8-2014

QUANTITATIVE CHARACTERIZATION OF PROTEINS AND POST-TRANSLATIONAL MODIFICATIONS IN COMPLEX PROTEOMES USING HIGH-RESOLUTION MASS SPECTROMETRY-BASED PROTEOMICS

Zhou Li
University of Tennessee - Knoxville, zli27@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Zhou Li entitled "QUANTITATIVE CHARACTERIZATION OF PROTEINS AND POST-TRANSLATIONAL MODIFICATIONS IN COMPLEX PROTEOMES USING HIGH-RESOLUTION MASS SPECTROMETRY-BASED PROTEOMICS." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Robert L. Hettich, Major Professor

We have read this dissertation and recommend its acceptance:

Mircea Podar, Steven W Wilhelm, Alison Buchan, Chongle Pan

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

QUANTITATIVE CHARACTERIZATION OF PROTEINS AND POST-TRANSLATIONAL MODIFICATIONS IN COMPLEX PROTEOMES USING HIGH-RESOLUTION MASS SPECTROMETRY-BASED PROTEOMICS

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

Zhou Li
August 2014

ACKNOWLEDGEMENTS

First, I would like to sincerely thank my advisors: Dr. Chongle Pan and Dr. Robert Hettich for mentoring in the past four years. This dissertation would have not been possible if they had not invested time and effort in guiding me through various scientific hurdles during the course of my Ph.D. study. Their scientific knowledge and the way of collaborating with people will be a reference point for the rest of my life. Secondly, I would like to thank my dissertation committee members: Dr. Alison Buchan, Dr. Steven Wilhelm, and Dr. Mircea Podar not only for their time spent in my committee meetings, but also for their constant encouragement, constructive criticisms, and insightful discussions about this work. Thirdly, I would like to thank my labmates in the Organic and Biological Mass Spectrometry Group at the Oak Ridge National Laboratory and collaborators from other institutions for helping me move my research forward. Fourthly, I would like to thank the University of Tennessee-Oak Ridge National Laboratory Graduate School of Genome Science and Technology for giving the opportunity to pursue my Ph.D. degree in the United States and to utilize various cutting-edge research facilities, such as high-performance mass spectrometer and supercomputer. Finally, I would like to thank my parents: Dingyi Li and Haiyue Zhou for their love and support.

ABSTRACT

Mass spectrometry-based proteomics is focused on identifying the entire suite of proteins and their post-translational modifications (PTMs) in a cell, organism, or community. In particular, quantitative proteomics measures abundance changes of thousands of proteins among multiple samples and provides network-level insight into how biological systems respond to environmental perturbations. Various quantitative proteomics methods have been developed, including label-free, metabolic labeling, and isobaric chemical labeling. This dissertation starts with a systematic comparison of these three methods, and shows that isobaric chemical labeling provides accurate, precise, and reproducible quantification for thousands of proteins. Based on these results, we applied this approach to characterizing the proteome of *Arabidopsis* seedlings treated with Strigolactones (SLs), a new class of plant hormones that modulate a range of developmental processes. Our study reveals that SLs regulate the expression of a range of proteins that have not been assigned to SL pathways, which provides novel targets for follow-up genetic and biochemical characterization of SL signaling. The same approach was also used to measure how elevated temperature impacts the physiology of individual microbial groups in an acid mine drainage (AMD) microbial community, and shows that related organisms differed in their abundance and functional responses to temperature. Elevated temperature repressed carbon fixation by two *Leptospirillum* genotypes, whereas carbon fixation was significantly up-regulated at higher temperature by a third member of this genus. Further, we developed a new proteomic approach that harnessed high-resolution mass spectrometry and supercomputing for direct identification and quantification of a broad range of PTMs from an AMD microbial community. We find that PTMs are extraordinarily diverse between different growth stages and highly divergent between closely related bacteria. The findings of this study motivate further investigation of the role of PTMs in the ecology and evolution of microbial communities. Finally, a computational approach has been developed to improve the sensitivity of phosphopeptide identification. Overall, the research presented in the dissertation not only reveals biological insights with existing quantitative proteomics methods, but also develops novel methodologies that open up new avenues in studying PTMs of proteins (e.g. PTM cross-talk).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
PRINCIPLES OF HIGH-RESOLUTION MASS SPECTROMETRY-BASED PROTEOMICS AND ITS APPLICATIONS IN SYSTEMS-LEVEL CHARACTERIZATION OF PROTEINS AND POST-TRANSLATIONAL MODIFICATIONS

1.1 Introduction to systems biology

One of the greatest scientific achievements in the 20th century was the discovery of the central dogma of molecular biology[1], which established the principle of biological information flow from DNA to RNA to protein in order to generate phenotypes. The establishment of the basic relationship among these biomolecules gave birth to the field of molecular biology, which is designed to address the fundamental question of where, when, and how the genetic information encoded in the genome of an organism is expressed and then translated into proteins to carry out biological functions, and how such biological information relay is spatially and temporally regulated?

Over the past decades, this molecular biology paradigm, based on the assumption of "one gene-one protein-one function"[2], has dominated biological research and provided a wealth of knowledge about the identity of molecular components and their functions. Despite its great success in the past, it is still very difficult to establish a direct link between genotype and phenotype[3]. Part of the reason is because a phenotype cannot be simply attributed to one single gene. It is becoming increasingly clear that molecular constituents of a cell do not function in isolation of one another; instead, they are organized as biological networks where the nodes of a network represent genes, RNAs, proteins, metabolites, or any other biomolecules and the edges between the nodes represent certain functional connection[4], such as enzyme-substrate interaction[5], or transcriptional factor-*cis* regulatory DNA interaction[6]. It is the structure and dynamics of these biological networks that collectively determine the phenotype of an organism.
Recently, a new paradigm in biology is emerging, which focuses on studies of the molecular components of a cell in the network perspective, referred to as systems biology[7,8]. Since its first introduction by Leroy Hood and colleagues in 2001, systems biology has embraced various branches of biological research and revolutionized our understanding of the molecular principles governing cellular functioning. In contrast to molecular biology, which usually focuses one or a

few gene(s) transcript(s), and protein(s) at a time, systems biology characterizes the entire biological molecules present in a cell or organism with ultimate goal of using such global information to built predicative mathematical model of a system[7].

Only limited and incremental progress in a scientific discipline would be made if there was no technology breakthrough. Just like the profound scientific impact of molecular biology would not have existed if polymerase chain reaction (PCR) technique had not been invented[9], the advancement of systems biology has been being powered by various omics-based technologies. The sequencing technologies created “Genomics”, which dissects an organism’s genetic blueprint and informs the potential functions that this organism is able to perform[10]. However, gene expression is dynamic and condition-dependent[11]. Just because a gene is encoded in a genome does not necessarily mean this gene would be expressed all the time (or at all). “Transcriptomics”, initiated by microarray technologies[12] and advanced by RNA-Seq technologies[13], profiles gene transcripts and provides insight into spatial-temporal gene expression dynamics.

Genomics and transcriptomics have significantly deepened our understanding of the molecular mechanism underlying the various biological processes. However, most biological functions and metabolic activities are carried out by proteins, and neither genomics nor transcriptomics would completely and accurately inform what proteins are present in a cell, which have closer connection to phenotypes than gene and RNA transcript. Furthermore, most protein activities are regulated by post-translational modifications, the chemical modification of proteins after translation[14]. The mass spectrometry technologies brought into being “Proteomics” which catalogs the entire proteins and their post-translational modifications present in a cell and provides a more direct measurement of molecular signatures underlying the phenotypes[15].

1.2 Mass spectrometry-based proteomics as a central approach in systems biology

1.2.1 Principles of mass spectrometry-based proteomics

The term, Proteomics, was first introduced by Wilkins and colleagues in 1996 with the goal of identifying all proteins present in a cell under a specific condition[16]. Despite a similar scale of ambition as genomics and transcriptomics, the technical challenges encountered by proteomics researchers were more formidable. Previous biotechnologies had already laid a solid foundation for genomics and transcriptomics for decoding the genome and transcriptome, respectively, primarily based upon the PCR which can be routinely used to amplify DNA fragments, so the sensitivity issue in the genome sequencing or gene chip-based RNA detection had been eliminated[17]. However, there was no such technology that could increase the copy number of proteins to benefit the protein measurement. Fortunately, the field of proteomics has been dramatically propelled forward by innovative advancements in new high-performance mass spectrometric technologies[15].

Nowadays, mass spectrometry has become an indispensable tool in proteomic measurement because it can be used to not only confidently identify proteins, but also accurately quantify their abundances. Such a pivotal role of mass spectrometry in proteomics was established due to the invention of two ionization techniques in 1980s: matrix assisted laser desorption/ionization (MALDI)[18] and electrospray ionization (ESI)[19]. Such technological breakthroughs eliminated the bottleneck of producing ions for large, non-volatile biomolecules, such as peptides and proteins, in the gas phase. Since the first large-scale proteome measurement of yeast by John Yates and colleagues in 2001[20], proteomics has undergone tremendous development, mainly driven by the continuous improvement on the performance of mass spectrometer, in terms of mass accuracy, speed, sensitivity, dynamic range, etc. Previously, it took ~68 hours to finish the first yeast proteome measurement with identification of 1484 proteins. Recently, with the latest generation of mass spectrometer-the Orbitrap Fusion mass spectrometer, Coon and colleagues were able to achieve much deeper measurement of the yeast proteome with identification of 3,977 proteins in 1.3 hours[21].

In a typical proteome measurement, whole proteins are extracted from cells. Because of the enormous complexity of the extracted protein sample and huge difference in the abundance between the most abundant protein and the least abundant protein (i.e. dynamic range), separation is normally required to reduce the sample complexity. In the early proteomic studies, two-dimensional gel electrophoresis (2D-GE)[16], which first separates protein by isoelectric focusing point and then by molecular size in a gel, had been commonly used to resolve the complex protein sample. Then, the gel is stained, and some selected spots that contain proteins are excised from the gel. Proteins within the gel spots can be identified by mass spectrometry. There are a few disadvantages in the 2D-GE-based approaches. Firstly, due to the complexity of proteome sample, SDS gel electrophoresis is unable to separate each individual protein species. Thus, a gel spot could contain a few different protein species. Such protein co-migration phenomenon would distort protein abundance quantification because the staining intensity of a spot is a composite of many different protein species. Secondly, this approach has poor separation for membrane proteins, limited dynamic range, and inability to identify proteins in all the spots. In addition, excising gel spot is time-consuming and label-intensive.

The instrumentation that couples liquid chromatography with mass spectrometry (LC-MS) has significantly accelerated the development of proteomics and made mass spectrometry a powerful analytical technology for proteome characterization. Various LC separation approaches have been developed, such as reverse phase liquid chromatography[20] that separates proteins or peptides based their hydrophobic interaction with C18 resins and hydrophilic interaction liquid chromatography[22] that separates proteins or peptide based on their hydrophilic interaction with anionic resins. The LC-MS configuration offers much better separation and higher throughput than the 2D-GE-based approach. Furthermore, this approach is highly automated, which requires much less human labor.

There are two types of proteomic approaches for protein characterization: top-down approach[23] and bottom-up approach (or shotgun approach)[24]. The top-down proteomics targets intact proteins. In this approach, various separations are usually carried out at the protein level to generate multiple fractions of a protein mixture. Then each fraction is introduced into mass spectrometry for protein sequence analysis. The advantage of the top-down approach is it can identify protein isoforms resulting from alterative splicing of mRNA and detect a combination of multiple PTMs present on the same protein molecule. The current high-throughput top-down proteomics is able to identify thousands of proteins, however, there are still significant challenges in this approach. First, it is much more difficult to separate proteins than peptides. Second, traditional separation methods still commonly used in the top-down approach, such as gel electrophoresis, result in limited recovery of proteins. Liquid chromatography-based separation offers less sample loss, but it works poorly for proteins with more than 500 amino acid residues. Third, identification of protein from mass spectrometric spectra requires fragmentation of protein to generate ladders of fragment ions for sequence interrogation. However, sequence-informative fragment ions needed for confident protein identification are often sparse in mass spectra.

The alternative approach that targets peptides is shotgun proteomics. Instead of separating and measuring intact proteins, proteins are digested into peptides that are separated by liquid chromatography and then sequenced by tandem mass spectrometry. The peptide sequences identified from tandem mass spectra are used to infer what proteins are present in the sample (Figure 1.1). The advantage of the shotgun approach over the top-down approach is that it is much easier to separate peptides than proteins. Because peptide is much smaller than a protein, it is also easier to generate a ladder of sequence-informative fragment ions for peptide sequence identification. However, because peptides are disconnected from their parent proteins during the enzymatic digestion in the shotgun proteomics approach, definitively inferring what proteins are present in a sample from their peptides is a non-trivial task because a peptide can be shared between multiple proteins[25]. This protein inference problem becomes even more severe when dealing with eukaryotic organisms that have undergone extensive gene duplication, resulting in high sequence redundancy. Furthermore, it is difficult to use shotgun proteomics approach to tease apart which protein isoforms resulting from alternative splicing are present in the sample, because identification of a peptide that is unique to a particular isoform may not always succeed. Despite these technical challenges, shotgun proteomics has been extremely successful for deep proteome characterization, not only qualitatively but also quantitatively. In 2011, Mann and colleague was able to identify and quantify ~10,000 proteins from a human cell line, demonstrating that the depth of proteome coverage by mass spectrometry-based proteomics was similar to that of transcriptome coverage by RNA-Seq[26].

Nowadays, it is routine to generate hundreds of thousands of tandem mass spectra within one day of measurement. Such magnitude of data precludes manual identification of peptides from spectra. However, database searching has automated the process of identifying peptides sequence from mass spectra[27]. During a database search, the protein sequence database predicted from the genome of an organism is usually input to a search algorithm, and each database protein is *in silico* digested into to characteristic peptides based on the cleavage rule of the protease used in an experimental digestion. Since the fragmentation pattern of a peptide is predictable, usually b- and y-ion series in collision-induced dissociation (Figure 1.2), the theoretical spectrum of each database peptide can be generated *in silico*. In order to identify the exact amino acid sequence from an experimental spectrum, a list of candidate theoretical spectra are scored to identify the best peptide-to-spectrum match. This spectrum-matching approach is the oldest and still the most

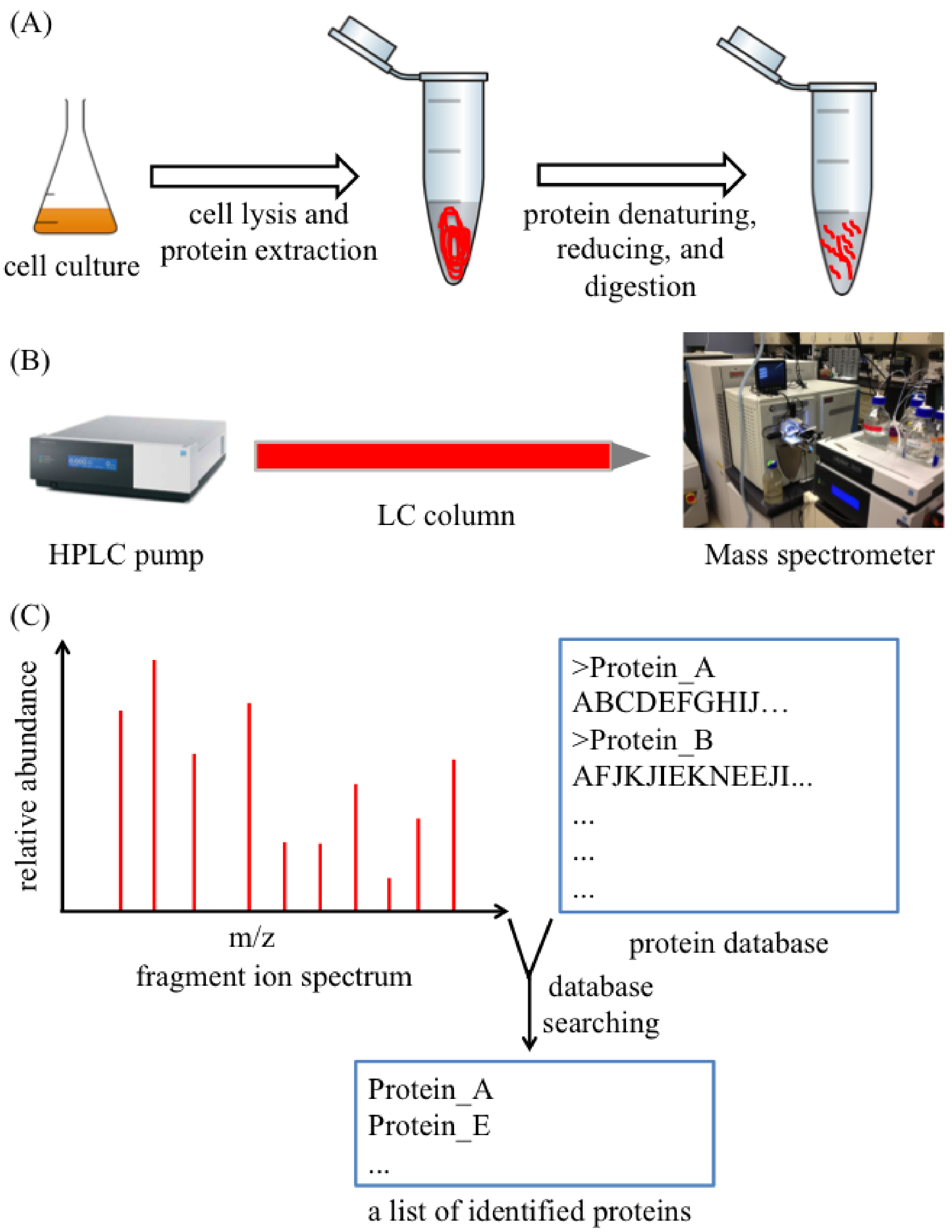


Figure 1.1 Workflow of shotgun proteomics. (A) Sample preparation. (B) Liquid chromatography-tandem mass spectrometry. (C) Database searching for protein identification.

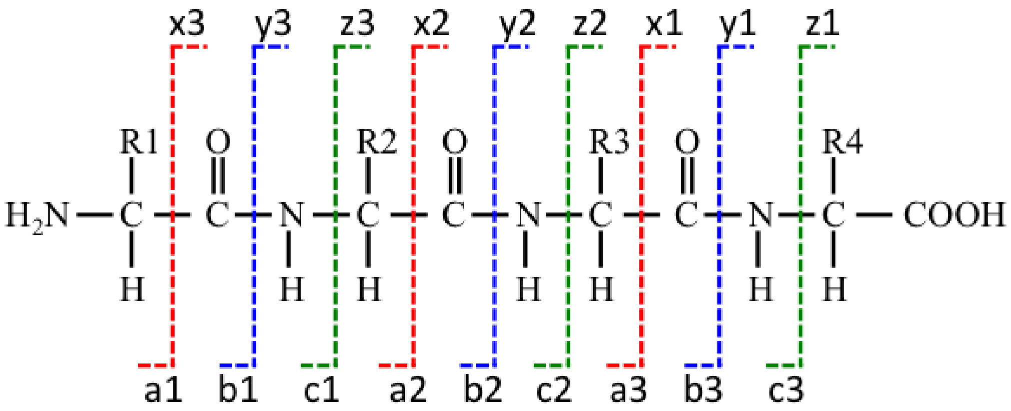


Figure 1.2 Nomenclature of fragment ions from a peptide. The fragment ions containing N-terminus are classified as a, b or c ions. The fragment ions containing C-terminus are classified as x, y or z ions. The number indicates the number of residues in a fragment ion.

widely used method for peptide identification. There are also a few other different database searching approaches, such as spectral library searching[28] and de novo sequencing[29].

Functional characterization of an organism requires not only cataloging what protein species are expressed in a cell but also quantifying the abundance change of these proteins in response to external and internal stimuli, because quantification of protein abundance changes could provide an insight into how an organism utilizes its protein machineries to deal with the stimuli. Mass spectrometry-based proteomics is a method of choice for quantitative characterization of proteome[30], because it can determine not only absolute copy number of protein in a cellular state, termed absolute quantification, but also relative abundance of each protein in different conditions, termed relative quantification, with high accuracy, precision, and reproducibility. There are two categories of quantitation methods in general: label-free methods and labeling-based methods. Label-free approaches correlate protein abundance with either the number of spectra that are identified for a protein (spectral counting) or mass spectrometric signal intensity of a protein. In contrast, with labeling-based approaches, different stable isotopes are incorporated into a series of samples, chemically, enzymatically, or metabolically, to generate different isotopic versions of the same organism. Then, the relative abundance of different versions of the same peptide, for example, ^{14}N-labeled peptide and ^{15}N-labeled peptide, are compared to quantify relative abundance change. In both label-free and labeling approaches, absolute quantification can be realized by spiking internal standards with known abundance. Then, the absolute abundance of a protein/peptide can be calculated by quantifying the relative abundance ratio between the protein/peptide and its internal standard.

1.2.2 Global characterization of PTMs by mass spectrometry-based proteomics

Cells are an information-processing unit that must quickly respond to various environmental stimuli in order to adapt to changing environmental conditions. Thus, an organism has developed complex, hierarchical regulatory systems that are able to adjust their molecular machineries for cell survival. For example, transcription factor binding to regulatory DNA sequence can modulate gene transcription[6]. RNA transcripts can undergo various post-transcriptional regulations before a protein can be made, such as alternative splicing which produces different mRNA isoforms from the same precursor[31], and RNA editing which can change final amino acid sequence of protein. MicroRNA binding to mRNA can affect the mRNA stability and protein translation[32]. After protein translation, post-translational modification can modulate various aspects of protein function[14]. These post-transcriptional and post-translational processing greatly enhances the functional potential of a genome. For example, the number of proteins encoded by the human genome surprisingly turned out to be about 20,000, which is not much bigger than that of the worm[33]. However, post-translational modifications vastly expand the chemical forms of a protein, which could be a molecular mechanism to compensate the relative paucity of protein-coding genes in a genome. For example, one protein can be modified with multiple types of PTMs and/or one type of PTM on multiple positions. Particularly, these protein modifications are combinatorial, meaning that the number of PTM isoforms of a protein would exponentially increase and different PTM isoforms could perform varied biological functions.

Mass spectrometry has been used to study post-translational modifications of purified proteins since 1974[34]. However, it was not until the early 2000's that proteome-wide PTM characterization became possible due to the development of various approaches for enriching PTM-containing peptides and the introduction of liquid chromatography into proteomics. Proteome-wide PTM characterization is more challenging than protein characterization because 1) PTM-containing peptides are usually low abundance. Thus, modified peptides have less chance to be sampled by mass spectrometer than unmodified peptides; 2) the presence of a chemical moiety on peptide creates a difficult in generating high quality spectra for confident peptide identification, because sequence-informative fragment ions are usually rare compared to unmodified peptides; 3) Identification of a modified peptide may not always guarantee the correct localization of PTM on the modified residue, because site-determining fragment ions are often missing in spectra.

Technical advancements in the past decade have begun to overcome these above-mentioned challenges. For example, various PTM enrichment techniques[35], such as antibody-based approaches and metal affinity-based approaches (e.g. TiO2 and IMAC), have been used to enrich low abundant modified peptides before the measurement. A range of fragmentation methods, such as neutral loss-triggered MS3[36], multistage activation[37], and higher-energy collisional dissociation[38], has been explored to generate information-rich fragment ion spectra for peptide sequencing. Sophisticated algorithms, such as phosphoRS[39] and Ascore[40], have been developed to localize PTM from tandem mass spectra. With the current high-performance mass spectrometry-based proteomics, it is now possible to identify tens of thousands of PTM events from a single study.

1.2.3 Proteomics in high-resolution mass spectrometry era

Currently, shotgun proteomics approach is still dominating the field. With this approach, a complex mixture of hundreds of thousands of peptide species with huge dynamic range is often measured in a few hours of chromatographic run. It is very common that hundreds of different peptide species could co-elute at a given chromatogram time point. Furthermore, a window with a few m/z wide is frequently isolated for fragmentation to determine peptide sequence. In such complex sample, it is highly likely that a few different co-eluting peptides with similar m/z are co-isolated and then co-fragmented to generate a multiplexed MS/MS spectrum. These multiplexed spectra are more prone to false positive unless the accurate mass measurement is used.

In the early 2000's, a proteomics researcher had to use 3-dimensional ion trap mass spectrometer with moderate sequencing speed, high sensitivity, but low mass resolving power. Such low-resolution mass spectrometer precluded accurate mass measurement, accurate charge state determination, and accurate quantification. In 2004, with introduction of hybrid linear ion trap-Fourier transform instrument (LTQ-FTICR-MS)[41], measuring proteomic sample with high resolution and high mass accuracy had become possible. This greatly boosted the data quality and increased the confidence of peptide identification. A few years later, a new generation of mass spectrometer, the LTQ-Orbitrap hybrid instrument[42], was introduced. Since then, proteome measurements have become quite common with a high-low approach: detection of precursor peptides occurs in Orbitrap with high resolution and high mass accuracy whereas detection of fragment ions occurs in LTQ with relative low resolution but high sequencing speed. Such a high-low approach is still the most widely used data acquisition method in proteome measurement.

In recent years, the high-high approach that measures both precursor peptides and fragment ions in Orbitrap with high resolution and high mass accuracy has begun to receive more attention. Mann and co-workers first demonstrated this high-high data acquisition approach for measuring post-translational modification on LTQ-Orbitrap platform using higher energy collisional dissociation[38]. One year later, a new generation of mass spectrometer, LTQ-Orbitrap Velos[43], was introduced, featuring a delicate higher energy collision cell. Then, the feasibility of large-scale phosphoproteomic study with the high-high approach on this new generation of mass spectrometer was demonstrated[44]. However, it took much longer time to acquire a high-resolution fragment ion spectrum in Orbitrap than recording a low-resolution spectrum in LTQ. The prolonged data acquisition time could adversely impact the depth of measurement, which led to the argument about whether the high-high approach would be suitable for the large-scale proteome or PTM characterization[45]. In 2012, the LTQ-Orbitrap Elite mass spectrometer was introduced with 4-fold increased scan rate of the Orbitrap mass analyzer[46]. Such an improvement made the duty cycle of the high-high approach comparable to that of the high-low approach, which allows similar depth of measurement between these two methods. However, the high-high approach offers much better specificity for database searching, which significantly decreases false positive identifications. For example, typically, 0.5 Dalton of fragment ion mass tolerance is allowed during the database searching of data acquired with the high-low approach. With the high-high approach, the mass tolerance can be specified as stringent as 0.01 Dalton, which is 50-fold more accurate when matching measured fragment ions with predicted ones.

1.3 Community proteomics of natural microbial consortia

1.3.1 Introduction to microbial community

Over the past centuries, microbiologists had focused on studying pure cultures using classic methodologies, such as microscopy, cell isolation and culturing, and recombinant DNA techniques. These approaches had tremendously advanced our understanding of microbial physiology and been frequently explored to benefit the human world, such as utilizing microbial fermentation for providing food and medicine, and harnessing microbial metabolism for environmental cleanup. It was not until the development of an approach to directly recover microbial DNA from natural environment that people began to realize centuries of culture-dependent microbial research only captured less than one percent of microbial species on the earth[47]. Fortunately, microbiology entered the systems biology era, which brought in various cultivation-independent approaches to study microorganisms in their natural settings. These approaches include metagenomics which sequences DNA directly recovered from free-living microorganisms[48], metatranscriptomics which measures *in situ* gene expression[49], and metaproteomics which characterizes expressed protein molecules[50].

The study of whole microbial communities is a daunting task, because of extraordinary biodiversity, and undefined structural and functional boundaries, of most natural communities. Thus, starting with a system with reduced complexity provides an opportunity to not only generate hypotheses that could be extrapolated to more complex systems, but also develop molecular methods that can be applied to characterizing those complex microbial communities.

Microbial biofilms growing in the Acid Mine Drainage (AMD) offer such a model system[51]. This natural environment contains high level of acid and toxic heavy metals, such as lead and arsenic, which creates a serious environmental problem. However, in such harsh environment, dozens of microbial species, mostly *Leptospirillum* group II bacteria, are able to thrive. They utilize Fe^{2+} released from pyrite dissolution to make energy and fix carbon for their growth. They also build microbial biofilms on the AMD solution-air interface. At the same time, the Fe^{3+} derived from the microbial oxidation of Fe^{2+} further drives pyrite dissolution, which continuously provides Fe^{2+} for the microbial growth. The life cycle of an AMD biofilm has three developmental stages. In the early developmental stage, carbon fixation carried out by bacteria is highly active and the biofilm begins to establish. As biofilm further grows, it will enter the late developmental stage where organic carbons synthesized in early stage are partitioned into heterotrophs that cannot fix carbon on their own. As biofilm further matures, it will be degraded by anaerobic archaea. Carbon fixed in the early development stage are respired and then released into environment as CO_2.

Decades of research on AMD microbial community has not only gleaned insight into the microbial evolution and ecology in nature, but also enabled the development of various culture-independent molecular techniques that have been adapted to study more complex microbial communities, particularly the community proteomics that allows functional characterization of microbial community by analysis of expressed proteins.

1.3.2 *in situ* functional analyses of microbial activities using community proteomics

Initial attempts to characterize proteins recovered directly from natural occurring microbial communities were hampered by the limited resolution of 2D-GE-based approach and lack of high-quality genomic databases that can be used to match experiment spectra with theoretical spectra. In 2004, the composite genomes of several co-existing microbial species in an AMD system were reconstructed through metagenomic sequencing[48]. This provided a high-quality genome database that allowed peptide and protein identification using automated database searching. Then, the first large-scale environmental proteomics study was made possible with liquid chromatography-mass spectrometry-based proteomics[50]. This study identified over two thousand proteins from five abundant organisms. Particularly, these identified proteins can be assigned to different organisms, permitting differentiation of functional activities among co-existing organisms. Since this pioneering study, community proteomics has been applied to studying microbial communities existing in other environments, such as soil[52], ocean[53], and human body[54].

It is now routine to identify thousands of proteins from low-to-medium complexity systems with community proteomics. However, it remains a significant challenge to achieve deep measurement of complex systems, such as soil environments, because 1) proteins are often co-extracted with soil contaminants, such as humic acids, which creates extra technical hurdles in the downstream sample processing and liquid chromatography-mass spectrometry measurement; 2) the enormous genetic diversity and redundancy within complex microbial community and short sequence reads generated from the current high-throughput sequencing technology often complicate genome assembling. This problem propagates into the proteomic database searching, which precludes peptide and protein identification even from high quality tandem mass spectra; 3) the size of a metagenomic database from a complex community could contain more than a million protein sequences. The database searching against such huge database is computationally time-consuming. Despite these technical hurdles, progress has been being made. For example, novel sample extraction approaches have been shown to overcome at least some of the difficulties in protein extraction from challenging environment[55]. Deep sequencing and new genome assembling algorithms would enable construction of high-quality metagenomic databases, which would aid in protein identification from mass spectrometry data[56]. High-performance computing would unblock the informatic bottleneck.

1.4 Objectives of dissertation

Various quantitative proteomics approaches have been developed, ranging from label free, metabolic labeling, to chemical labeling. Each approach has its own unique advantages and disadvantages. There have been a few studies that have compared different quantitative methods, however no study had compared all three methods simultaneously on the same analytical platform. Thus, the goal of the Chapter 3 is to head-to-head compare the identification and quantitation performance of these three above-mentioned approaches on the state-of-art Linear Ion Trap (LTQ) Orbitrap Velos mass spectrometer platform in order to provide guidance on how to choose an appropriate approach for a quantitative proteomic study[57].

Strigolactones are a new class of plant hormones. In addition to acting as a key inhibitor of shoot branching, they stimulate seed germination of root parasitic plants and promote hyphal branching and root colonization of symbiotic arbuscular mycorrhizal fungi. They also regulate many other aspects of plant growth and development. At the transcription level, strigolactones-regulated genes have been reported. However, nothing is known about the proteome regulated by this new class of plant hormones. Thus, the objective of the Chapter 4 is to apply isobaric chemical labeling-based quantitative proteomics approach to quantify the proteome regulated by the strigolactone in *Arabidopsis* seedlings and to identify potential target proteins for follow-up genetic and biochemical study[58].

Microbial communities are central to global carbon recycling and have direct feedbacks with the climate system. Nonetheless, their response to global change remains enigmatic largely due to the complex nature of their membership and metabolism. Our limited knowledge of microbial response to change, the difficulty of identifying and quantifying microbial function within complex samples, and the inability to link metabolisms directly to community members have proven to be major limitations in progressing integration of microbial carbon cycling into Earth System models. Thus, the objective of the Chapter 5 was to apply chemical labeling-based quantitative proteomics to determine how elevated temperature impacts the physiology of individual microbial groups in a community context, using a model microbial-based ecosystem.

Post-translational modifications play an important role in regulating protein function. Current global PTM studies by mass spectrometry-based proteomics generally rely on enrichment and can only target a specific type of PTM unless multiple enrichments are used. Also, the enrichment-based approach cannot directly quantify PTM fractional occupancy defined as the percentage of the copies of a protein that is modified with a specific PTM. Thus, the first objective of the Chapter 6 is to demonstrate a new proteomic approach that is able to simultaneously characterize a broad range of PTMs in microbial systems without enrichment. PTM-level regulation of protein activities in microbial community is largely unknown, thus the second objective of the Chapter 6 is to apply this demonstrated approach to profiling the diversity, dynamics, and divergence of PTMs in individual organisms within a natural AMD microbial community and to provide insight into the role of PTMs in microbial adaptation, evolution and ecology.

Protein phosphorylation is one of the most studied PTM type by mass spectrometry-based proteomics. Due to labile nature of phosphate group during fragmentation, fragment ions covering phosphorylated residue often suffer from various extent/form of neutral loss, resulting in a mass shift from their expected mass. Most database searching algorithms do not consider these neutral-loss fragment ions in their scoring functions, which could preclude identification of phosphopeptide when neutral loss fragment ions dominate the spectra. Thus, the objective of the Chapter 7 is to evaluate a neutral loss search algorithm for the improved sensitivity of phosphopeptide identification.

1.5 Overview of dissertation

The research in this dissertation demonstrates the feasibility of quantitative proteomic approaches that can be broadly used for characterization of proteins and post-translational modifications from complex proteome, focusing on acid mine drainage microbial community and *Arabidopsis*. This dissertation consists of following chapters: Chapter 2 describes experimental and bioinformatic methods. Chapter 3 compares three quantitative proteomics approaches: labeling free, metabolic labeling, and isobaric chemical labeling. Chapter 4 used isobaric chemical labeling-based quantitative approach to study how elevated temperature impacts organism's physiology in acid mine drainage biofilm growing in bioreactor. Chapter 5 also applied isobaric chemical labeling-based approach to profiling *Arabidopsis* proteome in response to the treatment with strigolactone. Chapter 6 demonstrated a new proteomic approach that combines high-performance mass spectrometry-based proteomics and high-performance for quantitative characterization of a broad range of PTM in both laboratory-grown *E. coli* and free-living microorganisms in acid mine drainage microbial community. Chapter 7 describes a bioinformatic approach for improving sensitivity of phosphopeptide identification in high-resolution mass spectrometry data. Chapter 8 serves as a conclusion of the research presented in the dissertation as well as the outlook of the future direction of mass spectrometry-base proteomics.

CHAPTER 2
EXPERIMENTAL AND BIOINFORMATIC APPROACHES

2.1 Overview of mass spectrometry-based approaches

In this dissertation, shotgun proteomics approach is used to characterize complex proteomes because 1) it is much easier to separate peptides in shotgun proteomics than proteins in top-down proteomics; 2) better separation allows improved identification of peptides from relatively low abundant proteins, which provides deeper proteome coverage; 3) tryptic digestion of proteins usually results in peptides with ideal length for generating more sequence-informative fragment ions during fragmentation, which enhances the confidence of peptide identification; and 4) there are more selection of bioinformatic tools for analyzing shotgun proteomic data.

The sample preparation starts with cell lysis to extract the whole protein, followed by denaturing, reducing, and digestion of the proteins. The resulting peptides are then loaded into back column off-line, separated with on-line two-dimensional liquid chromatography, and measured with tandem mass spectrometry. The raw mass spectrometric data are automatically searched for protein identification. Detailed experimental and bioinformatic procedures are described as follows.

2.2 Experimental approaches

2.2.1 Sample description

The general workflow of mass spectrometry-based proteomics starts with whole protein extraction from cells. As mass spectrometry is an extremely sensitive analytical technique and its performance can be adversely affected by various contaminants, such as salts, only a few micrograms of proteins with high purity (free of salts) are typically needed in one experiment. Because proteins and peptides are usually stable under low temperature, extracted proteins and digested peptides can be stored at -80 ^{o}C.

A variety of samples have been studied in this dissertation, including, *P. putida* F1, *Arabidopsis*, *E. coli* K12, and AMD microbial community. For the *P. putida* F1 samples in Chapter 3, cells were grown aerobically at 30 °C with vigorous shaking (200 rpm) in M9 minimal medium [2 mM $MgSO_4$, 0.1 mM $CaCl_2$, and 1X M9 salts (5X M9 salts contain per liter: 15 g KH_2PO_4, 2.5 g NaCl, 5 g NH_4Cl (normal NH_4Cl for unlabeled medium and 98%-enriched ^{15}NH4Cl for ^{15}N-labeled medium), 64 g $Na_2HPO_4 \cdot 7H_2O$)] supplemented with 50 mM (final concentration) of glucose. The M9 minimal medium was sterilized by autoclaving, and 1 M glucose stock solution was sterilized by passing through a 0.2 μm filter (Nalgene). Three cell cultures were grown identically. The unlabeled medium was used for cultures 1 and 2 and the ^{15}N-enriched medium was used for culture 3. Cells were harvested from the three cultures at the mid-log phase of growth (OD_{600} ~ 0.4).

For the *Arabidopsis* samples in the Chapter 4, wild-type Columbia-0 (*Col-0*) and mutants *max3-9* and *max3-12* (Salk_015785c) were obtained from the *Arabidopsis* Biological Resources Center (Columbus, Ohio). Seeds were surface sterilized by serial washing with 96% (volume / volume) ethanol, 20% (volume / volume) household bleach supplemented with 0.05% (volume / volume) Tween-20, and placed at 4°C for 2 days. Seeds were subsequently plated on ½ Murashige and Skoog (MS) medium supplemented with 1% (weight / volume) sucrose and 0.8% (weight / volume) agar, and germinated in 12 h/12 h photoperiod at 23°C, approx. 90 µmol photons m^{-2} s^{-1}. Four-day-old seedlings were transferred to fresh ½ MS supplemented with 1% (weight / volume) sucrose and 0.8% (weight / volume) agar plates with vertical growth orientation. A synthetic strigolactone GR24 was obtained from LeadGen Labs, LLC (Orange, CT) and a 10 mM stock solution was made in acetone. To monitor the time curve for the effect of GR24 treatment on root growth, four-day-old *Arabidopsis Col-0* seedlings were transferred to ½ MS supplemented with 1% (weight / volume) sucrose and 0.8% (weight / volume) agar plates plus 5 µM GR24 or acetone as mock treatment. Primary root length of the seedlings was measured at the beginning of the treatment and in 6 h intervals for 24 h in total. For the proteome analysis, 14-day-old seedlings were treated with 5 µM GR24 for 12 hours. Fifteen whole seedlings of vertically-grown *Col-0* and *max3-12* were picked from plates, transferred to 50 ml tubes and incubated in 20 ml liquid ½ MS supplemented with 1% (weight / volume) sucrose plus GR24 or acetone as mock treatment. Treatment was performed on a roller shaker in the darkness and was started at the beginning of the dark period of plant growth to potentially avoid the accumulation of large amount of light-regulated proteins. After incubation, plants were harvested, thoroughly washed with water, dried on filter paper, snap-frozen in liquid nitrogen and stored at -80°C until protein or RNA extraction. Three biological replicates were sampled for each treatment.

For the AMD microbial community samples in Chapter 5, biofilms were collected from the AB Muck Dam site at the Richmond Mine on 7/15/11, where pH is typically 0.85. For cultivation, biofilms were stored on ice for return to the laboratory. For community analyses of the biofilm inoculum used for cultivation, biofilms were flash-frozen on-site in a dry ice/ethanol bath and then transferred to –80 °C upon return to the laboratory. Biofilms were cultured in bioreactors using 9K-BR growth media. The flow rate of the bioreactors was approximately 200 µL/min. Incubator temperature was monitored using HOBO® Pendant Temperature Data Loggers. After four weeks of biofilm development at 40 °C, biofilms were regrown at 40ºC, 43 ºC, 46 ºC, and 49 ºC in separate reactors. Biofilms were harvested after three weeks and then reestablished before a second harvest five weeks later, representing two growth phases.

For the AMD community samples in Chapter 6, early growth stage and late growth stage biofilms were sampled the AB Muck Dam site at the Richmond Mine on 9/17/2010 (pH ~1, 39 °C). For the *E. coli* K-12 sample, cells were cultivated aerobically with constant agitation (250 rpm) at 37 °C in Luria-Bertani medium (pH 7.2). Cells were harvested when the culture reached an O.D. of 0.8.

2.2.2 Cell lysis and protein sample preparation

Since this dissertation covers different biological systems, such as microbial isolate, microbial communities, and plant, sample preparation methods have been tailored to specific sample in order to achieve maximum protein yield. For the research presented in Chapter 3, in label-free quantification, two standard samples were prepared using 1 g of cell pellet from culture 1 and 1 g of cell pellet from culture 2 of *P. putida* strain F1. For metabolic-labeling quantification, one standard sample was prepared by mixing 1 g of unlabeled cell pellet from culture 1 and 1 g of ^{15}N-labeled cell pellet from culture 3. For isobaric chemical labeling-based quantification, two peptide samples were prepared separately using 1 g of cell pellets from cultures 1 and 2, respectively, and were mixed after labeling. For each quantification method, all quantified proteins were expected to have an abundance ratio of 1:1. Cells were lysed by sonication with 2-minute duration at 20% amplitude (5 seconds on and 10 seconds off) in 6 M guanidine and 10 mM dithiothreitol (DTT). The extracted proteins were precipitated by chilled acetone. Protein pellets were obtained by centrifugation (21,000 g), air-dried, and then resolubilized in triethylammonium bicarbonate (TEAB) buffer. Protein concentration was measured by a bicinchoninic acid (BCA) (Thermo Scientific) assay, following the manufacturer's protocol. Sequencing grade trypsin (Promega, Madison, WI) was added at 1:100 (weight/weight) into proteins in TEAB buffer supplemented with 10 mM CaCl2 (final concentration). The first digestion was run overnight at 37 °C, and after adding additional trypsin at 1:100 (weight/weight) into proteins, the second digestion was run for 5 h at 37 °C. Finally, the samples were reduced with 10 mM DTT for 1 h at 60 °C and desalted using C18 solid-phase extraction (Sep-Pak Plus, Waters, Milford, MA). A BCA assay was conducted to determine peptide concentration in order to make sure that equal amount of peptides from each sample are labeled with an isobaric chemical labeling reagent as described later. BCA assay was originally designed for estimation of protein concentration based on the reduction of Cu^{2+} to Cu^{+} by amide bonds in proteins. Since amide bonds also exist in peptides, the BCA assay should be applicable to estimation of peptide abundance. All peptide samples were stored at -80 °C until ready to use.

For the research presented in Chapter 4, total proteins were extracted from frozen seedlings using a modified method for plant proteomes[59]. In brief, *Arabidopsis* seedlings were ground to powder in liquid nitrogen along with ~10 mg of polyvinylpolypyyrolidone and suspended in 2 ml chilled acetone containing 0.07% (volume/volume) beta-mercaptoethanol (β-ME) and 10% (weight/volume) trichloroacetic acid. The extract was kept at -20°C overnight followed by centrifugation at 21,000 g for 20 min. The resulting pellet was retained and washed with chilled acetone containing 0.07% β-ME three times with brief centrifugation (5 min, 21,000 g) between washes. The washed pellet was air dried, solubilized in 6 M guanidine HCl supplemented with 10 mM DTT and incubated at 60 °C for 3 h with intermittent mixing. The protein content of each sample was estimated using the reducing agent compatible/detergent compatible (RC/DC) protein assay kit (Biorad, Hercules, CA) as per the manufacturer's protocol. The total protein concentration was used as guideline for trypsin assisted proteolysis. Cysteines were blocked by adding 20 mM iodoacetamide for 15 min at room temperature. Samples were diluted six-fold using Tris-$CaCl_2$ buffer, pH 8.5 (50 mM Tris, 10 mM $CaCl_2$). Subsequently, the proteins were proteolysed by adding modified sequencing grade trypsin (Promega, Madison, WI) at 40 µg enzyme per mg protein and incubated overnight at 37 °C with gentle mixing. Samples were desalted using reverse-phase solid phase extraction (C18 SepPak, Waters, Milford MA) and solvent exchanged. All peptide samples were stored at -80 °C until ready to use.

For the research presented in Chapter 5, proteins were extracted from AMD biofilm using the sodium dodecyl sulfate (SDS)-boiling method. Between 500-750 mg of frozen biomass was resuspended in 1 mL SDS cell Lysis Buffer (5% SDS; 50 mM Tris-HCl, pH 8; 150 mM NaCl; 0.1 mM EDTA; 1 mM $MgCl_2$) and 10 µL of 5 M DTT. The biofilm was dispersed in the buffer by vigorously vortexing for 2-3 minutes. Samples were heated at about 100 °C for 15 minutes, followed by vigorous vortexing for 3 minutes. Cellular debris was pelleted by centrifugation at 21,000 g for 10 minutes at 4 °C. The supernatant was transferred to a fresh tube, 300 µl cold 100% trichloroacetic acid was added, and the proteins precipitated overnight at 4 °C. Precipitated proteins were centrifuged at 21,000 g for 20 minutes at 4 °C and the concentrated protein pellet washed three times with cold acetone. The pellet was resuspended in a guandinium chloride buffer (6 M guanidium chloride, 10 mM $CaCl_2$, 50 mM Tris pH 7.6) and reduced with 10 mM DTT. Total protein concentrations were estimated with the BCA assay. 50 µg of protein from each sample was further processed with the Filter-aided Sample Preparation (FASP) method following the manufacturer's protocol (Expedeon, CA) with a minor modification by substituting urea with triethylammonium bicarbonate (TEAB) buffer for sample washes to avoid the primary amine group-containing chemical that would interfere with isobaric chemical labeling. Each sample was digested with sequencing-grade trypsin (Promega, WI) in 500 mM TEAB buffer overnight in an enzyme: substrate ratio of 1:100 (weight: weight) at room temperature with gentle shaking, followed by a second digestion for 4 hours. Then the digested peptide samples were eluted off the filter by centrifugation and then were stored at -80 °C until ready to use.

For the research presented in Chapter 6, AMD biofilm samples and *E. coli* sample were prepared similarly, using the SDS-boiling/FASP-based methods as described for the Chapter 5. Biofilm samples were digested with trypsin (Promega), Lys-C (Roche), and Glu-C (Roche) in parallel. *E. coli* sample was digested with trypsin and Lys-C in parallel.

2.2.3. Isobaric chemical labeling

For research presented in Chapter 3, the two iTRAQ (isobaric Tag for Relative and Absolute Quantification) samples of *P. putida*, each containing 100 µg of peptides, were labeled using iTRAQ116 and iTRAQ117 following the manufacturer’s standard protocol. The two samples were then mixed, yielding the standard sample for iTRAQ. Similarly, the two TMT (Tandem Mass Tag) samples, each containing 100 µg of peptides, were labeled using TMT126 and TMT127 following the manufacturer’s protocol. The two samples were then mixed, yielding the standard sample for TMT.

For research presented in Chapter 4, 100 µg peptide from each set of four samples (i.e. *Col-0* without strigolactone treatment, *Col-0* with strigolactone treatment, *max3-12* without strigolactone treatment, and *max3-12* with strigolactone treatment) was labeled using iTRAQ114, iTRAQ115, iTRAQ116, and iTRAQ117 reagents according to the manufacturer's protocol, respectively, and then combined into one aliquot at a ratio of 1:1:1:1.

