford PK, et al. Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. Genome Research. 2011 Apr;21(4):634-41. PubMed PMID: ISI:000289067800014. English.
134. Delalande F, Carapito C, Brizard JP, Brugidou C, Van Dorsselaer A. Multigenic families and proteomics: extended protein characterization as a tool for paralog gene identification. Proteomics. 2005 Feb;5(2):450-60. PubMed PMID: 15627959. Epub 2005/01/04. eng.
135. Foston M, Hubbell C, Samuel R, Jung S, Fan H, Ding S-Y, et al. Chemical, ultrastructural and supramolecular analysis of tension wood in Populus tremula x alba as a model substrate for reduced recalcitrance. Energy & Environmental Science Journal. 2011.
136. Black DL. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell. 2000 Oct 27;103(3):367-70. PubMed PMID: 11081623. Epub 2000/11/18. eng.
137. Giannone RJ, Huber H, Karpinets T, Heimerl T, Kuper U, Rachel R, et al. Proteomic characterization of cellular and molecular processes that enable the

Nanoarchaeum equitans--Ignicoccus hospitalis relationship. PLoS One. 2011;6(8):e22942. PubMed PMID: 21826220. Pubmed Central PMCID: 3149612.
138. Zybailov BL, Florens L, Washburn MP. Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors. Mol Biosyst. 2007 May;3(5):354-60. PubMed PMID: 17460794. Epub 2007/04/27. eng.
139. Vogel C, Marcotte EM. Label-free protein quantitation using weighted spectral counting. Methods Mol Biol. 2012;893:321-41. PubMed PMID: 22665309.
140. Freund DM, Prenni JE. Improved Detection of Quantitative Differences Using a Combination of Spectral Counting and MS/MS Total Ion Current. J Proteome Res. 2013 Mar 12. PubMed PMID: 23445521.
141. Zhu W, Smith JW, Huang CM. Mass spectrometry-based label-free quantitative proteomics. J Biomed Biotechnol. 2010;2010:840518. PubMed PMID: 19911078. Pubmed Central PMCID: 2775274.
142. Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, Hettich RL, Samatova NF. Detecting differential and correlated protein expression in label-free shotgun proteomics. J Proteome Res. 2006 Nov;5(11):2909-18. PubMed PMID: 17081042.
143. Griffin NM, Yu J, Long F, Oh P, Shore S, Li Y, et al. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat Biotechnol. 2010 Jan;28(1):83-9. PubMed PMID: 20010810. Pubmed Central PMCID: 2805705.
144. Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol. 2005 Oct;1(5):252-62. PubMed PMID: 16408053. Epub 2006/01/13. eng.
145. Bantscheff M, Kuster B. Quantitative mass spectrometry in proteomics. Anal Bioanal Chem. 2012 Sep;404(4):937-8. PubMed PMID: 22825679. Epub 2012/07/25. eng.
146. Kolkman A, Dirksen EH, Slijper M, Heck AJ. Double standards in quantitative proteomics: direct comparative assessment of difference in gel electrophoresis and metabolic stable isotope labeling. Mol Cell Proteomics. 2005 Mar;4(3):255-66. PubMed PMID: 15632418.
147. Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc Natl Acad Sci U S A. 2000 Aug 15;97(17):9390-5. PubMed PMID: 10920198. Pubmed Central PMCID: 16874.
148. Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, et al. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. Anal Chem. 2003 Sep 15;75(18):4818-26. PubMed PMID: 14674459.
149. Pan C, Oda Y, Lankford PK, Zhang B, Samatova NF, Pelletier DA, et al. Characterization of anaerobic catabolism of p-coumarate in Rhodopseudomonas palustris by integrating transcriptomics and quantitative proteomics. Mol Cell Proteomics. 2008 May;7(5):938-48. PubMed PMID: 18156135.
150. Yao X, Freas A, Ramirez J, Demirev PA, Fenselau C. Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. Anal Chem. 2001 Jul 1;73(13):2836-42. PubMed PMID: 11467524.
151. Belnap CP, Pan C, Denef VJ, Samatova NF, Hettich RL, Banfield JF. Quantitative proteomic analyses of the response of acidophilic microbial communities to

