Development of MS-based proteomics approaches to examine metabolic pathways and protein:protein interactions in microbial systems

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

Chen Qian
August 2017

ACKNOWLEDGEMENTS

Standing at this ending point of my Ph.D. journey, I would like to express my most sincere gratitude to a lot of people. Without their support and guidance, I would not have been able to complete this dissertation.

Firstly, I would like to express my special appreciation and thanks to my advisor Prof. Robert Hettich for his continuous support, trust and guidance for my research work. He is a tremendous mentor for me. During the five years of my Ph.D. study, he did not only offer me a lot of chances to learn and grow, but also inspired me to become a real scientist. I learned the most precious scientific attitudes from him, always be passionate, persistent, and honest.

I also would like to thank my dissertation committee members: Dr. Steven Brown, Dr. Steven Wilhelm, Dr. Brynn Voy and Dr. Albrecht Von Arnim, for their encouragement and insightful comments for all of my dissertation work. Their expertise in microbiology, statistics, experimental designs, etc. helped me to improve my dissertation work to a more comprehensive and solid level.

My two major projects have given me the chance to collaborate with Dr. Jeffrey Becker and Dr. Baohua Gu. I would like to thank them for trusting me so much on my work and thank their group members for offering me a lot of advice on the projects. In particular, I would like to thank Dr. Melinda Hauser, Dr. Hongmei Chen and Dr. Xia Lu for their significant contributions on the projects. It was a very fantastic experience to work with them.

I also would like to thank all of my lab colleagues for making this lab so much like a big family. Dr. Richard Giannone and Dr. Paul Abraham are like two big brothers in this family, who are always there whenever we need help. In particular, I would like to thank Dr. Weili Xiong for all the discussion and guidance devoted to my dissertation work. I am really grateful to have her as my friend as well as my

colleague here. I would like to thank Genome Science and Technology program for providing me this valuable opportunity to build my early career here.

Last but not least, I would like to thank my family and my friends. I am always feeling so blessed to have the world's best parents, Liang Qian and Yueying Shen, who are always so supportive for all of decisions. I also thank my beloved husband, Mengqi Zhang, for standing with me all the time, offering me substantial love and support for all the critical time points, no matter in life or in work.

ABSTRACT

The proteome is perhaps the most functional operating machinery for almost all biological processes, serving as the bridge to link the genome and phenotypes. The proteome undergoes dynamic changes in terms of the abundance or interactions, responding to the environmental stimuli. Understanding this dynamic of protein alterations is the key to delineate critical biological mechanisms. Mass-spectrometry-based proteomics is a powerful tool to systematically monitor the heterogeneous alterations of the proteome, including the changes of abundance, modifications and interactions. In this dissertation, a research project was built upon current proteomics approaches to solve the issues regarding to the sample preparation and data analysis that would help propel this approach to better address certain questions in environmental microbiology and molecular biology. In the first study, we have designed an experimental method that efficiently removes humic acids prior to proteolytic peptide measurement, thus addressing one major challenge associated with soil microbiome proteome extraction and subsequent MS measurement. The second study was aimed at employing advanced proteomics approaches to better understand the microbial drivers of environmental mercury processes by using two model organisms: *Geobacter sulfurreducens* PCA and *Desulfovibrio desulfuricans* ND132. Our results elucidated the global proteome impacts caused by the deletion of mercury methylation essential genes *hgcAB* and revealed that deletion of *hgcAB* genes did not show significant impact on the microbial response to mercury addition. The third study focused on optimizing MS-based proteome approaches for characterizing protein-protein interactions. By coupling the rapid crosslinking procedure, affinity enrichments and high-performance MS measurements, protein-protein interaction can be captured and interrogated in a living cell by a time-resolved manner. Overall, the methods and results presented in this dissertation not only provides an enhanced sample preparation methodology for intractable samples (such as soils), but also demonstrates how this systems-biology approach can be utilized to characterize the basics of microbial physiology at a global proteome level as well as drilling down into specific proteins interactions.

The optimized methods and experimental/bioinformatics techniques described in this dissertation should be broadly extendable to proteome characterization/protein interaction examination in various systems.