For research presented in Chapter 5, a total of 12 samples were labeled by the TMT 6-plex reagent in the following scheme: for the 6 samples in the replicate A that included 40C-1A(note: the first number denotes the growth temperature of AMD biofilm and the second number denotes the growth phase), 43C-1A, 46C-1A, 40C-2A, 43C-2A, and 46C-2A, 50 μg of each was labeled with TMT126, TMT127, TMT128, TMT129, TMT130, and TMT131, respectively; and for the 6 samples in the replicate B that included 40C-1B, 43C-1B, 46C-1B, 40C-2B, 43C-2B, and 46C-2B, 50 μg of each was labeled with TMT126, TMT127, TMT128, TMT129, TMT130, and TMT131, respectively. After the labeling was finished, the 6 samples in the same replicate were combined into one aliquot at a ratio of 1:1:1:1:1:1.

2.2.4 Liquid chromatography

For all research presented in this dissertation, multi-dimensional liquid chromatography, first introduced by Yates and colleagues[20], was used to separate the complex peptide samples. This separation strategy couples strong cation exchange (SCX) as the first dimension with reverse phase as the second dimension. In SCX, positively charged peptides bind to negatively charged functional groups of SCX resins, typically sulfonate. Peptides are differentially eluted from SCX resins by increasing the concentration of salt, typically ammonium acetate, in mobile phase. The elution order of a peptide in SCX depends on its isoelectric point: the higher the isoelectric point is, the later it elutes. In reverse phase chromatography, peptides bind to C18 resins due to hydrophobic interaction. Differential elution of peptides is carried out by increasing organic solvent concentration, typically acetonitrile, in the mobile phase. The elution order of a peptide in reverse phase chromatography is based on its hydrophobicity: the higher the hydrophobicity is, the later it elutes.

Specifically, in each run, 10~25 μg of peptides were loaded offline into a 150-μm-I.D. 2-dimensional back column (Polymicro Technologies) packed with 3 cm of C18 reverse phase resin (Luna, Phenomenex) and 3 cm of strong cation exchange (SCX) resin (Luna, Phenomenex). The back column loaded with peptides was de-salted offline with 100% Solvent A (95% H_2O, 5% CH_3CN, and 0.1% formic acid), and washed with a 1-hr gradient from 100% Solvent A to 100% Solvent B (30% H_2O, 70% CH_3CN, and 0.1% formic acid) to move peptides from reverse phase resin to SCX resin. Then, the back column was connected to a 100-µm-I.D. front column (New Objective) packed in-house with 15 cm of C18 RP resin and placed in-line with a U3000 quaternary high-performance liquid chromatography pump (Dionex). Each run was configured with 11 SCX fractionations using 5%, 7%, 10%, 12%, 15%, 17%, 20%, 25%, 35%, 50%, and 100% of Solvent D (500 mM ammonium acetate dissolved in Solvent A). Each SCX fraction was separated by a 110-min reverse gradient from 100% Solvent A to 60% Solvent B. The LC eluent was directly nanosprayed (Proxeon) into mass spectrometer (Thermo Scientific) with a flow rate of ~300 nL/min.

2.2.5 Mass spectrometry

2.2.5.1 Analytical figures of merit

An mass spectrometer is composed of three indispensable parts: the ionization source which generates ions for analytes in gas phase and delivers them into mass spectrometer, the mass analyzer which measures the mass-to-charge ratio (m/z) of ions, and detector which records the measured m/z value of ions. There are many different types of mass spectrometers that use different ways to ionize, analyze, and detect analytes. However, when choosing an appropriate instrumentation for a proteomic measurement, the following analytical figures of merit should be considered[60]:

1: mass resolution: the ability to tease part different ions with similar m/z - calculated by the full width at half-height of a peak.

2: sensitivity: the amount of signal gain with increasing concentration of analyte.

3: dynamic range: the ability to measure the concentration difference between the most abundant and least abundant analytes.

4: speed: the amount of time that is taken to collect a spectrum.

5: precision: the variation in ion abundance when the same analyze is measured multiple times.

6: mass accuracy: the difference between the measured mass and calculated mass.

7: mass range: the m/z range that can be measured by a mass analyzer.

8: detection limit: the lowest quantity of an analyte needed to generate a signal larger than background noise level.

In this dissertation, LTQ-Orbitrap family mass spectrometers were used for proteomic measurements. Such versatile instrumentation allows measuring ions with high mass accuracy, high resolution, and smaller low-end of mass range in Orbitrap and with high speed, high sensitivity in LTQ ions (Table 2.1). In previous generations of LTQ-Orbitrap mass spectrometers, the speed of measuring ions in Orbitrap is lower than that in LTQ. However, since the LTQ-Orbitrap-Elite mass spectrometer has been introduced, the scanning speed of Orbitrap becomes comparable to that of LTQ.

2.2.5.2 Ionization techniques

The first step in mass spectrometric analysis of molecules is to generate ions from analytes. However, producing ions in gas phase for biomolecules had long been the bottleneck that impeded the wide application of mass spectrometry in biology. In the late 1980s', the development of the two soft ionization techniques: matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) that is able produce gas phase ions from proteins and peptides, paved the way for the rapid development of the proteomics in the early 2000s.

In MALDI ionization techniques, analytes of interest are spotted on a MALDI plate. Firing UV laser at the plate causes desorption and then ionization of matrix. Ionized matrix is thought to transfer a proton into analyte for ionization. Since the laser is fired intermittently and singly charged ions are often generated, MALDI is often coupled to pulsed mass analyzer with high mass range, such as Time-of-Flight.

ESI is more frequently used in the proteomic measurement. During ESI process, mobile phase containing volatile solvent and pre-charged analytes is sprayed into fine aerosol, which is driven by the high voltage (2 – 4 kV) applied between the emitter and the inlet of a mass spectrometer. As extensive solvent evaporation occurs, the size of droplets shrink until they reach the Rayleigh limit where Coulombic explosion occurs and droplets fall apart, producing smaller droplets that repeat this process and ions that enter mass spectrometer for mass analysis. ESI is especially suitable for analyzing large biomolecules, such as proteins and peptides, because it produced multiply charged ions with relatively small m/z that falls within the optimal mass range of some mass analyzer (e.g. LTQ).

In this dissertation, a variant of ESI-nanospray ionization that operates with flow rate of 200 nl/min was used. The use of such low flow rate produces smaller droplets and improves ionization efficiency and sensitivity.

Table 2.1 Performance metrics of the mass spectrometer used in this dissertation.

Instrument	Mass resolution	Sensitivity*	Dynamic Range	Speed (# of MS/MS per second)	Mass Accuracy	Mass Range	Detection Limit	Precision
LTQ Orbitrap Pro	7,500-100,000	s/n: 100:1	5,000 within a scan	10	<1 ppm	50 - 2,000	Attomole-Femtomole	high
LTQ Orbitrap Elite	15,000-240,000	s/n: 100:1	> 5,000 within a scan	12	< 1 ppm	50 - 2,000	Attomole-Femtomole	high

* evaluated with 2 µL of a 50 fg/µL solution of reserpine (100 femtograms total) injected at a flow of 500 µL/min. Some performance metrics were from http://planetorbitrap.com/

2.2.5.3 Mass analyzer

Each mass spectrometer must be equipped with a mass analyzer that measures the mass-to-charge ratio of ionized analytes. There are many different types of mass analyzers, but LTQ and Orbitrap have emerged as the most popular ones in proteomic measurement (Figure 2.1). LTQ can either operate stand-alone in LTQ mass spectrometer or be hybridized with Orbitrap to build LTQ-Orbitrap mass spectrometer (Figure 2.1A). Ions in LTQ are trapped radially by a two-dimensional radio frequency (RF) potential, and axially by a direct current (DC) potential applied to electrodes at the entrance and exit. Manipulation of ions, such as transmission, isolation, and activation can be realized by changing the RF and DC potential. LTQ offers fast scanning speed that is crucial for deep proteomic measurements and high sensitivity that are critical to sequencing low abundant peptides. However, since LTQ has limited mass accuracy and resolution, ambiguity in peptide identifications could arise when measuring complex mixture. Recently, the dual-pressure linear ion trap mass spectrometer-LTQ Velos was introduced[43]. Two ion traps with differential pressure are installed (Figure 2.1B). The first ion trap operates at high pressure, which allows more efficient ion trapping, isolation, fragmentation and the second one at low pressure, which enables faster spectra acquisition. The commercial introduction of Orbitrap mass analyzer in 2005 has revolutionized the mass analysis for proteomic measurement[42]. The Orbitrap consists of an outer barrel-like electrode and an inner spindle-like electrode, coaxial with each other. In contrast to linear ion trap, ions in Orbitrap are confined radially around the inner spindle-like electrode by electrostatic attractions, instead of using RF as in linear ion trap. Oscillation frequencies of ions are measured by acquisition of time-domain image current transients and then converted to mass spectra by fast Fourier transforms. Orbitrap features high resolution, high mass accuracy which significantly reduces the false positive identifications in proteomics. The lower mass cutoff in Orbitrap than LTQ detection makes it compatible with isobaric chemical labeling-based quantification. The disadvantage of Orbitrap was it requires longer time for spectra production than LTQ, which lowers the duty-cycle of measurement. However, the introduction of stand-alone Orbitrap (Q-Exactive)[61] and Orbitrap Elite[46] with significantly increased scanning speed made the sequencing speed comparable between these two mass analyzers.

For research presented in Chapter 3, 4, and 5, all precursor and fragment ion mass analyses were performed in LTQ and Orbitrap, respectively. Reporter ion mass analyses for

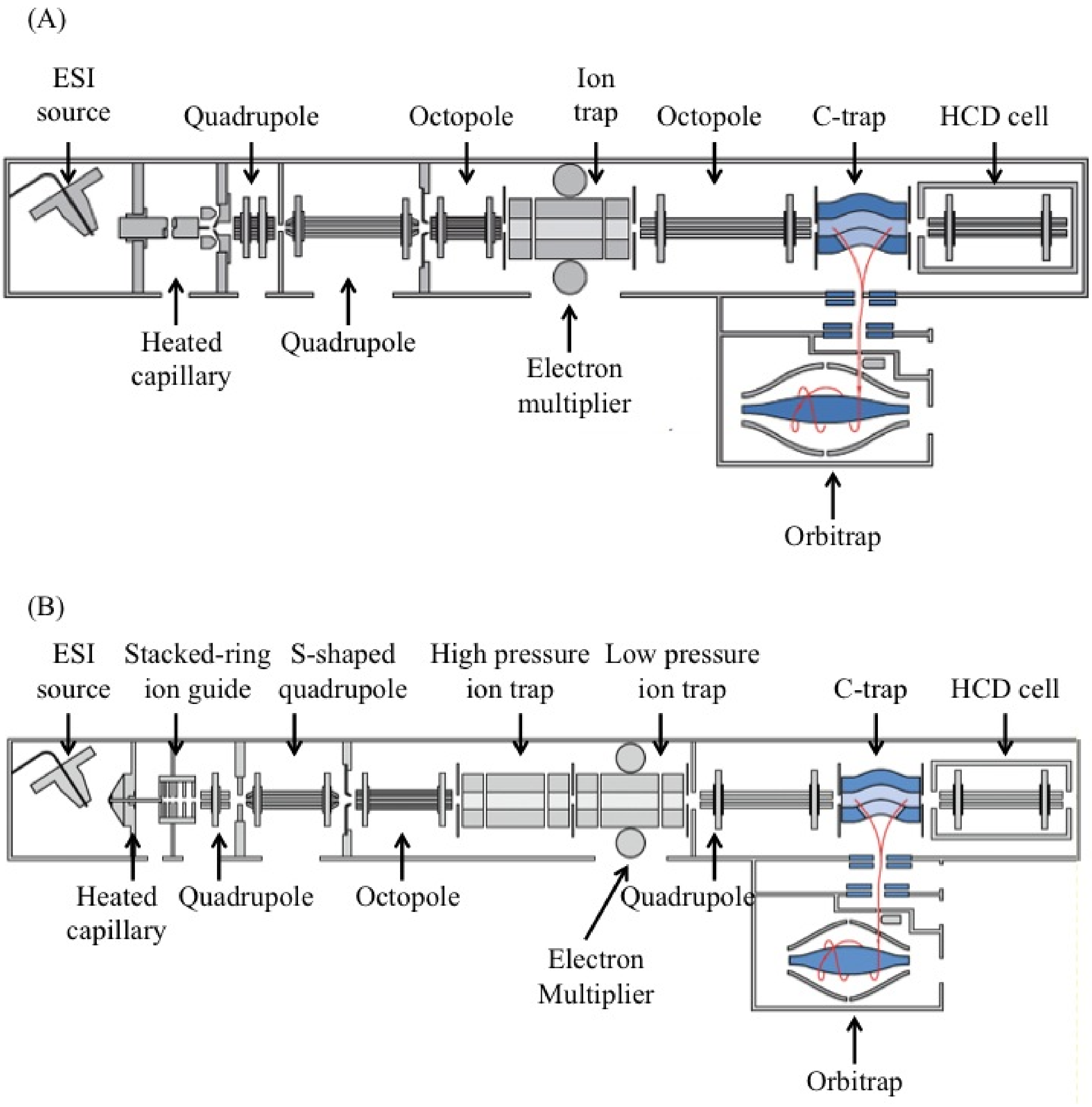


Figure 2.1 Schematics of LTQ Oribtrap XL mass spectrometer (A) and LTQ Orbitrap Velos/Pro/Elite mass spectrometer (B). The cartoons are adapted from http://planetorbitrap.com/

isobaric chemical labeling-based quantification were performed in Orbitrap. For research presented in Chapter 6, both precursor and fragment ion mass were analyzed in Orbitrap.

2.2.5.4 Fragmentation methods

Because of the huge complexity of proteome sample, precursor peptide mass detection is coupled with fragment ion mass measurement for more definitive peptide identification. In order to avoid repetitively fragmenting relatively high-abundant peptides, intelligent data-dependent acquisition (DDA) strategy has been devised, which utilizes the ion information in full scan to determine target ions for triggering MS/MS experiments. In a typical proteomic measurement, DDA is often realized by designing a Top N method where a full scan is followed by sequential isolation of top N (usually N = 10~20) most abundant peptide species for fragmentation. Such Top N method is often combined with the dynamic exclusion method where already fragmented peptide ions are temporarily excluded from MS/MS experiment for certain period of time (e.g. 60 seconds) in order to maximize the diversity of peptides being sequenced.

There are a few different types of fragmentation approaches used in proteomics[62], including collisional-induced dissociation (CID), higher-energy collisional dissociation (HCD), and electron transfer/capture dissociation (ETD/ECD).

In CID, peptide ions are accelerated in the electric field and collided with neutral gas, such as helium. The collision converts kinetic energy generated in the acceleration to internal energy, which results in chemical bond breakage, usually the amide bond in peptides. The fragmentation produces a ladder of fragment ions from a precursor peptide, which can be used to infer the amino acid sequence. CID is the most commonly used fragmentation method in proteomics. However, because of the 1/3 rule where fragment ions with m/z smaller than 1/3 of precursor ion m/z are not stable, this fragmentation method is not suitable for isobaric-chemical labeling-based quantification.

In HCD, precursor ions are isolated in ion trap and then injected, via the C-trap, to an octopole collision cell where precursor ions are collided with nitrogen gas to generate fragment ions. The fragment ions are then sent to Orbitrap via the C-trap for mass measurement. HCD is becoming more popular in proteomic measurement, not only for regular peptide identification, but also post-translational modification identification, because Orbitrap measures HCD-generated fragment ions with high resolution, high mass accuracy, and no constraint from the 1/3 rule and HCD spectra contain richer fragment ions, which provides more information content for peptide identification.

For the research presented in Chapter 3, MS data were acquired with following configurations: ten CID MS/MS scans per full scan for label free-based quantification; six CID MS/MS scans per full scan for metabolic labeling-based quantification; and four data-dependent CID-HCD dual MS/MS scans per full scan for isobaric chemical labeling-based quantification. In the CID-HCD dual scan, each selected parent ion was first fragmented by CID and then by HCD. Such configuration combines the benefit of using high-speed detection of CID-generated fragment ions in ion trap for peptide identification and Orbitrap detection of HCD-generated low mass reporter ions for quantification. The research presented in Chapter 4 and 5 also used the CID-HCD dual scan mode. The research presented in Chapter 6 used the HCD for fragmentation.

2.3 Bioinformatics

2.3.1 Peptide identification

The experimental part of mass spectrometry-based proteomics is to generate mass spectra which can be used for peptide identification. Typically, one day of measurement can produce hundreds of thousands of tandem mass spectra. Such magnitude of the data makes it impossible to decode amino acid sequence from spectra manually. However, numerous peptide identification software have been developed, such as Sequest[63], Mascot[64], Myrimatch[65], and Sipros[66-68]. By reversing target protein sequences and then appending the resulting decoy sequences to the original protein database for searching, false positive identifications can be controlled through adjusting the score threshold until the percentage of decoy identifications is reduced to a desirable level[69,70] (usually 1%). The false positive rate is usually calculated by
[2 x (the number of decoys)] / (the number of targets + the number of decoys).

In the research presented in Chapter 3, 4, and 5, Sequest was used for peptide identification. Specifically, all MS Raw files were converted into MS2 flat file using Raxport and then searched with following parameters: precursor mass tolerance of 3 Da; fragment ion mass tolerance of 0.5 Da; full enzymatic cleavage specificity required. Reverse sequences were appended into the protein sequence database for estimation of false positive rate (FDR). In the Chapter 3, all MS/MS spectra were searched against the *P. putida* F1 genome database (released in 2010) containing a total of 5251 predicted proteins and 44 common contaminants (trypsin, keratin, etc.). 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 second search was used to evaluate the labeling efficiency. In Chapter 4, all MS/MS spectra were searched against a *Arabidopsis* thaliana proteome database which contains 27,416 protein sequences from the January 2011 TAIR10 annotation (TAIR10_pep_20110103_representative_gene_model), 1,413 small proteins, 36 common contaminants, and 28,865 reverse sequences of all these proteins for a total of 57,730 entries. Cysteine blocking by iodoacetamide was specified as a static modification. Static modification at the N-terminus and dynamic modification at lysine residues by the iTRAQ labeling reagents were specified for peptide identification. An additional search with no iTRAQ modification specified was performed for the test run to estimate labeling efficiency. In the Chapter 5, all MS/MS spectra were searched against a database containing 79,633 proteins obtained from previous genomic characterizations of acid mine drainage biofilms sampled from the Richmond Mine AMD system. Static modification of cysteine by iodoacetamide and static modification of N-terminus, and dynamic modification of lysine by the TMT labeling reagent were specified for peptide identification.

For research presented in the chapter 6 and 7, Sipros was used for peptide identification. All Raw files were converted to FT1 and FT2 flat files using Raxport and then searched against either the *E. coli* K12 MG1566 genome database or the AMD metagenome database using Sipros (http://sipros.omicsbio.org). A broad range of PTMs were dynamically searched, including oxidation of methionine, hydroxylation of proline and lysine, deamidation of asparagine and glutamine, citrullination of arginine, mono-methylation of arginine, lysine, aspartic acid, and glutamic acid, di-methylation of arginine and lysine, tri-methylation of lysine, phosphorylation of serine, threonine, tyrosine, histidine, and aspartic acid, acetylation of lysine, s-nitrosylation of cysteine, nitration of tyrosine, methylthiolation of aspartic acid, and alkylation of cysteine by iodoacetamide. Although oxidation and hydroxylation introduce identical mass shift, so do deamidation and citrullination, these isobaric PTMs can be distinguish by localization: when methionine co-exists with proline or lysine on the same peptide, PTM localized to proline or lysine was considered as hydroxylation, otherwise it was ignored; when asparagine or glutamine co-exists with arginine on the same peptide, PTM localized to arginine was considered as citrillunation, otherwise it was ignored. Because some UBA-type peptides and 5wayCG-type peptides only differ in one single amino acid and some of these amino acid mutations introduce identical mass shift as PTM (e.g. aspartic acid/glutamic acid mutation is isobaric as mono-methylation), in order to be conservative, we selected the unmodified peptide when the top-rank modified peptide and unmodified peptide have tied scores for a spectrum, or selected the peptide with the fewest PTM when there were multiple top-rank modified peptides with tied scores. In order to handle the exponentially expanded search space, the broad-range PTM searches were conducted with scalable database searching algorithm Sipros 3.0 on the Titan, a supercomputer located in the Oak Ridge National Laboratory. A run of the trypsin-digested GS2 sample was initially searched against the complete AMD database. Based upon the preliminary search results, species that were not significantly detected in the sample were excluded in order to reduce computational search space. The final database contains 15,523 proteins from 7 organisms, including *Leptospirillum* group II UBA, *Leptospirillum* group II 5wayCG, *Leptospirillum* group III, and several archaea. All runs were searched with the following parameters: parent mass offsets of -1, 0, +1, +2, +3 Da; 0.03 Da and 0.01 Da of mass tolerance for parent ions and fragment ions, respectively; up to 3 missed cleavages; a maximum of 2 PTMs per peptides, and full enzyme specificity required. For the HCD run and CID run that were used to compare the identification performance of CID with HCD, only oxidation of methionine, and deamidation of asparagine and glutamine were dynamically searched, and alkylation of cysteine by iodoacetamide were statically searched on a desktop computer. The fragment ion mass tolerance for the CID run and HCD run was 0.5 Da and 0.01 Da, respectively. All the other search parameters were kept the same as mentioned above. Particularly for the phosphorylation search, we implemented neutral loss search algorithm in Sipros. For each predicted phosphopeptide, we generated three types of fragment ion spectra *in silico* to search: 1) phosphorylation without neutral loss, 2) phosphorylation with neutral loss of HPO_3 where the mass of the HPO_3 (79.966331 Da) was subtracted from fragment ions that contained modifiable residues, and 3) phosphorylation with neutral loss of HPO_3 and H_2O where the mass of the HPO_3 and H_2O (97.9769 Da) was subtracted from fragment ions that contained modifiable residues.

To evaluate the accuracy of PTM localization by Sipros, we also re-searched the first 20 libraries of synthetic peptide/phosphopeptide with known phosphorylation sites from a published study. Each library's MS/MS data was searched against its matched peptide database with concatenated reverse sequences. The database searching parameters were kept the same as the broad-range PTM search, but only oxidation of methionine and phosphorylation of serine, threonine, and tyrosine were dynamically searched. The identified phosphorylated spectra with DeltaP of 0 were ignored. Of the remaining modified spectra, 97% had correct PTM localization.

After the database searching is finished, the search result is filtered in order to control false positive identifications. In the research presented in the chapter 3, 4, and 5, DTAselect[71] was used for post-search filtering with the following parameters: minimum XCorr score of 1.8, 2.5, and 3.5 for charge states (z) = +1, z=+2, and z=+3 precursor peptide ions, respectively; a minimum DeltCN value of 0.08. For the research presented in the chapter 6 and 7, Sipros-Post was used for the post-search filtering to achieve 1% FDR at the peptide level.

2.3.2 Protein inference

In shotgun proteomics, proteins are unlinked from peptides due to enzymatic digestions. Thus, proteins present in a sample have to be inferred from a list of identified peptides. Protein inference is non-trivial task because sequence redundancy between proteins results in many non-unique peptides that are shared among some proteins, which makes it difficult to definitively infer which protein(s) are actually expressed in the sample. In the research presented in the Chapter 3, 4, and 5, we required a minimum of two distinct peptides for each identified protein (note: these two distinct peptides may not be unique to that identified protein). In the Chapter 6 and 7, we required a minimum of two peptides for each identified proteins and one of them must be unique to that identified protein. Due to sequence redundancy between different proteins, the same peptide could originate from multiple proteins. This creates ambiguity in inferring which protein(s) are actually present in the sample (Figure 2.2). Thus, these indistinguishable proteins were combined into protein groups, indicating that each protein or all proteins in a group are likely to exist in the sample.

2.3.3 Protein quantification

In this dissertation, various protein quantification approaches have been used. For the research presented in the Chapter 3, three different approaches (i.e. label-free, metabolic labeling, and isobaric chemical labeling) were used. For label-free quantification, the raw spectral counts calculated by DTASelect for identified proteins were normalized using the following formula:

$$N_i = R_i \cdot \frac{\bar{C}}{C_i}$$

where N*i* and R*i* are the normalized and raw spectral counts of a protein in run *i*, respectively; C*i* is the total spectral count of run *i*; and C□ is the averaged total spectral count of all the runs under comparison. The scaling factor, C□/C*i*, was used to normalize total spectral count of each run to the same to reduce run-to-run variability. ^{14}N/^{15}N metabolic labeling-based quantification was performed using the ProRata program[72]. ProRata re-constructed ion chromatograms for each identified peptides using high mass accuracy (0.03 Da mass tolerance) and detected its chromatographic peak. Both peak area and peak height for each quantified peptide were calculated, but the peak height was used to represent the peptide abundance in this study due to its lower run-to-run variability. For protein quantification, ProRata summed peak heights of all quantified unique peptides from a protein and used such total peak height for protein relative abundance estimation. For isobaric chemical labeling-based quantification, in-house Perl scripts were developed to process iTRAQ and TMT data sets for protein quantification. Briefly, 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 program. 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. The abundance ratio of a protein was estimated using the ratio between the protein's total intensities in different reporter ion channels. For research presented in chapter 4 and 5, isobaric chemical labeling-based quantifications were used. The quantification procedures were the same as described in Chapter 3. For research presented in chapter 6, ProRata was used to conduct intensity-based label free quantification. For quantification of protein abundance, ProRata used similar procedure as described for Chapter 3. For quantification of PTM fractional occupancy, ProRata calculated the total peak heights of all modified unique peptides that carried a specific PTM event and the total peak heights of both modified and unmodified unique peptides that covered the same residue. Fractional occupancy was calculated as the total peak

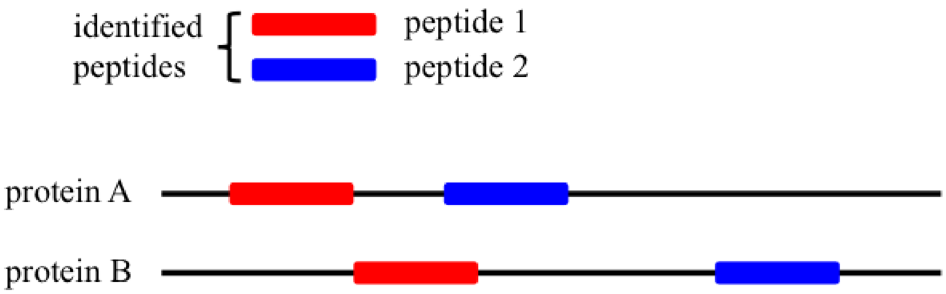


Figure 2.2 Illustration of indistinguishable proteins. Since Peptide 1 and peptide 2 are shared between protein A and protein B, it is likely that either protein A or protein B, or both protein A and protein B exist in the sample.

height of a modified site out of the total peak height of this site. In instances where only modified peptide was quantified for a given site, the abundance of unmodified site was assigned with “1”.

In summary, this dissertation used various cell lysis methods tailored for each biological system, including microbial isolate, microbial community, and plant. Extracted proteins were digested and analyzed with high-resolution mass spectrometry-based shotgun proteomic approaches. A range of quantitative proteomic results has been covered, such as label-free, metabolic labeling, and isobaric chemical labeling.

CHAPTER 3
SYSTEMATIC COMPARISON OF LABEL-FREE, METABOLIC LABELING, AND ISOBARIC CHEMICAL LABELING FOR QUANTITATIVE PROTEOMICS ON LTQ ORBITRAP VELOS

All of the data presented below has been adapted from

Zhou Li, Rachel Adams, Karuna Chourey, Gergory Hurst, Robert Hettich, 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), pp 1582–1590

Zhou Li's contributions include experimental design, MS measurement, data analysis, and writing the manuscript.

3.1 Introduction to quantitative proteomics

Quantitative proteomics measures abundance changes of many proteins among multiple samples in a high-throughput manner[30]. 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[73]. In proteomics analysis based on 2D-GE[74], 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. However, both identification and quantification are difficult for gel spots containing multiple co-migrating proteins[75]. 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[76].
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. 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[77] or the number of MS/MS spectra matched to peptides and proteins (spectral counting)[78]. 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[79], enzymatic labeling[80], and chemical labeling[81]. In metabolic labeling, stable heavy isotopes are incorporated into proteins by growing cells in controlled media containing a ^{15}N-enriched nitrogen source[82] (^{15}N labeling) or isotopically labeled essential amino acids (stable isotope labeling by amino acids in cell culture or SILAC[83]). 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. 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)[81]. After protein digestion, peptides can be labeled enzymatically at the C-terminus using $H_2{}^{18}O$[80]. Peptides can also be labeled on the primary amine group at the N-terminus and lysine side chain using reductive dimethylation (ReDi)[84]. 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)[85] and tandem mass tag (TMT)[86], 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, and the relative intensity of the reporter ions measures the relative abundance of the peptide labeled with that specific tag variant. 6-Plex TMT and 8-plex iTRAQ 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[87]. A comparison of ^{14}N /^{15}N metabolic labeling with spectral counting showed that spectral counting was less sensitive to detecting small fold changes[88]. iTRAQ was also compared to a label-free quantification method based on normalized chromatographic peak intensity[89]. 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, ^{14}N /^{15}N metabolic labeling, iTRAQ, and TMT were benchmarked using standard proteome samples prepared from a model microorganism, *Pseudomonas putida* F1 (Figure 3.1). *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[90]. Measurements for all four methods were performed using the LTQ Orbitrap Velos. 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.

3.2 Comparison of Protein and Peptide Identification Results

The results of protein identifications from label-free, metabolic labeling, and isobaric chemical labeling are summarized in Table 3.1. A total of 1980 unique proteins were identified

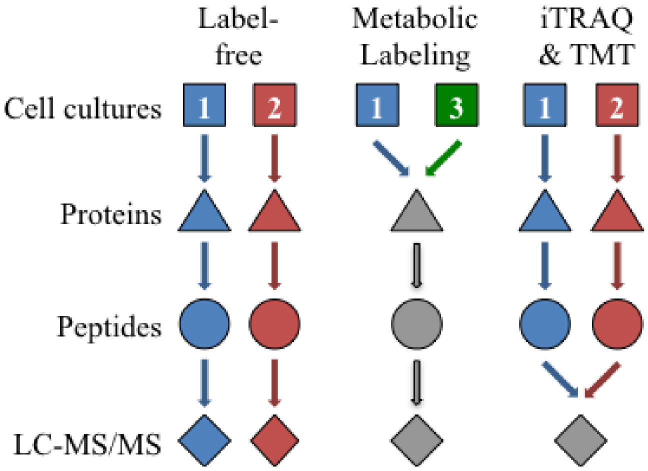


Figure 3.1 Experimental design. Three *P. putida* cultures were grown in parallel, except that the culture 3 was metabolically labeled with 15N. Proteins were extracted from cell cultures and digested into peptides, which were measured using LC–MS/MS. In the label-free method, the cultures 1 and 2 were prepared and measured separately. In metabolic labeling, the cultures 1 and 3 were mixed at the beginning. In isobaric chemical labeling, peptides from the cultures 1 and 2 were mixed after isobaric chemical labeling with TMT or iTRAQ.

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 (Figure 3.2A). A total of 1606 unique proteins were identified using the metabolic labeling method with 77% identification reproducibility between duplicate runs (FDR = 3%) (Figure 3.2B). 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 (Figure 3.2C) and 76% in TMT (Figure 3.2D). 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.

Different data acquisition schemes were used for label-free, metabolic labeling, and isobaric chemical labeling in the current study. Every full scan was followed by ten data-dependent CID MS/MS scans in the label-free analysis, which generated the highest number of identified peptides and proteins. Because in metabolic labeling proteins were quantified using full scans, six data-dependent CID MS/MS scans per full scan were configured to provide more frequent full scan acquisition and better reconstruction of chromatographic peaks of peptides. The sample complexity in full scans was doubled as a result of mixing an unlabeled proteome with a ^{15}N-labeled proteome. Because many peptides were identified redundantly in both isotopic variants, although more spectra were identified in the metabolic labeling analysis than in the label-free analysis, fewer peptides and proteins were identified, and the average sequence coverage of proteins was not increased. For iTRAQ and TMT analysis, every full scan was followed by four CID-HCD dual MS/MS 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 MS/MS 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[91]. Therefore, fewer peptides and fewer proteins were identified in isobaric

Table 3.1 Protein identification results from label-Free, metabolic labeling, iTRAQ, and TMT.

Method	Label-Free				Metabolic Labeling		iTRAQ		TMT	
Run	Culture1 Run 1	Culture2 Run 1	Culture1 Run 2	Culture2 Run 2	Run 1	Run 2	Run 1	Run 2	Run 1	Run 2
Spectrum count	58674	61440	43595	49389	52348	64972	29926	29328	35826	32897
Peptide count	12391	12727	11472	11184	9862	9618	7317	8248	6464	6795
Lys: Arg peptide ratio	0.64	0.66	0.68	0.69	0.67	0.57	0.68	0.76	0.66	0.70
Protein count	1687	1607	1598	1516	1447	1394	1202	1353	1239	1233
Average spectrum count per peptide	4.7	4.8	3.8	4.4	5.3	6.7	4.0	3.6	5.5	4.8
Average peptide count per protein	7.3	7.9	7.2	7.4	6.8	6.9	6.1	6.1	5.2	5.5
Average sequence coverage	24.6%	25.8%	23.6%	24.1%	22.3%	23.2%	19.9%	19.3%	16.5%	17.3%
Genome coverage	32.1%	30.6%	30.4%	28.9%	27.6%	26.6%	22.9%	25.8%	23.6%	23.5%

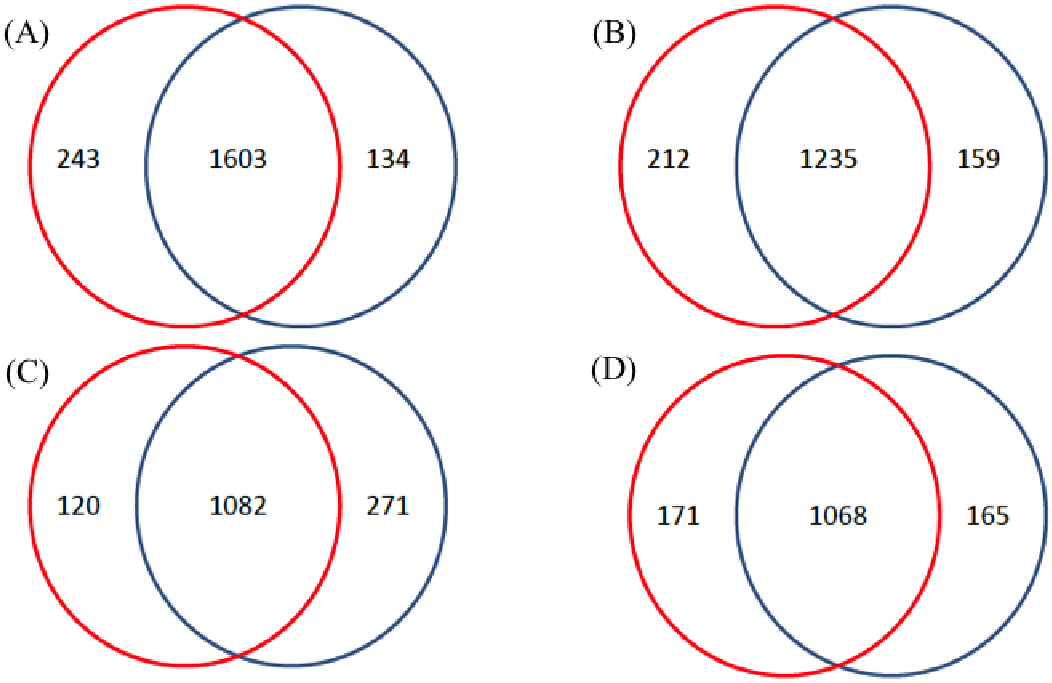


Figure 3.2 Protein identification reproducibility. The venn diagrams show the overlap of protein identifications between the duplicate runs (A: label-Free; B: metabolic labeling; C: iTRAQ; and D: TMT). The red circle and the blue circle represent proteins identified in run 1 and run 2, respectively. More than 70% of proteins were reproducibly identified between the duplicate runs.

chemical labeling than in label-free and metabolic labeling (Table 3.1). Similar protein identification results were observed between iTRAQ and TMT.

Because HCD spectra can be used for both peptide identification and quantification, TMT and iTRAQ samples can be analyzed using only HCD. We found that only 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[85]. 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 (Table 3.1). 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* digestion[92] of the *P. putida* F1 proteome. The observed Lys/Arg peptide ratios in all runs were higher than the expected ratio.

3.3 Comparison of Protein Quantification Results

Standard samples were prepared for the quantitative proteomics methods under comparison such that every protein was expected to have an abundance ratio of 1:1 (Figure 3.1). The measured abundance ratios of peptides and proteins were transformed to $\log_2$ scale ($\log_2$ ratio). Protein quantification results from each quantitative proteomics method are summarized in Table 3.2. Figure 3.3 shows that the majority of spectral counting variability stemmed from proteins with low spectral counts. Therefore, a minimum spectral count cutoff of four was used to filter out proteins with poor quantification precision[93]. As a result, although more proteins were identified using the label-free approach than labeling-based approaches, fewer proteins were precisely quantified.

For iTRAQ and TMT measurements, we examined the relationship between the reporter ion intensity and the quantification accuracy and precision of peptides. Log_2 ratios of peptides were plotted against reporter ion intensity in $\log_2$ scale ($\log_2$ intensity) (Figure 3.4A and B). For both iTRAQ and TMT, most peptides had reporter ion intensities greater than 2^{10} and were quantified accurately. The $\log_2$ ratios of peptides measured by iTRAQ have greater spread than those measured by TMT at $\log_2$ intensity below 10 (Figure 3.4A and B), indicating that the observed TMT ratios were slightly more precise. The median of peptide $\log_2$ ratios was slightly closer to 0 in the TMT runs than in the iTRAQ runs (Table 3.2), suggesting that TMT ratios were slightly more accurate. Therefore, TMT may have slightly better quantification performance than iTRAQ at the peptide level. However, there was little difference at the protein level (Table 3.2). To assess quantification accuracy and precision at different reporter ion intensity ranges, peptides were binned by their reporter ion intensities, and the median and median absolute deviation (MAD) of $\log_2$ ratios in each intensity bin were calculated (Figure 3.4C). The quantification precision as measured by MAD was consistently maintained at ~0.2 across the entire range of reporter ion intensities. Karp *et al*. observed that the quantification variability was higher at lower reporter ion intensities in iTRAQ measurements[94]. This discrepancy may be due to different instruments and data acquisition schemes used in the two studies. The quantification bias as measured by the deviation of the median from the expected value, 0, decreased as the reporter ion intensity increased. The quantification bias for low-intensity peptides could stem from the background noise in the detection of their reporter ions. Thus, peptides with higher reporter ion intensities should be given higher weight when used to calculate a protein's relative abundance. To be general to comparisons involving more than two samples, let us represent a protein's relative abundance in sample x as the percentage of the protein's quantity in sample x out of the protein's total quantity from all mixed samples, or x %. Suppose this protein has n quantified peptides. x % can be calculated as follows:

$$x\% = \sum_{i=1}^{n} \frac{x_i}{P_i} \cdot \frac{P_i}{T}$$

where x_j is the reporter ion intensity of peptide i at the reporter ion channel corresponding to sample x, P_j is the total reporter ion intensity of peptide i from all channels, and T is the sum of the total reporter ion intensities of all peptides from this protein. In this formula, the relative reporter ion intensity of a peptide at a channel, x_j/ P_j, is simply weighted by its total ion intensity, P_j, when it is pooled together with other peptides to calculate a protein's relative abundance. This abundance. This is mathematically equivalent to the summing method previously described[95]:

$$x\% = \sum_{i=1}^{n} \frac{P_i}{T} \cdot \frac{x_i}{P_i} = \frac{\sum_{i=1}^{n} x_i}{T}$$

In this study, abundance ratios of proteins were calculated using this approach for TMT and iTRAQ. As a result, the overall quantification accuracy and precision were significantly better for proteins than for peptides.

Quantification precision of proteins by the four quantitative proteomics methods was compared using MAD of protein $\log_2$ ratios and the percentage of proteins within 2-fold abundance change (Table 3.2). The performance metrics were highly reproducible between the two technical replicates of every method. To examine how the measured protein and peptide abundance ratios from each method were distributed, density plots were generated for the set of $\log_2$ ratios from each method, both at the protein level and at the peptide level (Figure 3.5). The distributions from iTRAQ and TMT experiments were narrowest, indicating the highest quantification precision. Together, our data demonstrates that iTRAQ and TMT provided the most precise measurements and will be more sensitive for detecting protein expression with small fold changes. Metabolic labeling was able to yield accurate quantification; however, the measurement variability was relatively wider than iTRAQ and TMT. Although the spectral counting method was the least precise among the compared methods, reasonable quantitative results can be still obtained.

We finally examined the quantification reproducibility of each method across technical replicates. Protein $\log_2$ ratios from duplicate measurements of each method were plotted on a two-dimensional histogram (Figure 3.6). Correlation between protein $\log_2$ ratios of the technical duplicates was also the lowest in the spectral counting analysis ($R^2 = 0.2$) (Figure 3.6A). Note

Table 3.2 Protein quantification results from label-Free, metabolic labeling, iTRAQ, and TMT.

Method	Label-free		Metabolic labeling		iTRAQ		TMT	
Run	Run 1	Run2	Run 1	Run 2	Run 1	Run 2	Run 1	Run 2
Median of protein $\log_2$ratio	0.07	0.10	0	0	-0.02	0.00	0.00	0.00
Median of peptide $\log_2$ratio	N/A	N/A	0.02	0.03	-0.13	-0.14	-0.07	-0.06
Median absolute deviation of protein $\log_2$ratio	0.40	0.43	0.3	0.3	0.17	0.17	0.17	0.05
Median absolute deviation of peptide $\log_2$ratio	N/A	N/A	0.37	0.35	0.22	0.24	0.20	0.18
Percentage of proteins with $\log_2$ratio within [-1,1]	87%	84%	94%	93%	99%	98%	99%	100%
Percentage of peptides with $\log_2$ratio within [-1,1]	N/A	N/A	86%	86%	98%	98%	99%	99%
Number of quantified proteins	1174	1116	1327	1300	1185	1338	1231	1215
Number of quantified spectra	N/A	N/A	23331	24300	20919	21447	24818	23147

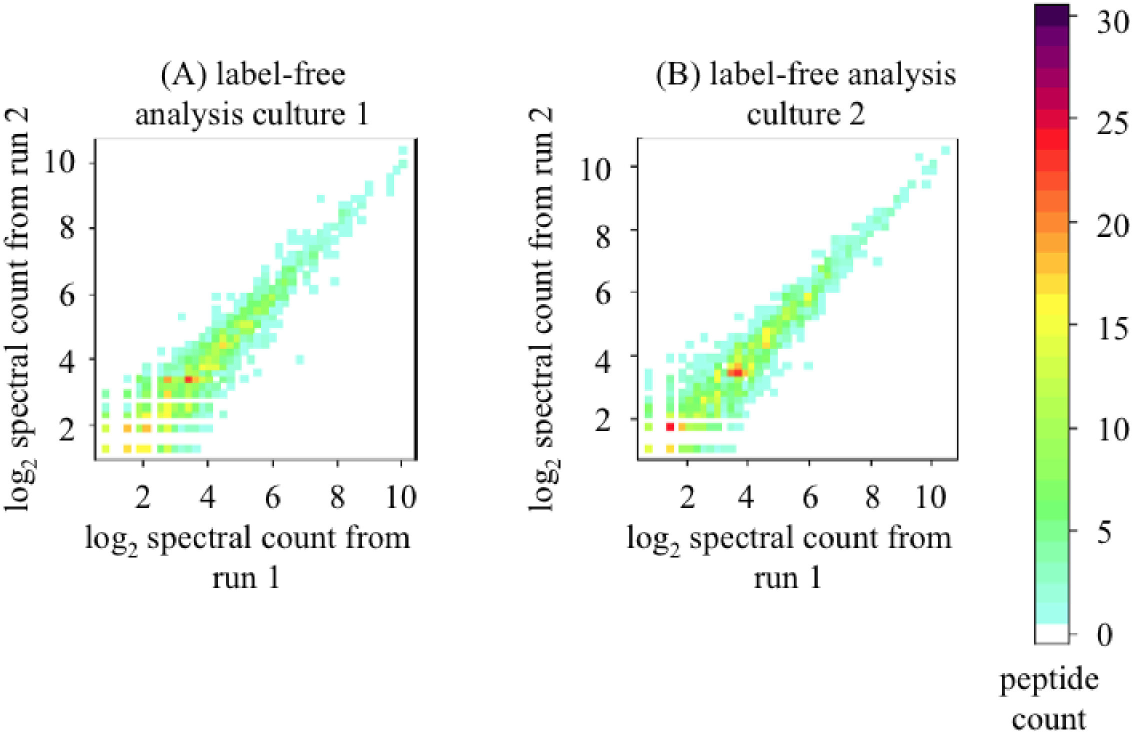


Figure 3.3 Reproducibility of the spectral counting method. 2-Dimensional histograms were constructed using $\log_2$ spectral counts of protein measured in the duplicate runs of culture 1 (A) and culture 2 (B). The color encodes protein frequency in the 2-dimensional histograms. Proteins with higher spectral counts have more similar spectral counts between the duplicate runs.

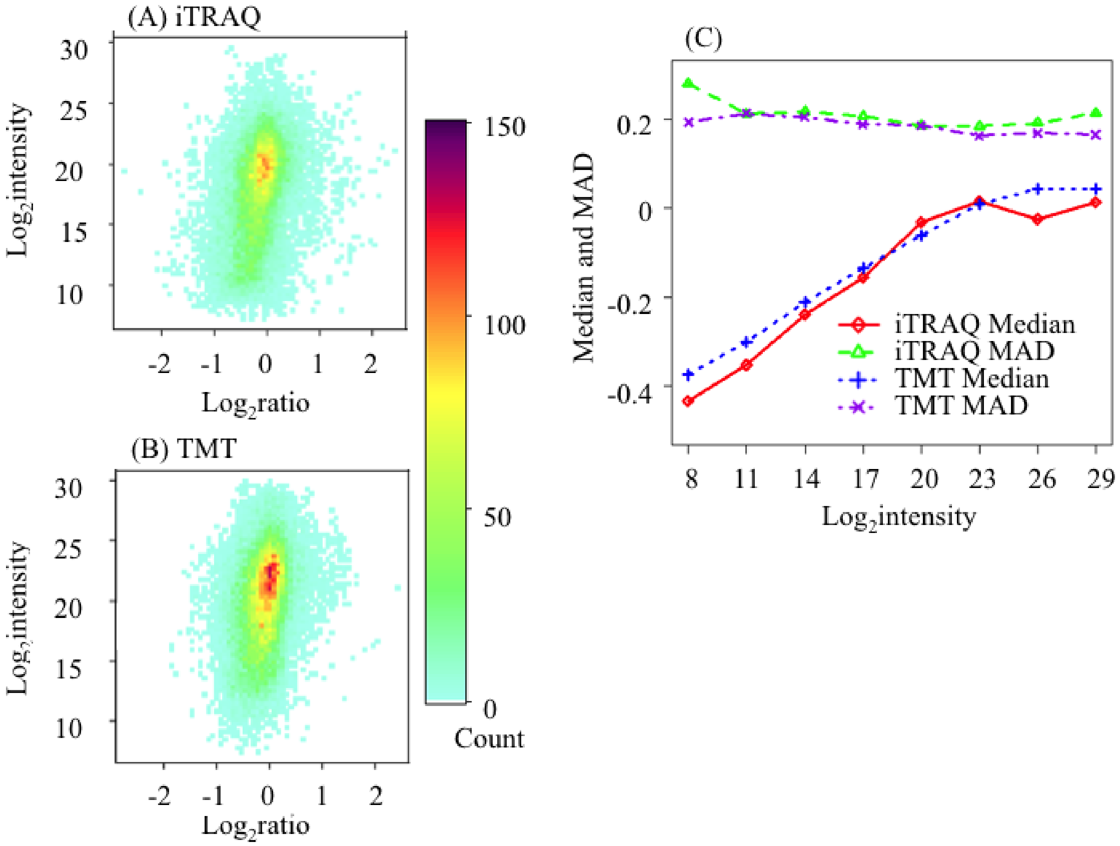


Figure 3.4 Peptide quantification results at different reporter ion intensities of iTRAQ and TMT. Panels A (iTRAQ) and B (TMT) show two-dimensional histograms of peptide log2 ratio versus the associated log2 intensity for reporter ions. The color encodes the frequency of peptides at a given log2 ratio and log2 intensity. Then, the entire intensity range was split into eight bins. Median and median absolute deviation were calculated and plotted for each bin (Panel C). As reporter ion intensity increased, quantification accuracy was improved. The value of MAD was independent of reporter ion intensity.