different pH conditions. The ISME journal. 2011 Jul;5(7):1152-61. PubMed PMID: 21228889. Pubmed Central PMCID: 3146278.
152. Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, Halfvarson J, et al. Shotgun metaproteomics of the human distal gut microbiota. The ISME journal. 2009 Feb;3(2):179-89. PubMed PMID: 18971961.
153. Tolonen AC, Haas W, Chilaka AC, Aach J, Gygi SP, Church GM. Proteome-wide systems analysis of a cellulosic biofuel-producing microbe. Mol Syst Biol. 2011 Jan 18;7:461. PubMed PMID: 21245846. Pubmed Central PMCID: 3049413.
154. Ong SE, Mann M. Stable isotope labeling by amino acids in cell culture for quantitative proteomics. Methods Mol Biol. 2007;359:37-52. PubMed PMID: 17484109.
155. Dayon L, Hainard A, Licker V, Turck N, Kuhn K, Hochstrasser DF, et al. Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags. Anal Chem. 2008 Apr 15;80(8):2921-31. PubMed PMID: 18312001.
156. Ow SY, Cardona T, Taton A, Magnuson A, Lindblad P, Stensjo K, et al. Quantitative shotgun proteomics of enriched heterocysts from Nostoc sp. PCC 7120 using 8-plex isobaric peptide tags. J Proteome Res. 2008 Apr;7(4):1615-28. PubMed PMID: 18290607.
157. Collier TS, Sarkar P, Franck WL, Rao BM, Dean RA, Muddiman DC. Direct comparison of stable isotope labeling by amino acids in cell culture and spectral counting for quantitative proteomics. Anal Chem. 2010 Oct 15;82(20):8696-702. PubMed PMID: 20845935.
158. Hendrickson EL, Xia Q, Wang T, Leigh JA, Hackett M. Comparison of spectral counting and metabolic stable isotope labeling for use with quantitative microbial proteomics. Analyst. 2006 Dec;131(12):1335-41. PubMed PMID: 17124542. Pubmed Central PMCID: 2660848.
159. Patel VJ, Thalassinos K, Slade SE, Connolly JB, Crombie A, Murrell JC, et al. A comparison of labeling and label-free mass spectrometry-based proteomics approaches. J Proteome Res. 2009 Jul;8(7):3752-9. PubMed PMID: 19435289.
160. Thompson DK, Chourey K, Wickham GS, Thieman SB, VerBerkmoes NC, Zhang B, et al. Proteomics reveals a core molecular response of Pseudomonas putida F1 to acute chromate challenge. BMC Genomics. 2010;11:311. PubMed PMID: 20482812. Pubmed Central PMCID: 2996968.
161. Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, et al. A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. Mol Cell Proteomics. 2009 Dec;8(12):2759-69. PubMed PMID: 19828875. Pubmed Central PMCID: 2816009.
162. Pichler P, Kocher T, Holzmann J, Mazanek M, Taus T, Ammerer G, et al. Peptide labeling with isobaric tags yields higher identification rates using iTRAQ 4-plex compared to TMT 6-plex and iTRAQ 8-plex on LTQ Orbitrap. Anal Chem. 2010 Aug 1;82(15):6549-58. PubMed PMID: 20593797. Pubmed Central PMCID: 3093924.
163. Zhang Y, Ficarro SB, Li S, Marto JA. Optimized Orbitrap HCD for quantitative analysis of phosphopeptides. J Am Soc Mass Spectrom. 2009 Aug;20(8):1425-34. PubMed PMID: 19403316.
164. Peng JM, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-

MS/MS) for large-scale protein analysis: The yeast proteome. Journal of Proteome Research. 2003 Jan-Feb;2(1):43-50. PubMed PMID: ISI:000180874400005. English.
165. Griffin TJ, Xie H, Bandhakavi S, Popko J, Mohan A, Carlis JV, et al. iTRAQ reagent-based quantitative proteomic analysis on a linear ion trap mass spectrometer. J Proteome Res. 2007 Nov;6(11):4200-9. PubMed PMID: 17902639. Pubmed Central PMCID: 2533114.
166. Pan C, Fischer CR, Hyatt D, Bowen BP, Hettich RL, Banfield JF. Quantitative tracking of isotope flows in proteomes of microbial communities. Mol Cell Proteomics. 2011 Apr;10(4):M110 006049. PubMed PMID: 21285414. Pubmed Central PMCID: 3069347.
167. Carvalho PC, Han X, Xu T, Cociorva D, Carvalho Mda G, Barbosa VC, et al. XDIA: improving on the label-free data-independent analysis. Bioinformatics. 2010 Mar 15;26(6):847-8. PubMed PMID: 20106817. Pubmed Central PMCID: 2832823.
168. Chakraborty AB, Berger SJ, Gebler JC. Use of an integrated MS--multiplexed MS/MS data acquisition strategy for high-coverage peptide mapping studies. Rapid Commun Mass Spectrom. 2007;21(5):730-44. PubMed PMID: 17279597.
169. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, et al. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol Cell Proteomics. 2005 Oct;4(10):1487-502. PubMed PMID: 15979981.

# VITA

Rachel Adams was born in Dallas, Texas and completed her high school education at Ursuline Academy in 2004. She received her Bachelor of Science degree in Bioinformatics with a minor in Chemistry at Baylor University from 2004-2008. During her college summer internships at Gradalis Inc. a biotechnology company in Dallas, she was introduced to personalized cancer therapeutics that relied heavily on cutting-edge, high-throughput analytical platforms including mass spectrometry. Since then, she has been pursuing a Ph.D. in the Genome Science and Technologies program that has joint affiliations with the University of Tennessee-Knoxville and Oak Ridge National Laboratory. She received additional opportunities for funding, training, and coursework in computational biology through an NSF IGERT fellowship, SCALE-IT (Scalable Computing and Leading Edge Innovative Technologies). She expects to receive her PhD in Life Sciences in May 2013.