TABLE OF CONTENTS

CHAPTER 1 Principles of mass-spectrometry-based proteomics and its applications for environmental microbiology and molecular biology 1

1.1 "omics" science: Understanding the cellular function in system-interrogation level 1

1.1.1 Understanding cellular function from DNA to protein 1

1.1.2 System level understanding of the celluar function from genome to proteome 2

1.2 Mass-spectrometry based proteomics 4

1.2.1 Origins of mass spectrometry-based proteomics 4

1.2.2 Different strategies for mass-spectrometry based proteomics measurement 7

1.2.3 Application of label-free shotgun proteomics in this dissertation 10

1.3 Applications of mass-spectrometry based proteomics on environmental microbial investigations 13

1.3.1 Environmental proteomics of single organisms and microbial communities 13

1.3.2 Metabolic characterization of mercury methylating bacteria 18

1.4 Affinity proteomics for characterizing protein interactions 21

1.5 Overview of this dissertation 25

CHAPTER 2 Experimental and informatics methodologies for MS-based microbial proteome investigations 26

2.1 General workflow of shotgun proteomics 26

2.2 Sample preparation 28

2.2.1 Cell lysis 28

2.2.2 Protein extraction and precipitation 30

2.2.3 Protein digestion and peptide clean up 30

2.3 Peptide separation by liquid chromatography 33

2.4 Mass spectrometry instrumentation 35

2.4.1 Ionization Methods 37

2.4.2 Mass Analyzers and Detectors 39

2.4.3 Tandem mass spectrometry 50

2.4.4 Data acquisition in proteome measurement 53

2.5 Bioinformatics for proteome interrogation 54

2.5.1 Database searching algorithms for peptide assignment 55

2.5.2 Protein inference 57

2.5.3 Protein quantification 58
2.5.4 Statistical methods for multiple comparisons in proteomics data analysis 61
CHAPTER 3 Development of An Optimized Extraction Method To Remove Humic Acid Interferences from Soil Samples Prior to Microbial Proteome Measurement 67
3.1 Introduction 67
3.2 Materials and Methods 70
3.3 Small amounts of humic acids in proteome samples adversely affect the peptide measurements 76
3.4 Humic acids can be efficiently removed by a combination of pH adjustment and filtering 83
3.5 Humic acids removal does not impact the peptide identification or abundance profiles 85
3.6 Conclusions 95
CHAPTER 4 Global Proteome Response to Deletion of Genes Related to Mercury Methylation and Dissimilatory Metal Reduction Reveals Changes in Respiratory Metabolism in *Geobacter sulfurreducens* PCA 96
4.1 Introduction 96
4.2 Materials and methods 98
4.3 Overall comparison of proteomes 102
4.3.1 Differential protein abundances among the three strains 102
4.3.2 Functional classification of differentially abundant proteins 106
4.3.3 Impact on membrane protein abundance 109
4.4 Biological processes change by deletion of *hgcAB* genes 111
4.4.1 Increased abundance of PilA-C and c-type cytochromes after deletion of *hgcAB* 111
4.4.2 Increased abundance of proteins annotated as methyltransferases after deletion of *omcBESTZ* 114
4.4.3 Impact of gene deletions on carbon metabolism in *G. sulfurreducens* PCA 114
4.5 Conclusions 117
CHAPTER 5 Quantitative Proteome Analysis of *Desulfovibrio desulfuricans* ND132 reveals an additional function of *hgcAB* genes 120
5.1 Introduction 120
5.2 Materials and methods 122
5.3 Global proteome comparison between wild type strain and Δ*hgcAB* strain 125
5.4 The expression level of HgcA and HgcB proteins are directly associated with mercury methylation 127
5.5 Functional Differences between wild type and mutant strains 129

5.6 Deletion of *hgcAB* genes significantly impacts metal transporters and electron transporter chain ……132
5.6.1 Deletion of *hgcAB* genes significantly impacts metal transporters 132
5.6.2 Deletion of *hgcAB* genes impacts extracellular solute-binding protein family 135
5.6.3 Deletion of *hgcAB* genes significantly impacts electron transport chain 137
5.7 *hgcAB* genes were involved in acetate formation pathways 138
5.8 Differential Protein Abundance during Hg Exposure 145
5.9 Conclusions 150
CHAPTER 6 Development of an MS approach for characterizing intracellular proteins interacting with a G protein-coupled receptor captured using laser-induced photo-activation 151
6.1 Introduction 151
6.2 Materials and methods 155
6.3 Optimization of laser activated cross-linking of [bioBpa]α-factor to BSA with respect to irradiation time and energy 161
6.4 Identification of Cross-Linked Products by HPLC-MS/MS 165
6.5 MS reveals that GnRH binds to BSA at the same region as [bioBpa]α-factor 168
6.6 Computational DOCKing analysis of BSA to ligands 170
6.7 Initial results on *in vivo* Ste2p crosslinking 173
6.8 Conclusions 176
CHAPTER 7 Conclusions and Future Perspectives 177
7.1 Overview 177
7.2 Conclusions and future perspectives on environmental proteomics 178
7.3 Conclusions and future perspectives on affinity proteomics 183
7.4 Overall concluding perspectives 185
REFERENCES 186
VITA 209