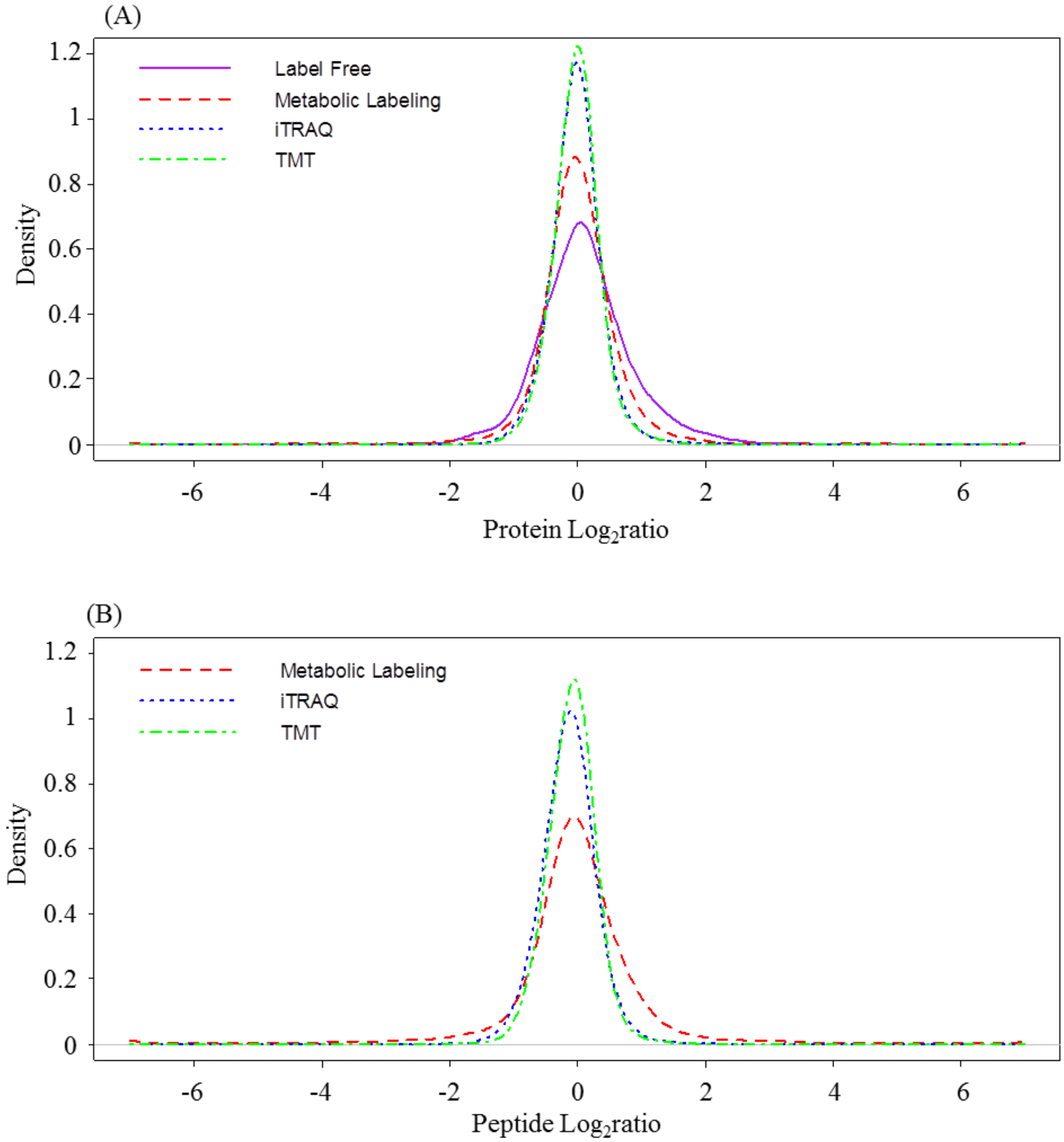


Figure 1.5 Distributions of quantified protein log2 ratios and peptide log2 ratios. Density plots were generated for log2 ratios quantified by each method at the protein level (A) and at the peptide level (B). iTRAQ and TMT produced narrower log2 ratio distributions than metabolic labeling and label-free at both the protein level and the peptide level, which indicates higher quantification precision.

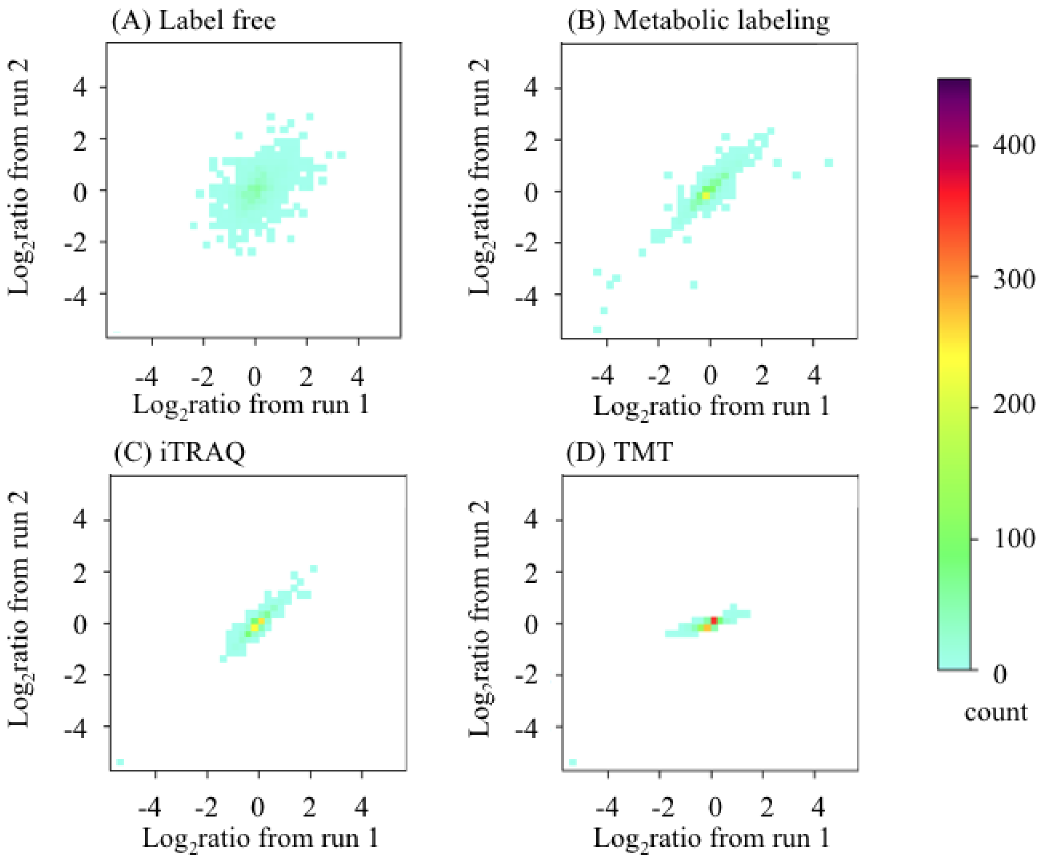


Figure 3.6 Quantification reproducibility. Two-dimensional histograms were plotted to represent log2 ratios measured from the two technical replicates of each method (A: label-free (R^2 = 0.2); B: metabolic labeling (R^2 = 0.77); C: iTRAQ (R^2 = 0.87); D: TMT (R^2 = 0.87)). The color encodes the frequency of proteins quantified at log2 ratios in the two replicates. Quantification reproducibility was significantly improved in the labeling-based approaches.

that spectral counts of proteins from two technical replicates of a culture are relatively reproducible: $R^2 = 0.86$ for culture 1 and $R^2 = 0.87$ for culture 2 (Figure 3.3). Quantification reproducibility was significantly improved in labeling-based approaches: $R^2 = 0.77$ for metabolic labeling (Figure 3.6B) and $R^2 = 0.87$ for iTRAQ and TMT (Figure 3.6C and D). Note that biological variability was observed to be more significant than technical variability in the comparison of different biological samples. Therefore, regardless of the quantification method used, it is important to use not only technical replication but also biological replication for statistical assessment in biological studies.

3.4 Considerations in method selection for a quantitative proteomics study

In label-free quantification, each sample of interest must be prepared and analyzed by LC−MS/MS separately. The semirandom-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[96]. Carvalho *et al.* extended this method and developed an algorithm to identify multiplexed spectra acquired with CID and electron transfer dissociation[97]. In the MS^E approach, a quadrupole time-of-flight mass spectrometer was used to fragment all precursor ions in an elevated-energy mode[98]. These data-independent methods will probably increase the reproducibility of label-free quantification. Alternative data analysis methods have also been developed to improve label- free quantification. For example, chromatographic peak areas of peptides, instead of spectral counts, can be used as the measure of protein abundance for quantification. The normalized spectral index (SI_N) method estimates protein abundance by combining spectral counts and total ion intensity of MS/MS spectra[99].

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 iTRAQ 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.

3.5 Conclusions

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.

CHAPTER 4
STRIGOLACTONE-REGULATED PROTEINS REVEALED BY ITRAQ-BASED QUANTITATIVE PROTEOMICS IN *ARABIDOPSIS*

All of the data presented below has been adapted from

Zhou Li, Olaf Czarnecki, Karuna Chourey, Jun Yang, Gerald Tuskan, Gregory Hurst, Chongle Pan, Jin-Gui Chen. "Strigolactone-regulated Proteins Revealed by iTRAQ-based Quantitative Proteomics in *Arabidopsis*". *J. Proteome Res.,* 2014, 13 (3), pp 1359–1372

Zhou Li's contributions include MS measurement, data analysis, and co-writing the manuscript.

4.1 Introduction

Plant architecture plays an important role in determining efficiency of light capture for photosynthesis and biomass yield. Shoot branching is a key determinant of plant architecture. The discovery of strigolactones (SLs) as a new class of plant hormones controlling shoot branching[100,101] was a recent breakthrough in the field of plant biology.

SLs are terpenoid lactones derived from carotenoids. They were originally isolated from plant root exudates and recognized as germination stimulants for root parasitic plants[102] and for hyphal branching of symbiotic arbuscular mycorrhizal fungi[103]. Recent studies demonstrate that SLs also regulate many other processes of plant growth and development including primary root growth, lateral root formation, adventitious root formation, root hair development, seed germination, photomorphogenesis, stress response and nodulation[104].

Several important genes involved in SL biosynthesis or signaling pathway have been identified through the analysis of branching mutants in *Arabidopsis* (*more axillary growth, max*), pea (*ramosus, rms*), rice (*dwarf, d; high tillering dwarf, htd*) and petunia (*decreased apical dominance, dad*). SL biosynthesis involves two carotenoid cleavage dioxygenases, CCD7 (encoded by *MAX3* gene in *Arabidopsis*) and CCD8 (encoded by *MAX4* gene in *Arabidopsis*), one cytochrome P450 monooxygenase and one novel iron-containing protein[105]. SL biosynthesis may also involve GRAS-type transcription factors NODULATION SIGNALING PATHWAY1 (NSP1) and NSP2 in *Medicago*[106]. The synthesis and exudation of SLs from roots are regulated by nutrient availability, in particular inorganic phosphate (Pi) deficiency[101]. For systemic signaling, SLs are transported through the xylem, partly mediated by ATP-binding cassette (ABC) transporters[107].

SL signaling involves *MAX2*, an F-box leucine-rich protein118[108], and DWARF14 (D14), a protein of the *α/β*-fold hydrolase superfamily[109]. Recent studies have provided strong evidence to support a function of D14 as a receptor for SLs. For example, rice *d14* is an SL-insensitive mutant that displays accelerated outgrowth of tiller[109]. Similarly, knockout of the *Arabidopsis D14* ortholog, *AtD14*, also conferred increased shoot branching and SL-insensitivity[110]. Computational-based structure analysis using homology modeling and molecular dynamic simulation and the analysis of crystal structure support the view that D14 functions as an important component of SL perception complex[111]. Recently, it was demonstrated that PhDAD2 (a petunia ortholog of D14) interacts with PhMAX2A (a petunia ortholog of MAX2) in a GR24, a synthetic SL analog[112], concentration-dependent manner[113]. Further, D14 can directly bind GR24. One potential downstream component in the SL signaling pathway is FINE CULM1 (FC1), a member of the TCP transcription factor family. It was shown that the pea TCP transcription factor PsBRC1, a homolog of the maize TEOSINTE BRANCHED1 and the *Arabidopsis* BRANCHED1 (AtBRC1) acts downstream of MAX2 to control shoot branching[114]. Recently, it has been shown that rice D53, a member of class I Clp ATPase protein family, is a substrate of the SCFD3 ubiquitination complex and that the degradation of D53 protein is promoted by GR24 and is dependent on D14 and D3 (a rice ortholog of MAX2)[115].

Plant hormones interact with each other in the regulation of plant growth and development[116]. There is substantial evidence that this is also the case for SLs. For example, SLs interact with auxin and cytokinin to regulate shoot branching and secondary growth[117], interact with auxin to regulate mycorrhizal symbiosis[118] and primary root growth, lateral root formation and adventitious root formation[119], interact with auxin and ethylene to regulate root hair elongation[120], and interact with gibberellin (GA) and abscisic acid (ABA) to regulate seed germination and photomorphogenesis[121]. Recently, it has been reported that BES1, a positive regulator in brassinosteroid (BR) signaling pathway, interacts with MAX2 and that the degradation of BES1 is dependent on MAX2 and is promoted by GR24[122].

At the transcription level, SLs regulate the expression of many genes. When *Arabidopsis max3* mutant seedlings that were defective in SL biosynthesis were treated with GR24, the expression of a total of 31 and 33 genes was significantly up- and down-regulated, respectively[123]. In tomato, a mutant deficient in SL production (*Sl-ORT1*) was used to investigate the SL-regulated transcriptome. It was found that GR24 induces the expression of a number of genes putatively involved in light harvesting. Consistent with this notion, the Sl-ORT1 mutant contained less chlorophyll and showed reduced expression of light-harvesting-associated genes[124].

However, despite this progress, essentially nothing is known about cellular proteome regulation by this new class of plant hormones. In this study, we treated *Arabidopsis* seedlings with GR24 and applied quantitative proteomics to determine the SL-regulated proteome.

4.2 Additional materials and methods

Genotyping and semi-quantitative RT-PCR: A SALK T-DNA specific primer (LBb1.3, 5'-ATTTTGCCGATTTCGGAAC-3') and *AtMAX3* gene-specific primers (RP, 5'-TATCGTTAAACCCAAGCAACG-3'; LP, 5'-AGCCCATAAACCATGAAAACC-3') were used for PCR genotyping. To examine the absence or presence of *AtMAX3* transcripts in the *max3-12* mutant, total RNA was extracted from 8-day-old seedlings using the Invisorb Spin Plant Mini Kit (Stratec Molecular). Two µg of total RNA were reversely transcribed in cDNA using Fermentas RevertAid reverse transcriptase (Thermo Scientific). For semi-quantification, *AtMAX3*-specific primers (sqP1: 5'-TATCGTTAAACCCAAGCAACG-3' and sqP2: 5'-CAATGTAACCATCGTCCTCT-3') spanning 636 bp of *AtMAX3* fragment were used (Figure 4.1B). PCR amplification of *AtTUA5* (At5g19780) using primers 5'-TGGTTCTGGATTGGGTTCTC-3 and 5'-ACAGCATGAAATGGATACGG-3 served as a control.

Quantitative RT-PCR: Quantitative Real-time PCR (qRT-PCR) was performed using a StepOnePlus (Applied Biosystems), Maxima SYBR Green/ROX qPCR Master Mix (Thermo Scientific) and cDNA corresponding to 80 ng RNA in a total volume of 25 µl. The following cycling conditions were used for PCR: 10 min at 95°C, 40 cycles of 15 s at 95 °C, 60 s at 60 °C, and 30 s at 72 °C. Calculation of expression levels in relation to *AtACT2* (At3g18780) expression was performed using the $2^{-\Delta\Delta Ct}$ method according to Livak and Schmittgen[125]. Gene-specific primers were designed using QuantPrime[126] or taken from Mashiguchi *et al*, and are listed in Table 4.1.

Data normalization: raw protein intensities were normalized by making the total intensity of each sample identical. The ratio of a protein's intensities between two conditions was $\log_2$-transformed and also converted to fold-change. The data were further normalized by centering the median of the distribution of these $\log_2$ratios between two conditions to 0. Rank product statistical analysis[127], a non-parametric statistical test based on calculating the rank products of protein's fold-changes from replicates, was used to detect differentially regulated proteins.

Calculation of molecular mass and isoelectric point: molecular mass and isoelectric point distributions were compiled for detected proteins and for all *Arabidopsis* proteins in the protein database used for Sequest searches. Molecular mass and isoelectric point for each protein were obtained by uploading the *Arabidopsis* protein sequences to the Compute pI/Mw tool on the ExPASy server at URL http://web.expasy.org/compute_pi/. Distributions were normalized to the largest bin to facilitate comparison of the detected proteome to the fully predicted proteome.

4.3 Experimental design

In addition to acting as a key inhibitor of shoot branching in *Arabidopsis*, SLs regulate diverse processes of plant growth and development in young seedlings, including primary root growth, lateral root formation, adventitious root formation, root hair development, seed germination, and photomorphogenesis. In this study, we specifically focus on the investigation of the proteome regulated by SLs in *Arabidopsis* seedlings. Because an SL microarray study has been previously conducted using 14 day old light-grown *Arabidopsis* whole seedlings[123], we wanted to use *Arabidopsis* seedlings at a developmental stage that is similar to the previous study (e.g., 14 day old light-grown seedlings) as our experimental materials. This would allow us to potentially compare the SL-regulated proteome with the SL-regulated transcriptome.

The response of *Arabidopsis* seedlings to different concentrations of the synthetic SL analog GR24 has been previously studied[119]. In general, both primary root growth and lateral root formation can be affected by GR24 at micromolar concentration ranges. These effective concentrations of GR24 were also consistent with the level of GR24 (e.g., 5 μM) required for a complete rescue of the branched phenotypes of *Arabidopsis* mutants defective in SL biosynthesis. Therefore, in our proteomics study, we treated 14 day old light-grown *Arabidopsis* whole seedlings with GR24 at 5 μM.

In a previous SL microarray using *Arabidopsis* whole seedlings, seedlings were treated with GR24 for 90 min. Because the proteome response typically required longer time than the transcriptome response due to the need of additional processes for de novo protein synthesis (at least for some proteins), in our proteomics study, we chose to treat *Arabidopsis* whole seedlings with 5 μM GR24 for 12 h. This selection was mainly based on proteomics studies using other plant hormones. For example, in a proteomics study using BR, no protein was detected to have differential expression after 0.5 or 3 h of brassinolide (BL) treatment, 6 BR-induced proteins were detected after 6 h of BL treatment, and 15 BR-induced and 3 BR-repressed proteins were detected after 12 h of BL treatment[128]. Furthermore, GR24-regulated growth response becomes significant after 12 h of treatment, as measured by the growth inhibition of primary root (Figure 4.1). Therefore, we expected that 12 h of GR24 treatment would allow the detection of a substantial number of proteins up- or down-regulated by GR24.

In light of these considerations, we used 14 day old light-grown *Arabidopsis* whole seedlings for GR24 treatment (5 μM GR24 for 12 h). Three biological replicates, each containing 50 individual seedlings, were used for each sample. Because SL-deficient mutants were used in previous transcriptomics studies in *Arabidopsis* and tomato and because a plant hormone deficient mutant may respond more potently to the corresponding hormone, we also included an *Arabidopsis max3* mutant in our study. MAX3 encodes one of the two carotenoid cleavage dioxygenases, CCD7, required for SL-biosynthesis. The widely used *max3* mutant allele, *max3-9*, is an ethyl methanesulfonate (EMS)-induced mutant allele harboring a recessive mutation in MAX3 in *Col-0* background, which had been backcrossed to *Col-0* three times. For our experiments, we sought to use a T-DNA insertion mutant allele in the *Col-0* background to avoid the potential effects of other EMS-induced mutations in the *max3-9* mutant background because three backcrosses were able to remove only 90% of EMS-induced mutations in the *max3-9* mutant background. By searching the collection of T-DNA insertion lines at SIGNAL (http://signal.salk.edu/cgi-bin/tdnaexpress), we selected SALK_015785, in which T-DNA was inserted in the fourth intron of MAX3 (Figure 4.2B and C). This mutant displayed identical branching phenotype to max3-9 (Figure 4.2A). RT-PCR analysis indicated that the full transcript of MAX3 was largely absent in this allele (Figure 4.2D), suggesting that this likely represents a loss-of-function allele of MAX3. To follow the existing nomenclature for max3 mutants, we named this mutant allele, *max3-12*. We prepared a total of 12 samples (two genotypes: *Col-0* and

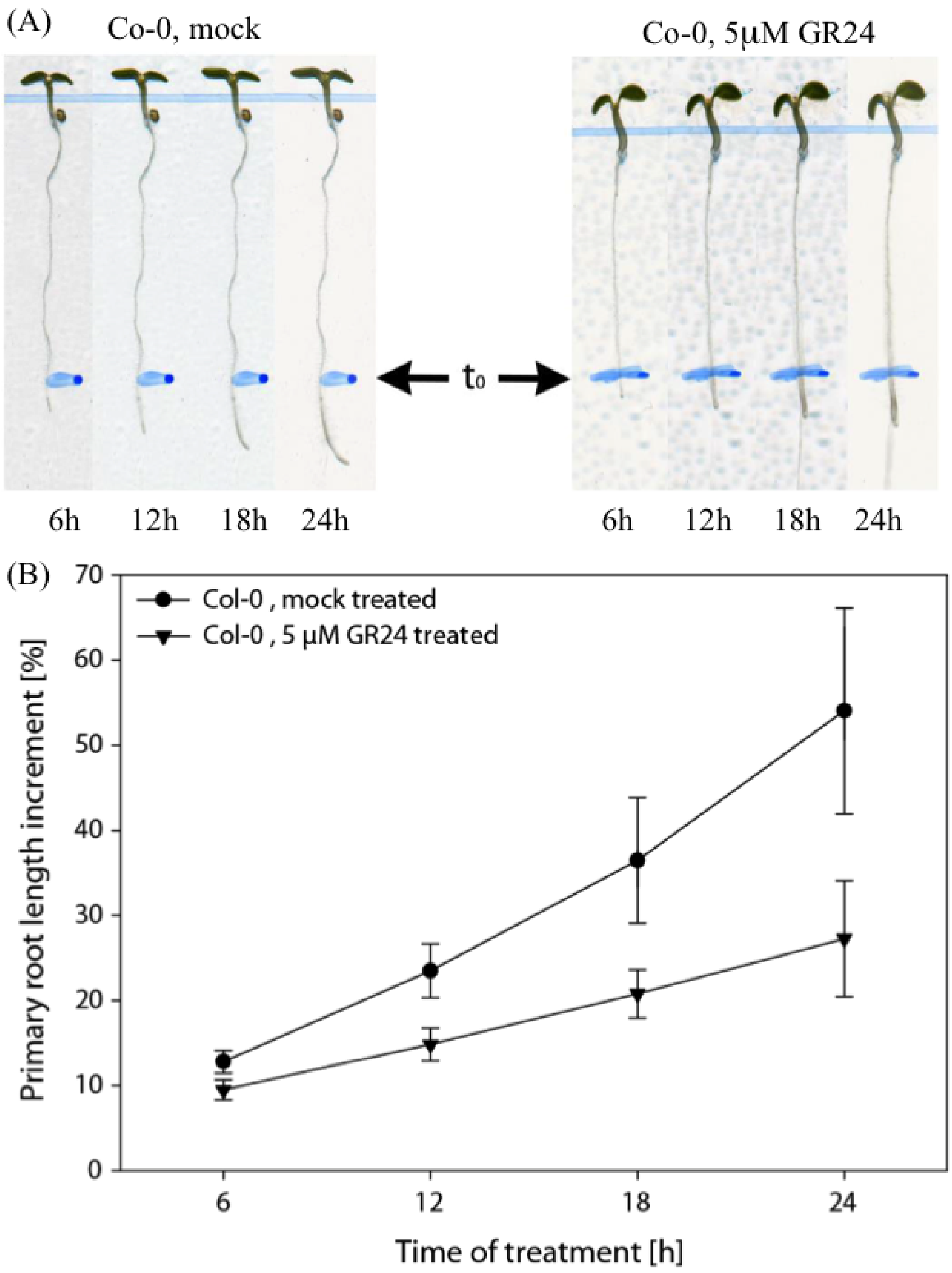


Figure 4.1 Response of *Arabidopsis* wild-type *Col-0* primary root growth to 5 μM GR24. Four-day-old seedlings were transferred to 1/2 MS supplemented with 1% (w/v) sucrose and 0.8% (w/v) agar plates plus 5 μM GR24 or acetone as mock treatment. (A) Primary root growth within 24 h of treatment. (B) Primary root length increment within 24 h of treatment. t0, beginning of treatment.

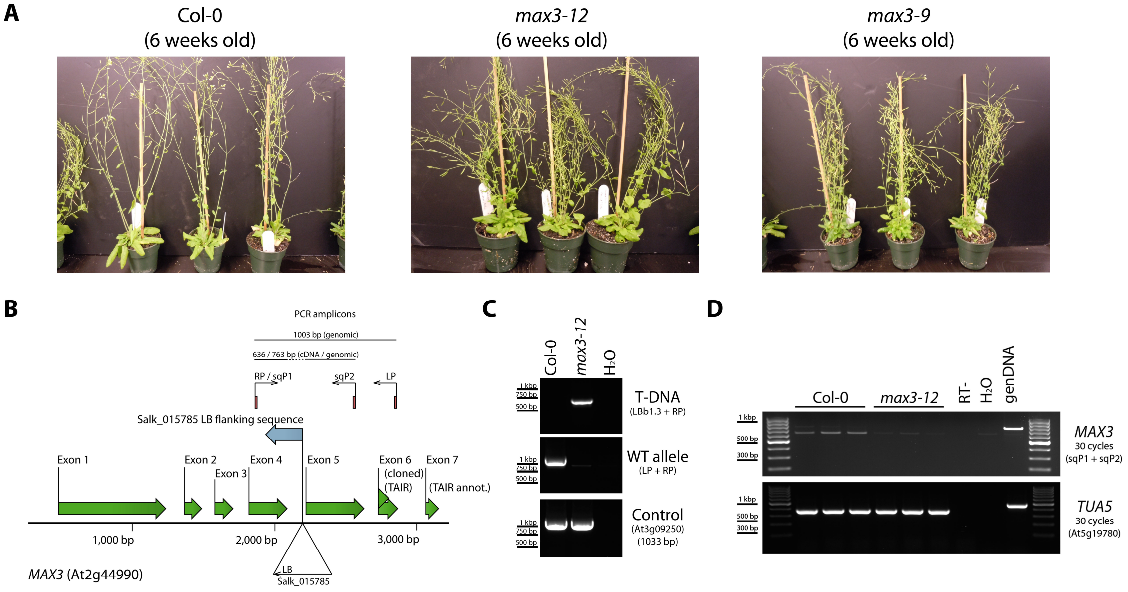


Figure 4.2 *Arabidopsis max3-12* mutant. (A) Phenotype of 6 weeks old soil-grown *Arabidopsis max3-9* (right panel) and *max3-12* (middle panel) compared with the wild-type *Col-0* (left panel). (B) Schematic presentation of the *AtMAX3* gene structure and the *max3-12* (Salk_015785) TDNA insertion site. Primers used for genotyping and expression analysis and the corresponding PCR amplicon sizes are indicated. It is noteworthy that the originally cloned *AtMAX3* coding sequence (Booker *et al*., 2004) encoded six exons, which is different from the current annotation in TAIR10. (C) Genotyping of *max3-12* mutant. A PCR product can be amplified using a T-DNA-specific primer (LBb1.3) and an *AtMAX3*-specific primer (RP) in *max3-12*, whereas the wild-type allele (primers RP and LP) can be amplified in *Col-0* only. A 1033bp PCR product amplified by using primers specific to At3g09250 served as a control. (D) Semi-quantitative RT-PCR analysis of *AtMAX3* in the *max3-12* mutant in comparison with *Col-0* wild type. A 636 bp amplicon specific for *AtMAX3* can be amplified from *Col-0* but not from *max3-12* cDNA (upper panel). The presence of equal amounts of template is shown by amplification of an AtTUA5 (At5g19780) amplicon (lower panel). Non-reverse-transcribed *Col-0* RNA (RT$^-$), H_2O, and genomic DNA served as PCR controls. Shown are the results of three biological replicates

Table 4.1 Primers used for quantitative RT-PCR analysis in this study.

LocusID	Foward and reverse primer sequences (5' – 3')
AtACT2	CCAGAAGGATGCATATGTTGGTGA; GAGGAGCCTCGGTAAGAAGA
AtSAND	AACTCTATGCAGCATTTGATCCACT; TGATTGCATATCTTTATCGCCATC
At1g18270.3	TCCAGCTGAAGTGACGAAAGATG; TCCTCAGCCTGAACCTCGTTTG
At1g75270.1	AGATATCTGCGTGAAGGTTGCC; GAACACGTTGGCTAAACGGACAG
At1g76680.2	TGGATGCACCGCTGAATAAGTACG; AATCGGTGTAACCGACGACTGG
At2g22450.1	TTACTGCGGTTGCGCCTATACC; TCCCATCTCCGATTTCTCCCTTG
At3g18600.1	ATGATGGCAGACGCAAGGTGAC; TTGCTCGGAACAACACAATAGCC
At4g01870.1	ATGCCTTGCTGGTCTCCGAAAG; ACGGCCGTGTTCTCTGGATTATG
At4g15760.1	TGTTTCCGGTTGTTCACTTAGCC; TCGCATCCAATCAGGACCTTGG
At5g13710.1	AGGATTAGTCGACGGTGGAAGGAG; ACGATGACCATCGCCATCTCTC
At5g57460.1	TGCAGCTGCTGAGGGAAATACAG; AGACTCCGGTTTCGTACCTGTG
At5g61820.1	ACCAATGGCGATGGATGCGTTG; AGCGCATATCCCTTCACCGTTCTC
At1g15580.1	GCTCTGCAAATTCTGTTCGGATGC; CACGATCCAAGGAACATTTCCCAAG
At2g40670.1*	CCGATTACTGTATGCCTGGA; TTTGAGCTCCACTCGCTAAA
At1g64380.1*	CTCCGGTGACGACAACTACT; TTCACTAGGGACCGAAACTG
At3g18550.1*	CCAGTGATTAACCACCATCG; TGCATGAGGTCTCTTGGTTT
At1g29440.1*	TGCTCTTTTCAACCACAAGA; TGAAATGATCTGTCTATCTAATCCA

* Primer sequences were taken from Mashiguchi, K; Sasaki, E.; Shimada, Y.; Nagae, M.; Ueno, K.; Nakano, T.; Yoneyama, K.; Suzuki, Y.; Asami, T. Feedback-regulation of strigolactone biosynthetic genes and strigolactone-regulated genes in *Arabidopsis.* Bioscience Biotechnology and Biochemistry 2009, 73, (11), 2460-2465.

max3-12; two treatments: with/without 5.0 µM GR24 for 12 h; three biological replicates) for proteome analysis.

4.4 iTRAQ Analysis

Total proteomes were extracted from the whole seedlings and compared by quantitative proteomics using an isobaric chemical labeling reagent, iTRAQ. A trial run with a test sample indicated high iTRAQ labeling efficiency, resulting in 14, 780 spectra identified with iTRAQ labeling on N-terminal or lysine residues and only 537 spectra identified with no iTRAQ labeling. Each of the four samples (*Col-0*, *Col-0* + GR24, *max3-12*, *and max3-12* + GR24) in each biological replicate was labeled by one of the four reagents of the iTRAQ 4-plex and then was combined into one aliquot. The three biological replicates were measured in technical duplicates on an LTQ Orbitrap Velos Pro mass spectrometer.

A total of 2095 proteins were identified across all samples with an average of about 1,700 proteins across the biological triplicates. Approximately 63% of the identified proteins were reproducibly detected across all three biological replicates (Figure 4.3). This is well within the range of reproducibility of other iTRAQ studies in plants. For example, in an iTRAQ study using *Arabidopsis* guard cells, the overlaps of proteins detected between any two biological replicates were from 49 to 71%[129]. As shown in Figure 4.4, the molecular weight distribution and isoelectric point range of identified proteins largely matched with predicted ranges of the fully annotated *Arabidopsis* proteome, although it appeared that our iTRAQ study detected fewer proteins with high isoelectric point. In general, our overall protein extraction and digestion protocols and subsequent proteomics measurements were not strongly biased with respect to these physical and chemical traits of the proteins.

Our primary objective was to identify GR24-regulated proteins. Specifically, we wanted to compare protein abundances between *Col-0* + GR24 and *Col-0* and between *max3-12* + GR24 and *max3-12*. As discussed in an iTRAQ study using *Arabidopsis* guard cells, stringency requirements to determine a protein as significantly changed in abundance have not been standardized for quantitative proteomics methods. Zhao *et al.* used fold-change ratios of protein abundance of <0.85 or >1.17 with P < 0.05. In our study, protein abundances were compared between *Col-0* + GR24 and *Col-0*, and between *max3-12* + GR24 and *max3-12* using the rank product test. We considered proteins with fold changes <0.75 or >1.25 with P < 0.01 as

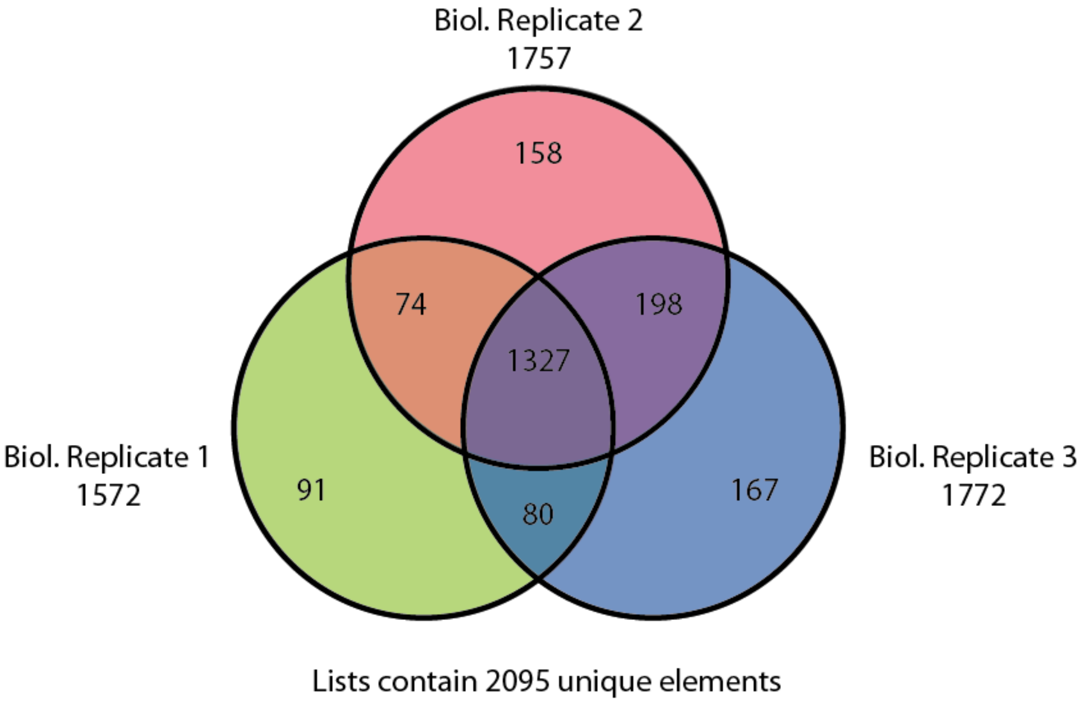


Figure 4.3 Venn diagram illustrating the number of proteins detected in all three biological replicates. Each circle represents a set containing 4 samples (*Col-0*, *Col-0* + GR24, *max3-12*, *max3-12* + GR24), each labeled with a different iTRAQ reagent.

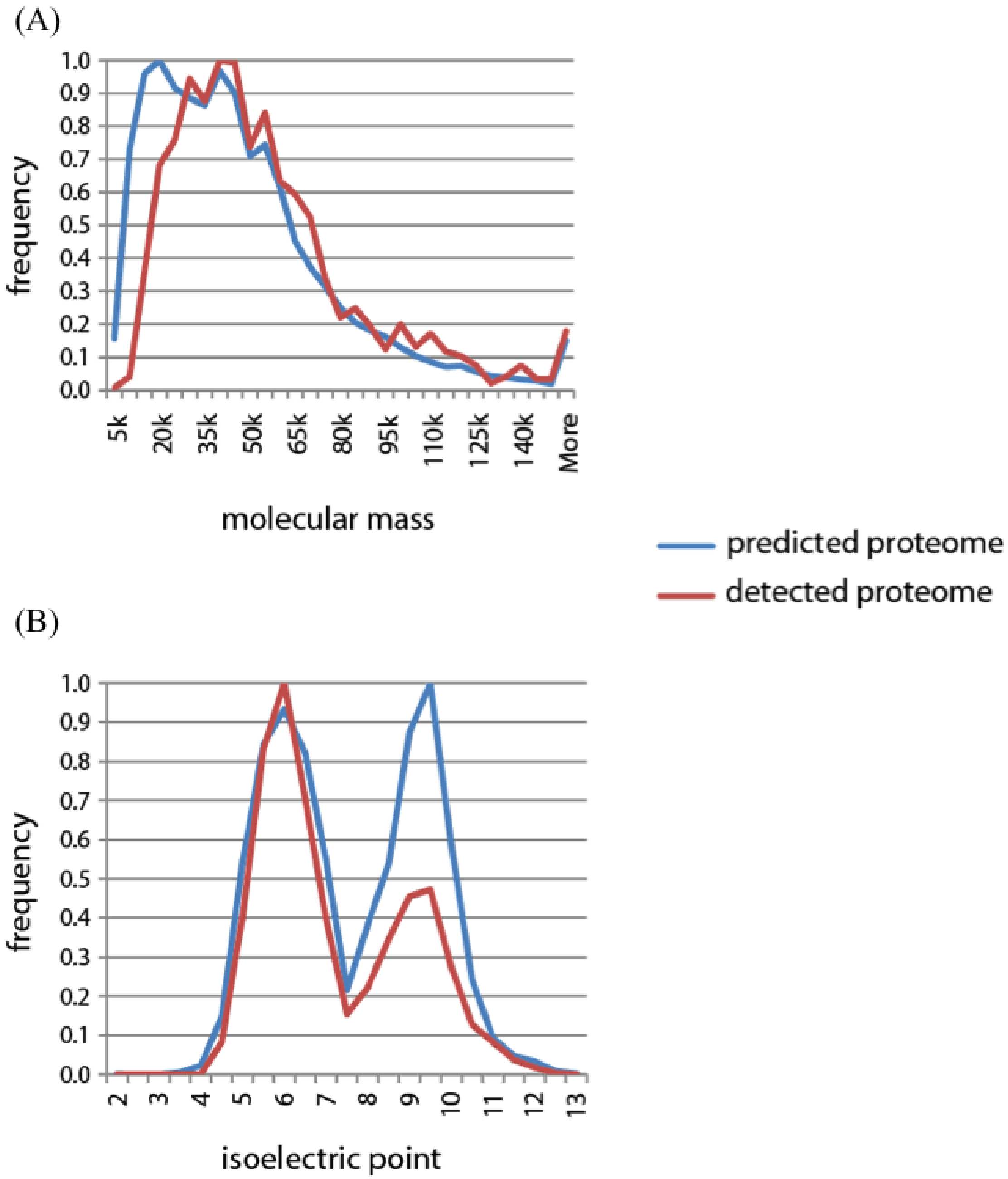


Figure 4.4 Comparisons in the molecular mass (A) and isoelectric point (B) frequency distributions between the fully predicted *Arabidopsis* proteome (blue) and proteins detected by iTRAQ-based quantitative proteomics in this study (red).

significantly changed in abundance. In addition, we applied FDR < 0.25. The FDR value (i.e. falsely identifying proteins with significant changes from a large number of proteins with no change) was estimated empirically using a permutation test. Biological replicates were randomly permutated between different conditions to create decoy comparisons with random variability and no biological difference. The described filtering criteria found no protein with statistically significant difference in such decoy comparisons, indicative of a minimum empirical FDR of our analysis.

4.5 Proteins up-regulated by SL

The distributions of $\log_2$-transformed ratios of protein abundance changes of all quantified proteins in *Col-0* and *max3-12* background were plotted (Figure 4.5). The set of $\log_2$ ratios in *Col-0* appeared to have slightly wider distribution than that in *max3-12*, suggesting that more proteins may be regulated in response to GR24 in *Col-0* than in *max3-12*. Consistent with this observation, the iTRAQ detected 19 proteins whose abundance was reproducibly increased by more than 1.25-fold and 18 proteins whose abundance was decreased by more than 25% in *Col-0* seedlings treated with GR24 compared with the mock control (Tables 4.2 and 4.3). In *max3-12*, the iTRAQ detected nine GR24-up-regulated and two GR24-down-regulated proteins.

Among 19 proteins that were reproducibly increased by more than 1.25-fold in *Col-0* seedlings treated with GR24 compared with the mock control (Table 4.2), the highest induction of protein abundance by GR24 was 2.52-fold. This is consistent with reports that less than 2.0-fold changes in protein abundance in the identification of plant proteins are common in iTRAQ studies[129-131]. Because it has been reported that fold changes in protein abundance determined by iTRAQ were relatively smaller than those determined by blue native gel and label-free methods[132], the fold changes here may be underestimated. We found that six proteins were commonly up-regulated by GR24 in *Col-0* and *max3-12* mutant background. We briefly discuss each of these six proteins below.

Among proteins whose abundance is up-regulated by GR24, OPR1 (12-oxophytodienoate reductase 1) showed the highest fold change (2.52 in *Col-0* and 2.42 in *max3-12*). OPR1 shares similarity with the Old Yellow Enzyme family and can transform explosive 2,4,6-trinitrotoluene (TNT) to yield nitro-reduced TNT derivatives[133]. The biological function of OPR1 in plants is unknown. OPR1 was found to be predominantly expressed in roots[134]. Previous studies showed

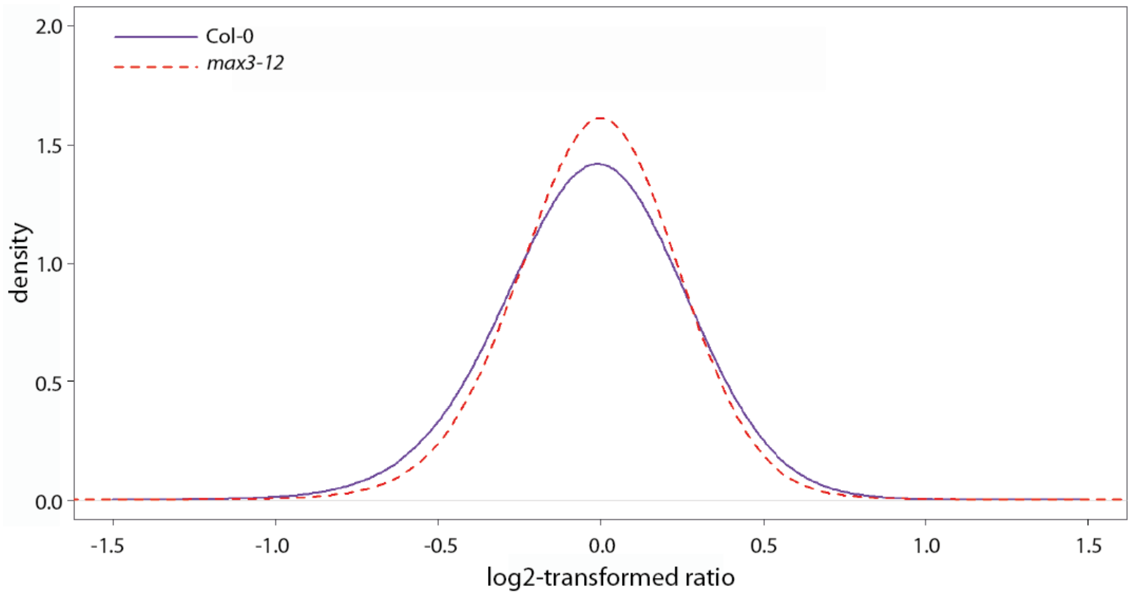


Figure 4.5 The distributions of GR24-induced changes in protein abundance (as $\log_2$-transformed ratios) of all quantified proteins in *Arabidopsis* wild-type *Col-0* (purple, solid) and the *max3-12* mutant (red, dashed) genetic backgrounds.

Table 4.2 List of proteins up-regulated in *Arabidopsis Col-0* wild-type background after 12 h of treatment with 5 μM GR24[a]

[a]Proteins were considered to be up-regulated if they met the following criteria: fold change ≥1.25, P < 0.01, and FDR < 0.25. Locus identifiers of genes selected for subsequent quantitative RT-PCR analysis are bold. FDR, false discovery rate. [b]Number of identified tryptic peptide ions including different charge states. [c]Identified peptides are not unique and can be attributed to both loci.