LIST OF TABLES

Table 2.1. Summary of the commonly used protein proteases……...32

Table 2.2. Types of mass analyzers and principles of separation……....41

Table 3.1. Overview of the mass spectrometry measurement results of standard peptide mixture & standard peptide mixture with 25 μg of humic acids added to each sample ……...……79

Table 3.2. Overview of the mass spectrometry measurement results of each sample in simple sample type in which only a microbial isolate was used, and in the moderately more complex sample type that a microbial mixture was employed…….87

Table 3.3. Overview of the mass spectrometry measurement results of each sample in lower biomass samples…….92

Table 4.1. Gene ontology (GO) enrichment analysis of the biological processes (BP) and molecular functions (MF) that are over-represented or under-represented in Δo*mcBESTZ* strain compared to WT strain……...…108

Table 4.2. Homologs of proteins with known function in the acetyl-CoA and related C1 metabolic pathways identified in *Geobacter sulfurreducens* PCA…….……116

Table 5.1. Summary of identification of electron carriers and the primary enzyme (dehydrogenase) that are involved in electron transport chain……..…139

Table 5.2. Detailed information of proteins that are involved in Figure 5.7……...…..146

Table 5.3. Mass spectrometry measurements of differentially abundant proteins between wild type strain and wild type strain exposed to mercury……...…149

Table 6.1. Peptides of BSA that are cross linked to α-factor as identified by StavroX…….……169

Table 6.2. MS measurement results of the Ste2p interacting complex which is captured by the assistance of crosslinking and further enriched by tandem affinity enrichments. ……..……..…….……175

LIST OF FIGURES

Figure2.1. The schematic illustration of the workflow typically employed for the bottom-up proteomics analysis……...……..27

Figure 2.2. The working scheme of the MudPIT experimental setup…….…….36

Figure 2.3. Schematic representation of the electrospray ionization process……..40

Figure 2.4. Schematic of the Orbitrap Elite Hybrid Mass spectrometer…….…….47

Figure 2.5. Ion fragmentation generated under different activation methods……..……52

Figure 3.1. A schematic illustration of the workflow to analyze the soil metaproteome by incorporating the newly developed humic acids removal approach……...……74

Figure 3.2. Depiction of how Sep-pack approach impacts protein/peptide identification……...77

Figure 3.3. The impact of humic acids on standard peptide identification was investigated as a function of their charge and hydrophobicity……...…….……..80

Figure 3.4. Humic acids can be efficiently removed by combination of pH adjustment and filtering by a molecular weight spin column……...…….…..……84

Figure 3.5. Protein identification and correlation of abundance profile among Control sample (microbial inoculant *Pseudomonas putida* F1), humic acids sample (microbial inoculant *Pseudomonas putida* F1 added with humic acids) and soil sample (natural topsoil spiked with microbial inoculant *Pseudomonas putida* F1)…… ……...…….……88

Figure 3.6. Numbers of proteins assigned to Clusters of Orthologous Groups (COG) among control runs, humic acids runs, and soil runs containing simple microbial isolate……..…….……89

Figure 3.7. Protein identification and correlation of abundance profile and COG analysis among control sample (four microbial mixture), humic acids sample (four microbial mixture added with humic acids) and soil sample (natural topsoil spiked with four microbial mixture)…….……90

Figure 3.8. For lower biomass samples, Pearson correlation coefficients were calculated between control samples and humic acids samples……...……..…….……..93

Figure 3.9. Distribution of ScanRanker scores of collected spectra from two soil samples which have been spiked with different biomass…….…...…….……94

Figure 4.1. (a) Venn diagram of the numbers of identified proteins and their overlap among different strains of *Geobacter sulfurreducens* PCA. (b) Pearson correlation coefficients were calculated for each two samples under R environment……..…….…..……103

Figure 4.2. Principal component analysis (PCA) for biological replicates of wild-type and mutant *Geobacter sulfurreducens* PCA cultures……...…..…….…….….……104

Figure 4.3. Spectral counts of all proteins were normalized by z-score and hierarchical clustering of differently expressed proteins was performed in Perseus……...…….…….105

Figure 4.4. Subcellular localization of differentially abundant proteins (DAP) between wild-type (WT) and the two mutant strains …….……..…….107

Figure 4.5. Top 10 significantly changed proteins in Δ*hgcAB* and Δ*omcBESTZ* compared to wild type (WT)……...…110

Figure 4.6. Outer membrane cytochrome analysis among the three strains……...…113

Figure 4.7. Methyltransferases analysis among the three strains……..…….115

Figure 4.8. The proposed function of HgcA and HgcB and linkages to the folate branch of the reductive acetyl-CoA (Wood-Ljungdahl) pathway……...…….…...…118