Locus identifier	*P* value	FDR value	(Col+GR 24)/Col fold change	Peptide count[b]	Sequence coverage	Description
AT1G76680	**0.0000**	**0.020**	**2.52**	**11**	**30%**	**OPR1, 12-oxophytodienoate reductase 1**
AT3G18740	0.0001	0.070	1.70	7	29%	Ribosomal L7Ae/L30e/S12e/Gadd45 family protein
AT5G61820	**0.0001**	**0.053**	**1.64**	**4**	**9%**	**molecular function unknown**
AT1G36240[c]	0.0007	0.179	1.55	6	29%	ribosomal L7Ae/L30e/S12e/Gadd45 family protein
AT1G77940[c]	0.0007	0.179	1.55	6	29%	ribosomal L7Ae/L30e/S12e/Gadd45 family protein
AT4G01870	**0.0009**	**0.192**	**1.45**	**9**	**18%**	**tolB protein-related**
AT1G60950	0.0015	0.211	1.43	3	31%	FED A, ATFD2, 2Fe-2S ferredoxin-like protein
AT1G56340	0.0013	0.196	1.42	8	16%	CRT1, CRT1a, AtCRT1a, calreticulin 1a
AT5G64040	0.0021	0.238	1.39	2	6%	PSAN, photosystem I reaction center subunit PSI-N
AT1G10960	0.0023	0.240	1.39	3	31%	ATFD1, FD1, ferredoxin 1
AT3G50820	0.0017	0.209	1.38	28	57%	PSBO2, PSBO-2, photosystem II subunit O-2
AT2G20260	0.0024	0.226	1.38	9	50%	PSAE-2, photosystem I subunit E-2
AT4G28750	0.0025	0.219	1.37	13	43%	PSAE-1, photosystem I reaction center subunit IV
AT4G15760	**0.0030**	**0.229**	**1.36**	**3**	**10%**	**MO1, monooxygenase 1**
AT3G02560	0.0034	0.232	1.35	3	13%	ribosomal protein S7e family protein
AT4G13180	0.0034	0.245	1.34	7	18%	NAD(P)-binding Rossmann-fold superfamily protein
AT2G40610	0.0009	0.166	1.34	3	13%	ATEXPA8, EXP8, ATEXP8, expansin A8
AT4G34620	0.0040	0.248	1.33	3	38%	SSR16, small subunit ribosomal protein 16
AT1G75270	**0.0038**	**0.247**	**1.33**	**5**	**23%**	**DHAR2, dehydroascorbate reductase 2**
AT4G19880	0.0030	0.244	1.30	2	7%	glutathione S-transferase family protein

Table 4.2 continued

Table 4.3 List of proteins down-regulated in *Arabidopsis Col-0* wild-type background after 12 h of treatment with 5 μM GR24[a]

[a]Proteins were considered to be down-regulated if they met the following criteria: fold change ≤0.75, P < 0.01, and FDR < 0.25. Locus identifiers of genes selected for subsequent quantitative RT-PCR analysis are bold. FDR, false discovery rate. [b]Number of identified tryptic peptide ions including different charge states.

Locus identifier	*P* value	FDR value	(Col+GR 24)/Col fold change	Peptide count[b]	Sequence coverage	Description
AT5G57460	**0.0001**	**0.130**	**0.43**	**2**	4%	**molecular function unknown**
AT3G03060	0.0018	0.246	0.48	4	7%	P-loop containing nucleoside triphosphate hydrolase
AT5G13710	**0.0002**	**0.097**	**0.53**	**2**	**4%**	**SMT1, CPH, sterol methyltransferase 1**
AT4G14160	0.0020	0.224	0.57	3	4%	Sec23/Sec24 protein transport family protein
AT1G21150	0.0002	0.110	0.60	2	2%	mitochondrial transcription termination factor
AT3G59780	0.0033	0.240	0.64	3	6%	rhodanese/cell cycle control phosphatase
AT4G14040	0.0021	0.221	0.65	9	17%	EDA38, SBP2, selenium-binding protein 2
AT4G30190	0.0034	0.221	0.65	6	8%	HA2, H(+)-ATPase 2
AT3G18600	**0.0009**	**0.175**	**0.65**	**2**	**2%**	**P-loop containing nucleoside triphosphate hydrolase**
AT1G18270	**0.0012**	**0.178**	**0.66**	**4**	**3%**	**ketose-bisphosphate aldolase class-II family protein**
AT2G22450	**0.0024**	**0.226**	**0.68**	**2**	**3%**	**putative riboflavin biosynthesis protein**
AT5G64120	0.0011	0.200	0.70	8	28%	peroxidase superfamily protein
AT3G09910	0.0004	0.128	0.70	3	11%	ATRAB18C, ATRABC2B, RAB GTPase homologue C2B
AT4G02520	0.0033	0.227	0.70	6	27%	ATGSTF2, GST2, glutathione S-transferase PHI 2
AT1G54270	0.0031	0.242	0.71	14	41%	EIF4A-2, eif4a-2
AT2G18730	0.0008	0.192	0.73	2	4%	ATDGK3, DGK3, diacylglycerol kinase 3
AT3G22060	0.0037	0.228	0.73	4	11%	receptor-like protein kinase-related family protein
AT1G64510	0.0026	0.226	0.73	3	5%	translation elongation factor EF1B/ribosomal protein S6

Table 4.3 continued

that the transcript level of OPR1 was transiently increased in response to abiotic stimuli, but no subsequent changes in protein levels were detected. It was proposed that post-transcriptional regulation may inhibit the generation of higher levels of OPR1 proteins in the *Arabidopsis* plants overexpressing OPR1. Because the transcript of OPR1 is highly induced by TNT and OPR1 may function in the detoxification of TNT and here we show that the protein abundance of OPR1 is increased by GR24 in *Arabidopsis* seedlings, this raises a question of whether the SL pathway may affect TNT detoxification.

Protein encoded by gene locus At4g01870 was up-regulated 1.45-fold in *Col-0* and 1.64-fold in *max3-12* by GR24. At4g01870 is annotated as tolB protein-related and categorized as a gene involved in stress responses. tolB is a bacterial protein that maintains outer membrane stability and integrity[135], which is important for protecting cells against antibacterial agents. However, the molecular function of tolB-related proteins in plants is unknown. The transcript of At4g01870 is highly induced by A1-phytoprostanes (PPA1) (20.1-fold induction) and PPB1, compounds that are structurally highly similar to 12-oxo-phytodienoic acid (OPDA), which is a precursor for the plant hormone jasmonic acid but may also act as a signal molecule regulating plant development and stress response[136].

Protein encoded by gene locus At5g61820 was up-regulated 1.64-fold in *Col-0* and 1.96-fold in *max3-12* by GR24. The molecular function of this protein is unknown. BLAST searching revealed that this protein shares similarity with a family of proteins implicated in nodule development in the legume *Medicago*, whose transcription is induced during nodulation[137]. It is interesting to note that SLs promote nodulation in pea[138].

The protein abundance of monooxygenase 1 (MO1) was increased by 1.36-fold in *Col-0* and 1.42-fold in *max3-12* by GR24. MO1 was identified as a gene whose transcript is preferentially up-regulated by *Alternaria brassicicola*, a ubiquitous plant pathogenic fungus, in a compatible *Arabidopsis* ecotype *Dijon G* but not in the incompatible ecotype *Col-0*. MO1 has similarity with monooxygenases that are known to degrade salicylic acid (SA), but the exact reaction catalyzed by MO1 is unknown. It has been postulated that MO1 may catalyze the production of SA-derived aromatic compounds that have signaling roles or that MO1 may be involved in the suppression of SA pathway[139].

The fifth protein whose abundance was enhanced by GR24 in both *Col-0* (1.34-fold increase) and *max3-12* (1.32-fold increase) is a protein encoded by gene locus At4g13180. This protein is predicted to be a member of NAD(P)-binding Rossmann-fold superfamily, but its biochemical function has not been experimentally demonstrated. At4g13180 was identified as one of the early phosphate starvation-responsive genes. It has been well-documented in several plant species that Pi deficiency stimulates SL production and exudation from roots[140]. Therefore, it will be interesting to investigate whether At4g13180 is involved in SL-mediated Pi deficiency response.

The sixth protein that was up-regulated both in *Col-0* (1.30-fold increase) and *max3-12* (1.33-fold increase) was a glutathione S -transferase (GST) family protein encoded by the gene locus At4g19880. Plant GSTs perform both catalytic functions, such as glutathione conjugation in the metabolic detoxification of herbicides and natural products, and nonenzymatic functions, such as binding plant hormones to facilitate their distribution and transport. Therefore, plant GSTs are considered to be a heterogeneous superfamily of multifunctional proteins. Little is known about the role of GTSs in SL-biosynthesis, transport, or signaling. Recently, it has been reported that GR24 can increase the level of glutathione in roots in an MAX2-dependent manner, implying an involvement of glutathione in SL-regulation of root architecture[141].

It was also interesting to note that several members of the photosystems I and II (PS I and II) were up-regulated by GR24 in *Col-0* (Table 4.1). These results suggest that SL may be involved in the regulation of some photosynthetic processes. This finding is consistent with the result of microarray experiments in tomato, where SLs are found to be positive regulators of light-harvesting genes.

4.6 Proteins down-regulated by SL

Among proteins identified by our proteome analysis, the abundance of a total of 18 proteins was reproducibly decreased by more than 25% in seedlings treated with GR24 compared with the mock control in *Col-0* (Table 4.3). At least 11 of these identified proteins are involved in enzymatic reactions in diverse pathways, including methyltransferase (SMT1, sterol methyltransferase 1), ATPase (HA2, H+-ATPase 2), hydrolase (P-loop containing nucleoside triphosphate hydrolases superfamily protein), aldolase (ketose-bisphosphate aldolase class-II family protein; aldolase-type TIM barrel family protein), proteins involved in riboflavin biosynthesis, peroxidase superfamily protein, GTPase (ATRABC2B, RAB GTPase homologue C2B), GST (ATGSTF2, glutathione S-transferase PHI2), diacylglycerol kinase (ATDGK3, diacylglycerol kinase 3), phosphatase (Rhodanese/cell cycle control phosphatase superfamily protein), and receptor-like protein kinase. The involvement of these proteins in SL pathways has not been previously studied, but these findings may imply that enzymatic reactions represent an important mechanism in SL-regulated processes.

4.7 Transcriptional response of selected SL-regulated proteins

To complement the proteomics results, we applied quantitative RT-PCR (qRT) analysis to examine transcript levels of selected proteins in the *Arabidopsis Col-0* seedlings treated with GR24. We selected a total of 10 proteins, five each for GR24 up- or down-regulated. Our selection covered both those strongly (e.g., >2.0-fold change) regulated by GR24 and those weakly regulated (e.g., ~1.3-fold change) as well as those with low (e.g., 0.02) or high FDR values (e.g., 0.25) (Tables 4.2 and 4.3).

In a test run of qRT, we first wanted to examine whether we could validate those GR24-responsive genes reported by Mashiguchi *et al*. Similar to the microarray experiment, we treated *Arabidopsis Col-0* seedlings with 1.0 μM GR24 for 90 min, except that we used 10 day old seedlings whereas 14 day old seedlings were used in the microarray experiment. For GR24-up-regulated genes reported by Mashiguchi *et al*., we selected ARR16, AP2, and BRC1 (not shown in the microarray experiment but shown in one of the qRT analyses). For GR24-down-regulated genes, we selected IAA5 and At1g29440. As shown in Figure 4.6, all three selected genes reported as GR24-induced genes by Mashiguchi *et al.* were also up-regulated by GR24 in our qRT experiment, although the up-regulation is generally less than 2.0-fold. This is consistent with the microarray data, where the highest GR24-upregulation in transcript among all 31 genes was 3.56-fold. It had also been noted by Mashiguchi et al. that up-regulation of BRC by GR24 was much lower or rarely observed when the seedlings were treated with GR24 for 90 min. Similar to the results from microarray experiment, the transcript levels of IAA5 and At1g29440 were down-regulated by GR24 in our experiment (Figure 4.6). Taken together, our results suggested that we could largely validate those GR24-responsive genes reported in the microarray experiment under our experimental conditions and that our GR24 induction experiment was reliable.

Subsequently, we examined the levels of transcripts of those 10 genes selected from our proteomic analysis in 14-day old *Arabidopsis Col-0* seedlings treated with 5.0 μM GR24 for 90 min (induction time used in the microarray experiment by Mashiguchi *et al.*) and 12 h (induction time used in our proteomic study) and compared their levels of transcripts with mock control. For normalization of the measured gene expression data, we used two reference genes. The first one was ACT2, one of three traditional housekeeping genes. The second one was AtSAND, one of the 10 most stably expressed genes for *Arabidopsis* thaliana ecotype *Col-0* identified by Czechowski *et al.*[142]. The qRT data normalized using AtSAND as reference gene are presented in Figure 4.7, and the qRT data normalized using ATC2 as reference gene are presented in Figure 4.8. As shown in Figure 4.7A, expression of none of those five selected genes whose products were down-regulated by GR24 in the proteome analysis were significantly different from the mock control, implying that post-transcriptional and post-translational modifications may be important for SL's action. All transcripts of those five selected genes whose products were up-regulated by GR24 in the proteome analysis were significantly increased upon GR24 treatment (Figure 4.7B). Over 10-fold increase in the transcript level was observed in At1g76680 (encoding OPR1), At5g61820 (encoding a protein with unknown function), and At4g01870 (encoding a tolB protein-related) when *Arabidopsis Col-0* seedlings were treated with 5.0 μM GR24 for 12 h.

4.8 Discussion

SLs are a new class of plant hormones. In this study, we explored the effects of SLs on the proteome by using quantitative proteomics to uncover SL-regulated proteins in *Arabidopsis* seedlings. Because transcription and translation are not always correlated well with each other, we expect that our proteomic analysis may reveal novel players in the SL pathways.

Our iTRAQ study identified 19 GR24-up-regulated and 18 GR24-down-regulated proteins in wild-type *Arabidopsis* and 9 GR24-up-regulated and 2 GR24-down-regulated proteins in the *max3-12* mutant. The fold changes of protein abundances between GR24 treatment and mock control ranged from 0.43 to 2.52. This detection range of protein abundance was consistent with previous iTRAQ studies using *Arabidopsis*. For example, in an iTRAQ study using *Arabidopsis* guard cells, protein abundance ratios varied from 0.6 to 2.8 ($P < 0.05$). In another iTRAQ study for analyzing early changes to the phosphoproteome during the defense response to *Pseudomonas syringae* pv tomato DC3000, the fold changes in protein abundance ranged from 0.5 to 3.0 ($p < 0.05$). In a third iTRAQ study of *Arabidopsis* chloroplast

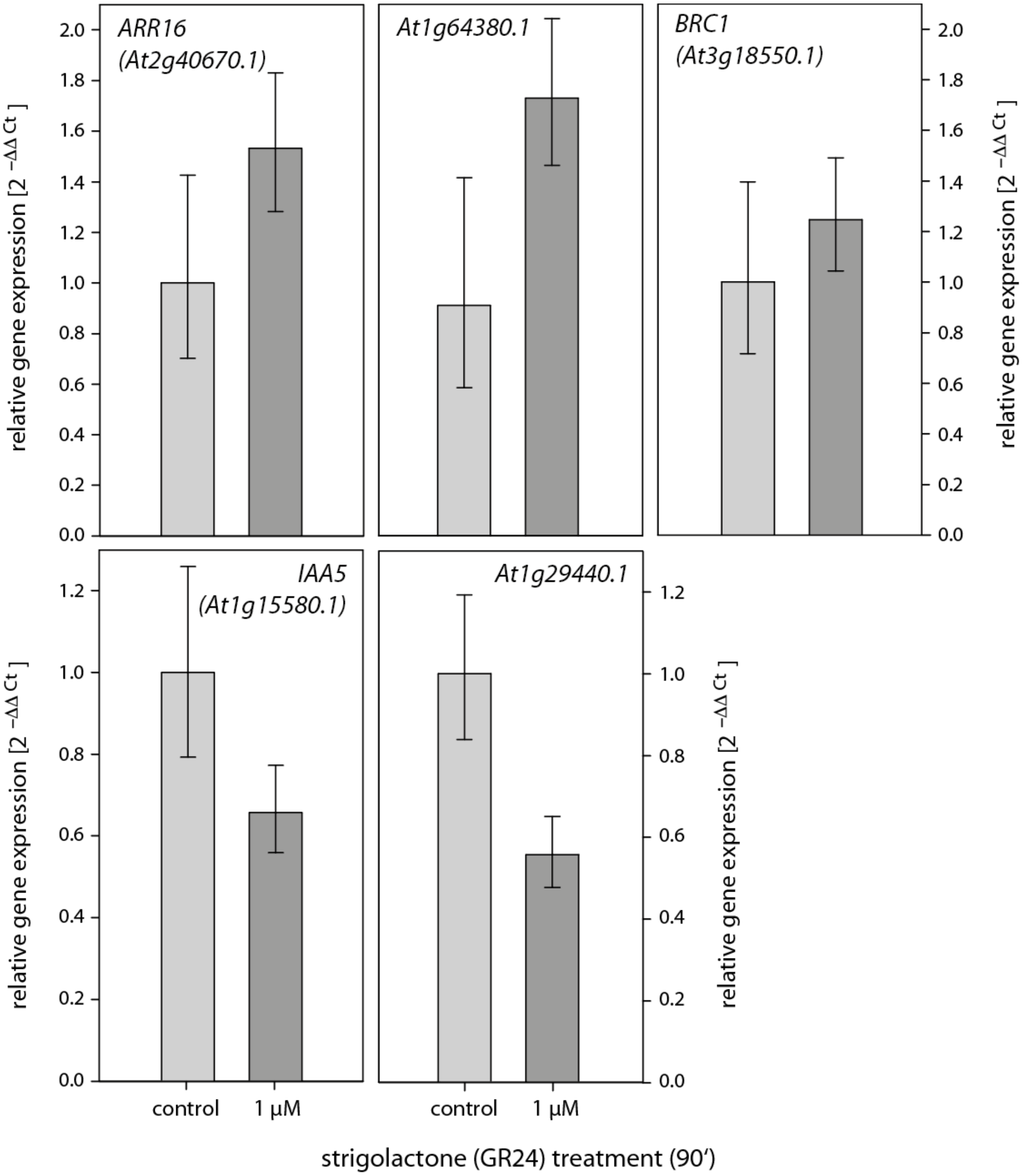


Figure 4.6 Quantitative RT-PCR analysis of GR24-responsive genes. Genes were selected from the microarray experiment previously performed by Mashiguchi *et al*. RNA was extracted from 10-day old *Arabidopsis Col-0* seedlings treated with or without 1µM GR24 for 90 min. Expression levels were quantified by real-time RT-PCR and calculated by the $2^{-\Delta\Delta Ct}$ method normalized against *AtACT2* expression. Expression data are compared to the untreated control and shown as means of at least three biological replicates ± S.D.

Figure 4.7 Quantitative RT-PCR results normalized using *AtSAND* as reference gene. A total of 10 proteins, 5 each for GR24-down- (A) or up-regulated (B), were selected for quantitative RT-PCR analysis. RNA was extracted from 14 day old *Arabidopsis Col-0* seedlings treated with or without 5 μM GR24 for 90 min and 12 h. Expression levels were quantified by real-time PCR and calculated by the $2^{-\Delta\Delta Ct}$ method normalized against AtSAND expression. Expression data are compared with the untreated control at both time points and shown as means of at least three biological replicates ± SE. *, significant difference (F-test followed by *t* test) from untreated control, $p < 0.05$.

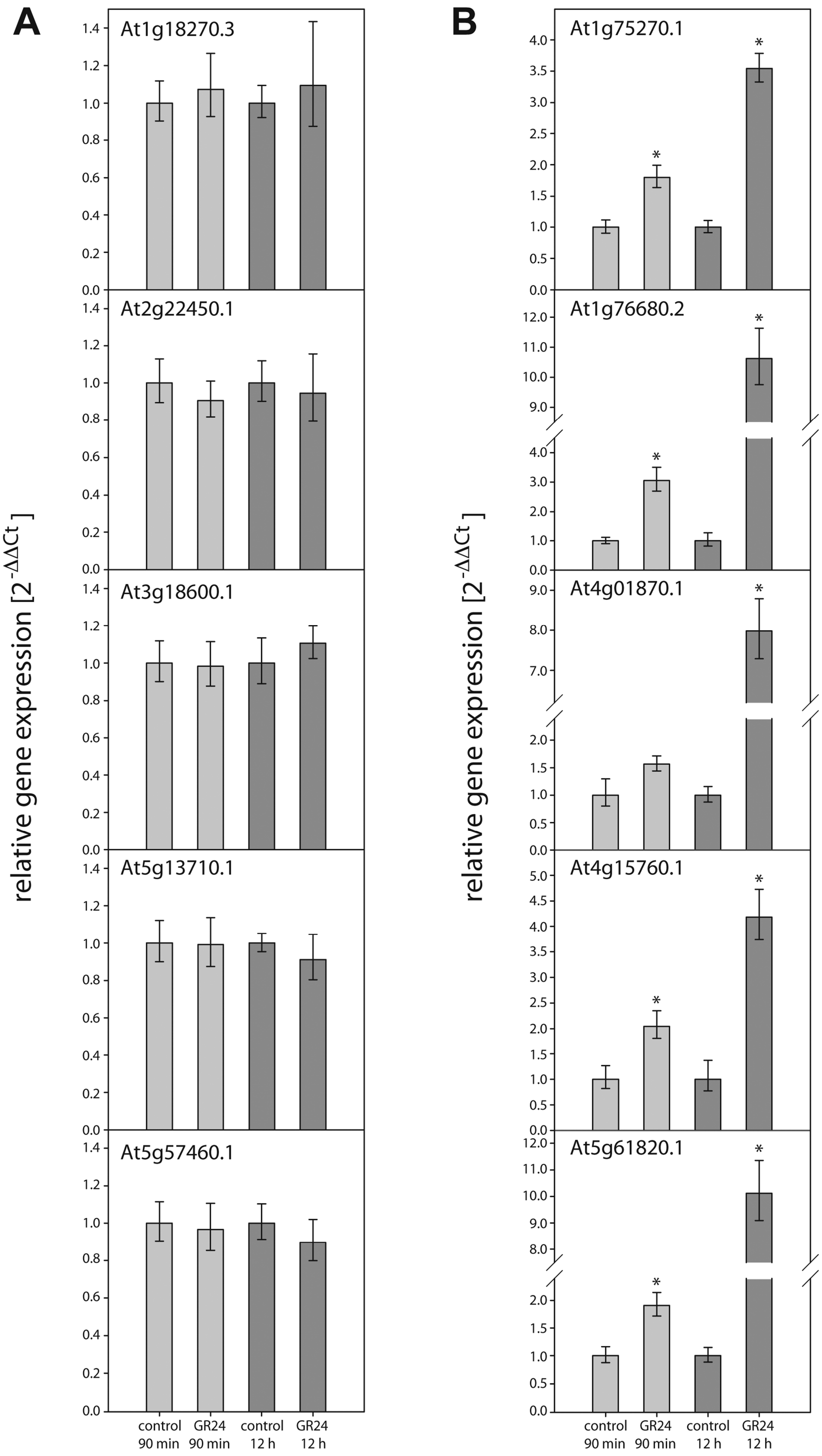


Figure 4.7 continued

Figure 4.8 Quantitative RT-PCR results normalized using *ATC2* as reference gene. A total of 10 proteins, five each for GR24-down- (A) or up-regulated (B), were selected for quantitative RT-PCR analysis. RNA was extracted from 14-days-old *Arabidopsis Col-0* seedlings treated with or without 5µM GR24 for 90 min and 12 hours. Expression levels were quantified by real-time PCR and calculated by the $2^{-\Delta\Delta Ct}$ method normalized against *AtACT2* expression. Expression data are compared to the untreated control at both time points and shown as means of at least three biological replicates ± S.E. *, significant difference from untreated control, $p<0.05$.

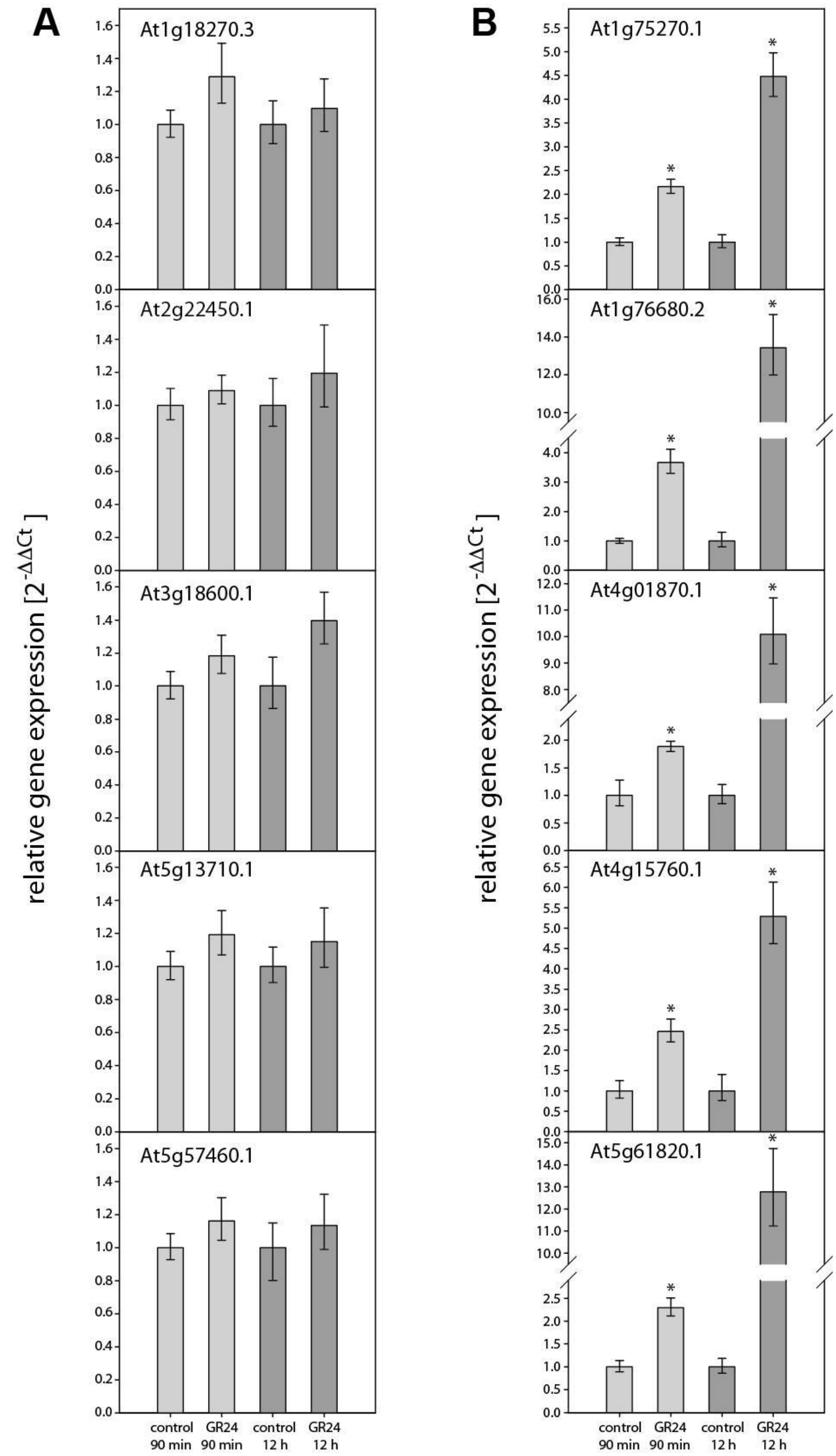


Figure 4.8 continued

SRP54 sorting mutant, fold changes of all proteins in leaves ranged from 0.59 to 2.07 ($p < 0.05$). Furthermore, our selection criteria of up- or down-regulated proteins with fold change >1.25 or <0.75 ($p < 0.01$) detected in at least two labeling experiments were similar to those imposed by previous iTRAQ studies using *Arabidopsis*. The same FDR value cutoff of 0.25 was used in the *Arabidopsis* SL microarray experiments.

In *Arabidopsis* seedlings, GR24 up-regulates the expression of 31 genes and down-regulates the expression of 33 genes. This is a relatively small set of genes compared with the number of genes regulated by other plant hormones, which typically range from several hundred to a couple of thousands (e.g., 791 IAA responsive genes and 2936 ABA-responsive genes)[143]. Furthermore, the magnitude of regulation of transcripts by GR24 is also relatively small. In the SL microarray experiment, the largest up-regulation was 3.56-fold and the largest down-regulation was 5.06-fold. In contrast, the maximum of induction observed by the plant hormone auxins could be as high as 2000-fold and the maximum suppression as high as 20-fold in *Arabidopsis* seedlings186. In another example, the maximal induction and suppression by plant hormone ABA are as high as 1666-fold and 50-fold, respectively. Therefore, compared with other plant hormones, the effect of global regulation of transcription by SLs is generally mild in *Arabidopsis* seedlings.

At the protein level, the present iTRAQ study identified a small set of GR24-responsive proteins (e.g., 19 GR24-upregulated and 18 GR24-down-regulated proteins) in *Arabidopsis* seedlings. The total number of proteins (37 proteins) with significant fold-change detected in our proteomic analysis is smaller than the total number of transcripts (64 transcripts) with significant fold-change detected in the microarray experiment. However, when considering the fact that our proteomics approach detected ∼1,700 proteins whereas typically over 10 000 transcripts can be detected in microarray studies using the *Arabidopsis* ATH1 genome array[144], the percentage of SL-responsive proteins may be actually much higher than the percentage of SL-responsive transcripts at the whole genome level, implying an important role of protein regulation in SL's action. The number of proteins detected with significant fold change in our study is comparable to previous proteomic studies using other plant hormones. For example, in a proteomics study using BR, 15 BR-up-regulated and 3 BR-down-regulated proteins were detected when *Arabidopsis* seedlings were treated with BR for 12 h.

In contrast with the microarray study in which the expression of many auxin-responsive genes was found to be repressed by GR24 (e.g., 76% GR24-repressible genes encoded auxin-inducible genes), we did not detect any of these types of proteins as affected by GR24. In fact, no proteins corresponding to GR24-responsive transcripts were detectable by iTRAQ. This could be due to a number of reasons. In the microarray experiment, 14-day-old seedlings were treated with 1 μM GR24 for 90 min, whereas in our iTRAQ experiment, seedlings were treated with 5 μM GR24 for 12 h. The concentration of GR24 used and duration of treatment may have partially contributed to these differences. It should also be noted that proteomics methods typically detect proteins that are relatively abundant; quantitative proteomics is therefore limited to identifying proteins of altered abundance from among these detected proteins. Transcriptomics is capable of characterizing genes even with low expression levels. More specifically, gene products of many auxin-inducible genes, such as AUX/IAA genes, are undergoing rapid degradation, and their abundance is at very low levels in plant cells[145]. Therefore, our proteomic analysis may have missed those auxin-responsive proteins whose transcripts were shown to be repressed by GR24 in the microarray study. Consistent with this view, the iTRAQ detected only two proteins (encoded by gene loci At3g07390 and At1g28130) annotated as auxin-responsive proteins, but neither of them showed significant fold changes upon GR24 treatment and neither of them were identified as GR24-repressible genes in the microarray study. This may also partially explain the fact that none of the SL-responsive genes identified by microarray experiment was identified in our proteomic study.

None of the genes encoding those proteins whose abundance was shown to be up- or down-regulated by GR24 in our proteomics study was shown to be GR24 responsive in the microarray study. Similar trends were found with BR proteomic studies where ~ 80% of the BR-responsive proteins were not identified in microarray studies. Furthermore, in our qRT-PCR test with five selected genes whose gene product was shown to be down-regulated by GR24, we did not detect corresponding down-regulation at the transcript level. Similarly, direct comparison between protein and RNA changes in BR mutants also revealed a very weak correlation. These results argue for an important role of post-transcriptional or post-translational process in SL pathways. Consistent with this view, it has been found that microRNA plays an important role in regulating shoot branching[146], a process where SLs are best known to function as a negative regulator. More importantly, MAX2 functions as a key signaling component in the SL pathway.

MAX2 is an F-box protein that forms a protein complex with the core SCF complex (Skp, Cullin, F-box-containing complex) subunits ASK1 and AtCUL1[147], indicating that ubiquitin-proteasome-mediated protein degradation plays an important role in SL signaling. Consistent with this view, recently, it has been shown that the degradation of D53 protein is dependent on D14 and D3 (a rice ortholog of MAX2). *Arabidopsis* orthologs of D53 (the SMAX1 and SMXL proteins[148]) were at undetectable levels in our proteomics studies.

Our proteomic analysis has revealed several proteins that could potentially fit in the SL pathways. For example, the protein encoded by gene locus At5g61820 has similarity with proteins implicated in nodule development in the legume Medicago, and it has been shown that SLs promote nodulation in pea. Another example is the NAD(P)-binding Rossmann-fold superfamily protein encoded by gene locus At4g13180, which was up-regulated by GR24 in our study. This gene was among one of the early phosphate starvation responsive genes. As previously noted, Pi deficiency can stimulate SL production and exudation from roots. Furthermore, several proteins that are known or predicted to have a role in photosynthesis were also up-regulated by GR24 in our proteomics study, consistent with the finding that SLs function as positive regulators of light-harvesting genes

Another outcome of our proteomic analysis was the identification of proteins that have not been previously known to have a role in SL pathways. For example, OPR1, whose abundance was up-regulated by GR24 at the highest level, was implicated in TNT detoxification. A tolB-related protein was also shown to be up-regulated by GR24 in our study. Although the molecular function of tolB-related protein in plants is still unknown, this protein was implicated in plant detoxification and stress responses. We also noticed that the action of both OPR1 and this tolB-related protein might potentially involve OPDA-related pathways because OPR1 is annotated as a 12-oxophytodienoate reductase, although it showed little activity with the naturally occurring OPDA isomer, whereas the transcript of this tolB-related protein was shown to be induced by compounds that are structurally highly similar to OPDA. Further proteome studies using more time points and different concentrations of SL as well as using mature plants may help reveal a more comprehensive view of SL-regulated proteins.

In summary, our iTRAQ studies have uncovered about three dozens of proteins that have not been previously known to have any roles in SL pathways. Further characterization of these SL-regulated proteins may provide new insights into the molecular mechanism of action of SLs.

CHAPTER 5
ELEVATED TEMPERATURE ALTERS CARBON CYCLING IN A MICROBIAL COMMUNITY

All of the data presented below has been adapted from

Annika Mosier, Zhou Li, Brian Thomas, Robert Hettich, Chongle Pan, and Jillian Banfield. “Elevated temperature alters carbon cycling in a microbial community”. Submitted to *the ISME Journal*

Zhou Li’s contributions include MS measurement, data analysis, and co-writing the manuscript.

5.1 Introduction

The impacts of elevated temperature on microbial communities will have direct implications for ecosystem- and global-scale processes. Many microbial community studies have evaluated the effect of warming on overall population structure and on specific metabolic processes such as respiration[149]. Far fewer studies have comprehensively assessed functional responses across the entire community (e.g., using “omic” approaches[150] or functional gene arrays[151]).

Individual microbial groups (e.g., genotypes, species, or functional groups) will likely have different functional responses to elevated temperature, and yet an organism’s response and adaptation to changing conditions in part relates to its behavior within a community. Thus, understanding the physiology and activity of individual microbial groups within a community context is essential for predicting the impact, resilience and response of ecological systems to changing conditions. This topic is relatively little studied, in part because it can be challenging to tease apart contributions of individual organisms from overall metabolic processes. Further, such investigations require a high level of taxonomic and functional resolution because closely related strains and species may respond very differently to temperature regime.

Quantitative proteomics can elucidate function of individual microbial groups within a community context by measuring protein abundance in a high-throughput manner. Both taxonomic and functional annotations are simultaneously assigned to unique proteins in the community proteome. Protein abundance can more accurately represent cellular activities than mRNA quantification, because mRNA abundance changes do not necessarily correlated with protein abundance change[79]. For example, some proteins may have long lifetimes, so that new production from mRNA is infrequently required. Conversely, cells may also have a low level of a protein with abundant mRNA expression because of protein degradation and post-transcription regulation.

Recent advances in protein quantification using tandem mass tags (TMTs) and isobaric tags for relative and absolute quantification (iTRAQ) have dramatically improved measurement precision, accuracy, and reproducibility, surpassing label-free quantification methods such as spectral counting[57]. TMT/iTRAQ-based quantitative proteomics can be used with complex samples, including biological systems that are not amenable to efficient metabolic labeling with stable isotopes. In isobaric chemical labeling, peptides from different samples are labeled separately with different isotopic variants of the labeling reagent and then combined for analysis using liquid chromatography coupled with tandem mass spectrometry (LC–MS/MS). Each isotopic variant has the same overall mass but contains a reporter ion with a unique molecular mass, thus enabling accurate overall quantification alongside precise measurement of the relative protein abundance between samples. Currently, TMT/iTRAQ-based quantitative proteomics enables multiplexing of up to 8-10 samples with deep proteome coverage.

The objective of this study was to determine the impact of elevated temperature on the physiology of individual microbial groups in a community. The experiments were conducted at temperatures between the average *in situ* temperature and the maximum growth temperature, which was established in this study. We compared the protein expression levels using a new approach that combined shotgun community proteomics analysis with TMT quantification. The analyses targeted laboratory-grown acid mine drainage (AMD) biofilms that represent natural AMD populations and have served as a model microbial community system in many prior studies. The current research shows the utility of quantitative proteomics for understanding ecological processes by highlighting differential expression of closely related organisms.

5.2 Additional materials and methods

Fluorescence *in situ* hybridization: Fluorescence *in situ* hybridization (FISH) was carried out on fixed (4% paraformaldehyde) AMD biofilm samples as described previously[152,153]. Oligonucleotide probes used in this study for identification of the dominant individual species and groups were as follows: EUBMIX (all Bacteria); ARC915 (all Archaea); EUKMIX (all Eukaryotes); LF655 (all *Leptospirillum* bacteria); LF1252 (*Leptospirillum* group III bacteria);

L2UBA353 (*Leptospirillum* group II UBA-genotype); L2CG353 (*Leptospirillum* group II 5-way genotype); and SUL230 (Sulfobacillus spp.). For estimation of abundance, cells were counted in three replicate fields of view (average of 468 total cells counted per probe per sample) and converted to a percentage of the total cell count found using the general nucleic acid stain 4',6-diamidino-2-phenylindole (DAPI). Cell counts from both growth phases were averaged.

Proteome-based community structure and function analyses: Hierarchical clustering was performed on protein abundance values normalized at the community-level with absolute intensities converted to percentages for each protein (the sum of the percentages for each protein is equal to 1). The clustering method used a Pearson correlation distance matrix and average linkage clustering (using Multi-experiment Viewer; MeV_4_8; http:// www.tm4.org/mev/)[154]. For further analyses, protein counts from both growth phases were summed. Community structure was evaluated by summing the total intensities of the proteins for each organismal group (e.g., *Leptospirillum* group III, archaea, etc.) and then dividing by the total sum of all proteins in a sample. Differentially expressed proteins were identified as those with normalized total intensity ratios (40 ºC : 46 ºC) >1.2 or <0.8 combined with a Rank Product p-value ≤0.05 (except where noted), similar to methods used in other studies[155]. Functional categories of significant proteins were assigned upon manual review of annotations in ggKbase (http://ggkbase.berkeley.edu/) [including Clusters of Orthologous Groups[156] assignment], as well as reciprocal blast searches against the KEGG database (conducted using the KAAS[157] server). Carbohydrate Active Enzymes (CAZymes) were predicted with the CAZymes Analysis Toolkit (http://mothra.ornl.gov/cgi-bin/cat.cgi)[158] with an e-value threshold of 0.0001 for Pfam searches and 0.000001 for orthology searches with "Domain consistent" and "Length Consistent" rules.

5.3 Growth of AMD biofilms at different temperatures

During periods of observed biofilm accumulation at the AB-muck site in the Richmond Mine over the last decade, the average *in situ* temperature of the AMD solution was 40.4 ºC (based on discrete measurements during sampling trips; Figure 5.1). Biofilms were never observed at temperatures above 47 °C.

AMD biofilms were grown in laboratory bioreactors in order to evaluate the effect of elevated temperature on community composition and function. Mature biofilms developed at 40 ºC, 43 ºC, and 46 ºC. This temperature range corresponds with the normal range of temperatures associated with biofilm growth in the field. There was no visible biofilm growth at 49 ºC after four weeks. Biofilm growth rates may have differed between the temperature treatments; however, we expect that any differences that might have occurred were likely a result of temperature since all other growth conditions were identical.

Tandem mass tag quantitative shotgun proteomics (TMT-proteomics) was used to determine protein abundance and inferred function in bioreactor-grown biofilms grown at 40 ºC, 43 ºC, 46 ºC. TMT-proteomics identified 1724-1916 proteins from the biofilm communities (across all samples, extraction replicates, and technical runs), 1596 of which could be uniquely assigned to one organism. Hierarchical clustering showed that the samples reproducibly clustered into two groups based on their protein abundance levels: biofilms grown at 40 ºC and 43 ºC clustered together, whereas those grown at 46 ºC clustered independently (Figure 5.2).

5.4 Community composition of AMD biofilms grown at different temperatures

Proteins were quantified from 23 different bacterial, archaeal, and eukaryal organisms (Table 5.1). We evaluated community composition based on FISH and TMT-proteomics measurements. FISH estimates indicated that archaea were very abundant in the field-collected and bioreactor biofilms, making up 35 to 51% of the communities (Figure 5.3). Proteins were identified from many different archaea: ARMAN I, ARMAN II, ARMAN IV, ARMAN V, Ferroplasma I, Ferroplasma II, A-plasma I, A-plasma II, C-plasma, D-plasma, E-plasma, G-plasma and I-plasma. While the overall abundance of archaea did not change significantly with increasing temperature, closely related organisms responded differently to temperature (Figure5.4, Table 5.2). For instance, ARMAN II abundance increased with temperature, but ARMAN IV decreased with temperature. Two other ARMAN types had similar abundance levels at 40ºC and 46ºC, but lower abundance at 43ºC.

The chemoautotrophic iron-oxidizing bacteria *Leptospirillum* Group II, which often dominate natural biofilms, were present in all of the biofilms. Elevated temperature differentially impacted the abundance of three distinct *Leptospirillum* Group II organisms referred to as the Type I (5-way), Type III (C75), and Type VI (UBA) genotypic groups[159-162]. The UBA and C75 genotypes increased in abundance from 40ºC to 46ºC, whereas the 5-way genotype abundance

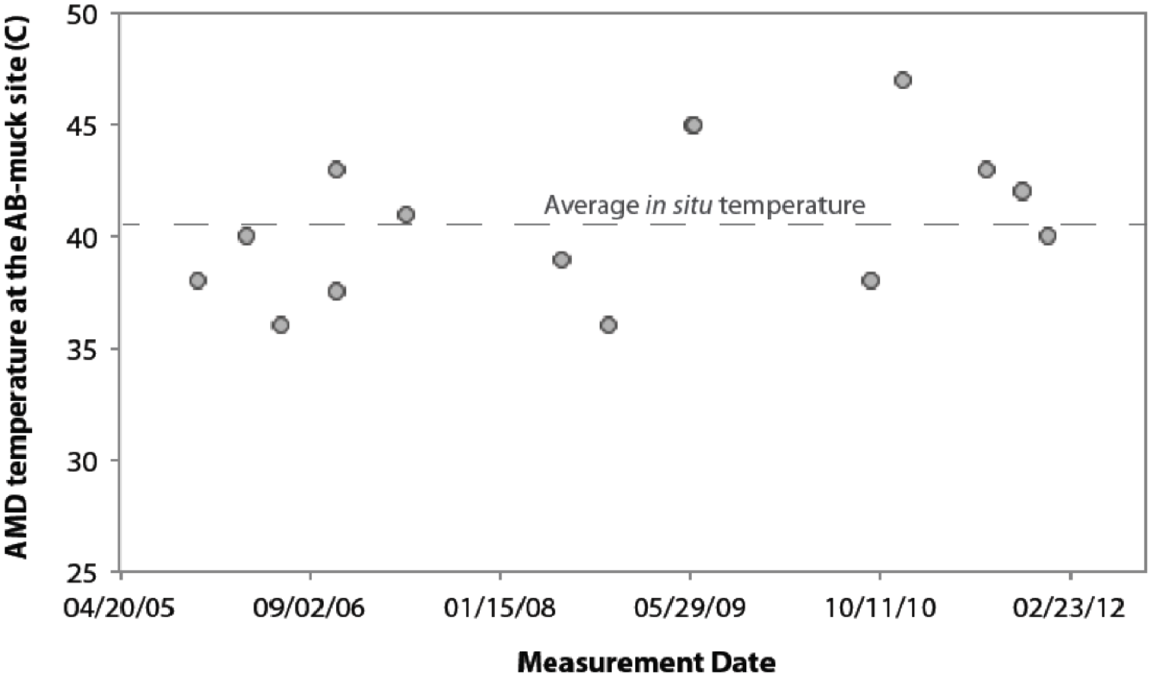


Figure 5.1 Temperature of the acid mine drainage solution at the AB-muck site in the Richmond Mine.

Figure 5.2 Hierarchical clustering of protein abundance values normalized at the community-level (A) and the number of proteins assigned to COGs that are significantly different between temperatures at the community-level (B). COG categories J: Translation, ribosomal structure and biogenesis; A: RNA processing and modification; K: Transcription; L: Replication, recombination and repair; B: Chromatin structure and dynamics; D: Cell cycle control, cell division, chromosome partitioning; Y: Nuclear structure; V: Defense mechanisms; T: Signal transduction mechanisms; M: Cell wall/membrane/envelope biogenesis; N: Cell motility; Z: Cytoskeleton; W: Extracellular structures; U: Intracellular trafficking, secretion, and vesicular transport; O: Posttranslational modification, protein turnover, chaperones; C: Energy production and conversion; G: Carbohydrate transport and metabolism; E: Amino acid transport and metabolism; F: Nucleotide transport and metabolism; H: Coenzyme transport and metabolism; I: Lipid transport and metabolism; P: Inorganic ion transport and metabolism; and Q: Secondary metabolites biosynthesis, transport and catabolism

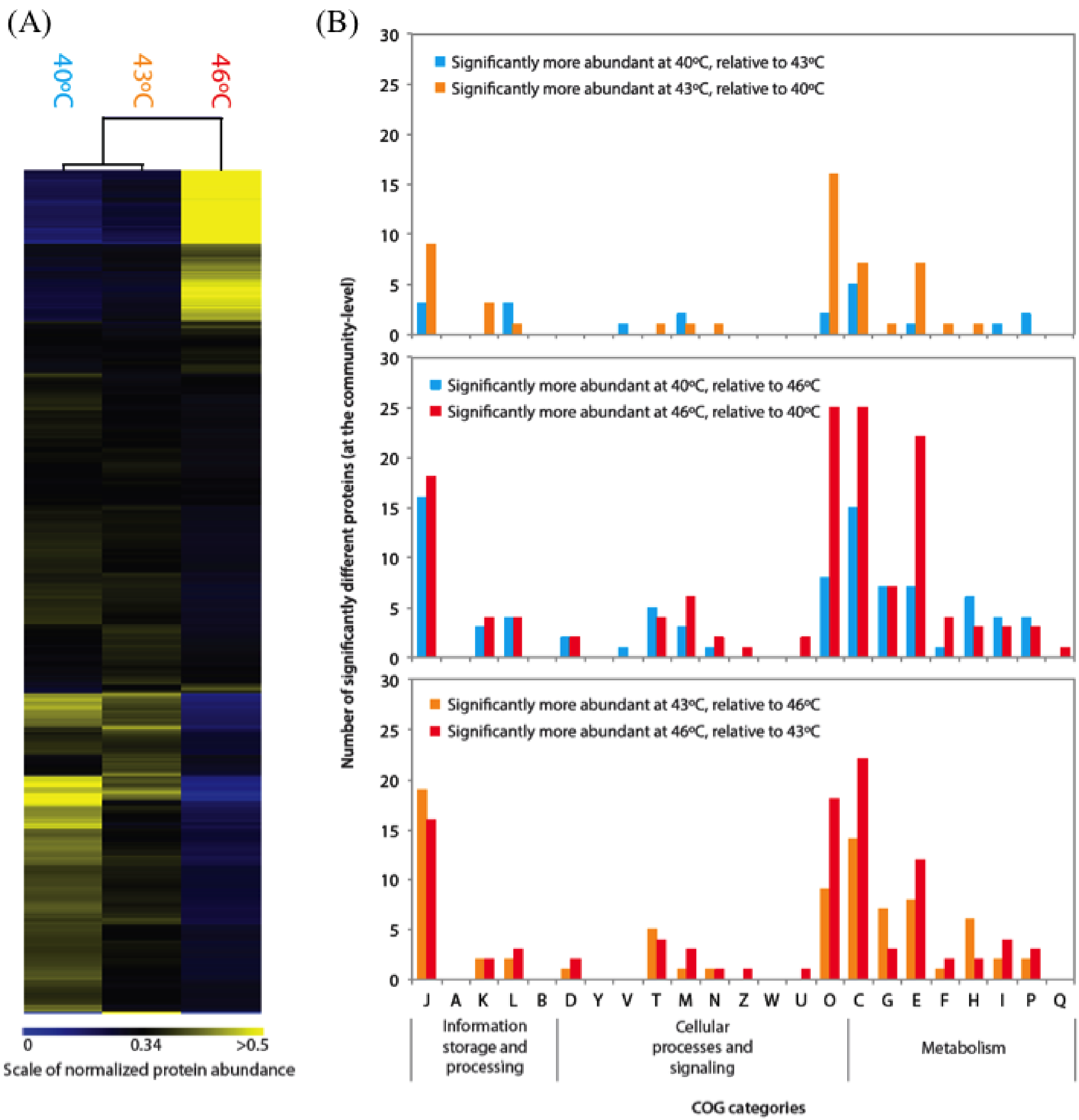


Figure 5.2 continued

Table 5.1 Number of proteins quantified per organism. Only proteins quantified with unique peptides were considered.

domain	Organism	No. of Proteins
Bacteria	*Leptospirillum* group II CG 5-way type	195
	Leptospirillum group II UBA type	197
	Leptospirillum group II C75 type	11
	Leptospirillum group III	695
	Leptospirillum group IV	27
	Actinobacteria I	7
	Actinobacteria II	7
	Firmicute	3
	Sulfobacillus III	6
Archaea	A-plasma I	74
	A-plasma II	1
	C-plasma	4
	D-plasma	6
	E-plasma	4
	G-plasma	206
	I-plasma	10
	Ferroplasma I	25
	Ferroplasma II	12
	Arman I	6
	Arman II	4
	Arman IV	11
	Arman V	6
Eukarya	*Acidomyces richmondensis*	28

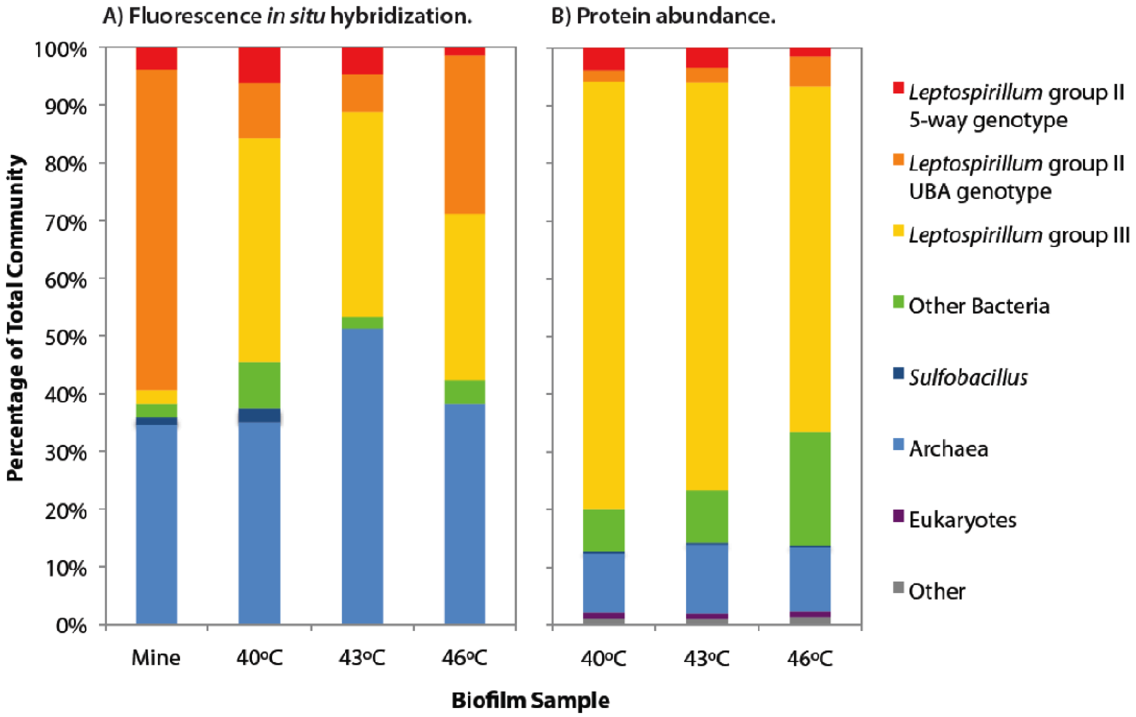


Figure 5.3 Community composition of natural and cultivated AMD communities based on A) fluorescence *in situ* hybridization and B) protein abundance as measured by TMT-proteomics.

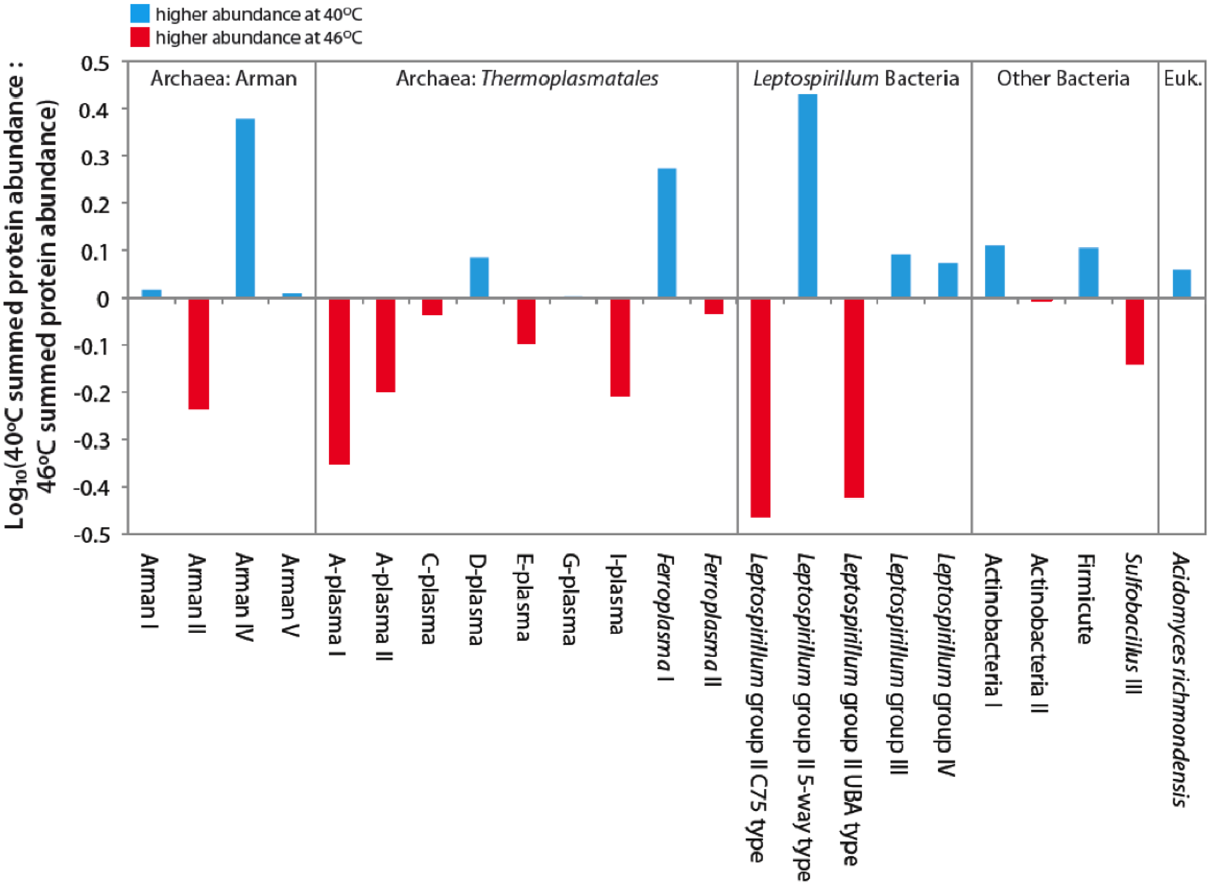
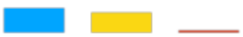
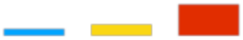
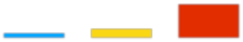
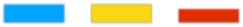
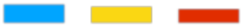
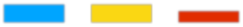
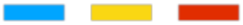
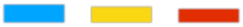
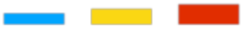
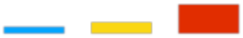
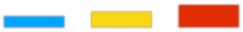
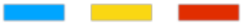
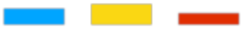
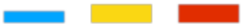
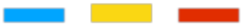
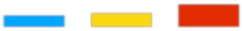
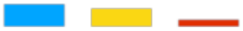
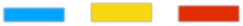
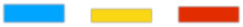
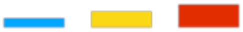
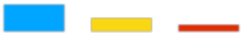
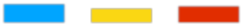
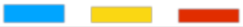


Figure 5.4 Relative abundance of archaea, bacteria, and eukarya in AMD biofilms grown at 40ºC and 46ºC (based on protein abundance as measured by TMT-proteomics).

Table 5.2 Relative abundance of each organism at 40ºC, 43ºC, and 46ºC (based on summing the total intensities of the proteins for each organism and then dividing by the total sum of all proteins in a sample). Abundance plots are scaled by normalizing to 100% for each organism.

Organism	Scaled Abundance Plot			Relative Abundance (%)		
Bacteria:	**40°C**	**43°C**	**46°C**	**40°C**	**43°C**	**46°C**
Leptospirillum group II CG 5-way type				3.893	3.484	1.450
Leptospirillum group II UBA type				1.952	2.469	5.172
Leptospirillum group II C75 type				6.544	8.451	19.020
Leptospirillum group III				74.120	70.660	59.952
Leptospirillum group IV				0.539	0.520	0.456
Actinobacteria I				0.070	0.067	0.054
Actinobacteria II				0.092	0.090	0.093
Firmicute				0.007	0.006	0.005
Sulfobacillus III				0.195	0.231	0.269
Archaea:						
A-plasma I				0.825	1.009	1.860
A-plasma II				0.000	0.001	0.001
C-plasma				0.051	0.055	0.055
D-plasma				0.050	0.056	0.041
E-plasma				0.006	0.007	0.007
G-plasma				8.129	9.631	8.103
I-plasma				0.136	0.157	0.220
Ferroplasma I				0.307	0.272	0.164
Ferroplasma II				0.175	0.204	0.190
Arman I				0.022	0.020	0.021
Arman II				0.156	0.224	0.269
Arman IV				0.519	0.322	0.217
Arman V				0.014	0.012	0.013
Eukarya:						
Acidomyces richmondensis				1.147	1.035	1.000

decreased. *Leptospirillum* group III bacteria were very abundant at all temperatures in the bioreactors. As seen previously[163], the bioreactor biofilms had a much higher percentage of *Leptospirillum* group III bacteria than that seen in the *in situ* mine biofilm (31 to 41% compared to only 2% in the mine). *Leptospirillum* Group III has been shown to dominate AMD biofilms in solutions with low Fe(II)/Fe(III) ratios (Spaulding *et al*., unpublished data).

5.5 Community function at low and high temperatures

Protein expression was further evaluated to determine if elevated temperature impacted function. Protein abundance was first normalized at the community-level to determine each protein's abundance compared with all proteins in the sample (normalizing to account for biomass differences between samples, but not accounting for differences in each organism's abundance).

In a COG-based functional analysis (Figure 5.2), the greatest number of significantly different proteins occurred when comparing biofilms grown at 40 ºC and 46 ºC. There were fewer significantly different proteins in the 40 ºC: 43 ºC and 43 ºC: 46 ºC comparisons (78 significantly different proteins between 40 ºC and 43 ºC; 191 significantly different proteins between 43 ºC and 46 ºC; and 239 significantly different proteins between 40 ºC and 46 ºC).

Overall, increasing temperature led to an increasing number of significantly different proteins in the COG functional categories (E) amino acid transport and metabolism, (C) energy production and conversion, and (O) posttranslational modification, protein turnover, chaperones. In particular, more than three times as many proteins involved in the metabolism and transport of amino acids (COG E) were significantly more abundant at 46ºC than at 40ºC. Nearly twice as many proteins involved in energy production and conversion (COG C) were significantly more abundant at 46ºC compared to 40°C. Additionally, there were 3.1 times as many proteins in the functional category of posttranslational modification, protein turnover, and chaperones (COG O) that were significantly more abundant at 46ºC than at 40ºC.

5.6 Function of individual organisms in biofilms growing at 40 ºC and 46 ºC

Protein abundance was evaluated at the organism-level by normalizing individual proteins to the total protein abundance from each specific organism, allowing for evaluation of protein abundance for individual organisms. Organisms representing ≥10% of the total proteins were analyzed, including three closely related *Leptospirillum* bacteria, as well as G-plasma archaea. In the significance analysis (based on fold-change and Rank Product p-value), proteins that are considered as up-regulated in one condition are concomitantly considered as down-regulated under the other condition. Protein expression at the organism level was only analyzed between the 40ºC and 46ºC conditions because very few proteins were significantly between 40ºC and 43ºC (ranging from 2 to 18 proteins per organism-level comparison). Additionally, the 40ºC and 43ºC community-level proteomes were similar (based on hierarchical clustering and community COG analysis).

5.7 Function of *Leptospirillum* bacteria in biofilms growing at 40 ºC and 46 ºC

Protein abundance was evaluated at the organism-level for three closely related *Leptospirillum* bacteria: *Leptospirillum* group II UBA genotype, *Leptospirillum* group II 5way genotype, and *Leptospirillum* group III. Overall, 144 proteins were significantly different between 40 ºC and 46 ºC for the three organisms, spanning a broad range of functions.

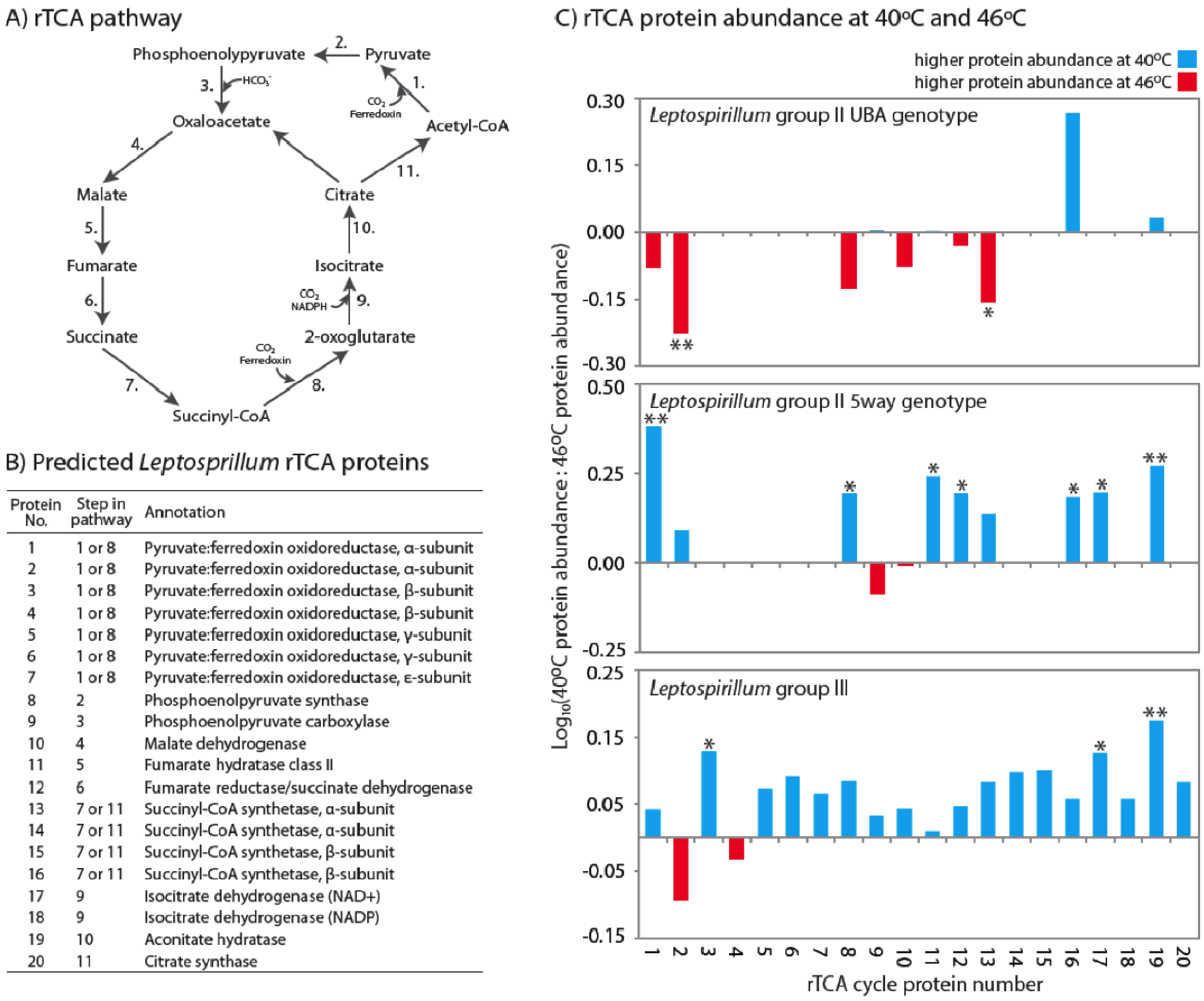
(I) Protein folding, sorting and degradation: several proteins involved in protein degradation were significantly up-regulated at 46 ºC for *Leptospirillum* group III and the group II UBA genotype. Of the proteins with the highest total intensities in the entire dataset (top 5% for each of the 3 organisms at 40 ºC and 46 ºC), 13% were chaperones. DnaK and ClpB chaperones were significantly up-regulated at 40 ºC for the group II 5way genotype and at 46 ºC for the UBA genotype. *Leptospirillum* group III bacteria had GroEL and HscA (p=0.06) chaperones and a chaperonin that were significantly up-regulated at 46ºC. Two trigger factors (ribosome associated chaperones) were up-regulated at 40ºC (p=0.01, p=0.09).

(II) Carbon cycling: Carbon fixation by the *Leptospirillum* group II (UBA and 5way genotypes) and III bacteria responded strongly to temperature (Figure 5.5). These organisms are believed to fix carbon via the reductive tricarboxylic acid (rTCA) cycle[164]. Of the 60 different *Leptospirillum* proteins predicted to be involved in rTCA, 41were detected and quantified. Many of these proteins had very high total intensities: 11 were ranked in the top 5% highest total intensities. The majority of rTCA proteins from *Leptospirillum* group III and the group II 5way genotype were more abundant at 40ºC than at 46ºC. Many of these proteins were significantly up-regulated at 40ºC relative to 46ºC (7 proteins for 5way and three for group III). Conversely, rTCA proteins for the *Leptospirillum* group II UBA genotype had the opposite abundance pattern. The majority of the rTCA proteins had higher intensities at 46ºC, two of which were significantly up-regulated relative to 40ºC.

Twenty-five Carbohydrate Active Enzymes (CAZymes) were predicted amongst the quantified *Leptospirillum* proteins (additional CAZymes are found within the complete *Leptospirillum* genomes but were not measured here). CAZymes are classified as families of structurally-related enzymes that degrade, modify, or create glycosidic bonds. Only one CAZyme was predicted for the *Leptospirillum* group II 5way genotype (GH57) and four for the *Leptospirillum* group II UBA genotype (CBM13, GH109, and two GH57s). *Leptospirillum* group III had 20 predicted CAZymes (Figure 5.6). Most (7 out of 8) of the carbohydrate esterases (CEs; hydrolysis of carbohydrate esters) and glycosyltransferases (GTs; biosynthesis of saccharides) had higher total intensities at 46 ºC and one was significantly up-regulated (GT2). Two glycoside hydrolases (GH3 with ß-N-acetylhexosaminidase activity and GH109 with an oxidoreductase domain) were also significantly up-regulated at 46 ºC. Half of the *Leptospirillum* group III CAZymes were related to GH families (GH13 and GH57) acting on substrates containing α-glucoside linkages including starch, glycogen, and α-maltose: 8/10 of these proteins had higher total intensities at 40 ºC, three of which were significantly up-regulated relative to 46ºC. The *Leptospirillum* group II UBA genotype also had one GH57 protein that was also significantly up-regulated at 40 ºC. The *Leptospirillum* group II 5way genotype had one predicted GH57 protein that had a higher total intensity at 46ºC but was not significantly different than 40 ºC.

(III) Energy production: The proposed mechanism of iron oxidation by *Leptospirillum* bacteria suggests that Cytochrome572 (Cyt572) functions as the Fe(II) oxidase, oxidizing Fe(II) on the surface of cells and transferring electrons to Cytochrome579 (Cyt579)[165-167]. However, further studies suggest additional routes of iron oxidation, including the possibility that both Cyt572 and Cyt579 act as Fe(II) oxidases or c-type cytochromes act as initial electron acceptors from Cyt572[168].

For *Leptospirillum* group III, abundance of iron oxidation proteins at 40ºC and 46ºC was decoupled between the initial steps of electron transfer and last steps converting oxygen to water and generating ATP. Three cytochromes were more abundant at 46 ºC (2 of which were significantly up-regulated: Cyt579 p = 0.05; cytochrome C p = 0.03), whereas 15 out of 20 downstream proteins were more abundant at 40 ºC (including 3 significant proteins: a cytochrome C oxidase and 2 ATP synthase subunits p = 0.003, 0.03, 0.02). It is unclear how the



B) Predicted *Leptosprillum* rTCA proteins

Protein No.	Step in pathway	Annotation
1	1 or 8	Pyruvate:ferredoxin oxidoreductase, α-subunit
2	1 or 8	Pyruvate:ferredoxin oxidoreductase, α-subunit
3	1 or 8	Pyruvate:ferredoxin oxidoreductase, β-subunit
4	1 or 8	Pyruvate:ferredoxin oxidoreductase, β-subunit
5	1 or 8	Pyruvate:ferredoxin oxidoreductase, γ-subunit
6	1 or 8	Pyruvate:ferredoxin oxidoreductase, γ-subunit
7	1 or 8	Pyruvate:ferredoxin oxidoreductase, ε-subunit
8	2	Phosphoenolpyruvate synthase
9	3	Phosphoenolpyruvate carboxylase
10	4	Malate dehydrogenase
11	5	Fumarate hydratase class II
12	6	Fumarate reductase/succinate dehydrogenase
13	7 or 11	Succinyl-CoA synthetase, α-subunit
14	7 or 11	Succinyl-CoA synthetase, α-subunit
15	7 or 11	Succinyl-CoA synthetase, β-subunit
16	7 or 11	Succinyl-CoA synthetase, β-subunit
17	9	Isocitrate dehydrogenase (NAD+)
18	9	Isocitrate dehydrogenase (NADP)
19	10	Aconitate hydratase
20	11	Citrate synthase

Figure 5.5 Abundance of proteins involved in CO_2 fixation for *Leptospirillum* group II (UBA and 5way genotypes) and group III. (A) Reductive tricarboxylic acid (rTCA) pathway, as proposed by Aliaga Goltsman *et al*., 2009. (B) *Leptospirillum* proteins predicted to be involved in rTCA (Aliaga Goltsman *et al*., 2009). (C) Abundance of rTCA proteins at 40ºC and 46ºC. Stars indicate proteins that are significantly different at a given temperature (fold-change >1.2 or <0.8; one star p≤0.1; two stars p≤0.05).

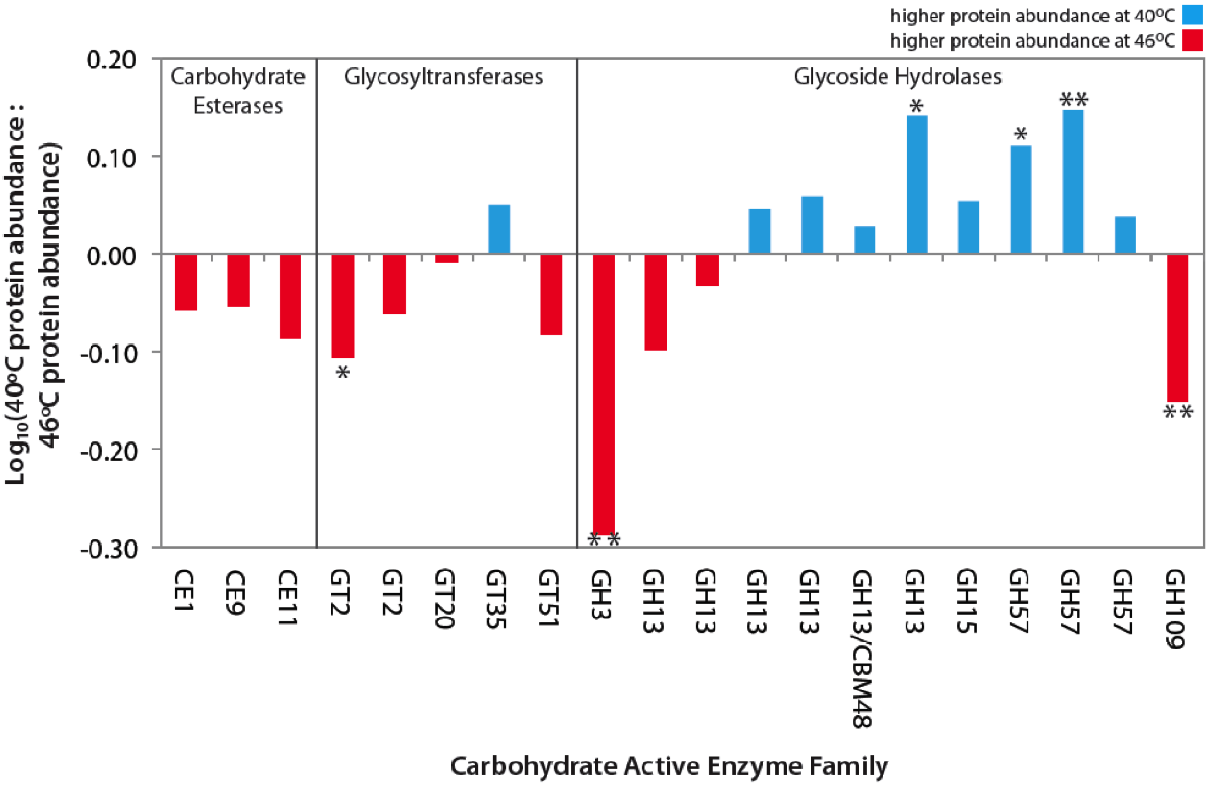


Figure 5.6 Abundance of predicted Carbohydrate Active Enzymes at 40ºC and 46ºC for *Leptospirillum* group III. Stars indicate proteins that are significantly different at a given temperature (fold-change >1.2 or <0.8; one star p≤0.1; two stars p≤0.05).

uncoupling of iron oxidation, electron transfer and ATP generation affects energy generation within the cells.

(IV) Amino acid metabolism: at higher temperatures, the *Leptospirillum* bacteria (the group II 5way and UBA genotypes and Group III) increased expression of proteins involved in amino acid metabolism. Twelve amino acid biosynthesis and degradation proteins were significantly up-regulated at 46ºC, whereas only 1 was up-regulated at 40 ºC. Among those proteins up-regulated at 46ºC were those involved in the biosynthesis of alanine, lysine, glutamate, cysteine, isoleucine, and tryptophan. Additionally, 4 separate proteins in the histidine biosynthesis pathway were up-regulated at 46ºC for the *Leptospirillum* group II 5way genotype (p = 0.01-0.07).

(V) Genetic information processing: of proteins with the highest total intensities (top 5% for each of the three *Leptospirillum* bacteria), 31% were involved in genetic information processing functions. More than 3.4 times as many proteins involved in genetic information processing were significantly up-regulated at 40 ºC relative to 46 ºC (24 at 40 ºC versus 7 at 46 ºC) including functions of replication, recombination and repair; transcription; translation; and nucleotide transport and metabolism. More ribosomal proteins were significantly up-regulated at 40 ºC relative to 46 ºC for all three *Leptospirillum* bacteria, but most striking was *Leptospirillum* group III which had 13 ribosomal proteins significantly up-regulated at 40 ºC and only one at 46 ºC.

(VI) Chemotaxis and Stress: Methyl-accepting chemotaxis sensory transducer proteins were significantly up-regulated at 40 ºC for each of the three *Leptospirillum* bacteria ($p \leq 0.05$ except for the UBA genotype with p-values of 0.09 and 0.08). The *Leptospirillum* bacteria exhibited various stress responses: an oxidative stress protein was significantly up-regulated at 46 ºC for the UBA genotype; an osmotic stress protein was significantly up-regulated at 40 ºC for the 5way genotype; and a metal stress protein was significantly up-regulated at 40 ºC for *Leptospirillum* group III. One phage integrase protein was up-regulated at 40 ºC for *Leptospirillum* group III. The *Leptospirillum* group III genome contains a cluster of Cas genes, which are CRISPR-associated genes involved in viral defense. Five of these Cas proteins were up-regulated at 46 ºC (p = 0.013-0.095; one with an abundance ratio of 0.84). Two phage proteins were also up-regulated at 46 ºC for the 5way genotype (phage shock protein A p = 0.01;

phage integrase p = 0.08). One viral protein was detected and quantified in the dataset (AMDVIR_10150G0005).

5.8 Function of G-plasma archaea in biofilms growing at 40 ºC and 46 ºC

G-plasma (of the *Thermoplasmatales* order of *Euryarchaea*) protein abundance was also evaluated at the organism-level. Overall, only 24 proteins were significantly different between 40 ºC and 46 ºC. Functions of these significant proteins included chaperones, amino acid metabolism, genetic information processing, and transport.

A total of 34 proteins were quantified with predicted function in carbon cycling process[169], including the Entner-Doudoroff pathway, glycolysis, pyruvate dehydrogenase complex, TCA cycle, and beta oxidation (Figure 5.7). Of these, 29 proteins (85%) were more abundant at 40 ºC than at 46 ºC, though only two were significantly different.

5.9 Discussion

5.9.1 Experimentation on acid mine drainage biofilms

Here, AMD biofilms were used to test an approach to determine how elevated temperature regulates physiology of individual microbial groups in a community context. These communities have a level of complexity suitable for ecological experiments and are tractable for testing new proteomic methods. The biofilms contain organisms that represent all three domains of life (Bacteria, Archaea, and Eukaryotes; as well as viruses); span multiple trophic levels; and carry out many steps of the carbon cycle, including autotrophic carbon fixation, heterotrophic carbon consumption, and turnover of fixed carbon during degradation.

5.9.2 Use of TMT-proteomics to study microbial communities

Prior proteomics studies using TMT- or iTRAQ-based isobaric chemical labeling have been applied exclusively to human tissues or cultured isolates[170]. Here, we show that TMT-based quantitative proteomics can provide mechanistic insights into enzymes and pathways of individual microbial groups in microbial communities and define their functional response to temperature change. By multiplexing our samples, we were able to obtain accurate, precise, and reproducible quantification of proteins from three treatments with two sample replicates and two technical replicates per treatment in just four LC-MS/MS runs. Across our samples, we

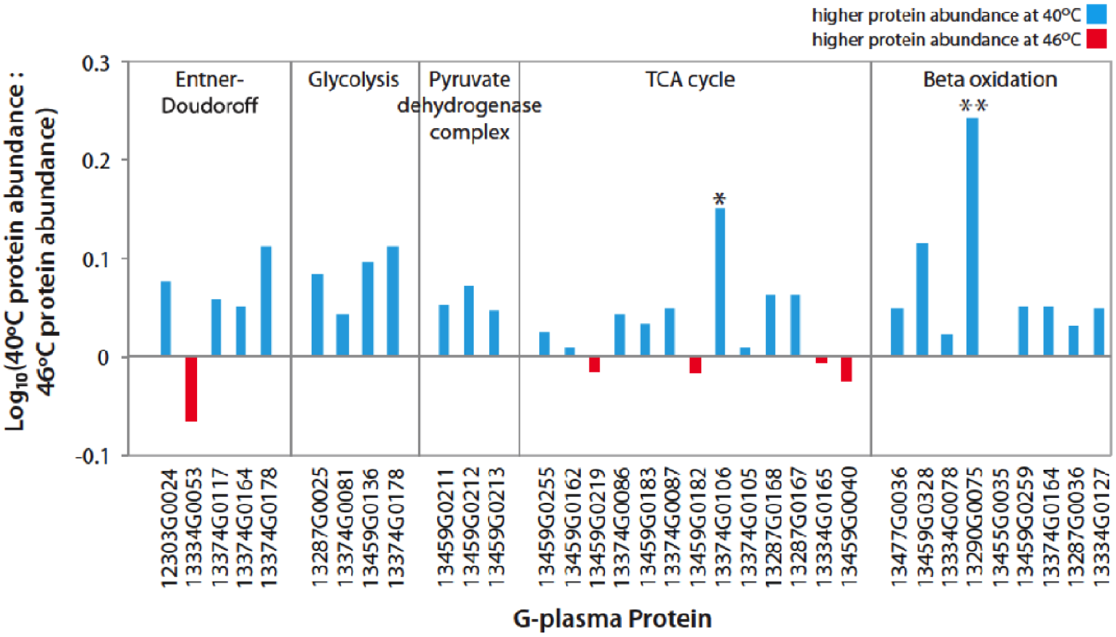


Figure 5.7 Abundance of G-plasma carbon cycling proteins (predicted by Yelton *et al*., 2013) at 40 ºC and 46 ºC. Stars indicate proteins that are significantly different at a given temperature (fold-change > 1.2 or < 0.8; one star p ≤ 0.1; two stars p ≤ 0.05).

identified an average of 1799 proteins from 25 different organisms including Bacteria, Archaea, Eukaryotes, and viruses.

5.9.3 Effect of warming on community structure

We found that a thermal shift from 40 ºC to 46 ºC caused a dramatic change in community composition (as reflected within the community proteome), as has been reported in other warming studies in soils, oceans, and freshwater[171-173]. Nearly a quarter of the organisms had a greater than two-fold change in abundance between temperatures. It has been previously suggested that *Leptospirillum* group III favors environmental conditions with lower stress, including lower temperature[174]. Here, we found that the overall abundance of *Leptospirillum* group III, the dominant organism in the cultivated biofilms, decreased by 14% from 40 ºC to 46 ºC (based on protein abundance) and ribosomal proteins were significantly down-regulated at 46 ºC, suggestive of reduced cell growth at elevated temperature.

The lack of visible biofilm growth at 49 ºC suggests that persistent temperatures above 46ºC alter community structure and/or function in such a way that biofilm formation and development are hindered. Previous culture studies have shown that while some *Leptospirillum* isolates are capable of growth up to 45 °C, many others are unable to grow at that temperature or higher[175,176]. Thus, it may be the case here that the colonizing *Leptospirillum* bacteria have a maximum growth temperature around 46 ºC, thereby preventing the initial stages of biofilm formation at higher temperatures. Studies such as these enable observations of growth in a community context where organisms experience competition for resources and interactions with other organisms, compared to static culture conditions.

Communities made up of both specialists and generalists are likely more productive and more stable over time under environmental fluctuations. Amongst groups of closely related organisms in the AMD biofilms, there appear to be a subset that are specialists in terms of their temperature optima, as well as generalists able to grow over a wider range of temperature. For instance, for the ARMAN archaea, ARMAN IV was more abundant at 40 ºC, ARMAN II was more abundant at 46 ºC, and two other ARMAN types had similar abundance levels at both temperatures.

5.9.4 Effect of warming on *Leptospirillum* function

Elevated temperature differentially affected carbon fixation by closely related bacterial genotypes: high temperature repressed carbon fixation by *Leptospirillum* group III and the group II 5way genotype, whereas carbon fixation was significantly up regulated at higher temperature by the *Leptospirillum* group II UBA genotype. Functional overlap of three *Leptospirillum* genotypes (iron-oxidation coupled with carbon fixation) with different temperature responses may provide ecological insurance for community function under heterogeneous environments. Niche differentiation of *Leptospirillum* allows for asynchronous responses to fluctuating conditions and assists in preserving function within the community across changing environments.

Increasing temperatures have been shown to enhance the decomposition of organic matter and the extracellular release of carbohydrates in seawater[177,178]. Here, 20 different Carbohydrate Active Enzymes (CAZymes) were quantified for *Leptospirillum* group III bacteria. CAZymes with different functionality were up-regulated under different conditions. For instance, GT2 involved in biosynthesis of carbohydrates was up-regulated at 46 ºC, whereas GH57 involved in hydrolysis of carbohydrates was up-regulated at 40 ºC. The extracellular polymeric substance (EPS) in AMD biofilms from the Richmond mine has been shown to contain abundant carbohydrates, including galactose, glucose, heptose, rhamnose, and mannose[179]. Thus, *Leptospirillum* group III might act both as a source and sink to the carbohydrate pool in the biofilm matrix.

Leptospirillum group III bacteria may have been subject to increased viral stress at elevated temperature (CRISPR-associated proteins were up-regulated at 46 ºC). Viral
induced mortality impacts not only the abundance and composition of microbial communities but also system-level nutrient cycling. Viral lysis releases the contents of the host cell (including cytoplasmic and structural material) into the environment, thereby liberating a fraction of the organic matter pool and shifting nutrients from the particulate to dissolved states. Dissolved organic carbon (and other nutrients including phosphorus and nitrogen) released by viral lysis can stimulate the growth of non-infected populations, increase community respiration, and decrease the efficiency of carbon transfer to higher trophic levels[180]. Thus, increased susceptibility to viral stress at elevated temperature, as shown here, will likely lead to greater carbon turnover and altered community structure.

The expression of proteins involved in amino acid metabolism was up-regulated at higher temperatures both at the community-level and for each *Leptospirillum* genotype (group II 5way and UBA genotypes and Group III). It has been shown that some amino acids are thermo-labile, and thus can have a reduced frequency in thermophilic proteomes[181,182]. AMD organisms may be increasing expression of amino acid biosynthesis proteins at 46 ºC to increase the size of the amino acid pool available for making other cellular proteins that may be inactivated at higher temperature.

Temperature has been shown to affect bacterial movement via impacts on both chemotaxis and flagellar assembly[183,184]. Here, methyl-accepting chemotaxis sensory transducer proteins were significantly up-regulated at 40ºC for each of the three *Leptospirillum* bacteria. Previous reports also showed that chemotaxis was strongly inhibited by high temperature in *Escherichia coli*[185-187]. Structural studies of AMD biofilms show *Leptospirillum* group II at the base of mature biofilms and *Leptospirillum* group III as dispersed cells and microcolonies within the interior regions[188]. Chemotaxis may be critical for positioning the bacteria within areas of the biofilm that are best suited for optimal growth. Decreased activity of chemotaxis proteins at elevated temperature may subject these bacteria to unfavorable geochemical conditions such as lower oxygen concentrations at the base of the biofilm or less nutrient availability in the interior. Nutrient limitation resulting from decreased chemotaxis at elevated temperature may be compensated for by enhanced nutrient scavenging, as indicated by up-regulation of two nutrient assimilation proteins at 46 ºC for the group II 5way genotype (NifA and a periplasmic phosphate binding protein).

5.10 Conclusions

The current research shows the utility of quantitative proteomics for studies of ecological of phenomena such as niche differentiation. The approach provided information about differential expression of thousands of proteins involved in diverse functions including metabolism, growth, signaling, and stress response. It enabled protein analysis at the level of individual microbial groups within a community context across the bulk community, and is broadly applicable to experimental studies that target microbial communities and more complex natural ecosystems.

CHAPTER 6
DIVERSE AND DIVERGENT PROTEIN POST-TRANSLATIONAL MODIFICATIONS IN TWO GROWTH STAGES OF A NATURAL MICROBIAL COMMUNITY

All of the data presented below has been adapted from

Zhou Li, Yingfeng Wang, Qiuming Yao, Nicholas Justice, Tae-Hyuk Ahn, Dong Xu, Robert Hettich, Jillian Banfield, Chongle Pan. "Diverse and divergent post-translational modification of proteins of closely related bacteria in two growth stages of a natural microbial community". Submitted to *Nature Communications*

Zhou Li's contributions include experimental design, MS measurement, data analysis, and writing the manuscript.

6.1 Introduction

Microbial communities populate and shape diverse ecological niches within natural environments[189]. The physiology of organisms in natural consortia and their responses to changing environmental conditions have been studied by measuring the abundance changes of proteins across a series of field samples using community proteomics[50,52,190]. However, microorganisms in a community regulate their metabolic processes not only by changing the copy numbers of proteins, but also by modulating the specific activities of expressed protein molecules.

Post-translational modifications (PTMs) are one of the most important mechanisms for activating, changing, or suppressing proteins' functions[14]. Phosphorylation cascades are the mechanism for many signal transductions, such as two-component systems[191] and cell-division cycle control[192]. Acetylation preferentially targets large protein complexes and regulates key metabolic pathways[193], including glycolysis and gluconeogenesis in bacteria[194]. Proteome-wide interactions between phosphorylation and acetylation have been observed in bacterium *Mycoplasma pneumonia*[195]. Methylation is actively involved in mediating protein-protein interactions through methylation-dependent binding domains[196]. Phosphorylation and methylation were found to collectively regulate signaling during bacterial chemotaxis[191]. S-nitrosylation and nitration, caused by reactive nitrogen species, are the major signaling mechanisms under nitrosative and oxidative stress[197,198]. Citrullination of arginine plays critical roles in regulation of gene expression by changing DNA-protein interaction[199]. Hydroxylation has been primarily found in collagen[200] and beta-methylthiolation has been mainly observed in bacterial ribosomal proteins[201], but their scope and significance in various complex biological processes remain unclear.

PTMs can be identified using a shotgun proteomics approach by searching for modified peptides in liquid chromatography-tandem mass spectrometry (LC-MS/MS) data. Previous in-depth PTM studies have generally targeted specific types of PTMs by enriching modified peptides via affinity purification prior to LC-MS/MS analysis. Although enrichment reduces the sample complexity, the measurements do not provide a comprehensive PTM profile of proteins that may carry many types of PTMs unless a separate enrichment is used for each PTM type[202,203]. Furthermore, enrichment does not permit direct quantification of PTM fractional occupancy, which is the percentage of the copies of a protein that is modified with a specific PTM event[204,205]. To overcome these limitations of enrichment-based approaches, we used an optimized global shotgun proteomics approach for broad-range PTM identification and quantification. The eight types of PTMs described above were simultaneously measured in two samples of a natural microbial biofilm community growing in an acid mine drainage (AMD) environment[48]. The proteomic PTM profiles were compared between two biofilm growth stages to uncover the dynamics of PTMs during community succession. These biofilms are ideal for such a study because there is an essentially comprehensive, curated, genome-resolved database of predicted protein sequences representing all dominant organisms.

Because of the functional roles of PTMs in regulating biological activities, PTMs may be conserved across orthologous proteins, and the divergence of PTMs may contribute to phenotypic diversity[206]. Recently, phosphorylation, acetylation, and ubiquitination were compared between different eukaryotic species, and divergent PTM patterns were revealed between their orthologs[207-211]. However, the conservation of PTMs in bacteria and archaea, especially between closely related, co-evolving microorganisms, is largely unknown. Previously, community genomics and proteomics were used to study ecological differentiation of the UBA and 5wayCG variants of *Leptospirillum* group II, which typically dominates AMD biofilms[212]. These *Nitrospirae* phylum bacteria share 99.7% 16S rRNA gene sequence identity and exist as a continuum of genotypes that have undergone varying degrees of recent large-scale homologous recombination[213]. In the current study, we found divergent PTM patterns on many orthologous proteins of *Leptospirillum* group II bacteria, further underlining metabolic distinction of these closely related organisms[212].

6.2 Additional materials and methods

PTM localization: Sipros assigned PTMs to all modifiable residues in a peptide. Each candidate with the same PTM on different positions (PTM isoforms) was scored to identify the top-rank peptide for a spectrum. For every modified spectrum, Sipros calculated the DeltaP score, which is the score difference between the top-rank modified peptide and its next lower-ranked PTM isoform[214]. In modified spectra that had a DeltaP greater than 0, PTMs were localized on the modified residues of the top-rank peptides by Sipros. In modified spectra that had a DeltaP equal to 0, PTMs cannot be localized, because the top two candidates with different modified residues cannot be differentiated by any fragment ions. Only modified peptides with DeltaP greater than 0 were used for the PTM dynamics and divergence analysis, unless noted otherwise.

PTM divergence analysis: Orthologous protein pairs between the UBA and 5wayCG *Leptospirilli* was obtained from our previous studies[212,215]. Sequence alignment was performed using EMBOSS Needle[216](http://www.ebi.ac.uk/Tools/psa/emboss_needle/). The PTM divergence analysis used the orthologous protein pairs satisfying the following requirements: both proteins in a pair must be identified; and at least one protein in a pair must have organism-specific localized PTMs.

Statistical Analysis : Rank product test[217] was used to calculate the *p* values for the changes of PTM fractional occupancy and the changes of protein abundances between GS1 and GS2. For the dynamic PTM and COG category enrichment analysis, *p* values were calculated using a two-tailed Fisher's exact test and corrected for multiple comparisons using the Benjamini–Hochberg method.

Protein Structure Prediction**:** MUFOLD[218] was used to predict the structures and solvent accessibilities of Cas proteins and rTCA enzymes. UBA-type sequences of rTCA enzymes were used as the input for structure prediction. The top five structural templates in PDB were selected for each protein together with the optimal target-template alignment. The best model was then comprehensively determined by a composite score of the identity score, template coverage, and model quality assessment scores. PyMOL was then used to display the protein structures and mark the modified sites by ball structures.

6.3 Identification of diverse PTMs in community proteomes

Large intact sheets of AMD biofilm were sampled at the AB Muck Dam site in the Richmond Mine at Iron Mountain, CA, on 9/17/2010 (pH ~1, 39 ^{o}C). Two samples were collected only centimeters apart: a thin, early growth-stage biofilm community (GS1), and a late growth-stage biofilm community represented by a thicker portion of the biofilm (GS2) (Figure 6.1). Consistent with previous observations[212], the GS1 biofilm had low diversity and was comprised predominantly of *Leptospirillum* bacteria, whereas the GS2 biofilm harbored a more diverse community with increased abundance of archaeal species.

Three protein samples were prepared in parallel for each AMD biofilm sample, digested with trypsin, Lys-C, or Glu-C. Each protein digest sample was analyzed in technical duplicate by 22-hour LC-MS/MS using an LTQ Orbitrap Elite mass spectrometer. We optimized the high-resolution MS/MS method using higher-energy collisional dissociation (HCD)[38]. Spectral quality and identification results were significantly improved by using the mass-to-charge ratio (m/z) cutoff of 180 in MS/MS scans, normalized collision energy of 30% for HCD, and minimum ion threshold of 1,000 for MS/MS triggering. We compared the optimized high-resolution MS/MS method with a conventional low-resolution MS/MS method based on collision-induced dissociation (CID) using a test AMD biofilm sample. The two methods identified comparable numbers of peptide-spectrum matches (PSMs) and peptides (Table 6.1). However, the high-resolution MS/MS method identified 71% of the acquired fragment ion spectra and provided high mass accuracy (<0.01 Da mass error) on matched fragment ions, which allowed searching a much larger sequence space with a low false discovery rate (FDR).