Figure 5.1. (a) Pearson correlation coefficients were calculated for each two samples in Perseus; (b) PCA plot shows that mutant strain and wild type strain can be well separated……..……...…126

Figure 5.2: Pearson correlation coefficients were calculated for three different time points in log phase of the growth of *Desulfovibrio desulfuricans* ND132 wild type……128

Figure 5.3. (a) Protein rank of all the proteins based on their LFQ intensity. (b) LFQ intensity of the HgcA and HgcB proteins in different time points of log phase. (c) Mercury methylation rate was measured in the same three different times points of log phase. ……..…130

Figure.5.4.(a) GO enrichment on differentially abundant proteins; (b) 2D annotation distribution…..… 131

Figure 5.5. Comparison of the predicted metal transporters…….…….136

Figure 5.6. Comparison of all the electron carriers and dehydrogenase between wild type strain and mutant strain……...……140

Figure 5.7. Pathway and metabolites analysis. (a) Overview of the central carbon metabolism; (b) Growth curves of the wild type and *hgcAB* mutant strains and analyses of the production or consumption of added electron donors and acceptors (e.g., pyruvate, acetate, fumarate, succinate); (c) Overview of the reductive acetyl-CoA pathway……...…141

Figure 6.1. Peptide structures for: (a) α-factor (b) [bioBpa]α-factor (c) GnRH……...…..…153

Figure 6.2. BSA – [bioBpa]α-factor cross-linking as a function of time and laser output…….…...…163

Figure 6.3. Cross-linking as a function of time and energy delivered……...…….……...….164

Figure 6.4. Various laser conditions (power density, irradiation times) were evaluated, both with controls (single BSA) and mixtures ([bioBpa]α-factor bound to BSA)……..……..…...…166

Figure 6.5. Identification of cross-linked products between biotinylated GnRH and BSA, irradiated with 30% power at 50s……...…….…...…171

Figure 6.6. DOCK-predicted binding mode of α-factor to BSA……..…….…….……172

Figure 6.7. The schematic illustration of the workflow employed for Ste2p interaction analysis…...….174

CHAPTER 1

Principles of mass-spectrometry-based proteomics and its applications for environmental microbiology and molecular biology

1.1 "omics" science: Understanding cellular function at the system level

1.1.1 Understanding cellular function from DNA to protein

Cells are the structural and functional unit of all the living organisms, and thus understanding the molecular basis of how cells function is a fundamental goal of molecular biology. The most remarkable property of living cells is their precise genetic machinery to reproduce themselves with incredible fidelity. This accurate reproduction is controlled by the genetic material, deoxyribonucleic acid, DNA.[1] In the twentieth century, the three-dimensional structure of DNA was elucidated,[2] leading to more clear understanding of this near-perfect replication mechanism. In general, the DNA is composed of a linear polymer of four different monomeric subunits, which are deoxyribonucleotides that are arranged in a preordered linear sequence. The information in DNA is encoded in its linear sequence of deoxyribonucleotide subunits. Two of the polymeric strands which have complementary deoxyribonucleotides are twisted together to form the DNA double helix.[3-4] The genetic information carried by the linear sequence of deoxyribonucleotide subunits is further transferred (through an intermediary, RNA) to the production of a protein with a corresponding linear sequence of amino acids. This flow of genetic information has been known for very long time as the central dogma of biology.[5-6] Once formed, the generated protein folds into a particular three-dimensional structure which is determined by its amino acids sequence, and is usually stabilized by noncovalent interactions. This three-dimensional structure is called the native conformation of the protein and is essential for its proper functions. Regardless of the origin of the proteins from the simplest to the most complicated form of life, they are composed of the same ubiquitous approximately 20 amino acids. Each amino acid has a side chain of distinctive chemical properties, and they are covalently linked together to form proteins with a

wide range of sizes. The different combinations and sequences of various amino acids further result in different properties and activities of the proteins. Based on these building blocks, different organisms make widely diverse protein products, such as enzymes, transporters, hormones, antibodies, etc. Single proteins can further interact with other macromolecules (nucleic acids, lipids, proteins) to form supramolecular complexes, such as chromosomes, ribosomes, and membranes. These self-assembled supramolecular complexes constitute the basic scaffold of cells. Therefore, unlike the DNA which is generally static unless mutation occurs, proteins are dynamic and are the final products that carry the genetic information from DNA, mediating virtually all the processes occurred in the cell. Protein analysis in terms of primary sequence, post-translational modifications (PTMs), protein structure and protein interactions, has been a major research topic in biochemistry for very long time, in order to understand life at the molecular level.[7]