In addition to the eight biological PTMs described above, we also searched for three PTMs commonly resulting from proteome sample preparation, including oxidation of methionine, deamidation of asparagine and glutamine, and alkylation of cysteine. Because simultaneous consideration of these types of PTMs vastly expanded the sequence space for database searching, the database searching was performed with a scalable Sipros[219,220] algorithm on a supercomputer, Titan, using up to 35,000 central processing unit (CPU) cores across thousands of compute nodes. Identification results were organized in a hierarchical structure with five levels: organisms, proteins/protein groups, PTM events, peptides, and PSMs. An identification at a given level generally comprised multiple identifications at the next lower level. A PTM event was defined as a specific type of PTM on a specific residue of a protein. As

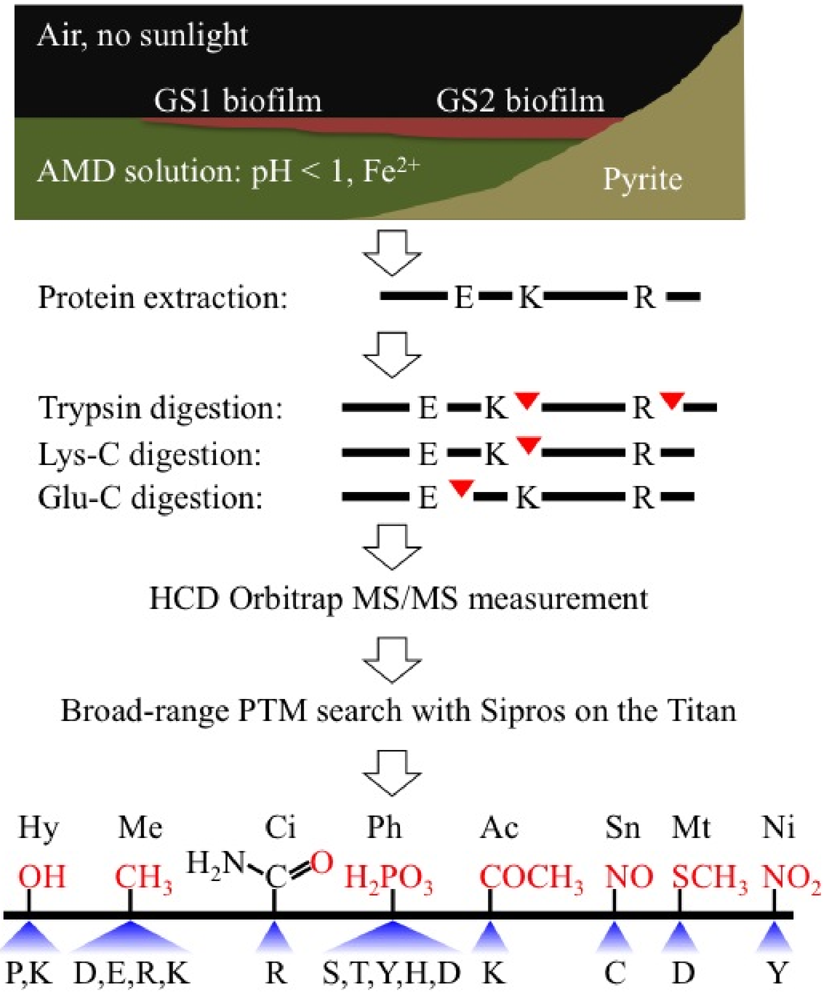


Figure 6.1 Study overview. AMD biofilms representing two growth stages (GS1 and GS2) were collected from a site within the Richmond Mind. Proteome samples were digested using three proteases in parallel and analyzed by HCD MS/MS. The following biological PTMs were searched: hydroxylation (Hy), methylation (Me), citrullination (Ci), phosphorylation (Ph), acetylation (Ac), S-nitrosylation (Sn), methylthiolation (Mt), and nitration (Ni). The chemical formula (red) and modifiable amino acids are listed for each type of PTM.

Table 6.1 Comparison of the high-resolution HCD method with the low-resolution CID method.

Method	HCD (FDR)	CID (FDR)
Measurement time	22 hours	22 hours
# of acquired fragment ion spectra	335,633	467,943
# of PSMs	239,502 (0.3%)	221,806 (0.3%)
Spectra identification rate	71%	47%
# of distinct peptides	45,947 (1%)	43,948 (1%)
# of proteins/protein groups	3,824 (0.3%)	4,013 (0.8%)

identification results were pooled across trypsin, Lys-C, and Glu-C digests of a sample, many PTM events were covered by multiple peptides.

We identified 765,202 and 743,413 PSMs, 78,539 and 104,893 non-redundant peptides, 3,599 and 3,723 unique PTM events, and 4,259 and 5,055 proteins/protein groups from 7 major organisms in the GS1 and GS2 biofilm, respectively (Table 6.1). Approximately 76% of the PSMs and 46% of the unique PTM events were identified from the UBA or 5wayCG *Leptospirilli*. The FDR, estimated by searching concatenated reverse sequences[221], was ~0.34% at the PSM level, ~1.0% at the peptide level, ~2.0% at the PTM event level, and ~3.5% at the protein/protein group level. The FDRs in the identification hierarchy increased because upper-level identifications from reverse sequences often comprised far fewer lower-level identifications than those from forward sequences. Adding longer Lys-C and Glu-C peptides to tryptic peptides nearly tripled the number of identified PTM events.

The PTM events identified from the GS1 and GS2 samples were categorized by PTM types and organisms (Figure 6.2). Because of the ~95% average amino acid identity between orthologous proteins of the UBA and 5wayCG *Leptospirilli*, ~2,700 PTM events unique to the *Leptospirillum* group II cannot be resolved to a specific organism (shared *Lepto* II, Figure 6.2). Hydroxylation, methylation, and citrullination were the most commonly identified PTM types. In the *Leptospirillum* group II proteomes, ~29% of the identified proteins were modified with at least one unique PTM event in a biofilm sample. Of these modified proteins, ~43% carried multiple types of unique PTM events. Some extensively modified proteins included chaperone proteins, such as DnaK, GroEL and ClpB, enzymes in the reductive tricarboxylic acid (rTCA) cycle, and proteins involved in iron oxidation and electron transport, such as NADH dehydrogenase and several cytochromes.

Because confident identification of a PTM event requires placement of that PTM on the correct residue, we separately estimated the accuracy of PTM localization by re-analyzing the HCD high-resolution MS/MS data of synthetic peptides with known phosphorylation sites from a recently published study[222]. The phosphorylation sites were assigned to the correct residues for more than 97% of the modified spectra with PTM localization (i.e. DeltaP > 0 for site differentiation) by Sipros. We then evaluated the depth and breadth of our broad-range PTM identification approach with a model organism, *E. coli*. 5,005 unique PTM events were found from 966 proteins out of a total of 2,082 identified proteins (Table 6.3). Hydroxylation,

Table 6.2 Summary of identification results at the PSM, peptide, PTM event, and protein/protein group levels.

Sample	Protease	# of PSMs (FDR)	# of peptides (FDR)	# of unique PTM event* [non-unique event**] (FDR)	# of proteins/ protein groups (FDR)
AMD GS1	Trypsin	356,233 (0.34%)	40,825 (1.0%)	1,329 [2,931] (1.4%)	3,423 (1.5%)
	Lys-C	294,856 (0.26%)	30,325 (1.0%)	1,905 [3,041] (1.1%)	2,871 (1.4%)
	Glu-C	114,113 (0.29%)	13,709 (1.0%)	601 [963] (1.6%)	1,749 (0.85%)
	Total	765,202 (0.30%)	78,539 (1.0%)	3,599 [5,970] (2.0%)	4,259 (3.1%)
AMD GS2	Trypsin	328,989 (0.37%)	50,699 (1.0%)	1,240 [2,336] (1.0%)	3,937 (1.9%)
	Lys-C	273,566 (0.37%)	37,720 (1.0%)	1,768 [2,794] (1.2%)	3,318 (1.8%)
	Glu-C	140,858 (0.43%)	23,683 (1.0%)	743 [1,080] (1.6%)	2,702 (1.2%)
	Total	743,413 (0.38%)	104,893 (1.0%)	3,723 [5,595] (1.9%)	5,055 (3.9%)

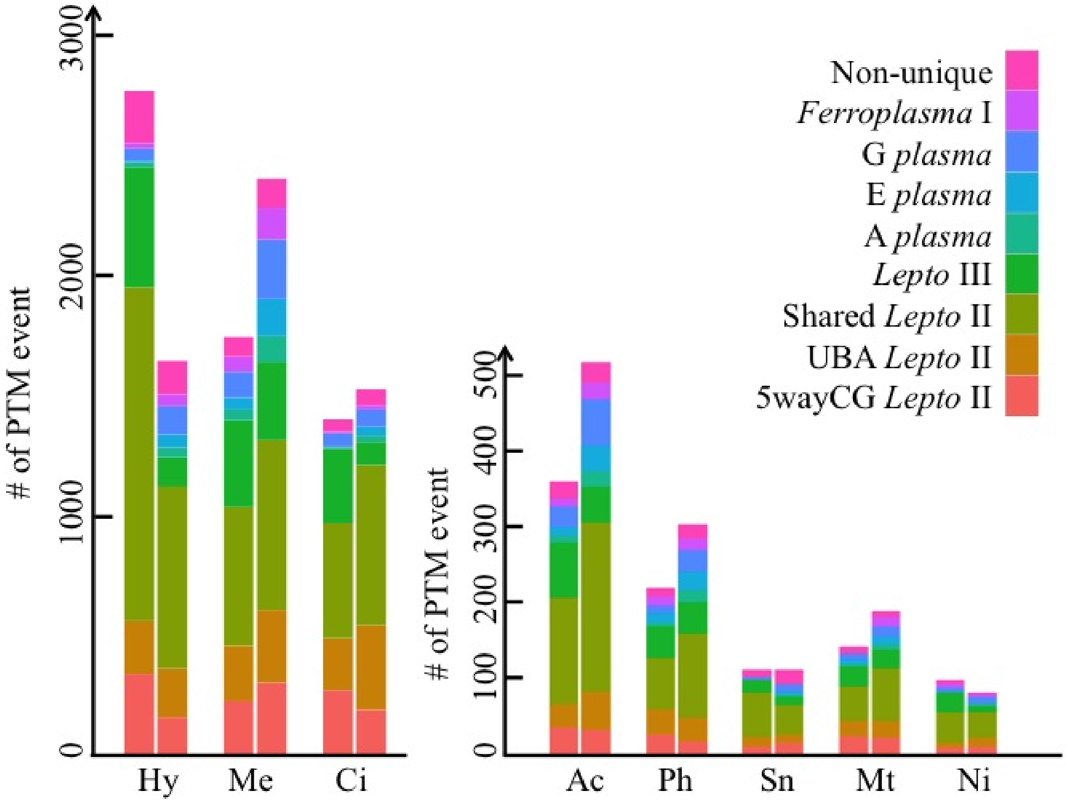


Figure 6.2 Histograms of identified PTM events in the AMD community. For each PTM type, the numbers of PTM events were compared among major community members (color-coded sections) between GS1 (left bar) and GS2 (right bar). Shared *Leptospirillum* (*Lepto*) II represents PTM events that can be only assigned to the *Leptospirillum* group II, but cannot be resolved to a specific organism.

citrullination, and methylation were the three most abundant PTM types in laboratory-grown *E. coli*, which was consistent with the results of the AMD microbial communities. Previous enrichment-based measurements identified ~150 phosphorylation events[223] and ~1,070 acetylation events[224] in *E. coli*. Here, our approach simultaneously identified 284 unique phosphorylation events and 470 unique acetylation events. While the PTM identification results were not directly comparable between different *E. coli* samples used in these measurements, our approach was successful in finding extensive modifications of proteins by many different types of PTMs without using multiple enrichments. For example, the transcriptional factor OxyR was recently discovered as a master regulator of S-nitrosylation in *E. coli* under anaerobic respiration on nitrate[197]. Here, we identified 176 unique S-nitrosylation events from 159 proteins under aerobic growth condition, including OxyR. In addition to the S-nitrosylation, OxyR was simultaneously modified with phosphorylation, methylation, and acetylation.

6.4 PTM dynamics of the community during ecological succession

We compared the unique and localized PTM events in UBA and 5wayCG *Leptospirilli* between the GS1 and GS2 samples. While ~75% of the proteins were identified in both samples, only ~17% of the organism-specific PTMs were maintained between the two samples (Figure 6.3). PTMs that differed between samples were significantly enriched (p value <0.05) in citrullination and methylation. These PTMs may be involved in regulation of pathway activities. For instance, *Leptospirillum* group II was previously found to have different chemotaxis and motility activities between the two growth stages, possibly reflecting the diversified community membership and increased competition for nutrients in GS2[225]. The PTM patterns were compared between GS1 and GS2 for the chemotaxis gene clusters in the *Leptospirillum* group II and III (Figure 6.4A and B)[226]. While all proteins from the two gene clusters were identified, they had distinct PTM patterns in the two growth two gene clusters were identified, they had distinct PTM patterns in the two growth stages. More methylation events were identified in GS2 than in GS1 on the chemotaxis scaffolding protein CheW (locus ID: CGL2_11277G0245 and UBAL2_8241G0195, where CGL2 and UBAL2 denote 5wayCG- and UBA-type orthologs, respectively), histidine kinase CheA (CGL2_11277G0248 and UBAL2_8241G0198), response regulator CheY (CGL2_11277G0249 and UBAL2_8241G0199), and methyl-accepting chemotaxis protein MCP (CGL2_11277G0246 and UBAL2_8241G0196) from *Leptospirillum*

Table 6.3 The number of PTM events identified from an *E. coli* sample. *E. coli* runs were searched and filtered with the same settings as the AMD runs. The FDR estimated at the PTM level was 3.2%.

PTM type	# of unique events	# of non-unique events
Hydroxylation	1,094	71
Methylation	1,611	134
Citrullination	877	25
Acetylation	470	32
Phosphorylation	284	44
S-nitrosylation	176	16
Methylthiolation	278	24
Nitration	215	12

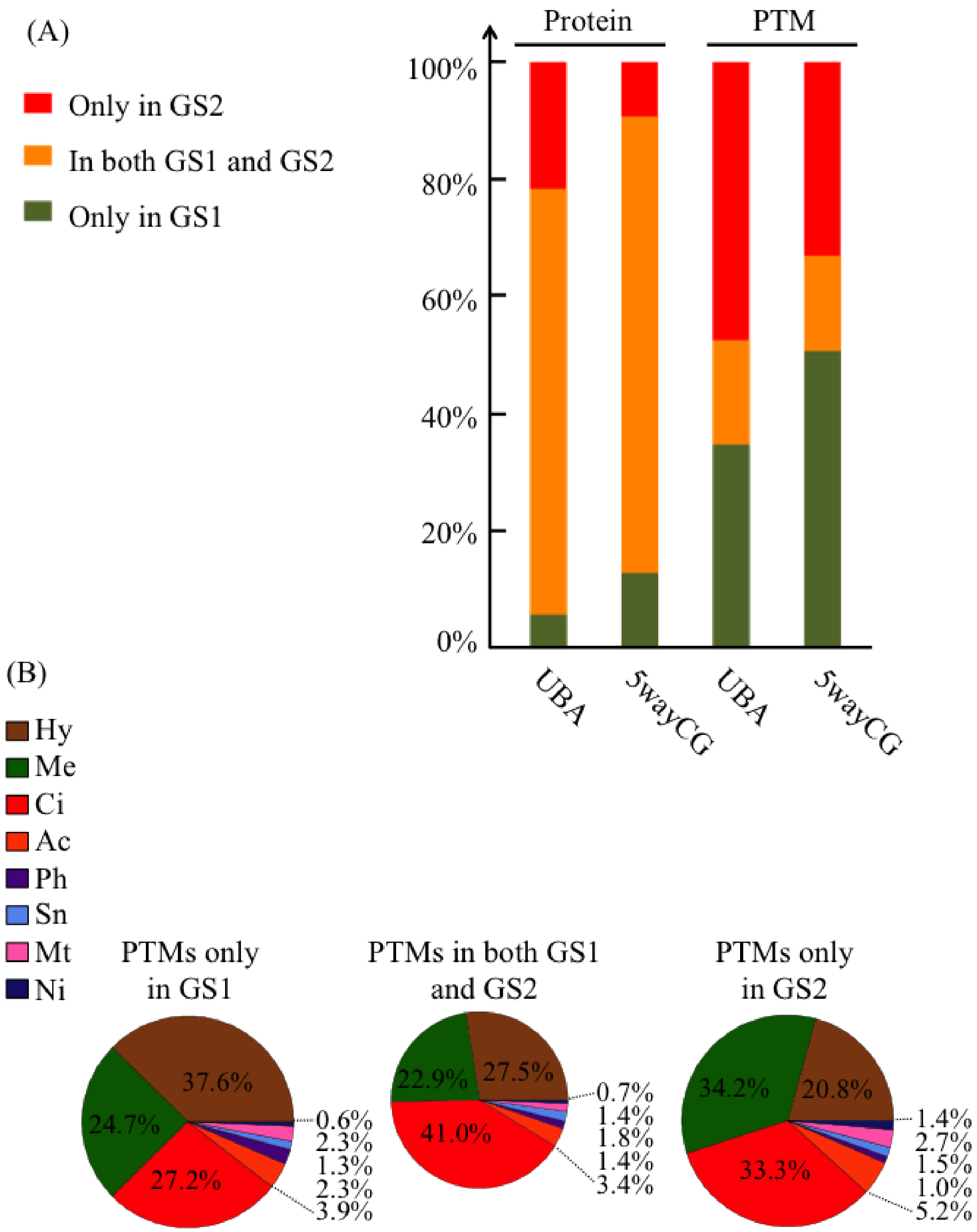


Figure 6.3 (A) Dynamics of proteins and organism-specific PTM events between the two growth stages in the UBA and 5wayCG *Leptospirilli.* The three sections in each bar represent identifications only in GS1 (green), in both GS1 and GS2 (orange), and only in GS2 (red). (B): The distributions of the organism-specific PTM events that were identified only in GS1, in both GS1 and GS2, and only in GS2 were shown in the pie charts, respectively. Only PTM events that can be localized to a specific residue were considered.

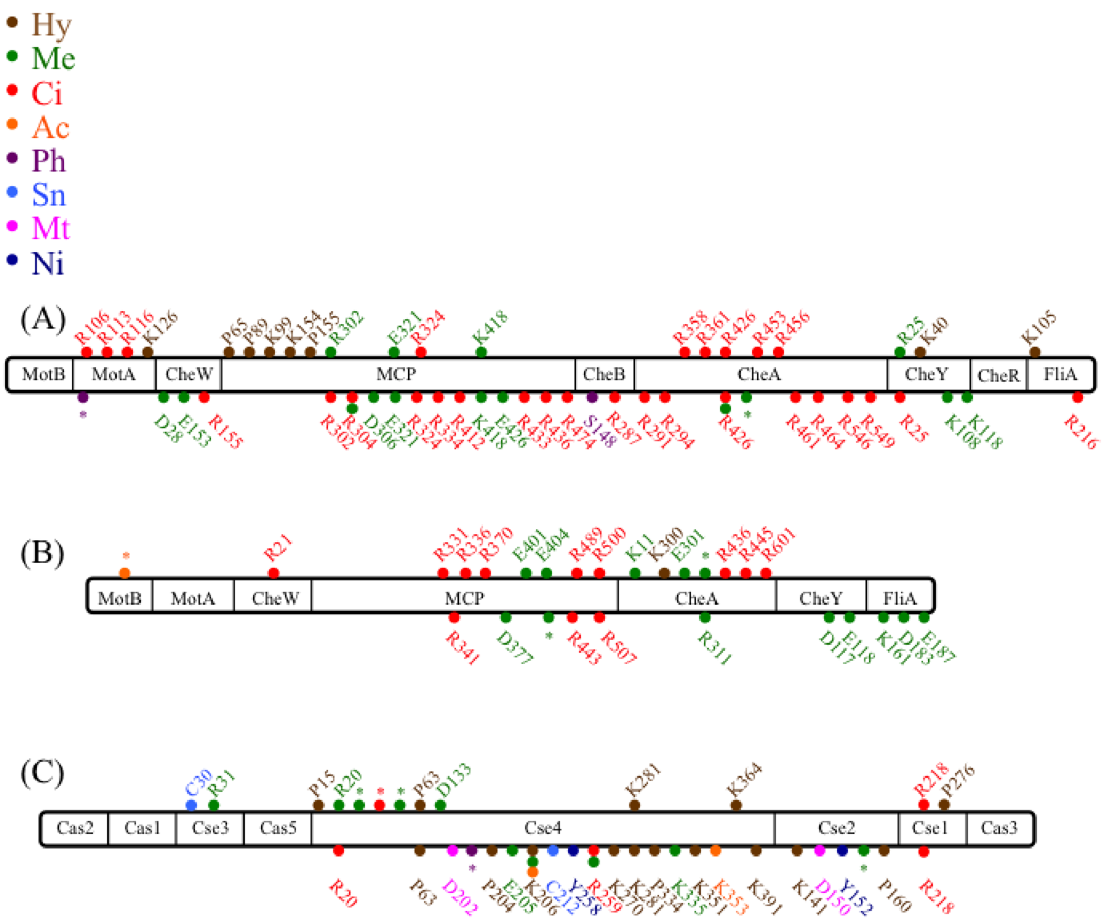


Figure 6.4 Changes of PTM patterns from GS1 to GS2 in key pathways. (A) Chemotaxis gene cluster in *Leptospirillum* group II. (B) Chemotaxis gene cluster in *Leptospirillum* group III. (C) Cas gene cluster in *Leptospirillum* group II. Different PTM patterns were identified in the GS1 sample (above the bar) and the GS2 sample (below the bar). PTM events were color-coded by PTM types and marked with amino acid types and residue positions. PTMs that cannot be localized to a specific residue were marked with *.

group II and on the CheY (UBAL3_8063G0048) and flagellar sigma factor FliA (UBAL3_8063G0049) from *Leptospirillum* group III. Citrullination was extensively found across many proteins. A cluster of hydroxylation events was only identified on the *Leptospirillum* II MCP in GS1. Two phosphorylation events were identified on the motility protein A MotA (CGL2_11277G0244 and UBAL2_8241G0194) and chemotaxis methylesterase CheB (CGL2_11277G0247 and UBAL2_8241G0197) from *Leptospirillum* group II in only GS2. An acetylation event was identified on the motility protein B (MotB; UBAL3_8063G0043) from *Leptospirillum* group III in GS1.

The clustered regularly interspaced short palindromic repeats (CRISPR) and associate proteins (Cas) provide bacteria and archaea with resistance to phage invasion[227,228]. The CRISPR/Cas locus of *Leptospirillum* group II is encoded on one recombined sequence block that is now common across the genotypic series. Consistent with the calculated recent timing of this event[213], the Cas proteins are identical among UBA, 5wayCG, and other variants of *Leptospirillum* II. Cas proteins showed very distinct PTM patterns between these two growth stages (Figure 6.4C). Citrullination on Cse1 (CGL2_11386G0024 and UBAL2_8241G0432) was maintained across the two growth stages. Cse2 (CGL2_11386G0025 and UBAL2_8241G0431) was modified with four different types of PTMs only in GS2. PTMs on Cse3 (CGL2_11386G0028 and UBAL2_8241G0427) were only identified in GS1. Cse4 (CGL2_11386G0026 and UBAL2_8241G0430) was extensively modified in both GS1 and GS2, but PTMs clustered on distinct regions of the protein in the two different biofilm growth stages. The structures of Cse1, Cse2, and Cse3, predicted based on homolog modeling, suggest that identified PTMs are localized on the surface residues of the structures (Figure 6.5 and Table 6.4). Because of the semi-stochastic nature of MS/MS acquisition in shotgun proteomics, some low-abundance PTM events were identified with technical variability between individual runs. However, as each sample was measured after digestion using multiple proteases and in technical replicates, the identified PTM events aggregated across the six runs of a sample were reproducible and very few new PTM events were found with additional runs after the first three measurements (Figure 6.6). Moreover, the percentage of organism-specific PTMs maintained between GS1 and GS2 was almost unchanged after the first two measurements as the PTM coverage increased. This indicates that the difference between aggregated PTM events identified in the two samples was not a result of run-to-run variability and the repeated measurements

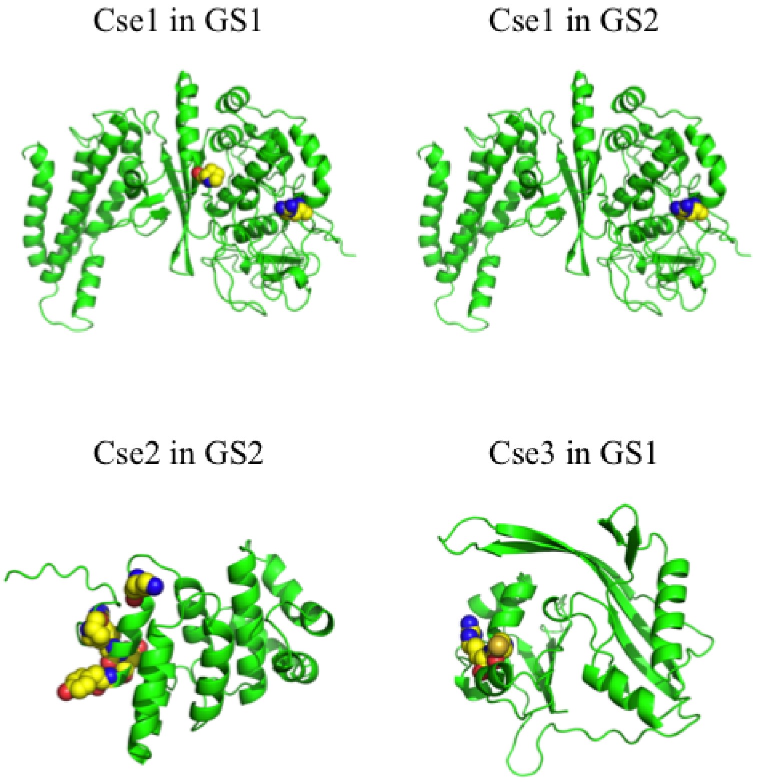


Figure 6.5 Predicted structures of the Cse1, Cse2, and Cse3 proteins. Modified residues were highlighted with the ball structure. Color code of the balls on the protein structures: green: backbone; yellow: carbon; blue: nitrogen; red: oxygen; orange: sulfur.

Table 6.4 The identity score between the query and template in Mufold prediction

Locus ID	Protein	Template	Identity
UBAL2_8241G0432	CRISPR-associated protein, Cse1	4F3E_A	30.91
UBAL2_8241G0431	CRISPR-associated protein, Cse2	2ZCA_A	38.60
UBAL2_8241G0427	CRISPR-associated protein, Cse3	3QRP_A	36.24
UBAL2_8241G0532	PFOR-beta	2UZA_A	19.22
UBAL2_8062G0194	PEP synthase	2OLS_A	52.32
UBAL2_8241G0507	PEP carboxylase	3ODM_C	29.32
UBAL2_8027G0031	Malate dehydrogenase	3TL2_A	46.88
UBAL2_8135G0106	Fumarate hydratase	3TV2_A	60.08
UBAL2_7931G0249	fumarate reductase	1YQ3_A	29.07
UBAL2_7931G0248	Succinyl-CoA synthetase-beta	1JLL_B	50.51
UBAL2_8241G0537	PFOR-alpha(second copy)	2UZA_A	20.69
UBAL2_8524G0125	Isocitrate dehydrogenase	2D1C_A	50.89
UBAL2_7931G0253	Aconitate hydratase	1ACO_A	37.29
UBAL2_7931G0252	Probable citrate synthase	1IOM_A	24.91

approached the detection limit of our methodology.

Many modified peptides were identified along with their unmodified versions in a sample, which indicated partial modification of those peptides. The fractional occupancies of 3,287 and 3,207 unique PTM events were estimated in the GS1 and GS2 samples, respectively, with an average standard deviation of 6.5% (Figure 6.7). The reproducibility of these results was comparable to that of a previous study[204]. The quantified PTM events were separated into three ranges based on their fractional occupancy: low (<20%), medium (20%-80%), and high (>80%), and the frequency of each range was compared for each type of PTM (Figure 6.8). The majority of PTM events had low fractional occupancy, which is in agreement with the fractional occupancy of phosphorylation from yeast[204]. However, most S-nitrosylation events had high fractional occupancy. The percentages of the PTM events in the low occupancy range decreased from the GS1 to GS2 for all PTM types.

Changes in PTM fractional occupancy were generally not correlated with changes in protein abundances, suggesting independent regulation of PTM abundances from protein abundances. For example, the fractional occupancy of the methylation at Glu35 of a pyruvate:ferredoxin oxidoreductase (PFOR; UBAL3_7952G0038) from the *Leptospirillum* group III increased from 3% in the GS1 to 88% in the GS2 (*p* value = 6.9×10^{-4}), but the protein abundance decreased by 8-fold (*p* value = 1.8×10^{-2}). Different PTM events on the same protein can also have different fractional occupancy changes. For example, the fractional occupancy of the citrullination at Arg113 of a 5wayCG-type chaperonin GroEL (CGL2_10776G0010) increased from 0.9% in GS1 to 9.6% in GS2 (*p* value = 2.7×10^{-2}), but the fractional occupancy of the hydroxylation at Lys92 of the same protein decreased from 3.2% to 0.4% (*p* value = 4.7×10^{-2}).

6.6 PTM divergence between closely related co-existing bacteria

Although the UBA and 5wayCG *Leptospirilli* share 95% average amino acid identity between their orthologs, we identified ~18,000 and ~25,000 organism-specific peptides that covered the positions with single amino acid polymorphisms (SAAPs) in GS1 and GS2, respectively. Amongst these organism-specific peptides, 1,373 and 1,457 PTM events were unambiguously assigned to the orthologs of specific organisms in GS1 and GS2, respectively. The conservation of PTM events between orthologs was analyzed by mapping these organism-

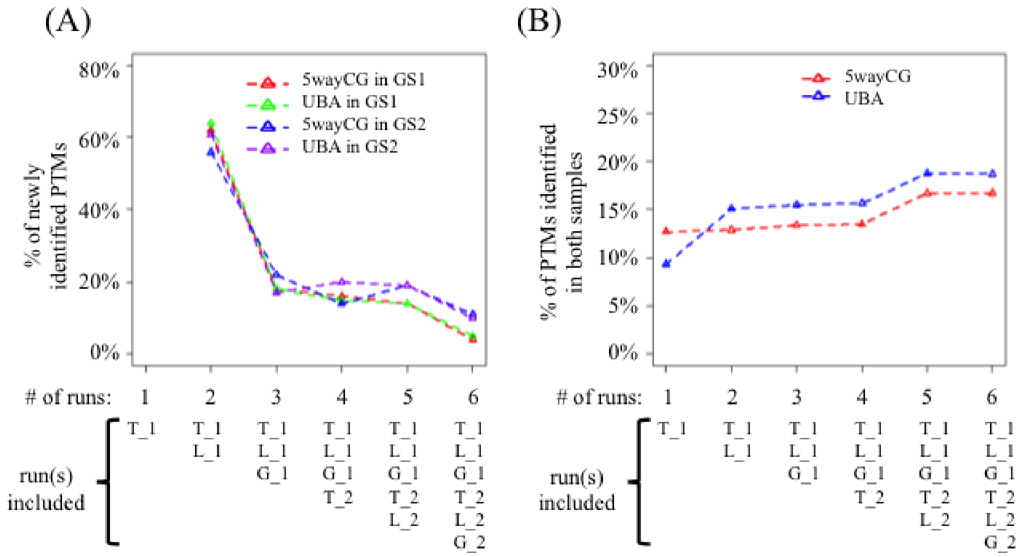


Figure 6.6 (A) Percentage of newly identified PTM events as additional runs were acquired in each organism, calculated by [(P_i-P_{i-1})/P_i] where P_i is the total number of PTM events identified in the first *i* runs and P_{i-1} is the total number of PTM events identified in the first (*i*-1) run(s). *i* = 2,3,4,5,6. (B) Percentage of PTM events identified in both GS1 and GS2 as additional runs were acquired, calculated by [P_iGS^{12}/(P_iGS^1+P_iGS^{12}+P_iGS^2)] where P_iGS^{12} is the total number of PTM events identified in both GS1 and GS2, P_iGS^1 is the total number of PTM events exclusively identified in GS1, and P_iGS^2 is the total number of PTMs exclusively identified in GS2 in the first *i* run(s). *i* = 1,2,3,4,5,6. Only organism-specific PTM events were considered in both (A) and (B). T_1: first run of the tryptic digest; L_1: first run of the Lys-C digest; G_1: first run of the Glu-C digest; T_2: second run of the tryptic digest; L_2: second run of the Lys-C digest; G_2: second run of the Glu-C digest

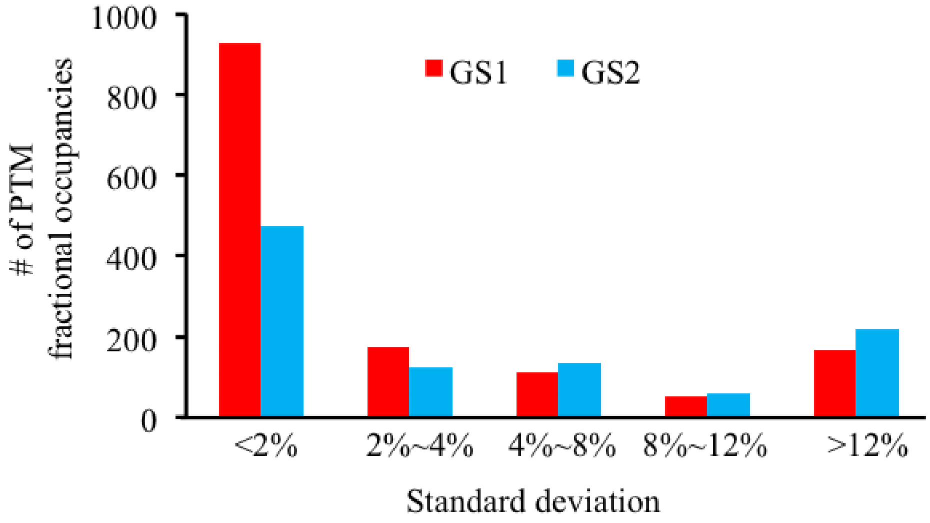


Figure 6.7 Distribution of standard deviations of the quantified PTM fractional occupancy from replicate measurements.

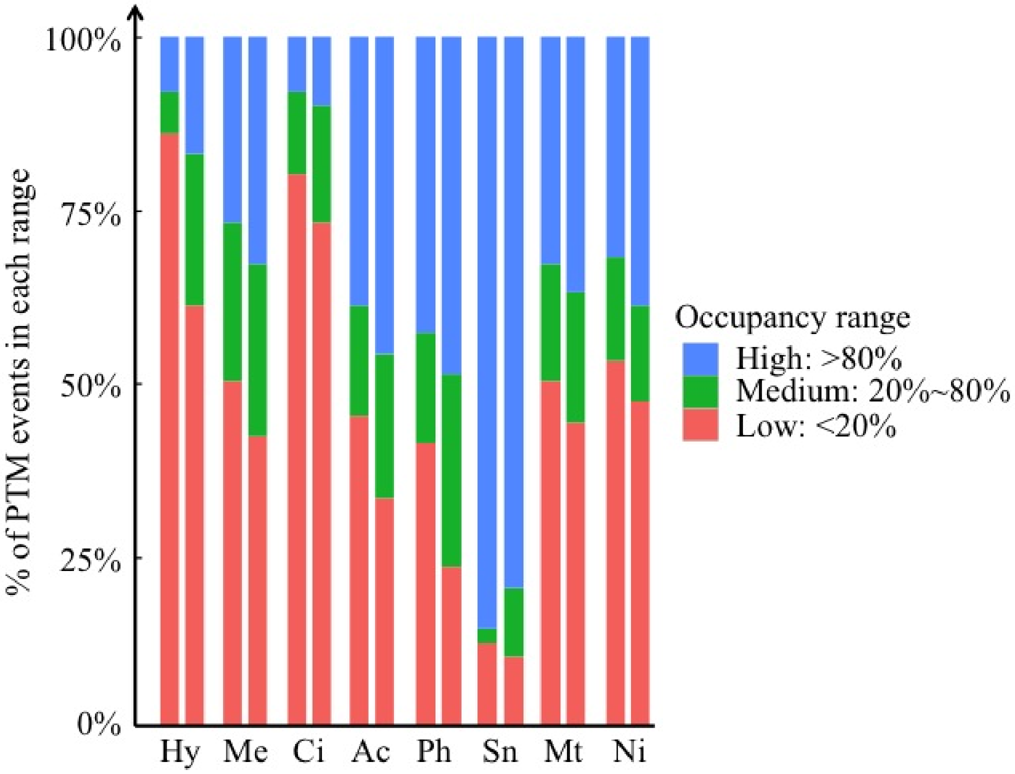


Figure 6.8 Frequency of each fractional occupancy range for each type of PTM. PTM events in GS1 (left bars) and GS2 (right bars) were separated into three ranges: high occupancy of >80% (blue), medium occupancy of 20% to 80% (green), and low occupancy of <20% (red).

specific PTM events onto aligned orthologous sequences. Of those PTM events that can be localized to specific residues, only 30% and 21% of them were conserved in GS1 and GS2, respectively (Figure 6.9A). The results were consistent between replicate measurements.

Of these divergent PTM events, a total of 774 and 880 from the two replicates occurred on the aligned positions with a conserved amino acid between orthologous proteins of the UBA and 5wayCG *Leptospirilli* in GS1 and GS2, respectively (Figure 6.9B). There were ~100 divergent PTM events that occurred on the SAAP position in a sample (Figure 6.9C). Many SAAP events changed the amino acid type of a residue in one organism to another amino acid that cannot carry the same PTM in the other organism. As an example, the 815th residue of a transaldolase was an acetylated lysine in the 5wayCG-type ortholog (CGL2_11067G0037) but it was substituted by a glutamic acid that cannot be acetylated in the UBA-type ortholog (UBAL2_8692G0154) (Figure 6.10A). Some residues on the SAAP positions were modified with different PTMs. For example, there was a K267R substitution in a pair of orthologs annotated as "outer membrane efflux protein", where the lysine of the 5wayCG-type ortholog (CGL2_11111G0096) was tri-methylated, but the arginine of the UBA-type ortholog was citrullinated (UBAL2_8241G0570) (Figure 6.10B). COG enrichment of the divergent PTMs was compared between GS1 and GS2. Proteins with divergent PTMs were significantly enriched (*p* value <0.05) in the COG categories of "Translation, ribosomal structure and biogenesis" in GS1 and "Amino acid transport and metabolism", "Replication, recombination and repair", "Secondary metabolites biosynthesis, transport and catabolism", and "Transcription" in GS2.

Conserved PTM events and divergent PTM events were compared in terms of fractional occupancy (Figure 6.11). The majority of conserved PTM events and divergent PTM events belonged to the low fractional occupancy range, but significantly more divergent PTM events than the conserved PTM events were found in the high fractional occupancy range. This observation was consistent for both GS1 and GS2.

Within the AMD biofilm, *Leptospirillum* group II likely uses the rTCA cycle for CO_2 fixation[226]. We identified every protein in the rTCA cycle from both UBA and 5wayCG *Leptospirilli*, with average sequence coverage of 82%. These proteins were extensively modified in both GS1 and GS2 samples. The PTM events were mapped onto the proteins' predicted structures (Figure 6.12A and Table 6.4). Most of the PTM events were localized on surface residues. We focused on the organism-specific PTM events to study their dynamics between the

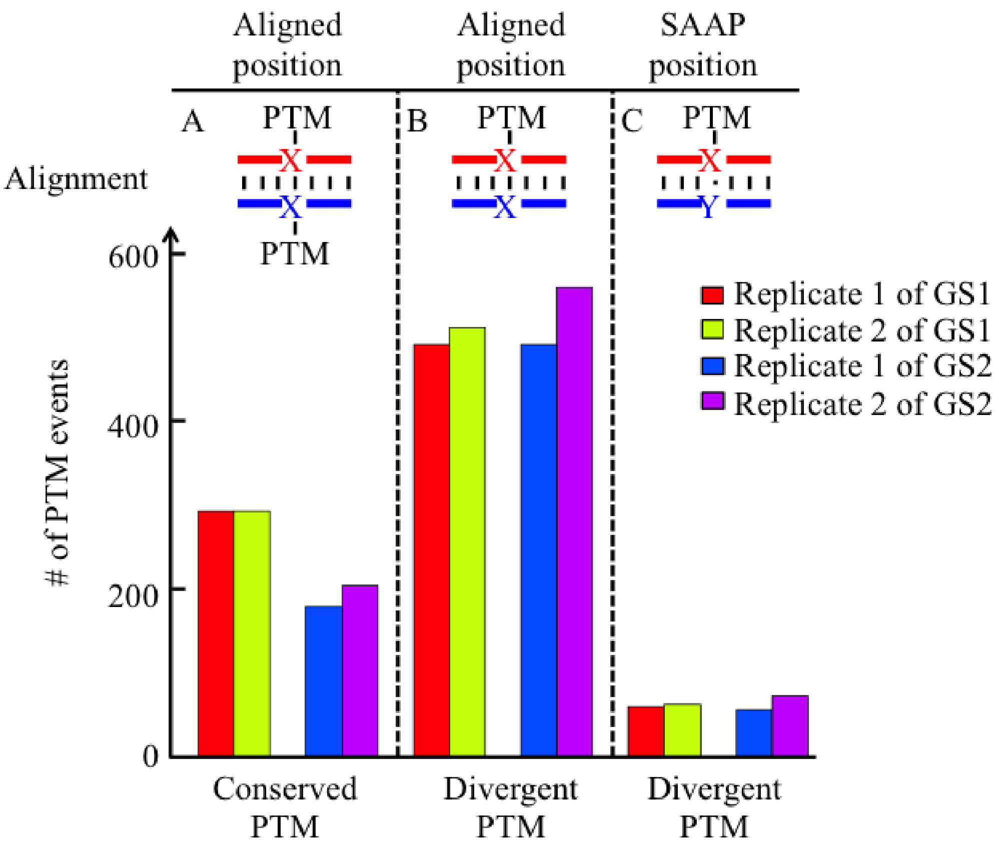
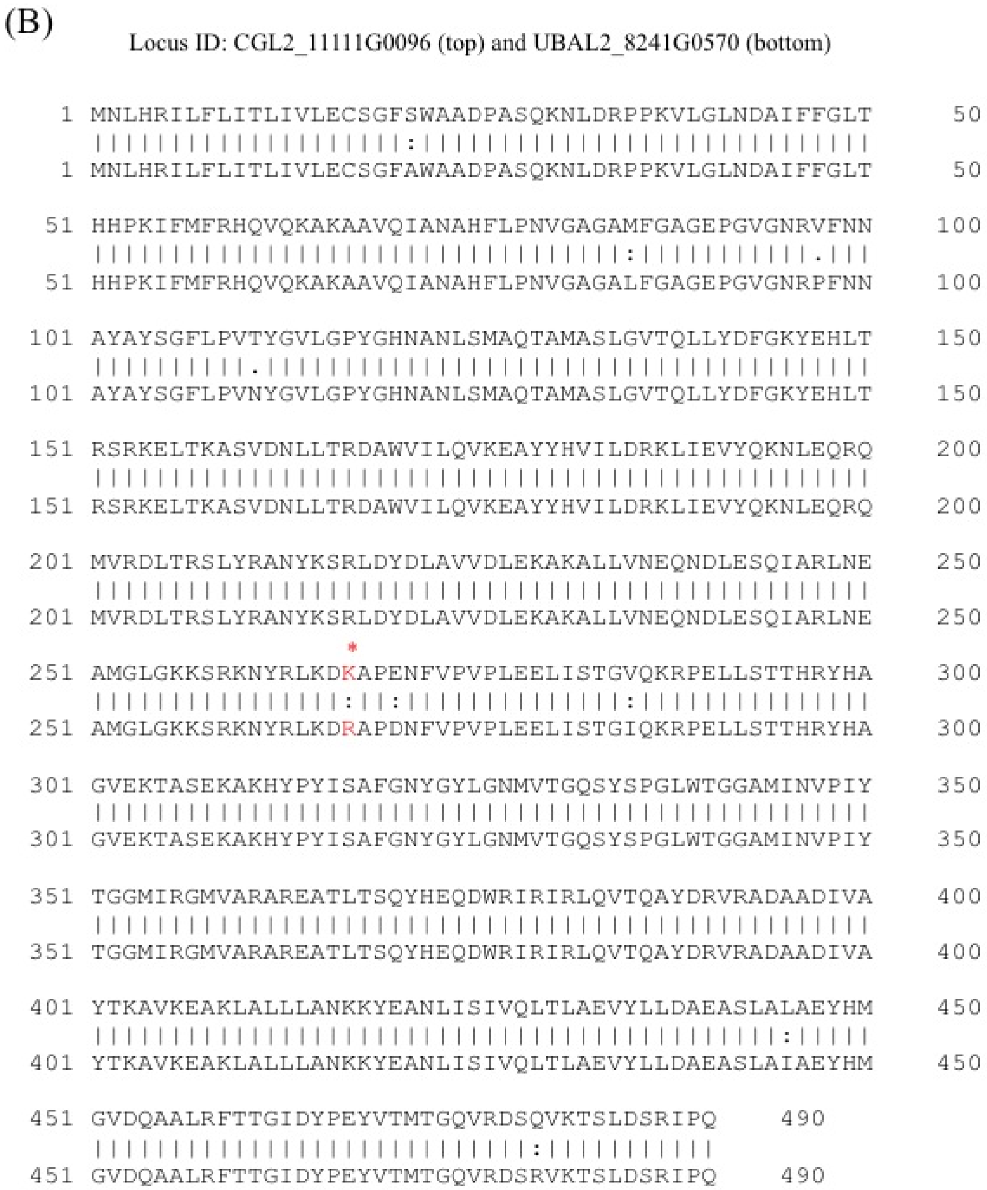


Figure 6.9 Organism-specific PTM conservation and divergence between the UBA and 5wayCG *Leptospirilli* and their frequency. Orthologous proteins were aligned and PTMs were mapped to the modified residues. (A): PTM conservation at the aligned position with a conserved amino acid. (B): PTM divergence at the aligned position with a conserved amino acid, (C): PTM divergence at the position with SAAP. The bar graph shows the frequency of each case. X and Y represent an amino acid residue.