1.1.2 System level understanding of the cellular function from genome to proteome

Over the last three decades, a focus in the biomolecular sciences has been experiencing a shift from the desire to understand how a gene function to how all genes / gene products of a cell function together.[8] "Omics" technologies, including genomics, transcriptomics, proteomics and metabolomics, which offer the capability to measure genes, mRNA, proteins and metabolites on a large scale, are employed to interrogate the cell functions in a system level. DNA sequencing was first carried out by Sanger sequencing in the late 1970's and has rapidly developed since then,[9-10] culminating in 1995 with the completion of first whole genome sequencing and assembly of the bacterium *Haemophilus influenza* Rd.[11] The success of the first whole-genome sequencing further led to the completion of the draft of the first human genome in 2001.[12] With the constantly refined technologies, genomes of different organisms have continued to be sequenced and annotated at an exponentially growing pace.[13] The tremendous amount of information contained in the sequenced genome has enabled scientists to explore the physiology, evolution and genotypes of different organisms (the field of genomics). Subsequently, the availability of a large number of sequenced genomes further leads to the spawning of functional genomics

research, which aims at revealing the relationship between genotype and phenotype. Functional genomics studies represent a new phase of genome analysis and refer to the development and application of global experimental approaches to assess the range of gene functions possible from the static DNA sequence information. More specifically, functional genomics focuses on studying gene transcription, translation and protein-protein interactions on a global scale (genome-wide or system-wide), to expand the scope of biological investigations from studying single gene/protein to study all the genes/proteins at once in the system level.[14]

Transcriptome studies serve as the natural starting point to survey gene expression, because only a selected set of genes are transcribed into mRNA molecular, and these expressed genes carry out specific functions. Unlike the genome, the transcriptome not only informs about what genes are present but also what genes are currently being transcribed and translated into proteins in a quantitative way. Microarray measurements or next generation sequencing of transcripts (RNA-Seq) are state of the art techniques employed for transcriptome studies. Despite the success of microarray technique for elucidating and interrogating the transcript patterns within cells, the advent of next-generation sequencing method provides more possibilities. Compared with microarray approach, next-generation sequencing method presents clear advantages in terms of better dynamic range, low technical variance, more complete genome coverage and more accurate quantification.[15-17] Although transcriptomics analyses provide valuable gene expression information, they are not able to deliver information about the protein concentration, activity, modifications and interactions, which more directly connect to the phenotype. In this regard, proteomics, which is the composite set of protein expression profile, represents better for the functional actives at the cellular level.[18-19]

The proteome, which is the complete suite of proteins present, carry out their functions at specific locations and times in living cells. The dynamic range of the proteins in each cell can be very large. For example, a proliferating *Schizosaccharomyces pombe* cell is known to have about 60 million protein molecules, with their abundances ranging from a few copies to around 1.1 million copies per expressed

gene.[20] Across various species, proteins are found to constitute around 50% of the dry mass of a single cell, and can reach a remarkable total concentration of 100–300 mg/ml. The extensive proteome network of the cells adapts quickly and dynamically to internal (genetic) or external perturbations (such as environmental changes) and thus determines the cell's functions and phenotypes. Therefore, understanding the whole proteome qualitatively (which proteins are present and when) and quantitatively (protein expression level), along with its structure, function and interaction is regarded as a central and fundamental question of biology.[21]

Metabolomics studies aim at profiling all the metabolites present in a specific biological sample. Metabolites are referring to the small molecules that are the substrates or products of metabolism which drive various cellular functions, such as cell growth, signal transduction and energy production. Mass spectrometry-based metabolomics is the main methodology for metabolite identification, either in targeted manner or global manner. The identified metabolites can be used to establish the linkage between their physiological roles and involvement in metabolic networks, and also to determine how metabolites change in response to different phenotypic outcomes.[22] Among the four "omics" technologies described above, metabolomics is the only one that does not rely on the genome information. Although spectrometry-based metabolomics studies are still associated with significant amount of challenges, they have started to play important roles in different fields, such as cancer research, drug discovery, early disease diagnostics and food analysis. [23-25]

1.2 Mass-spectrometry based proteomics

1.2.1 Origins of mass spectrometry-based proteomics

Two strategies are commonly used for studying the proteome and molecular mechanisms that it mediates. Conventionally, single proteins are isolated and further analyzed with respect to their function and structure by established biochemical or biophysical methods. More recently, it also has become feasible to interrogate the whole proteome in a large scale and systematic level, with the assistance of

computational bioinformatics analyses. Both strategies have been transformed fundamentally with the development of powerful mass spectrometry technologies, based primarily on its capability to identify and quantify almost any protein that is expressed. Additionally, mass spectrometry is also utilized for protein interaction and protein structure determination.[21]