(A)

Locus ID: CGL2_11067G0037 (top) and UBAL2_8692G0154 (bottom)

```
501 VCILGVDTTDPETIVRLAEEAGLGGEGGAPPSLRIVVSSQSGSTLEVSSL    550
    |||||||||||||||||||||||||||||||.||||||||||||||||||
501 VCILGVDTTDPETIVRLAEEAGLGGEGGAPPPLRIVVSSQSGSTLEVSSL    550

551 YSYFRALLEKKGVRAAGEYFWALTDPSSSLESLAIQEKFGRIIRNTPGIP    600
    ||||||||||||||||||||||||||||:|||||::||||||||||||||
551 YSYFRALLEKKGVRAAGEYFWALTDPSSALESLAMREKFGRIIRNTPGIP    600

601 GRFSAFSVPVLLAASLLKGPRGLDQAMNASRLAYEALQKNGYDGRGLSLA    650
    ||||||||||||||||||||:||||||||||:||||||||||||||||||
601 GRFSAFSVPVLLAASLLKGPQGLDQAMNASRVAYEALQKNGYDGRGLSLA    650

651 VFLGAGYRTGRDKVLLFGPPKLSGWFEQLLAEGTGKSGTGWIPSGPGNAD    700
    |||||||||||||||||.||||||||||||||||||||||||||||||||
651 VFLGAGYRTGRDKVLLFAPPKLSGWFEQLLAEGTGKSGTGWIPSGPGNAD    700

701 DRSEKEKPSDRLLLEIAFPESPDMDSLSRISRMEEAGESSFFFTVNSEAD    750
    ||||||||||||||||.|||||||||||||||.|:||||||.|||.|..|
701 DRSEKEKPSDRLLLEIGFPESPDMDSLSRISRTEDAGESSFSFTVTSAED    750

751 IYSFFLDAMVGVSLAARDRGVNPFEAPDVALPKEKTRDILDRFAQVGSRV    800
    ||||||||||||||||:|||||||||||||||||||||||||||.|||||
751 IYSFFLDAMVGVSLAAKDRGVNPFEAPDVALPKEKTRDILDRFALVGSRV    800
                  *
801 WEEPQTKTYLKPVFKNEDVSIFAEGFSPKISPSGDLSGSVHSLLDDLVQV    850
    |||||.|||.||.|:||||||||||||||||.|||||.||:|||||||:|
801 WEEPQAKTYRKPTFENEDVSIFAEGFSPKISSSGDLSASVYSLLDDLVKV    850

851 RKKSPYLVLQSWLPSPEKEEGRLLCWGKWVSRRYSLPVMVVRGPAFLHMV    900
    |||||||||||||||.||.||||||||||||.||||||||||||||||||
851 RKKSPYLVLQSWLPSSEKVEGRLLCWGKWVSNRYSLPVMVVRGPAFLHMV    900

901 GQVHKGGPAEGLFLQFSVKDMMDLPVPGQPYGFATLFRSQQLGDFFALSQ    950
    |||.|||||||||||||||||||||||||||||||||||||||:||||||
901 GQVQKGGPAEGLFLQFSVKDMMDLPVPGQPYGFATLFRSQQLGEFFALSQ    950

951 LGHPVAELRFSSRTEAERFLDSSLKQV    977
    |.|||||||||||||||||||||||||
951 LRHPVAELRFSSRTEAERFLDSSLKQV    977
```



(B)

Locus ID: CGL2_11111G0096 (top) and UBAL2_8241G0570 (bottom)

```
  1 MNLHRILFLITLIVLECSGFSWAADPASQKNLDRPPKVLGLNDAIFFGLT     50
    ||||||||||||||||||||:|||||||||||||||||||||||||||||
  1 MNLHRILFLITLIVLECSGFAWAADPASQKNLDRPPKVLGLNDAIFFGLT     50

 51 HHPKIFMFRHQVQKAKAAVQIANAHFLPNVGAGAMFGAGEPGVGNRVFNN    100
    ||||||||||||||||||||||||||||||||||:|||||||||||.|||
 51 HHPKIFMFRHQVQKAKAAVQIANAHFLPNVGAGALFGAGEPGVGNRPFNN    100

101 AYAYSGFLPVTYGVLGPYGHNANLSMAQTAMASLGVTQLLYDFGKYEHLT    150
    ||||||||||.|||||||||||||||||||||||||||||||||||||||
101 AYAYSGFLPVNYGVLGPYGHNANLSMAQTAMASLGVTQLLYDFGKYEHLT    150

151 RSRKELTKASVDNLLTRDAWVILQVKEAYYHVILDRKLIEVYQKNLEQRQ    200
    ||||||||||||||||||||||||||||||||||||||||||||||||||
151 RSRKELTKASVDNLLTRDAWVILQVKEAYYHVILDRKLIEVYQKNLEQRQ    200

201 MVRDLTRSLYRANYKSRLDYDLAVVDLEKAKALLVNEQNDLESQIARLNE    250
    ||||||||||||||||||||||||||||||||||||||||||||||||||
201 MVRDLTRSLYRANYKSRLDYDLAVVDLEKAKALLVNEQNDLESQIARLNE    250
                     *
251 AMGLGKKSRKNYRLKDKAPENFVPVPLEELISTGVQKRPELLSTTHRYHA    300
    ||||||||||||||||:||:||||||||||||||:|||||||||||||||
251 AMGLGKKSRKNYRLKDRAPDNFVPVPLEELISTGIQKRPELLSTTHRYHA    300

301 GVEKTASEKAKHYPYISAFGNYGYLGNMVTGQSYSPGLWTGGAMINVPIY    350
    ||||||||||||||||||||||||||||||||||||||||||||||||||
301 GVEKTASEKAKHYPYISAFGNYGYLGNMVTGQSYSPGLWTGGAMINVPIY    350

351 TGGMIRGMVARAREATLTSQYHEQDWRIRIRLQVTQAYDRVRADAADIVA    400
    ||||||||||||||||||||||||||||||||||||||||||||||||||
351 TGGMIRGMVARAREATLTSQYHEQDWRIRIRLQVTQAYDRVRADAADIVA    400

401 YTKAVKEAKLALLLANKKYEANLISIVQLTLAEVYLLDAEASLALAEYHM    450
    ||||||||||||||||||||||||||||||||||||||||||||:|||||
401 YTKAVKEAKLALLLANKKYEANLISIVQLTLAEVYLLDAEASLAIAEYHM    450

451 GVDQAALRFTTGIDYPEYVTMTGQVRDSQVKTSLDSRIPQ    490
    ||||||||||||||||||||||||||||:|||||||||||
451 GVDQAALRFTTGIDYPEYVTMTGQVRDSRVKTSLDSRIPQ    490
```

Figure 6.10 Sequence alignment of orthologous transaldolase (A) and outer membrane efflux protein (B). The sequence on the top is from the 5wayCG-type ortholog and the sequence on the bottom is from the UBA-type ortholog. SAAP positions with divergent PTM were marked with * and modified residues were highlighted in red. Partial sequences were shown for the transaldolase due to space constraint.

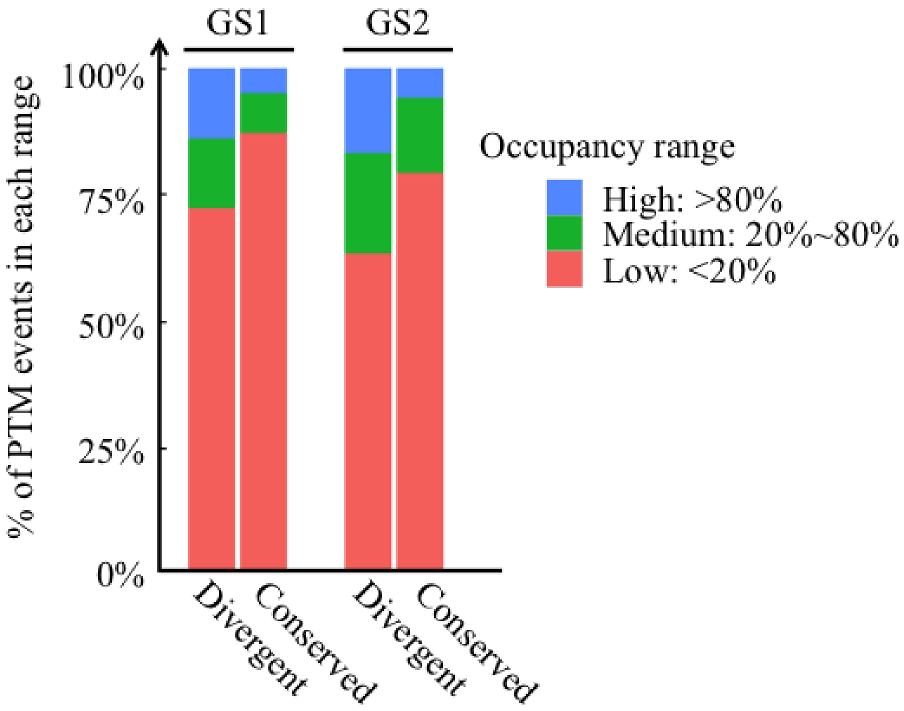


Figure 6.11 Comparison of divergent PTM event and conserved PTM events in terms of the fractional occupancy range.

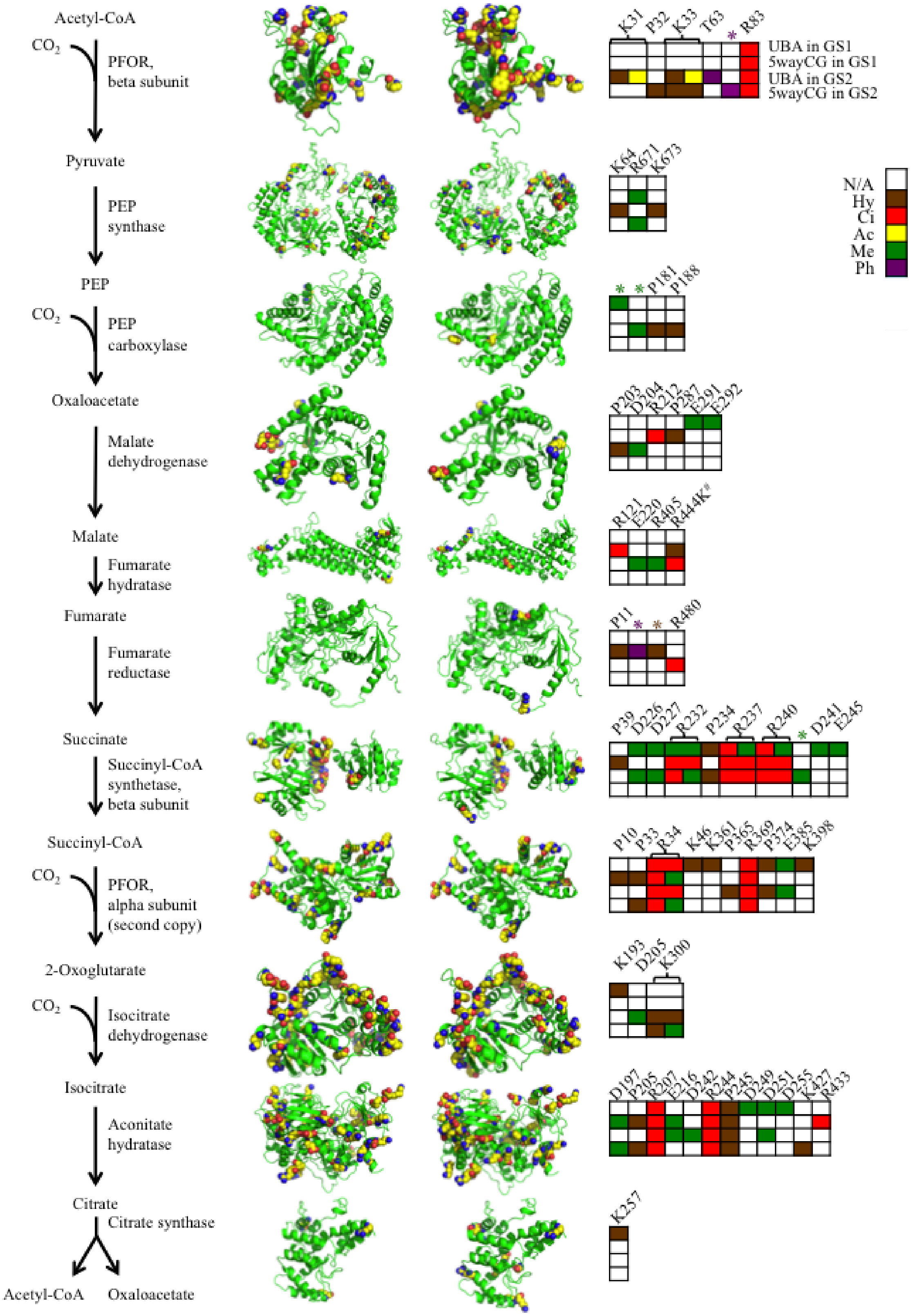
two growth stages and divergence between organisms on the rTCA cycle (Figure 6.12A and B). There were PTM events that occurred only in a specific organism in one sample [e.g. the PTMs on the fumarate reductase (GCL2_11068G0116 and UBAL2_7931G0249)] and PTM events that occurred across both organisms in both samples [e.g. the hydroxylation at the P245 and citrullination at the R207 and R244 on the aconitate hydratase (CGL2_11068G0120 and UBAL2_7931G0253)]. Some PTM events were specific to a growth stage and conserved across organisms [e.g. the hydroxylation on the alpha subunit of succinyl-CoA synthetase (CGL2_11068G0122 and UBAL2_7931G0255) in GS1, Figure 6.12B], while some PTM events were organism-specific in both samples [e.g. a cluster of methylations at the D226, D227, and R232 on the UBA-type succinyl-CoA synthetase's beta subunit (UBAL2_7931G0248)].

6.7 Discussion

In this study, we optimized a shotgun proteomic approach for identification and quantification of a broad range of PTMs. Combining multiple proteases with the optimized HCD method significantly increased the sequence coverage of proteins, which allowed identification of more PTM events and estimation of their fractional occupancy. High-resolution MS/MS provided parts per million-level mass accuracy on every matched fragment ion of the identified peptides, which was essential for controlling the FDR of PTM identification. High-performance computing enabled searching an enormous sequence space with these many types of PTMs. In comparison to enrichment-based approaches, our new approach has the advantages of simultaneous detection of multiple types of PTMs and direct quantification of the fractional occupancy of PTM events. Furthermore, our approach only consumes micrograms of proteins, whereas the enrichment-based approaches typically require milligrams of proteins, which are not available for many environmental samples. Using this approach, we identified a large number of PTM events in laboratory-grown *E. coli* and in the dominant bacteria associated with two growth stages of a natural biofilm community in acid mine drainage. This model community has a well-curated protein database for effective database searching and the extensive prior work provided the ecological and evolutionary context for our results.

PTM profiles of many proteins were substantially different between the two growth stages of the AMD community, indicating that dynamic PTMs may regulate the metabolic activities of organisms under different environmental conditions. *Leptospirillum* spp. are the

(A)


Acetyl-CoA
CO2
PFOR, beta subunit
Pyruvate
PEP synthase
PEP
CO2
PEP carboxylase
Oxaloacetate
Malate dehydrogenase
Malate
Fumarate hydratase
Fumarate
Fumarate reductase
Succinate
Succinyl-CoA synthetase, beta subunit
Succinyl-CoA
CO2
PFOR, alpha subunit (second copy)
2-Oxoglutarate
CO2
Isocitrate dehydrogenase
Isocitrate
Aconitate hydratase
Citrate
Citrate synthase
Acetyl-CoA
Oxaloacetate
K31 P32 K33 T63 * R83
UBA in GS1
5wayCG in GS1
UBA in GS2
5wayCG in GS2
K64 R671 K673
N/A
Hy
Ci
Ac
Me
Ph
* * P181 P188
P203 D204 R212 P287 E291 E292
R121 E220 R405 R444K#
P11 * * R480
P39 D226 D227 R232 P234 R237 R240 * D241 E245
P10 P33 R34 K46 K361 P365 R369 P374 E385 K398
K193 D205 K300
D197 P205 R207 E216 D242 R244 P245 D249 D251 D255 K427 R433
K257

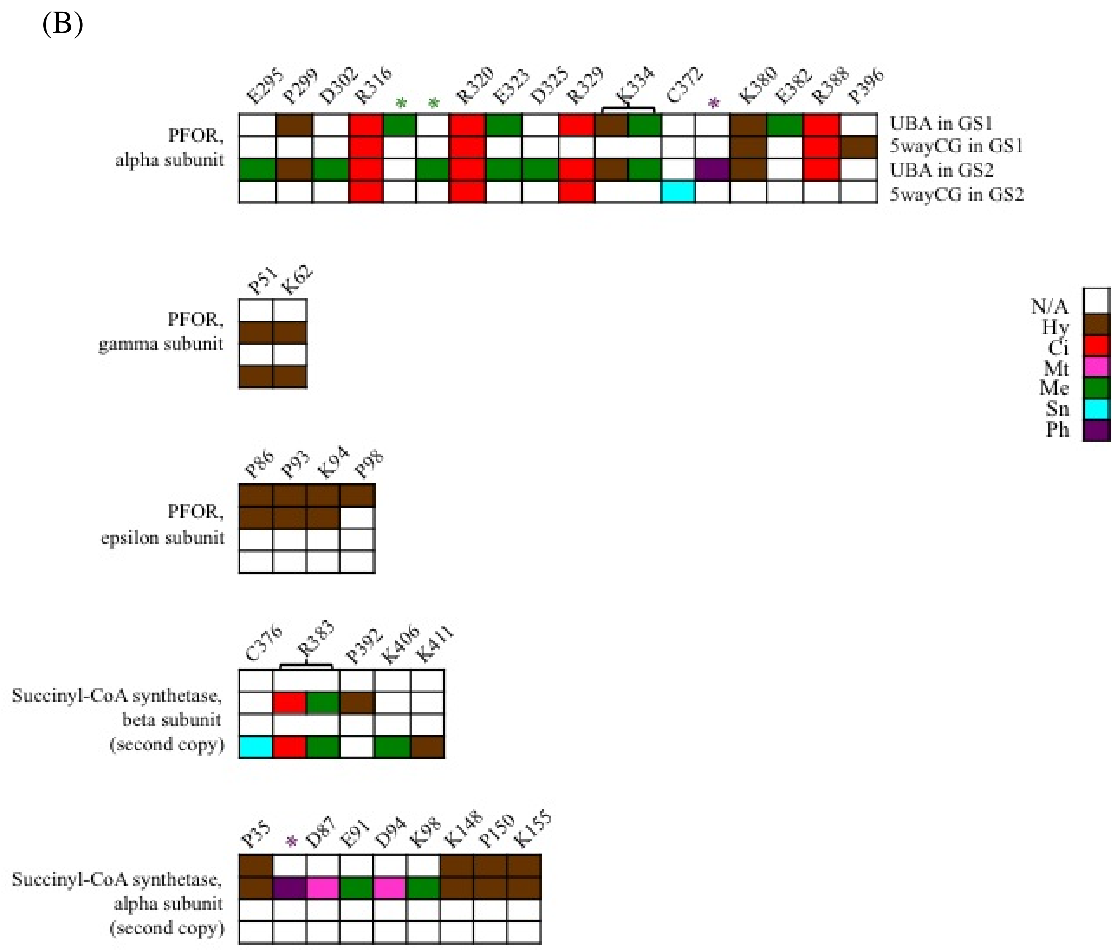


Figure 6.12 Dynamics and organism-specific divergence of the PTM patterns in the rTCA cycle of the *Leptospirillum* group II. UBA-type protein sequences were used as input for structure prediction. Modified residues that carried PTMs unique to the *Leptospirillum* group II (GS1 on the left and GS2 on the right) were highlighted with ball representation in the predicted structures. Only one subunit was showed for multi-subunit enzymes. Residues with organism-specific PTMs were showed in grids marked with residue positions and color-coded by PTM types. The four rows in each grid represent UBA-type ortholog in GS1, 5wayCG-type ortholog in GS1, UBA-type ortholog in GS2, and 5wayCG-type ortholog in GS2. PEP: phosphoenolpyruvate. PTMs that cannot be localized to a specific residue were marked with *. Position with SAAP was marked with #. Color code of the balls on the protein structures: green: backbone; yellow: carbon; blue: nitrogen; red: oxygen; orange: sulfur.

primary producers in the AMD system, using the rTCA cycle for carbon fixation. Proteomic stable isotope probing has shown scant protein production and little net growth in GS2 biofilms, which suggests decreased carbon fixation[220]. However, we found no significant protein abundance changes for most enzymes in the rTCA cycle, as has been observed previously[225]. Particularly, the key carbon fixation enzyme, PFOR, was highly abundant in both GS1 and GS2 (total spectral counts of 14,514 in GS1 and 11,160 in GS2). We believe the discrepancy between the expected decrease in rTCA cycle activity and the lack of corresponding protein abundance changes for rTCA enzymes can be explained by the PTM changes on these enzymes between the two growth stages. We hypothesize that, in GS2, *Leptospirillum* II bacteria may modulate the rTCA cycle activity through concerted PTM changes on the rTCA enzymes, while maintaining the protein stocks of these enzymes to be able to quickly respond to favorable conditions for growth and rapidly meet the demand for carbon fixation. This example shows the importance of taking PTM regulation into account when inferring the activity changes of enzymes from their abundance changes in microbial ecology studies.

In the mature biofilms, *Leptospirillum* group II bacteria have been shown to increase the abundances of chemotaxis proteins in response to diminishing availability of nutrients[225]. Here we observed similar results, with a 3.7-fold increase in the protein abundance of MCP and 4.1-fold for CheA from GS1 to GS2. We additionally observed a number of PTM changes on chemotaxis-associated proteins that could alter environmental sensing and signal transduction. For example, a series of hydroxylations on MCP were observed only in GS1 within its predicted extracellular ligand-binding domain[229]. These changes could have profound effects on the ligand binding activities and may generate distinct environmental sensory responses between the two growth stages. Methylation and demethylation of chemotaxis proteins is a well-known mechanism for regulating organisms' mobility in response to different attractants and repellents[230]. Here, we observed a number of methylation events of the chemotaxis proteins that differ between the two growth stages. Overall, these results suggest that, during the ecological succession of the AMD biofilm, *Leptospirillum* group II bacteria not only increase protein abundances to achieve higher degree of mobility, but also may use PTM changes to alter their chemotaxis behaviors for environmental sensing.

Viral defense is essential for natural communities. Here, the Cas proteins of *Leptospirillum* II bacteria were highly abundant in both growth stages (total spectral counts of

3,008 in GS1 and 5,271 in GS2), but the Cas proteins were not detected in the *E. coli* laboratory culture. The abundances of Cas proteins may reflect the risk level of virus attack in the organisms' environments, ranging from virtually none in the laboratory condition for *E. coli*, to medium risk in GS1 for *Leptospirillum* II bacteria, and to high risk in GS2 with elevated stress and no significant growth. We believe the activity of the CRISPR/Cas system may also be regulated by dynamic PTMs between the two growth stages to handle different levels or types of viral stresses. These PTMs were found to be located on the surface residues of proteins and, therefore, may exert regulatory effects by altering the protein-protein interactions or protein-nucleic acid interactions in the CRISPR/Cas complex. To the best of our knowledge, this is the first report of PTMs on the CRISPR/Cas system. The biotechnology applications of engineered Cas proteins need to consider potential structural changes and regulatory implications of PTMs[231]. The finding of extensive modifications in this study may provide the foundation for further biochemistry studies to determine the biological effects of PTMs in Cas proteins.

Remarkable PTM divergence was found between the UBA and 5wayCG *Leptospirilli*. SAAPs between orthologous proteins in these organisms can directly contribute to PTM divergence on the SAAP position by substituting a modifiable residue with an unmodifiable residue. There was also a large portion of divergent PTMs that occurred in the vicinity of the SAAP positions. Such divergent PTMs could be caused by SAAPs on the motif that might alter modification enzyme-substrate interactions[5]. Because of regulatory roles of PTMs in protein activities, differential modifications may contribute to subtle functional variations between orthologous proteins and may play an important role in ecological and evolutionary divergence between closely related organisms.

The dynamics and organism-specific divergence of PTMs may be interpreted using a *trans*/*cis* model adapted from gene regulation[206]. Changes in transcriptional factor (*trans*-effects) or in regulatory DNA sequence (*cis*-effects) could cause variations in gene expression between closely related organisms[232]. Similarly in protein post-translational modification, a modification enzyme (*trans*-element) recognizes a motif on the protein (*cis*-element) to carry out a modification reaction. The dynamics of PTMs between different conditions in the same organism is probably due to changes in the modification enzymes' activities (*trans*-effects). On the other hand, the divergence of PTMs between different organisms can be caused by polymorphism(s) on/around the target residue (*cis*-effects) or a combination of *cis*-effects and *trans*-effects.

Further study on model organisms will be needed to validate the *cis*-effects and *trans*-effects in the regulation of PTMs.

In conclusion, our new proteomic approach revealed a broad range of PTMs on proteins from coexisting microorganisms in a natural biofilm community. The prevalence and variety of PTMs greatly expands the structural diversity and the function promiscuity[231] of proteins. We believe dynamic PTMs are widely used in many ecological processes as a way of modulating enzyme activities in response to changes in environmental conditions. Closely related, but ecologically distinct, bacteria harbored notably divergent PTM patterns between orthologous proteins, which may contribute to their ecological divergence[212]. The findings of this study motivate further study of the role of PTMs in the ecology and evolution of microbial communities.

CHAPTER 7
SEARCHING FOR NEUTRAL LOSS FRAGMENT IONS BOOSTS SENSITIVITY OF PHOSPHO-SPECTRUM IDENTIFICATION FROM HIGH-RESOLUTION MS/MS DATA

7.1 Introduction to MS-based approaches in phosphopeptide analysis

Protein phosphorylation/de-phosphorylation is one of the most important signal transduction switches that control a myriad of molecular and cellular processes[233]. Abnormal addition or removal of a phosphate group on a protein can erroneously turn on or turn off signaling pathway, which is implicated in the molecular basis of various diseases, such as cancer. Technical advancement of mass spectrometry-based proteomics in the past decade has overcome many challenges in proteome-wide characterization of the phosphoproteome. These include developing various enrichment techniques[234], such as immobilized metal affinity chromatography (IMAC), antibodies, or strong cation exchange to increase in a complex peptide mixture the relative abundance of phosphopeptides that normally exist at low stoichiometry in cellular environment, and implementing PTM site localization algorithms[40] to pinpoint the modified residue on a peptide.

Fragmentation of phosphopeptide with ion trap CID typically results in MS/MS spectra that are dominated by precursor ions with the neutral loss of phosphate group[235]. The relatively paucity of sequence informative fragment ions in phospho-spectra (i.e. low quality spectra) often precludes phosphopeptide identification. To resolve this issue, various fragmentation methods have been tailored, such as multi-stage activation[236], or neutral loss-triggered MS3[237]. In the multi-stage activation method, a precursor phosphopeptide is first isolated and activated, followed by an activation of the neutral loss precursor. The fragment ions from both the primary activation and the secondary activation are mass-analyzed and recorded in the same spectrum. In contrast, in the neutral loss-triggered MS3 method, the primary isolation and activation of a precursor phosphopeptide is followed by a secondary isolation and activation of the neutral loss precursor. Only fragment ions from the secondary isolation and activation are mass-analyzed. Fragment ions from the primary activation are not retained. Although these two methods have been demonstrated to obtain more sequence-informative fragment ions from neutral loss precursor, which improves the success of phosphopeptide identification, collection of MS3 spectra in the neutral loss-triggered MS3 method or pseudo-MS3 spectra in the multi-stage activation requires extra data acquisition time, which may adversely impact overall identification performance in large-scale phosphoproteomic study. Thus, Gygi and colleagues contended that the conventional MS2 method implemented on high mass accuracy instrument platform can obviate the acquisition of MS3 or pseudo-MS3 because accurate precursor mass measurement already provides discriminating power for confident peptide identification, even though fragment ions are sparse[238].

HCD has become increasingly popular for phosphoproteomic studies. Compared with ion trap CID, it generates higher quality MS2 spectra with richer fragment ions that are usually detected in Orbitrap with high resolution and high mass accuracy. With Orbitrap HCD, collection of MS2 spectra is usually sufficient for phosphopeptide identification[44]; however, overwhelming neutral losses are also observed with both methods[239]. In either CID or HCD fragmentation, typically three types of outcome exist for fragment ion that originally bears a phosphate-group (Figure 7.1): Type I: without neutral loss, meaning phosphate group is still attached to the fragment ion; Type II: with neutral loss of HPO_3 group, meaning the mass of fragment ion bearing modified residue is reduced by 79.966331 Da; Type III: with neutral loss of H_3PO_4, meaning the mass of fragment ion bearing modified residue is reduced by 97.9769 Da. Most database searching algorithms only assume the Type I fragmentation pathway for phosphopeptide and ignore fragment ions from the Type II and Type III pathway. Since the Type II and III pathway are more prevalent than the Type I pathway during ion trap CID and HCD fragmentation and the identification success depends on the number of matched fragment ions and matched ion intensities, we propose that consideration of all three pathways during the database searching would improve the score of phosphopeptides for more confident identification. Particularly, such score improvement may boost the sensitivity of phosphopeptide identification in large-scale phosphoproteomic studies because it could recuse the identification of phosphopeptides that would not be identified due to the lower number of matched fragment ions and lower matched ion intensities in the Type I pathway.

7.2 Implementation of the neutral loss search module improves phospho-spectrum identification

During the development of Sipros 3.0[67], a neutral loss search module has been incorporated, which predicts neutral loss fragmentation pathway for a precursor ion. In this case,

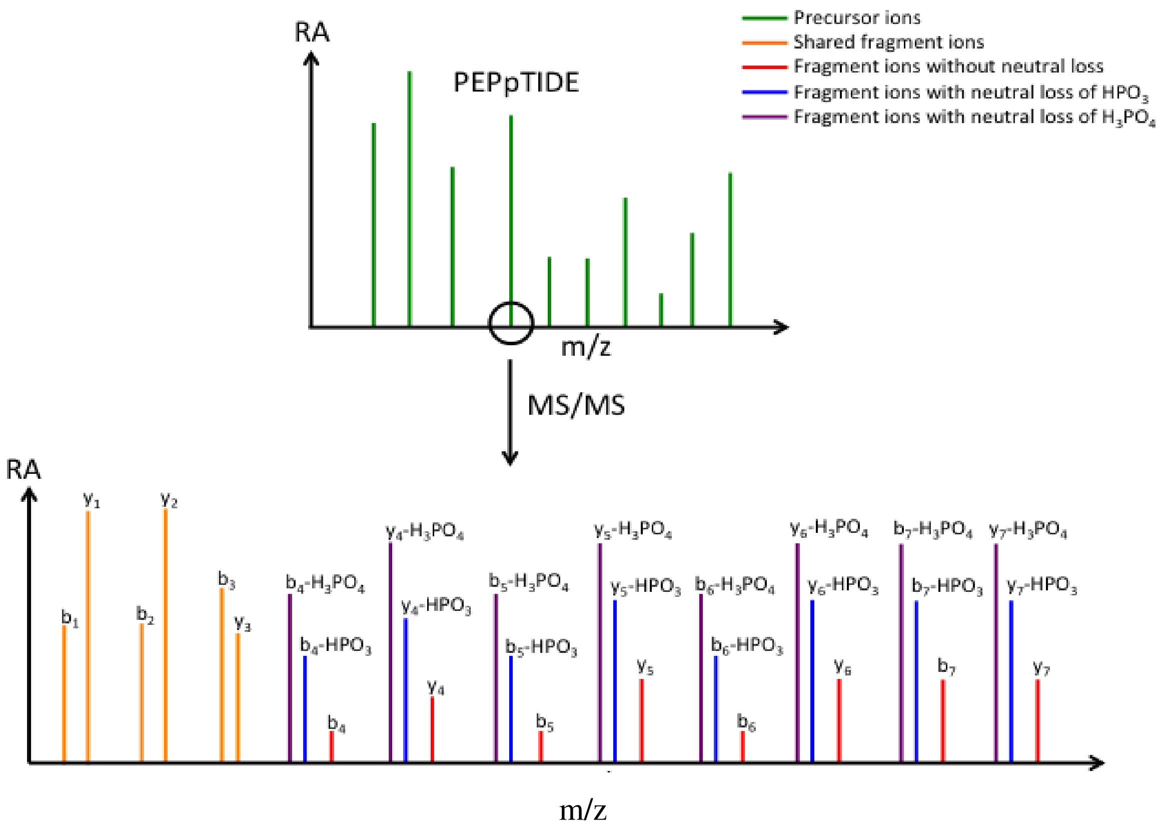


Figure 7.1 Illustration of three fragmentation pathways of phosphopeptide. RA: relative abundance

the user needs to specify the elemental compositions or the mass of neutral loss group and amino acid residue where the neutral loss occurs. Such flexible configuration design allows searching any types of neutral loss on any amino acid residue as long as the elemental compositions or the mass of neutral loss group and affected residue are provided to the program.

In the case of the neutral losses on phosphopeptide (Figure 7.1), an elemental composition of HPO_3 corresponding to the Type II fragmentation pathway were specified on serine, threonine, and tyrosine. An elemental composition of H_3PO_4 corresponding to the Type III fragmentation pathway was specified on serine and threonine residue. As a control, fragmentation without neutral loss (the Type I pathway) was specified on serine, threonine, and tyrosine. The performance of the neutral loss search module was evaluated by using a large synthetic peptide/phosphopeptide reference library from a published study. This library contains 96 sub-libraries, with each sub-library containing up to 1,200 synthetic peptides and 1,200 synthetic phosphopeptides, totaling more than 100,000 peptides and 100,000 phosphopeptides.

Since the sequence of these synthetic peptides/phosphopeptide are known, we used these peptide sequences to construct the database to search without further *in silico* enzymatic digestion. To evaluate whether the neutral loss search module can identify more phospho-spectra, we carried out two types of database searching against the first twenty sub-libraries, one with implementing neutral loss search module where all three fragmentation pathways are predicted for each phosphopeptides and the other without implementing the module where only the Type I fragmentation pathway is predicted. As shown in the Table 7.1, the implementation of the neutral loss search module increased the number of identified phospho-spectra for most sub-libraries (16 out of 20) with an overall improvement of 13%. The percentage increase for individual sub-library was up to 91% (i.e. sub-library #4 in Table 7.1). However, the implementation of the module decreased the number of identified phospho-spectra for some sub-libraries (3 out of 20). While it is unknown why these three sub-libraries (sub-library #9, #16, and #17 in Table 7.1) suffered a decrease in phospho-spectrum identifications, it is notable that the number of identified phospho-spectra for the sub-library #16 and #17 was considerably lower than other libraries, even without the module implemented. This raises concerns about the quality of peptide synthesis and or of mass spectrometric measurement for these two libraries. Overall, implementation of the module improved the identification of phospho-spectra by 13%. Since Sipros and some other search algorithms, such as MyriMatch, score candidate peptides

Table 7.1 Comparison of the number of identified phospho-spectra with or without neutral loss search module implemented.

Sub-library	# of identified phospho-spectra with NL implemented	# of identified phospho-spectra without NL implemented	% of increase (positive value) or decrease (negative value) with NL implemented
1	2897	2036	42%
2	1932	1700	14%
3	1285	798	61%
4	484	253	91%
5	2098	1877	12%
6	2238	1374	63%
7	3019	2803	8%
8	460	271	70%
9	1572	1898	-17%
10	1706	1511	13%
11	1199	653	84%
12	2191	1875	17%
13	1881	1706	10%
14	NA	NA	NA
15	2323	2110	10%
16	264	887	-70%
17	29	668	-96%
18	2347	2192	7%
19	1994	1871	7%
20	2343	2108	11%
total	32262	28591	13%

based on the number of matched fragment ions and matched ion intensities, the improved phospho-spectral identification can be explained by the fact that the Type II or Type III pathway generates more matched fragment ions and/or higher matched ion intensities, which increase the score of the phosphopeptides. As an example shown in Figure 7.2, there are three abundant peaks with m/z of 467.24, 667.35, and 780.44 that match to y4, y6, and y7 ions, respectively, from the Type III fragmentation pathway (Figure 7.2 A and C). In contrast, there are three noise-level peaks that also match to y4, y6 and y7 ions from the Type I fragmentation pathway (Figure 7.2D). Because the matched ion intensities from the Type III pathway are much higher than those from the Type I pathway, the neutral loss version of this phosphopeptide ALLSLHpSNK received a higher score (36.74) than its counterpart without neutral loss (24.27).

In summary, we designed a neutral loss search module in Sipros 3.0, which improves the sensitivity of phosphopeptide identification. We believe this module is broadly applicable to searching many other types of PTM-containing peptides that frequently suffer from neutral loss of PTM moiety during fragmentation, such as acetylation, methylation, glycosylation, etc. The deeper PTM identification enabled by this approach would allow new discoveries of regulatory mechanisms and functional roles of PTMs in many molecular and cellular processes.

Figure 7.2 Annotated spectrum for phosphopeptide ALLSLHpSNK's Type III (A and C) and Type I (B and D) fragmentation pathway. The MS2 spectrum for this peptide was separated into two sub-spectra due to space constraint: (A) and (B) cover the m/z range of 0-500 for the Type III and Type I fragmentation pathway, respectively, and (C) and (D) cover the m/z range of 500-1000 for the Type III and Type I fragmentation pathway, respectively. (A) and (C) constitute the annotated spectrum for the type III fragmentation pathway. (B) and (D) constitute the annotated spectrum for the Type I fragmentation pathway.

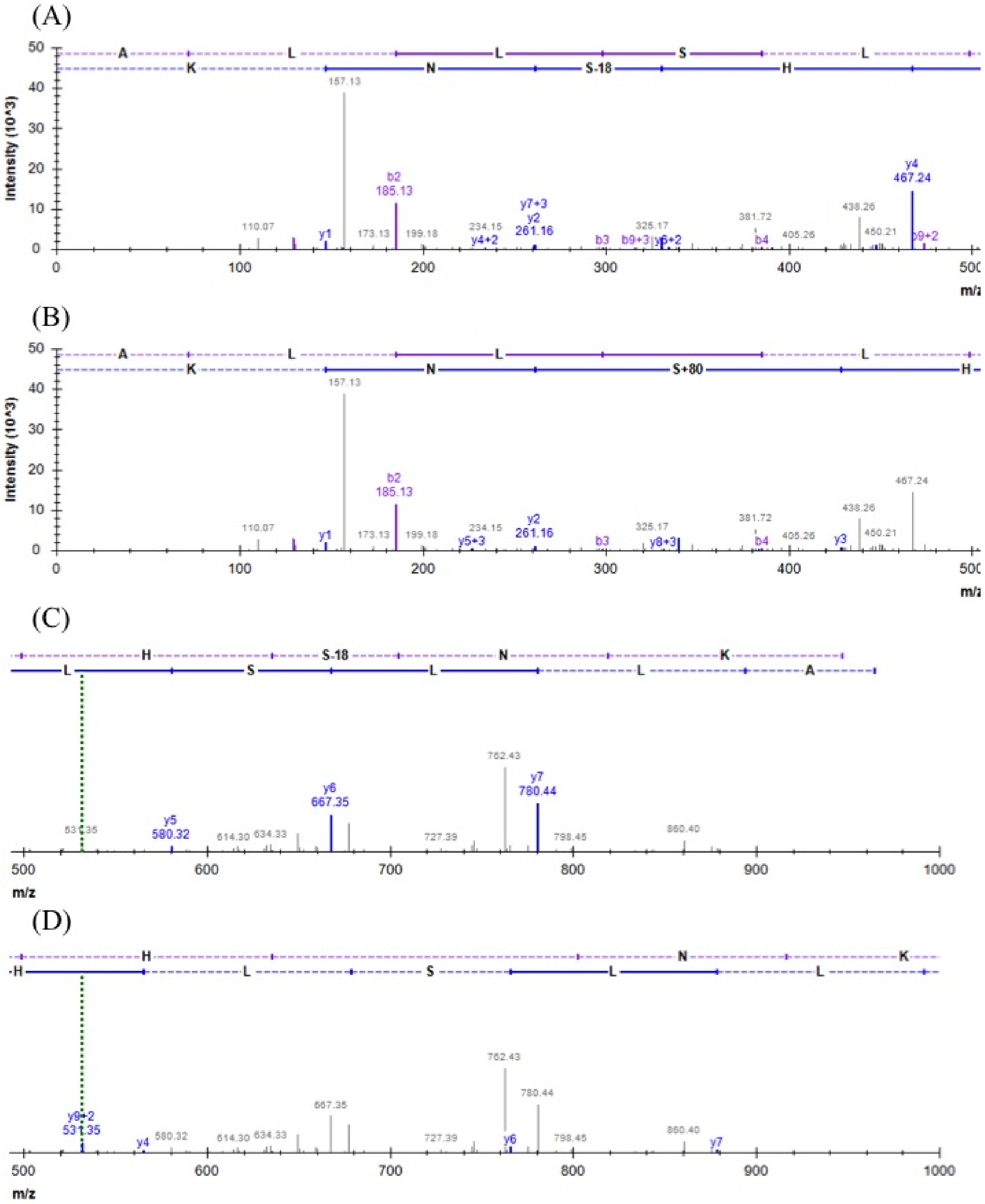


Figure 7.2 continued

CHAPTER 8
CONCLUSIONS AND FUTURE OUTLOOK

8.1 Scientific impact of this dissertation research

The capability to identify and quantify thousands of proteins and their PTM signatures enabled by mass spectrometry-based proteomics has proven its pivotal role in systems-level functional characterization of a cell, organism, or community. By utilizing existing proteomic approaches, and to develop new proteomic approaches, five goals were outlined in the Chapter 1: 1) head-to-head compare the identification and quantitation performance of label-free, metabolic labeling, and isobaric chemical labeling; 2) quantify the proteome of *Arabidopsis* seedlings in response to the strigolactone treatment; 3) characterize how elevated temperature impacts the physiology of individual microbial groups in a laboratory-grown AMD microbial community; 4) develop a new proteomic approach to profile a broad range of PTMs in individual organisms in a natural AMD microbial community; and 5) evaluate the utility of a neutral loss search module for improving the sensitivity of phosphopeptide identification.

As described in Chapter 3, the label-free method provides the deepest proteome coverage for identification, and metabolic labeling and isobaric chemical labeling are capable of accurate, precise, and reproducible quantification. On the basis of the unique advantages of each method, we provide guidance for selection of the appropriate method for a quantitative proteomics study. As described in Chapter 4, strigolactone regulates the expression of about three dozen proteins that have not been previously assigned to strigolactone pathways. These findings provide new targets for follow-up biochemical and genetic studies to further investigate the molecular mechanism of strigolactone signaling.

As described in Chapter 5, elevated temperature repressed carbon fixation by two *Leptospirillum* genotypes, whereas carbon fixation was significantly up-regulated at higher temperature by a third member of this genus. These results highlight the utility of proteomics-enabled community-based physiology studies, an approach that can be applied to more complex ecosystems.

As described in Chapter 6, a new proteomic approach has been developed for simultaneous characterization of a broad range of PTMs in microbial systems. Evaluation of this approach with an *E. coli* proteome revealed unexpected depth and breath of non-enriched PTMs in this model organism and provided valuable resource to study the regulatory mechanisms and functional implications of PTMs in prokaryotic system. Application of this approach to AMD microbial community reveal diverse, dynamic, and divergent PTM patterns during an ecological succession. The findings of this study motivate further study to unravel the role of PTMs in microbial adaptation, evolution and ecology.

As described in Chapter 7, implementation of the neutral loss search module improves the sensitivity of the phosphopeptide identification, evaluated with a synthetic phosphopeptide library. We believe this module should be broadly applicable to other PTM analysis by mass spectrometry-based proteomics to increase the depth of the PTM identification and to unravel novel biological insight into PTM regulations.

8.2 Future outlook

Although mass spectrometry-based proteomics has been able to provide robust proteome measurement, especially for model organisms, such as yeast[21] and human cell line[26], and microbial isolates[240], technical developments are still needed to advance the field in the following aspects:

(1) Direct proteome measurement without multi-dimensional liquid chromatography-based separation: the instrumentation that couples multi-dimensional liquid chromatography to mass spectrometry has been dominating the field of Proteomics over the past decade. While application of multi-dimensional separation has alleviated some major technical issues associated with analyzing complex proteome sample with mass spectrometry, such as ionization suppression and dynamic range, it also significantly prolongs the measurement time in order to achieve decent deep proteome coverage. However, the sequencing speed of current state-of-art mass spectrometers is becoming faster and faster. For example, the next-generation mass spectrometer, LTQ Orbitrap Fusion, is able to perform 20 MS/MS experiments in ion trap or 15 MS/MS experiments in Oribtrap within 1 second[21]. Assuming the average liquid chromatographic elution time of peptide is ~30 seconds, this high-speed mass spectrometer is able to sample up to 600 different peptide ions (20x30) within this time frame. Such unprecedented sampling depth would be able to sequence most, if not all, peptides delivered to the mass spectrometer. This would significantly shorten the proteome analysis time because extensive, time-consuming multi-dimensional separation would become unnecessary. Indeed, the current high-throughput mass spectrometry has begun to obviate the multi-dimensional

separation of peptides from relatively simple proteome[238] , such as yeast where nearly 4,000 proteins can be identified in ~1.3 h of measurement time with one-dimensional reverse phase-based separation. Although one-dimensional 4-hour separation has been applied to measuring human proteome[241], the depth of human proteome coverage is far from complete compared to that of the yeast proteome coverage (50% of the expressed proteins in human proteome vs. ~100% of the expressed proteins in yeast proteome). However, with further development of sequencing speed of mass spectrometer, it is likely that the complete human proteome analysis can be realized within an hour of measurement.

(2) Increasing multiplexing capability of proteomic measurement: nowadays, mass spectrometry-based proteomics has become readily accessible to many biological researchers; however, the cost of performing proteome measurement is still relatively high, compared to DNA and RNA sequencing. One solution to reduce the cost is to expand the multiplexing capability of proteomic measurement where multiple proteome samples are simultaneously measured instead of analyzing one sample at a time, so that the cost per sample can be reduced significantly. Isobaric chemical labeling-based quantitative proteomics is an ideal approach to expand the measurement throughput because it allows not only identification but also quantification of thousands of proteins from multiple samples in one experiment. Recently, TMT-10 plex has been commercially available. With this reagent, ten proteome samples can be individually labeled and then combined to analyze together. Another elegant way of enhancing multiplexing capability takes advantage of subtle mass difference due to differential neutron-binding energy between ^{13}C and ^{15}N isotopes[242,243]. Similar to isobaric chemical labeling reagent, a set of structurally identical but isotopically different chemical labels with tiny mass differences of only a few mDa are used to label multiple samples individually and then combined prior to the LC-MS/MS measurement. Due to such small mass differences between different chemical labels, peptides labeled with these isotopic variants (i.e. isotopologue-embedded peptide) are almost indistinguishable with modest high-resolution mass spectrometry analysis (30,000-60,000). However, these isotopologue-embedded peptides can be clearly resolved from each other with ultrahigh resolution mass spectrometry analysis (>200,000). The signal intensity of each isotopologue-embedded peptide can be used to quantify their relative abundance. Since regular ^{12}C and ^{14}N atoms of the chemical label can be replaced with ^{13}C and ^{15}N isotope at any position with any quantity, the number of structurally identical but isotopic distinct chemical label can be

expanded tremendously. Such concept can also be applied to isobaric chemical labeling reagent to further "plex" the reagent[239]. Overall, with the continuous development of multiplexed chemical labeling reagents, it would be routine to qualitatively measure and quantitatively compare dozens of proteome samples or even more within one experiment in future.

(3) Proteome-wide analysis of PTM cross-talk: mass spectrometry-based proteomics has been very successful for characterization of a single type of PTM at a time. However, there are a few hundred different types of PTMs documented so far, and their prevalence, dynamics, and function in an organism are largely unknown. Furthermore, many proteins, such as histone, tumor suppressor p53, and tublin, are modified with multiple types of PTMs[244]. These proteins' activities are often regulated by the cross-talk among different PTMs. To study how different PTMs collectively regulate protein functions at proteome-wide would require a method that is able to simultaneously identify a broad range of PTMs. The current enrichment methods have limited application to studying proteome-wide PTM cross-talk because only one type of PTM can be enriched at a time and peptides that carry other different types of PTMs are lost. Recently, tandem enrichment has been developed to enrich multiple types of PTMs[202,203]. However, these approaches consume a large quantity of materials that are often not available for many biological systems. Moreover, with the current enrichment approaches, it is difficult to quantify PTM abundance, especially the fractional occupancy because unmodified peptides are discarded, precluding the calculation of the ratio between modified version and unmodified version of a peptide.

In this dissertation, we demonstrated a broad-range PTM identification and quantification approach that can be readily applied to studying proteome-wide PTM cross-talk. Although this approach has been tested in microbial systems because of the relatively simplicity of their proteome, we believe that with improving sequencing speed of mass spectrometer, it would be likely to realize this approach in complex eukaryotic organisms. Enrichment-free approach is appealing because it consumes less protein materials and offers an opportunity to comprehensively capture multiple types of PTMs at once and to quantify PTM fractional occupancy.

(4) Environmental proteomics of complex microbial communities: environmental proteomics has been extremely successful for charactering microbial communities where enough biomass is relatively easy to obtain[245]. When dealing with recalcitrant communities, such as soil,

extraction of high quality protein sample becomes a major technical bottleneck. Thus, advanced protein preparation that is capable of obtaining clean, sufficient proteins for mass spectrometric analysis needs to be developed. Further, a complex microbial community likely contains thousands of species. Assuming each species may express 2,000 proteins in its cells, a complex metaproteome could contain up to 2 million (1,000 species x 2,000 proteins per species) different proteins. Since a typical mass spectrometric run can only identify thousands of proteins, a tiny fraction of the expressed proteins in a complex community can be captured with the current instrumentation.