In the twentieth century, mass spectrometry was already the cornerstone of analytical chemistry; however, it was limited for extensive protein analyses due to its inability to ionize and vaporize large and labile molecules. This situation was dramatically changed by the introduction of two "soft" ionization techniques: matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI), which are able to ionize large biomolecules without introducing further fragmentations. The Nobel Prize in Chemistry 2002 was awarded equally to John B. Fenn and Koichi Tanaka for "their development of soft desorption ionization methods for mass spectrometric analyses of biological macromolecules", which refers to the ESI and MALDI techniques.[26] These emerging technologies for soft ionization further stimulated the development of mass spectrometry instrumentation to enhance two types of desired information regarding to proteins: accurate molecule mass measurements and amino acid sequence information by tandem mass spectrometer (MS/MS).

In 1992, the first large scale analysis of the MHC class I-bound peptides was carried out in Donald Hunt's group, who demonstrated the capability of using the online liquid chromatography-tandem mass spectrometry (LC–MS/MS) to sequence femtomole amounts of peptides from real immune systems.[27] Compared with conventional chemical peptide-sequencing methods such as Edman degradation,[28] which was widely used for sequencing at that time, LC-MS presented much more promise in terms of speed, sensitivity, and throughput. Nineteen peptides were partially sequenced in this study, but there were strong hints at the existence of hundreds more. From that time forward, this electrospray nano-LC–MS/MS platform attracted significant attention due to its unparalleled potential to identify and sequence proteins/peptides in a systematic level; however, the proper data interpretation method to read out all the information hidden behind the complicated spectra was still lacking. To address this issue, an

efficient database searching algorithm was introduced shortly afterwards by the John Yates group in 1994,[29] which further propelled this electrospray nano-LC–MS/MS platform. This technique still remains at the heart of MS-based proteomics after so many years of development, although every aspect of the analytical performance of LC-MS has been improved dramatically and is continuously improving.[30]

Once the online liquid chromatography (LC)–MS platform was established for protein identification from biological samples, another big challenge was the protein separation, since almost every biological system contains a complex protein mixture. In the early stage of the mass spectrometry based proteomics research, whole proteome separations were routinely carried out by gel electrophoresis. Andrej Shevchenko from the Matthias Mann group was the first to develop an "in-gel digestion protocol", which could efficiently extract the clean-enough trypsinized peptides from the background polyacrylamide matrix, and was compatible with subsequent mass spectrometry analysis.[31-32] This gel-based separation method is not routinely used in the global proteome measurement nowadays, due to known drawbacks, such as time consuming, limited dynamic range, etc. Another gel-free separation method was introduced later by John Yates, termed Multidimensional protein identification technology (MudPIT).[33] Instead of separating proteins in the 2-dimensional gel, this technology separates proteins by 2-dimensional liquid chromatography which can be directly coupled with MS. Nowadays, LC-MS is the preferred platform for mass spectrometry based proteomics analysis, and offers an unprecedented insights for comprehensive characterization of proteome in terms of composition, structure, function and interaction, and further sheds light on the complex biological processes and phenotypes. At the same time, tremendous advancements of mass spectrometers, such as the technical breakthrough invention of the Orbitrap mass analyzer, have significantly improved all the merits of MS instruments, further leading the field towards more advanced and comprehensive proteomics measurements.

1.2.2 Different strategies for mass-spectrometry based proteomics measurement

Investigating living cells at the protein level continuously provide valuable insight to biological functions and regulations across the three domains of life. Mass-spectrometry based proteomics has fundamentally impacted the way to interrogate biological and medical questions. Since it can measure thousands of proteins or post translational modifications (PTM) in parallel, the proteomics technique enables the investigation of proteins at all levels of biological complexities, ranging from the protein complexes to whole proteome of single cell or community.