Most environmental samples, if not all, have been measured with the data-dependent acquisition approach. However, due to the currently un-measurably huge dynamic range of proteins in a complex metaproteome, data-dependent acquisition would inevitably miss sampling of low-abundant peptides. Recently, a data-independent acquisition (DIA) technique has become an attractive alternative to the DDA approach[246]. Instead of sequentially isolating and fragmenting the Top N (usually Top 10 - 20) most abundant peptide ions in the DDA approach and missing the remaining low abundant ions (Figure 8.1A), the mass spectrometer cycles through a series of precursor isolation windows (e.g. 25 m/z wide) and fragments all peptide ions within each window, until the entire m/z range is covered (Figure 8.1B). In the next sampling cycling, the series of same isolation window within the same m/z range will be isolated and fragmented again. Since the DIA approach sequences peptides independent of their intensities and is able to sample the entire m/z range which most eluting peptides fall within, theoretically almost all peptides that are delivered into mass spectrometer can be sequenced, as long as the cycling time is shorter than a peptide's elution time. While DIA approach may not provide much benefit when measuring a single organism because the well-established DDA approach has been able to identify most expressed proteins for some model organisms, this approach would be especially appealing to environment proteomics because the potential capability that sequences all eluting peptides is able to alleviate the huge dynamic range and sample complexity issues associated with environmental proteome samples and is likely to provide much deeper metaproteome coverage than the DDA approach. Since multiple peptides are likely to exist within an isolation window in the DIA approach, all peptide ions will be co-fragmented to generate complex, multiplexed spectra, which creates a significant bioinformatic challenge in sequence identification. Although exhaustive peptide sequence identification from these

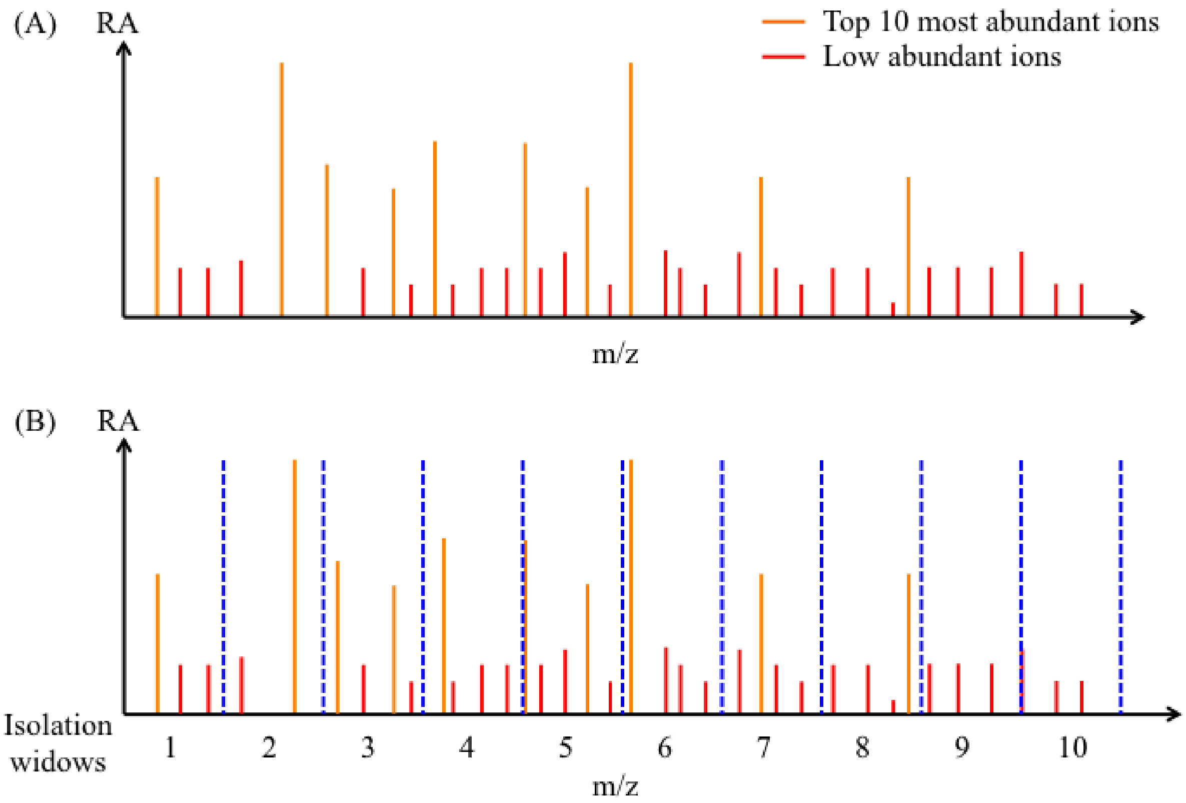


Figure 8.1 Illustration sampling difference between DDA approach (A) and DIA approach (B). In DDA approach, top 10 most abundant ions represented by orange peaks are sequenced and low-abundant ion represented by red peaks are often missed. In DIA approach, the entire m/z range is divided into a series of small isolation windows (spaced by blue dash line). Mass spectrometer sequentially cycle through each isolation windows until the entire m/z range is covered. All ions in each isolation window are fragmented and the resulting fragment ions are recorded in high-resolution tandem mass spectra.

multiplexed spectra is risky with elevated propensity to false positive, we believe future algorithm development would be able to resolve this informatic challenge. Alternatively, as the sequencing speed of mass spectrometer is becoming faster, the width of isolation window in the DIA may be shortened to the one used in the DDA approach (usually 2~3 m/z wide), while maintaining the cycling time within the chromatographic elution time of peptide. This would reduce the number of co-fragmented ions and generate less complex fragment ion spectrum. With such relatively clean spectra, database searching algorithms developed for the DDA approach might be readily applied to searching the spectra acquired with the DIA approach.

8.3 Concluding remarks

Although challenges exist, such as those in metaproteomic characterization of complex microbial communities described above, it is reasonable to expect a new wave of technological innovations to resolve these challenges. With integration with other systematic measurements and advanced computational and bioinformatic approaches, powerful mass spectrometry-based proteomics will undoubtedly not only keep propelling forward our understanding of the genotype-phenotype relationships, but also spark more biotechnological applications, such as drug discovery, bioremediation, and bioenergy production.

LIST OF REFERENCES

1 Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).

2 Beadle, G. W. & Tatum, E. L. Genetic Control of Biochemical Reactions in Neurospora. *Proceedings of the National Academy of Sciences of the United States of America* **27**, 499-506 (1941).

3 Cazier, J. B. & Tomlinson, I. General lessons from large-scale studies to identify human cancer predisposition genes. *The Journal of pathology* **220**, 255-262, doi:10.1002/path.2650 (2010).

4 Bensimon, A., Heck, A. J. & Aebersold, R. Mass spectrometry-based proteomics and network biology. *Annual review of biochemistry* **81**, 379-405, doi:10.1146/annurev-biochem-072909-100424 (2012).

5 Linding, R. *et al.* Systematic discovery of in vivo phosphorylation networks. *Cell* **129**, 1415-1426, doi:10.1016/j.cell.2007.05.052 (2007).

6 Borneman, A. R. *et al.* Divergence of transcription factor binding sites across related yeast species. *Science* **317**, 815-819, doi:10.1126/science.1140748 (2007).

7 Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annual review of genomics and human genetics* **2**, 343-372, doi:10.1146/annurev.genom.2.1.343 (2001).

8 Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662-1664, doi:10.1126/science.1069492 (2002).

9 Bartlett, J. M. & Stirling, D. A short history of the polymerase chain reaction. *Methods in molecular biology* **226**, 3-6, doi:10.1385/1-59259-384-4:3 (2003).

10 Kuska, B. Beer, Bethesda, and biology: how "genomics" came into being. *Journal of the National Cancer Institute* **90**, 93 (1998).

11 Nicolas, P. *et al.* Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. *Science* **335**, 1103-1106, doi:10.1126/science.1206848 (2012).

12 Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470 (1995).

13 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57-63, doi:10.1038/nrg2484 (2009).

14 Walsh, C. T., Garneau-Tsodikova, S. & Gatto, G. J., Jr. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angewandte Chemie* **44**, 7342-7372, doi:10.1002/anie.200501023 (2005).

15 Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207, doi:10.1038/nature01511 (2003).

16 Wilkins, M. R. *et al.* From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/technology* **14**, 61-65 (1996).

17 Cox, J. & Mann, M. Quantitative, high-resolution proteomics for data-driven systems biology. *Annual review of biochemistry* **80**, 273-299, doi:10.1146/annurev-biochem-061308-093216 (2011).

18 Karas, M., Bachmann, D. & Hillenkamp, F. Influence of the Wavelength in High-Irradiance Ultraviolet-Laser Desorption Mass-Spectrometry of Organic-Molecules. *Analytical Chemistry* **57**, 2935-2939, doi:Doi 10.1021/Ac00291a042 (1985).

19 Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64-71 (1989).

20 Washburn, M. P., Wolters, D. & Yates, J. R., 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature biotechnology* **19**, 242-247, doi:10.1038/85686 (2001).

21 Hebert, A. S. *et al.* The one hour yeast proteome. *Molecular & cellular proteomics : MCP* **13**, 339-347, doi:10.1074/mcp.M113.034769 (2014).

22 Boersema, P. J., Mohammed, S. & Heck, A. J. Hydrophilic interaction liquid chromatography (HILIC) in proteomics. *Analytical and bioanalytical chemistry* **391**, 151-159, doi:10.1007/s00216-008-1865-7 (2008).

23 Catherman, A. D., Skinner, O. S. & Kelleher, N. L. Top Down proteomics: Facts and perspectives. *Biochemical and biophysical research communications*, doi:10.1016/j.bbrc.2014.02.041 (2014).

24 Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C. & Yates, J. R., 3rd. Protein analysis by shotgun/bottom-up proteomics. *Chemical reviews* **113**, 2343-2394, doi:10.1021/cr3003533 (2013).

25 Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics : MCP* **4**, 1419-1440, doi:10.1074/mcp.R500012-MCP200 (2005).

26 Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* **7**, 548, doi:10.1038/msb.2011.81 (2011).

27 Sadygov, R. G., Cociorva, D. & Yates, J. R., 3rd. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nature methods* **1**, 195-202, doi:10.1038/nmeth725 (2004).

28 Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655-667, doi:10.1002/pmic.200600625 (2007).

29 Pan, C. *et al.* A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. *BMC bioinformatics* **11**, 118, doi:10.1186/1471-2105-11-118 (2010).

30 Ong, S. E. & Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nature chemical biology* **1**, 252-262, doi:10.1038/nchembio736 (2005).

31 Modrek, B. & Lee, C. A genomic view of alternative splicing. *Nature genetics* **30**, 13-19 (2002).

32 Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215-233, doi:10.1016/j.cell.2009.01.002 (2009).

33 Olsen, J. V. & Mann, M. Status of large-scale analysis of post-translational modifications by mass spectrometry. *Molecular & cellular proteomics : MCP* **12**, 3444-3452, doi:10.1074/mcp.O113.034181 (2013).

34 Stenflo, J., Fernlund, P., Egan, W. & Roepstorff, P. Vitamin K dependent modifications of glutamic acid residues in prothrombin. *Proceedings of the National Academy of Sciences of the United States of America* **71**, 2730-2733 (1974).

35 Zhao, Y. & Jensen, O. N. Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* **9**, 4632-4641, doi:10.1002/pmic.200900398 (2009).

36 Olsen, J. V. & Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 13417-13422, doi:10.1073/pnas.0405549101 (2004).

37 Schroeder, M. J., Shabanowitz, J., Schwartz, J. C., Hunt, D. F. & Coon, J. J. A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Anal Chem* **76**, 3590-3598, doi:10.1021/ac0497104 (2004).

38 Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nature methods* **4**, 709-712, doi:10.1038/nmeth1060 (2007).

39 Taus, T. *et al.* Universal and confident phosphorylation site localization using phosphoRS. *Journal of proteome research* **10**, 5354-5362, doi:10.1021/pr200611n (2011).

40 Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature biotechnology* **24**, 1285-1292, doi:10.1038/nbt1240 (2006).

41 Syka, J. E. *et al.* Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *Journal of proteome research* **3**, 621-626 (2004).

42 Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *Journal of mass spectrometry : JMS* **40**, 430-443, doi:10.1002/jms.856 (2005).

43 Olsen, J. V. *et al.* A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Molecular & cellular proteomics : MCP* **8**, 2759-2769, doi:10.1074/mcp.M900375-MCP200 (2009).

44 Nagaraj, N., D'Souza, R. C., Cox, J., Olsen, J. V. & Mann, M. Feasibility of large-scale phosphoproteomics with higher energy collisional dissociation fragmentation. *Journal of proteome research* **9**, 6786-6794, doi:10.1021/pr100637q (2010).

45 Jedrychowski, M. P. *et al.* Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics. *Molecular & cellular proteomics : MCP* **10**, M111 009910, doi:10.1074/mcp.M111.009910 (2011).

46 Michalski, A. *et al.* Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Molecular & cellular proteomics : MCP* **11**, O111 013698, doi:10.1074/mcp.O111.013698 (2012).

47 Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734-740 (1997).

48 Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43, doi:10.1038/nature02340 (2004).

49 Shi, Y., Tyson, G. W. & DeLong, E. F. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459**, 266-269, doi:10.1038/nature08055 (2009).

50 Ram, R. J. *et al.* Community proteomics of a natural microbial biofilm. *Science* **308**, 1915-1920, doi:10.1126/science. 1109070 (2005).

51 Denef, V. J., Mueller, R. S. & Banfield, J. F. AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *The ISME journal* **4**, 599-610, doi:10.1038/ismej.2009.158 (2010).

52 Benndorf, D., Balcke, G. U., Harms, H. & von Bergen, M. Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. *The ISME journal* **1**, 224-234, doi:10.1038/ismej.2007.39 (2007).

53 Sowell, S. M. *et al.* Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. *The ISME journal* **5**, 856-865, doi:10.1038/ismej.2010.168 (2011).

54 Verberkmoes, N. C. *et al.* Shotgun metaproteomics of the human distal gut microbiota. *The ISME journal* **3**, 179-189 (2009).

55 Chourey, K. *et al.* Direct cellular lysis/protein extraction protocol for soil metaproteomics. *Journal of proteome research* **9**, 6615-6622, doi:10.1021/pr100787q (2010).

56 Wrighton, K. C. *et al.* Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**, 1661-1665, doi:10.1126/science.1224041 (2012).

57 Li, Z. *et al.* Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on LTQ Orbitrap Velos. *Journal of proteome research* **11**, 1582-1590 (2012).

58 Li, Z. *et al.* Strigolactone-Regulated Proteins Revealed by iTRAQ-Based Quantitative Proteomics in *Arabidopsis*. *Journal of proteome research* **13**, 1359-1372, doi:10.1021/pr400925t (2014).

59 Damerval, C., De Vienne, D., Zivy, M. & Thiellement, H. Technical improvements in two□dimensional electrophoresis increase the level of genetic variation detected in wheat□seedling proteins. *Electrophoresis* **7**, 52-54 (1986).

60 McLuckey, S. A. & Wells, J. M. Mass analysis at the advent of the 21st century. *Chemical reviews* **101**, 571-606 (2001).

61 Michalski, A. *et al.* Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Molecular & cellular proteomics : MCP* **10**, M111 011015, doi:10.1074/mcp.M111.011015 (2011).

62 Guthals, A. & Bandeira, N. Peptide identification by tandem mass spectrometry with alternate fragmentation modes. *Molecular & cellular proteomics : MCP* **11**, 550-557, doi:10.1074/mcp.R112.018556 (2012).

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

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

65 Tabb, D. L., Fernando, C. G. & Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of proteome research* **6**, 654-661 (2007).

66 Pan, C. *et al.* Quantitative tracking of isotope flows in proteomes of microbial communities. *Molecular & Cellular Proteomics* **10**, M110. 006049 (2011).

67 Wang, Y., Ahn, T.-H., Li, Z. & Pan, C. Sipros/ProRata: a versatile informatics system for quantitative community proteomics. *Bioinformatics* **29**, 2064-2065 (2013).

68 Hyatt, D. & Pan, C. Exhaustive database searching for amino acid mutations in proteomes. *Bioinformatics* **28**, 1895-1901 (2012).

69 Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. 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* **2**, 43-50 (2003).

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

71 Tabb, D. L., McDonald, W. H. & Yates, J. R. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *Journal of proteome research* **1**, 21-26 (2002).

72 Pan, C. *et al.* ProRata: A quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation. *Analytical chemistry* **78**, 7121-7131 (2006).

73 Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry* **389**, 1017-1031, doi:10.1007/s00216-007-1486-6 (2007).

74 O'Farrell, P. H. High Resolution Two-Dimensional Electrophoresis of Proteins. *The Journal of biological chemistry* **250**, 4007 (1975).

75 Kolkman, A., Dirksen, E. H., Slijper, M. & Heck, A. J. Double standards in quantitative proteomics direct comparative assessment of difference in gel electrophoresis and metabolic stable isotope labeling. *Molecular & Cellular Proteomics* **4**, 255-266 (2005).

76 Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y. & Aebersold, R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proceedings of the National Academy of Sciences* **97**, 9390-9395 (2000).

77 Wang, W. *et al.* Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical Chemistry* **75**, 4818-4826 (2003).

78 Zybailov, B. *et al.* Statistical Analysis of Membrane Proteome Expression Changes in Saccharomyces c erevisiae. *Journal of proteome research* **5**, 2339-2347 (2006).

79 Pan, C. *et al.* Characterization of anaerobic catabolism of p-coumarate in Rhodopseudomonas palustris by integrating transcriptomics and quantitative proteomics. *Molecular & cellular proteomics : MCP* **7**, 938-948, doi:10.1074/mcp.M700147-MCP200 (2008).

80 Yao, X., Freas, A., Ramirez, J., Demirev, P. A. & Fenselau, C. Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem* **73**, 2836-2842 (2001).

81 Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature biotechnology* **17**, 994-999 (1999).

82 Belnap, C. P. *et al.* Quantitative proteomic analyses of the response of acidophilic microbial communities to different pH conditions. *The ISME journal* **5**, 1152-1161 (2011).

83 Ong, S.-E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics* **1**, 376-386 (2002).

84 Tolonen, A. C. *et al.* Proteome□wide systems analysis of a cellulosic biofuel□producing microbe. *Molecular systems biology* **7** (2011).

85 Ross, P. L. *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics* **3**, 1154-1169 (2004).

86 Thompson, A. *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical chemistry* **75**, 1895-1904 (2003).

87 Collier, T. S. *et al.* Direct comparison of stable isotope labeling by amino acids in cell culture and spectral counting for quantitative proteomics. *Analytical chemistry* **82**, 8696-8702 (2010).

88 Hendrickson, E. L., Xia, Q., Wang, T., Leigh, J. A. & Hackett, M. Comparison of spectral counting and metabolic stable isotope labeling for use with quantitative microbial proteomics. *Analyst* **131**, 1335-1341 (2006).

89 Patel, V. J. *et al.* A comparison of labeling and label-free mass spectrometry-based proteomics approaches. *Journal of proteome research* **8**, 3752-3759 (2009).

90 Thompson, D. K. *et al.* Proteomics reveals a core molecular response of Pseudomonas putida F1 to acute chromate challenge. *BMC genomics* **11**, 311 (2010).

91 Pichler, P. *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. *Analytical chemistry* **82**, 6549-6558 (2010).

92 Abraham, P. *et al.* Defining the boundaries and characterizing the landscape of functional genome expression in vascular tissues of Populus using shotgun proteomics. *Journal of proteome research* **11**, 449-460 (2011).

93 Old, W. M. *et al.* Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Molecular & cellular proteomics* **4**, 1487-1502 (2005).

94 Karp, N. A. *et al.* Addressing accuracy and precision issues in iTRAQ quantitation. *Molecular & Cellular Proteomics* **9**, 1885-1897 (2010).

95 Griffin, T. J. *et al.* iTRAQ reagent-based quantitative proteomic analysis on a linear ion trap mass spectrometer. *Journal of proteome research* **6**, 4200-4209 (2007).

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

97 Carvalho, P. C. *et al.* XDIA: improving on the label-free data-independent analysis. *Bioinformatics* **26**, 847-848 (2010).

98 Chakraborty, A. B., Berger, S. J. & Gebler, J. C. Use of an integrated MS–multiplexed MS/MS data acquisition strategy for high□coverage peptide mapping studies. *Rapid communications in mass spectrometry* **21**, 730-744 (2007).

99 Griffin, N. M. *et al.* Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nature biotechnology* **28**, 83-89 (2010).

100 Gomez-Roldan, V. *et al.* Strigolactone inhibition of shoot branching. *Nature* **455**, 189-194 (2008).

101 Umehara, M. *et al.* Inhibition of shoot branching by new terpenoid plant hormones. *Nature* **455**, 195-200 (2008).

102 Cook, C., Whichard, L. P., Turner, B., Wall, M. E. & Egley, G. H. Germination of witchweed (Striga lutea Lour.): isolation and properties of a potent stimulant. *Science* **154**, 1189-1190 (1966).

103 Akiyama, K., Matsuzaki, K.-i. & Hayashi, H. Plant sesquiterpenes induce hyphal branching in arbuscular mycorrhizal fungi. *Nature* **435**, 824-827 (2005).

104 de Saint Germain, A., Bonhomme, S., Boyer, F.-D. & Rameau, C. Novel insights into strigolactone distribution and signalling. *Current opinion in plant biology* **16**, 583-589 (2013).

105 Beveridge, C. A. & Kyozuka, J. New genes in the strigolactone-related shoot branching pathway. *Current opinion in plant biology* **13**, 34-39 (2010).

106 Liu, W. *et al.* Strigolactone biosynthesis in Medicago truncatula and rice requires the symbiotic GRAS-type transcription factors NSP1 and NSP2. *The Plant Cell Online* **23**, 3853-3865 (2011).

107 Kretzschmar, T. *et al.* A petunia ABC protein controls strigolactone-dependent symbiotic signalling and branching. *Nature* **483**, 341-344 (2012).

108 Stirnberg, P., van De Sande, K. & Leyser, H. O. MAX1 and MAX2 control shoot lateral branching in *Arabidopsis*. *Development* **129**, 1131-1141 (2002).

109 Arite, T. *et al.* d14, a strigolactone-insensitive mutant of rice, shows an accelerated outgrowth of tillers. *Plant and cell physiology* **50**, 1416-1424 (2009).

110 Waters, M. T. *et al.* Specialisation within the DWARF14 protein family confers distinct responses to karrikins and strigolactones in *Arabidopsis*. *Development* **139**, 1285-1295 (2012).

111 Gaiji, N., Cardinale, F., Prandi, C., Bonfante, P. & Ranghino, G. The computational-based structure of Dwarf14 provides evidence for its role as potential strigolactone receptor in plants. *BMC research notes* **5**, 307 (2012).

112 Besserer, A., Bécard, G., Jauneau, A., Roux, C. & Séjalon-Delmas, N. GR24, a synthetic analog of strigolactones, stimulates the mitosis and growth of the arbuscular mycorrhizal fungus Gigaspora rosea by boosting its energy metabolism. *Plant physiology* **148**, 402-413 (2008).

113 Hamiaux, C. *et al.* DAD2 is an α/β hydrolase likely to be involved in the perception of the plant branching hormone, strigolactone. *Current Biology* **22**, 2032-2036 (2012).

114 Braun, N. *et al.* The pea TCP transcription factor PsBRC1 acts downstream of strigolactones to control shoot branching. *Plant physiology* **158**, 225-238 (2012).

115 Jiang, L. *et al.* DWARF 53 acts as a repressor of strigolactone signalling in rice. *Nature* (2013).

116 Durbak, A., Yao, H. & McSteen, P. Hormone signaling in plant development. *Current opinion in plant biology* **15**, 92-96 (2012).

117 Brewer, P. B., Dun, E. A., Ferguson, B. J., Rameau, C. & Beveridge, C. A. Strigolactone acts downstream of auxin to regulate bud outgrowth in pea and *Arabidopsis*. *Plant Physiology* **150**, 482-493 (2009).

118 Foo, E., Yoneyama, K., Hugill, C. J., Quittenden, L. J. & Reid, J. B. Strigolactones and the regulation of pea symbioses in response to nitrate and phosphate deficiency. *Molecular Plant* **6**, 76-87 (2013).

119 Ruyter-Spira, C. *et al.* Physiological effects of the synthetic strigolactone analog GR24 on root system architecture in *Arabidopsis*: another belowground role for strigolactones? *Plant Physiology* **155**, 721-734 (2011).

120 Kapulnik, Y. *et al.* Strigolactones interact with ethylene and auxin in regulating root-hair elongation in *Arabidopsis*. *Journal of experimental botany* **62**, 2915-2924 (2011).

121 Toh, S. *et al.* Thermoinhibition uncovers a role for strigolactones in *Arabidopsis* seed germination. *Plant and Cell Physiology* **53**, 107-117 (2012).

122 Wang, Y. *et al.* Strigolactone/MAX2-Induced Degradation of Brassinosteroid Transcriptional Effector BES1 Regulates Shoot Branching. *Developmental cell* **27**, 681-688 (2013).

123 Mashiguchi, K. *et al.* Feedback-regulation of strigolactone biosynthetic genes and strigolactone-regulated genes in *Arabidopsis*. *Bioscience, biotechnology, and biochemistry* **73**, 2460-2465 (2009).

124 Mayzlish-Gati, E. *et al.* Strigolactones are positive regulators of light-harvesting genes in tomato. *Journal of experimental botany* **61**, 3129-3136 (2010).

125 Livak, K. J. & Schmittgen, T. D. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2< sup>– ΔΔCT</sup> Method. *methods* **25**, 402-408 (2001).

126 Arvidsson, S., Kwasniewski, M., Riaño-Pachón, D. M. & Mueller-Roeber, B. QuantPrime–a flexible tool for reliable high-throughput primer design for quantitative PCR. *BMC bioinformatics* **9**, 465 (2008).

127 Breitling, R., Armengaud, P., Amtmann, A. & Herzyk, P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters* **573**, 83-92 (2004).

128 Deng, Z. *et al.* A proteomics study of brassinosteroid response in *Arabidopsis*. *Molecular & Cellular Proteomics* **6**, 2058-2071 (2007).

129 Zhao, Z., Stanley, B. A., Zhang, W. & Assmann, S. M. ABA-regulated G protein signaling in *Arabidopsis* guard cells: a proteomic perspective. *Journal of proteome research* **9**, 1637-1647 (2010).

130 Jones, A. M., Bennett, M. H., Mansfield, J. W. & Grant, M. Analysis of the defence phosphoproteome of *Arabidopsis* thaliana using differential mass tagging. *Proteomics* **6**, 4155-4165 (2006).

131 Rutschow, H., Ytterberg, A. J., Friso, G., Nilsson, R. & van Wijk, K. J. Quantitative proteomics of a chloroplast SRP54 sorting mutant and its genetic interactions with CLPC1 in *Arabidopsis*. *Plant physiology* **148**, 156-175 (2008).

132 Majeran, W. *et al.* Consequences of C4 differentiation for chloroplast membrane proteomes in maize mesophyll and bundle sheath cells. *Molecular & Cellular Proteomics* **7**, 1609-1638 (2008).

133 Beynon, E. R. *et al.* The role of oxophytodienoate reductases in the detoxification of the explosive 2, 4, 6-trinitrotoluene by *Arabidopsis*. *Plant physiology* **151**, 253-261 (2009).

134 Biesgen, C. & Weiler, E. Structure and regulation of OPR1 and OPR2, two closely related genes encoding 12-oxophytodienoic acid-10, 11-reductases from *Arabidopsis* thaliana. *Planta* **208**, 155-165 (1999).

135 Llamas, M. A., Ramos, J. L. & Rodríguez-Herva, J. J. Mutations in Each of the tol Genes ofPseudomonas putida Reveal that They Are Critical for Maintenance of Outer Membrane Stability. *Journal of bacteriology* **182**, 4764-4772 (2000).

136 Mueller, S. *et al.* General detoxification and stress responses are mediated by oxidized lipids through TGA transcription factors in *Arabidopsis*. *The Plant Cell Online* **20**, 768-785 (2008).

137 Gamas, P., de Carvalho Niebel, F., Lescure, N. & Cullimore, J. V. Use of a subtractive hybridization approach to identify new Medicago truncatula genes induced during root nodule development. *MPMI-Molecular Plant Microbe Interactions* **9**, 233-242 (1996).

138 Foo, E. & Davies, N. W. Strigolactones promote nodulation in pea. *Planta* **234**, 1073-1081 (2011).

139 Mukherjee, A. K., Lev, S., Gepstein, S. & Horwitz, B. A. A compatible interaction of Alternaria brassicicola with *Arabidopsis* thaliana ecotype DiG: evidence for a specific transcriptional signature. *BMC plant biology* **9**, 31 (2009).

140 Hammond, J. P. *et al.* Changes in gene expression in *Arabidopsis* shoots during phosphate starvation and the potential for developing smart plants. *Plant Physiology* **132**, 578-596 (2003).

141 Moons, A. Regulatory and functional interactions of plant growth regulators and plant glutathione S-transferases (GSTs). *Vitamins & Hormones* **72**, 155-202 (2005).

142 Czechowski, T., Stitt, M., Altmann, T., Udvardi, M. K. & Scheible, W.-R. Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. *Plant physiology* **139**, 5-17 (2005).

143 Nemhauser, J. L., Hong, F. & Chory, J. Different plant hormones regulate similar processes through largely nonoverlapping transcriptional responses. *Cell* **126**, 467-475 (2006).

144 Redman, J. C., Haas, B. J., Tanimoto, G. & Town, C. D. Development and evaluation of an *Arabidopsis* whole genome Affymetrix probe array. *The Plant Journal* **38**, 545-561 (2004).

145 Worley, C. K. *et al.* Degradation of Aux/IAA proteins is essential for normal auxin signalling. *The Plant Journal* **21**, 553-562 (2000).

146 Wang, L., Mai, Y.-X., Zhang, Y.-C., Luo, Q. & Yang, H.-Q. MicroRNA171c-targeted SCL6-II, SCL6-III, and SCL6-IV genes regulate shoot branching in *Arabidopsis*. *Molecular Plant* **3**, 794-806 (2010).

147 Woo, H. R. *et al.* ORE9, an F-box protein that regulates leaf senescence in *Arabidopsis*. *The Plant Cell Online* **13**, 1779-1790 (2001).

148 Stanga, J. P., Smith, S. M., Briggs, W. R. & Nelson, D. C. SUPPRESSOR OF MORE AXILLARY GROWTH2 1 controls seed germination and seedling development in *Arabidopsis*. *Plant physiology* **163**, 318-330 (2013).

149 Zogg, G. P. *et al.* Compositional and functional shifts in microbial communities due to soil warming. *Soil Science Society of America Journal* **61**, 475-481 (1997).

150 Luo, C. *et al.* Soil microbial community responses to a decade of warming as revealed by comparative metagenomics. *Applied and environmental microbiology*, AEM. 03712-03713 (2013).

151 Tu, Q. *et al.* GeoChip 4: a functional gene□array□based high□throughput environmental technology for microbial community analysis. *Molecular ecology resources* (2014).

152 Amann, R. I., Ludwig, W. & Schleifer, K.-H. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiological reviews* **59**, 143-169 (1995).

153 Bond, P. L., Druschel, G. K. & Banfield, J. F. Comparison of acid mine drainage microbial communities in physically and geochemically distinct ecosystems. *Applied and Environmental Microbiology* **66**, 4962-4971 (2000).

154 Saeed, A. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374 (2003).

155 Jain, S., Graham, C., Graham, R. L., McMullan, G. & Ternan, N. G. Quantitative proteomic analysis of the heat stress response in Clostridium difficile strain 630. *Journal of proteome research* **10**, 3880-3890 (2011).

156 Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).

157 Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research* **35**, W182-W185 (2007).

158 Park, B. H., Karpinets, T. V., Syed, M. H., Leuze, M. R. & Uberbacher, E. C. CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology* **20**, 1574-1584 (2010).

159 Lo, I. *et al.* Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446**, 537-541 (2007).

160 Simmons, S. L. *et al.* Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS biology* **6** (2008).

161 Denef, V. J. *et al.* Proteomics□inferred genome typing (PIGT) demonstrates inter□population recombination as a strategy for environmental adaptation. *Environmental microbiology* **11**, 313-325 (2009).

162 Denef, V. J. & Banfield, J. F. *In situ* evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science* **336**, 462-466 (2012).

163 Mosier, A. C. *et al.* Metabolites Associated with Adaptation of Microorganisms to an Acidophilic, Metal-Rich Environment Identified by Stable-Isotope-Enabled Metabolomics. *mBio* **4**, e00484-00412 (2013).

164 Goltsman, D. S. A. *et al.* Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing "*Leptospirillum rubarum*"(Group II) and "*Leptospirillum* ferrodiazotrophum"(Group III) bacteria in acid mine drainage biofilms. *Applied and environmental microbiology* **75**, 4599-4615 (2009).

165 Jeans, C. *et al.* Cytochrome 572 is a conspicuous membrane protein with iron oxidation activity purified directly from a natural acidophilic microbial community. *The ISME journal* **2**, 542-550 (2008).

166 Singer, S. W. *et al.* Characterization of cytochrome 579, an unusual cytochrome isolated from an iron-oxidizing microbial community. *Applied and environmental microbiology* **74**, 4454-4462 (2008).

167 Bonnefoy, V. & Holmes, D. S. Genomic insights into microbial iron oxidation and iron uptake strategies in extremely acidic environments. *Environmental microbiology* **14**, 1597-1611 (2012).

168 Singer, S. W. *et al.* Posttranslational modification and sequence variation of redox-active proteins correlate with biofilm life cycle in natural microbial communities. *The ISME journal* **4**, 1398-1409 (2010).

169 Yelton, A. P. *et al.* Comparative genomics in acid mine drainage biofilm communities reveals metabolic and structural differentiation of co-occurring archaea. *BMC genomics* **14**, 485 (2013).

170 Westman, J. O., Taherzadeh, M. J. & Franzén, C. J. Proteomic analysis of the increased stress tolerance of Saccharomyces cerevisiae encapsulated in liquid core alginate-chitosan capsules. *PloS one* **7**, e49335 (2012).

171 Deslippe, J. R., Hartmann, M., Simard, S. W. & Mohn, W. W. Long□term warming alters the composition of Arctic soil microbial communities. *FEMS microbiology ecology* **82**, 303-315 (2012).

172 Lindh, M. V. *et al.* Consequences of increased temperature and acidification on bacterioplankton community composition during a mesocosm spring bloom in the Baltic Sea. *Environmental microbiology reports* **5**, 252-262 (2013).

173 Scheibner, M. *et al.* Impact of warming on phyto□bacterioplankton coupling and bacterial community composition in experimental mesocosms. *Environmental microbiology* (2013).

174 Mueller, R. S. *et al.* Ecological distribution and population physiology defined by proteomics in a natural microbial community. *Molecular systems biology* **6** (2010).

175 Coram, N. J. & Rawlings, D. E. Molecular relationship between two groups of the genus *Leptospirillum* and the finding that *Leptospirillum* ferriphilum sp. nov. dominates South African commercial biooxidation tanks that operate at 40 C. *Applied and Environmental Microbiology* **68**, 838-845 (2002).

176 Zhang, R.-Y. *et al.* A new strain< i> *Leptospirillum* ferriphilum</i> YTW315 for bioleaching of metal sulfides ores. *Transactions of Nonferrous Metals Society of China* **20**, 135-141 (2010).

177 Wohlers, J. *et al.* Changes in biogenic carbon flow in response to sea surface warming. *Proceedings of the National Academy of Sciences* **106**, 7067-7072 (2009).

178 Engel, A. *et al.* Effects of sea surface warming on the production and composition of dissolved organic matter during phytoplankton blooms: results from a mesocosm study. *Journal of Plankton Research* **33**, 357-372 (2011).

179 Jiao, Y. *et al.* Characterization of extracellular polymeric substances from acidophilic microbial biofilms. *Applied and environmental microbiology* **76**, 2916-2922 (2010).

180 Suttle, C. A. Viruses in the sea. *Nature* **437**, 356-361 (2005).

181 Russell, R. J., Ferguson, J. M., Hough, D. W., Danson, M. J. & Taylor, G. L. The crystal structure of citrate synthase from the hyperthermophilic archaeon Pyrococcus furiosus at 1.9 Å resolution. *Biochemistry* **36**, 9983-9994 (1997).

182 Hickey, D. A. & Singer, G. Genomic and proteomic adaptations to growth at high temperature. *Genome biology* **5**, /2004/2005/2010/2117-/2004/2005/2010/2117 (2004).

183 Schneider, W. & Doetsch, R. Temperature effects on bacterial movement. *Applied and environmental microbiology* **34**, 695-700 (1977).

184 Turner, L., Samuel, A. D., Stern, A. S. & Berg, H. C. Temperature Dependence of Switching of the Bacterial Flagellar Motor by the Protein CheY< sup> 13DK106YW</sup>. *Biophysical journal* **77**, 597-603 (1999).

185 Morrison, R. B. & McCAPRA, J. Flagellar changes in Escherichia coli induced by temperature of the environment. (1961).

186 Adler, J. & Templeton, B. The effect of environmental conditions on the motility of Escherichia coli. *Journal of general microbiology* **46**, 175-184 (1967).

187 Li, C., Louise, C., Shi, W. & Adler, J. Adverse conditions which cause lack of flagella in Escherichia coli. *Journal of bacteriology* **175**, 2229-2235 (1993).

188 Wilmes, P. *et al.* Natural acidophilic biofilm communities reflect distinct organismal and functional organization. *The ISME journal* **3**, 266-270 (2009).

189 Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034-1039, doi:10.1126/science.1153213 (2008).

190 Sowell, S. M. *et al.* Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *The ISME journal* **3**, 93-105, doi:10.1038/ismej.2008.83 (2009).

191 Stock, A. M., Robinson, V. L. & Goudreau, P. N. Two-component signal transduction. *Annual review of biochemistry* **69**, 183-215, doi:10.1146/annurev.biochem.69.1.183 (2000).

192 Holt, L. J. *et al.* Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* **325**, 1682-1686, doi:10.1126/science.1172867 (2009).

193 Choudhary, C. *et al.* Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **325**, 834-840, doi:10.1126/science.1175371 (2009).

194 Wang, Q. *et al.* Acetylation of metabolic enzymes coordinates carbon source utilization and metabolic flux. *Science* **327**, 1004-1007, doi:10.1126/science.1179687 (2010).

195 van Noort, V. *et al.* Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium. *Molecular systems biology* **8**, 571, doi:10.1038/msb.2012.4 (2012).

196 Erce, M. A., Pang, C. N., Hart-Smith, G. & Wilkins, M. R. The methylproteome and the intracellular methylation network. *Proteomics* **12**, 564-586, doi:10.1002/pmic.201100397 (2012).

197 Seth, D., Hausladen, A., Wang, Y. J. & Stamler, J. S. Endogenous protein S-Nitrosylation in *E. coli*: regulation by OxyR. *Science* **336**, 470-473, doi:10.1126/science.1215643 (2012).

198 Radi, R. Nitric oxide, oxidants, and protein tyrosine nitration. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4003-4008, doi:10.1073/pnas.0307446101 (2004).

199 De Ceuleneer, M., Van Steendam, K., Dhaenens, M. & Deforce, D. In vivo relevance of citrullinated proteins and the challenges in their detection. *Proteomics* **12**, 752-760, doi:10.1002/pmic.201100478 (2012).

200 Yang, C. *et al.* Comprehensive mass spectrometric mapping of the hydroxylated amino acid residues of the alpha1(V) collagen chain. *The Journal of biological chemistry* **287**, 40598-40610, doi:10.1074/jbc.M112.406850 (2012).

201 Strader, M. B. *et al.* A proteomic and transcriptomic approach reveals new insight into beta-methylthiolation of Escherichia coli ribosomal protein S12. *Molecular & cellular proteomics : MCP* **10**, M110 005199, doi:10.1074/mcp.M110.005199 (2011).

202 Swaney, D. L. *et al.* Global analysis of phosphorylation and ubiquitylation cross-talk in protein degradation. *Nature methods* **10**, 676-682, doi:10.1038/nmeth.2519 (2013).

203 Mertins, P. *et al.* Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nature methods* **10**, 634-637, doi:10.1038/nmeth.2518 (2013).

204 Wu, R. *et al.* A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nature methods* **8**, 677-683, doi:10.1038/nmeth.1636 (2011).

205 Olsen, J. V. *et al.* Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Science signaling* **3**, ra3, doi:10.1126/scisignal.2000475 (2010).

206 Moses, A. M. & Landry, C. R. Moving from transcriptional to phospho-evolution: generalizing regulatory evolution? *Trends in genetics : TIG* **26**, 462-467, doi:10.1016/j.tig.2010.08.002 (2010).

207 Tan, C. S. *et al.* Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Science signaling* **2**, ra39, doi:10.1126/scisignal.2000316 (2009).

208 Beltrao, P. *et al.* Systematic functional prioritization of protein posttranslational modifications. *Cell* **150**, 413-425, doi:10.1016/j.cell.2012.05.036 (2012).

209 Beltrao, P. *et al.* Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS biology* **7**, e1000134, doi:10.1371/journal.pbio.1000134 (2009).

210 Weinert, B. T. *et al.* Proteome-wide mapping of the Drosophila acetylome demonstrates a high degree of conservation of lysine acetylation. *Science signaling* **4**, ra48, doi:10.1126/scisignal.2001902 (2011).

211 Boekhorst, J., van Breukelen, B., Heck, A., Jr. & Snel, B. Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome biology* **9**, R144, doi:10.1186/gb-2008-9-10-r144 (2008).

212 Denef, V. J. *et al.* Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 2383-2390, doi:10.1073/pnas.0907041107 (2010).

213 Denef, V. J. & Banfield, J. F. *In situ* evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science* **336**, 462-466, doi:10.1126/science.1218389 (2012).

214 Savitski, M. M. *et al.* Confident phosphorylation site localization using the Mascot Delta Score. *Molecular & Cellular Proteomics* **10**, M110. 003830 (2011).

215 Lo, I. *et al.* Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446**, 537-541 (2007).

216 Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG* **16**, 276-277 (2000).

217 Breitling, R., Armengaud, P., Amtmann, A. & Herzyk, P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters* **573**, 83-92, doi:10.1016/j.febslet.2004.07.055 (2004).

218 Zhang, J. *et al.* MUFOLD: A new solution for protein 3D structure prediction. *Proteins* **78**, 1137-1152, doi:10.1002/prot.22634 (2010).

219 Wang, Y., Ahn, T. H., Li, Z. & Pan, C. Sipros/ProRata: a versatile informatics system for quantitative community proteomics. *Bioinformatics* **29**, 2064-2065, doi:10.1093/bioinformatics/btt329 (2013).

220 Pan, C. *et al.* Quantitative tracking of isotope flows in proteomes of microbial communities. *Molecular & cellular proteomics : MCP* **10**, M110 006049, doi:10.1074/mcp.M110.006049 (2011).

221 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* **4**, 207-214, doi:10.1038/nmeth1019 (2007).

222 Marx, H. *et al.* A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nature biotechnology* **31**, 557-564, doi:10.1038/nbt.2585 (2013).

223 Soares, N. C., Spat, P., Krug, K. & Macek, B. Global dynamics of the *Escherichia coli* proteome and phosphoproteome during growth in minimal medium. *Journal of proteome research* **12**, 2611-2621, doi:10.1021/pr3011843 (2013).

224 Zhang, K., Zheng, S., Yang, J. S., Chen, Y. & Cheng, Z. Comprehensive profiling of protein lysine acetylation in Escherichia coli. *Journal of proteome research* **12**, 844-851, doi:10.1021/pr300912q (2013).

225 Mueller, R. S. *et al.* Proteome changes in the initial bacterial colonist during ecological succession in an acid mine drainage biofilm community. *Environmental microbiology* **13**, 2279-2292, doi:10.1111/j.1462-2920.2011.02486.x (2011).

226 Goltsman, D. S. *et al.* Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing "*Leptospirillum* rubarum" (Group II) and "*Leptospirillum* ferrodiazotrophum" (Group III) bacteria in acid mine drainage biofilms. *Applied and environmental microbiology* **75**, 4599-4615, doi:10.1128/AEM.02943-08 (2009).

227 Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712, doi:10.1126/science.1138140 (2007).

228 Andersson, A. F. & Banfield, J. F. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320**, 1047-1050, doi:10.1126/science.1157358 (2008).

229 Alexander, R. P. & Zhulin, I. B. Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors. *Proceedings of the National Academy of Sciences* **104**, 2885-2890 (2007).

230 Wadhams, G. H. & Armitage, J. P. Making sense of it all: bacterial chemotaxis. *Nature reviews. Molecular cell biology* **5**, 1024-1037, doi:10.1038/nrm1524 (2004).

231 Nobeli, I., Favia, A. D. & Thornton, J. M. Protein promiscuity and its implications for biotechnology. *Nature biotechnology* **27**, 157-167 (2009).

232 Tirosh, I., Reikhav, S., Levy, A. A. & Barkai, N. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* **324**, 659-662, doi:10.1126/science.1169766 (2009).

233 Olsen, J. V. *et al.* Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635-648 (2006).

234 Villén, J. & Gygi, S. P. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nature protocols* **3**, 1630-1638 (2008).

235 Boersema, P. J., Mohammed, S. & Heck, A. J. Phosphopeptide fragmentation and analysis by mass spectrometry. *Journal of mass spectrometry* **44**, 861-878 (2009).

236 Schroeder, M. J., Shabanowitz, J., Schwartz, J. C., Hunt, D. F. & Coon, J. J. A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Analytical chemistry* **76**, 3590-3598 (2004).

237 Olsen, J. V. & Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 13417-13422 (2004).

238 Villén, J., Beausoleil, S. A. & Gygi, S. P. Evaluation of the utility of neutral-loss-dependent MS3 strategies in large-scale phosphorylation analysis. *Proteomics* **8**, 4444-4452 (2008).

239 Michalski, A., Neuhauser, N., Cox, J. r. & Mann, M. A systematic investigation into the nature of tryptic HCD spectra. *Journal of proteome research* **11**, 5479-5491 (2012).

240 Seraphin, B. & Hettich, R. Microbial proteomics: the quiet revolution. *Current opinion in microbiology* **15**, 348-350, doi:10.1016/j.mib.2012.05.004 (2012).

241 Pirmoradian, M. *et al.* Rapid and deep human proteome analysis by single-dimension shotgun proteomics. *Molecular & cellular proteomics : MCP* **12**, 3330-3338, doi:10.1074/mcp.O113.028787 (2013).

242 Hebert, A. S. *et al.* Neutron-encoded mass signatures for multiplexed proteome quantification. *Nature methods* **10**, 332-334, doi:10.1038/nmeth.2378 (2013).

243 Hebert, A. S. *et al.* Amine-reactive neutron-encoded labels for highly plexed proteomic quantitation. *Molecular & cellular proteomics : MCP* **12**, 3360-3369, doi:10.1074/mcp.M113.032011 (2013).

244 Yang, X. J. Multisite protein modification and intramolecular signaling. *Oncogene* **24**, 1653-1662, doi:10.1038/sj.onc.1208173 (2005).

245 Hettich, R. L., Pan, C., Chourey, K. & Giannone, R. J. Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. *Anal Chem* **85**, 4203-4214, doi:10.1021/ac303053e (2013).

246 Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics* **11**, O111. 016717 (2012).

VITA

Zhou Li was born in Changsha, Hunan, China on October 22nd, 1984. He graduated from the Changsha 15th High School in 2003. He earned his bachelor of engineering degree in Plant and Animal Quarantine from Hunan Agricultural University in 2007. He enrolled in the graduate school of Central South University where he studied microbiology from 2007 to 2009. He enrolled in the graduate school of Genome Science and Technology at the University of Tennessee in 2009 and expects to receive his Ph.D. in Life Sciences in August of 2014.