Top-down proteomics and bottom-up proteomics are two fundamental and complementary strategies for proteome identification and characterization.[34] The top-down approach can directly analyze the intact proteins, encompassing all forms of genetic variation, alternative splicing of RNA transcripts, and post-translational modifications (PTMs), thus preserving the labile structural information of proteins.[35] Modifications at different locations can also be simultaneously determined in one spectrum and any correlation between those modifications can be observed. Theoretically, top-down is the most straightforward approach to study the proteome; however, it still faces many technical challenges for the application of whole proteome analysis, including the difficulty in efficient solubilization and separation of protein mixtures prior to the MS analysis and in measuring the protein molecules with high molecular mass.[36]

In contrast, the bottom-up proteomics strategy, (also termed as shotgun proteomics), which measures the proteolytic peptides digested from proteins, is more favorable for large-scale proteomics investigation. In this method, proteins are first digested into peptides by a specific protease (eg. trypsin, chymotrypsin, pepsin, etc). The peptide mixtures then are loaded onto a high-performance liquid chromatography system (HPLC) for separation and further subjected to MS for measurement. The spectra generated from all detectable peptides in a sample are assigned with peptide sequences by database searching algorithm. Finally, the identified peptides are computationally assembled back to their

respective proteins by bioinformatics tools. From an analytical perspective, this method conquers most of the drawbacks in the top-down proteomics strategy, as the peptides are much easier to work with than intact proteins since they can be well solubilized and separated prior to MS, and they are easily disassociated to produce useful fragments that are required for peptide assignment. Additionally, the sensitivity, detection limits, and throughput of the mass spectrometer are better for peptides than for proteins.[37-38] However, due to the peptide-centric nature of bottom-up proteomics approach, some drawbacks are still inevitable: 1) since only a fraction of total peptides can be recovered, information regarding to the alternative splice variants and modifications in specific locations might be lost; 2) for higher order eukaryotes or microbial communities, sequence homologies among proteins are very common. Proteolytic digestion of these homologous proteins results in many shared peptides, confounding the protein inference from these peptides.

While proteomic analyses can be used to provide the qualitative identification information for thousands of proteins that are present in cells or any other biological samples, quantitative information of these proteins is also desirable. In proteomics studies, sensitivity is critical for identifying and quantifying low abundant proteins in the sample, thus impacting the availability of proteins for quantification. Strategies to improve the sensitivity and scope of proteomics analysis normally require large quantities of sample materials and multi-dimensional fractionations, which normally sacrifice the throughput. Alternatively, in order to improve the sensitivity and throughput of protein quantification, the number of features that can be monitored is limited. In this regard, proteomics research is divided into two categories: discovery and targeted proteomics. Discovery proteomics strategies optimize the protein identification by spending more time for each sample and thus reduce the number of samples to be analyzed. The primary goal is to identify as many proteins as possible across a broad dynamic range. In order to identify the low abundant proteins, it sometimes requires the depletion of very abundant proteins, enrichment of relevant components or proteins of interest, and fractionation to reduce the sample complexity. Quantitative

discovery proteomics adds a further challenge that proteins need to be quantified across multiple fractionated samples.

In targeted proteomics analysis, the mass spectrometer is programmed to analyze a preselected group of proteins/peptides, thus reducing the time required for analyzing each sample. Typically, less than one hundred proteins are set to be quantified simultaneously with high precision, sensitivity, specificity and throughput in one sample. This can be achieved using a technology called selected reaction monitoring (SRM), whereby the m/z (mass/charge) transitions of precursor ions (unique diagnostic peptides representing each of the targeted proteins) to its corresponding fragment ions are monitored on a triple quadrupole MS instrument. SRM has been proved to be highly sensitive, quantitative, accurate and highly reproducible.[39] While discovery proteomics is largely used to identify inventory proteins in a sample or detect differences in the abundance of proteins among multiple samples, targeted proteomics approaches are increasingly used in pharmaceutical and diagnostic applications to provide more accurate quantification of proteins and metabolites in complex samples. Additionally, targeted proteomics strategy is also applied in combination with discovery proteomics, to quantitate specific proteins of interest that are found during discovery screening.[40]

Mass spectrometry is not inherently quantitative, as it exhibits different responses (ionization efficiency and/or detectability) to analytes with different physicochemical properties, thus sparking the development of methods to determine relative and absolute abundance of peptides/proteins in samples. Despite this bias for different analytes, the intensities recorded by MS for the same analyte across different samples correlate very well with their relative abundance in each sample, which is the basis for the relative quantification.[41] Relative quantification strategies compare the abundance of individual peptides/proteins across multiple samples according to spectral counting or the intensity of precursor ions/fragmented ions. Relative quantification can be achieved by a label-free or isotopically-labeled approach. A typical relative quantification experiment which uses the isotopic labels entails labeling peptides or proteins from two different conditions with isotopically heavy and light atom. Since the heavy

and light forms of the individual peptides are chemically identical, they co-elute at the same retention time in the LC and thus are detected simultaneously in MS. The peak intensities of the heavy and light peptides (differ in m/z) are then compared and the abundance differences in different samples can be determined from this intensity change.[42] Another quantification approach is termed as "absolute quantification", which relies on the usage of isotopically labeled standards with known concentration. The standards refer to the chemically synthesized target peptides containing isotopically heavy atom. The target peptide in the experimental samples is compared to the standards, and back calculated to the initial concentration of the standards according to the pre-determined standard curve to yield the absolute quantification.[43]

The label-free quantification approach is the most convenient, simple and low-cost alternative to other quantification strategies.[44] It compares the abundance of individual peptides across multiple samples according to spectral counting or the intensity of precursor ions/fragmented ions. This method is ideal for screening or discovery based research. While it works well at capturing large changes of protein abundance, it is less reliable for measuring small changes, which can also be caused by experimental bias generated form sample preparation or MS measurement variance across different runs.[45]

1.2.3 Application of label-free shotgun proteomics in this dissertation

Label-free shotgun proteomics approach is used throughout this dissertation to interrogate different biological systems, ranging from simple lab cultured organisms to complex environmental microbial community. In recent years, the environmental proteomics field has experienced the shift from the proteomics analysis of single organisms to environmental microbial communities. At the early stage of environmental proteomics studies, the proteomics approach was largely applied to help understand the physiology, metabolism and ecology of model organisms, which have important environmental impacts such as dehalogenation, methanogenesis, denitrification, sulfate reducing or high tolerance to radiation and toxic compounds, etc. Proteomics has played an important role in helping uncover the pathways

behind these critical cellular processes and has become a very mature technique for these applications. However, application of proteomics of model organisms to answer environmental questions still suffers from several challenges: 1) the specific experimental design; 2) the choice of statistical analysis methods; 3) analytical incompleteness (under one specific growth condition, the proteome coverage and the rate of identification of proteins-of-interest, are typically much less than 100%).[46] These three challenges have yet to be fully addressed, but are largely alleviated by the development of more robust and more sophisticated techniques for experimental design and statistical analysis, along with the advanced improvement of MS instruments and bioinformatics tools. To be more specific, the first challenge is associated with the capability to design rigorous, environmentally relevant, and feasible experiments in the lab. It is well known that many factors, such as timing, intensity and duration of the experimental treatments, etc., can potentially impact the experimental results. However, the time and expense required for proteomics studies often requires focusing on a few time points or experimental conditions, thereby escalating the requirement for well-designed experiments.[47] For the second challenge, statistical methods applied for the proteomics data analysis have attracted significant attentions these years as they have profound effects on the biological conclusions drawn from the proteomics data. The level of statistical sophistication in the field of proteomics has been dramatically improved in recent years, by the introduction of various statistical methods, such as FDR-controlling techniques in univariate analyses, PCA, clustering analysis for system-level analysis, etc.[48] The third challenge refers to the concerns with the degree of proteome coverage and the proportion of proteins-of-interest (i.e., those that are regarded as significant by statistical criteria) that were identified by mass spectrometry.[49] With the development of MS and subsequent bioinformatics tools, the proteome coverage of single organism can reach around 50% to 80% of the entire predicted proteome,[50] which is already very encouraging given not all the genes are expressed under one growth condition. However, caution need to be taken to deal with the portion of the proteins that meet the criteria for statistical significance. Therefore, how to effectively apply proteomics to environmental questions to reach sound biological conclusions based on model organisms requires careful consideration regarding all the aspects of the experimental workflow, from the experimental

design to the proteomics analysis. A detailed proteomics analysis of two model organisms to help answer the questions regarding to mercury uptake, mercury methylation and methylmercury export, will be discussed in chapters 4 and 5.

Compared with the proteomics analysis of model organisms, metaproteomics analyses of microbial communities collected from their natural environment poses more technical challenges in almost all the aspects, including the sample preparation, MS measurement and bioinformatics analysis. While the protein extraction is very straightforward for lab cultured isolates, environmental samples face more problems such as limited biomass and abundant interferences, which prevents the efficient cell lysis and protein extraction directly from the environmental samples. Another challenge is the sample complexity with a huge dynamic range, which impacts the protein/peptide separation and impedes the identification of low abundance proteins. To this end, different fractionation methods have been developed to reduce the sample complexity and high-resolution MS with fast scanning speed to assist better resolution of the peptides and to reach unprecedented depths of the proteome measurement. While these technical improvements help generate more comprehensive datasets that contain rich information, the success of the metaproteomics also relies on the quality of the metagenome and the subsequent bioinformatics tools to interpret the large and complicated datasets.[51] The metagenome quality is largely impacted by the sample complexity and the sequencing technologies. For metagenome sequencing of environmental microbial community, it is much more difficult to get the finished sequence than the genome sequencing of single isolated organism. Therefore, the metaproteome databases used for metaproteomics analysis are the draft sequence most of the time. The choice of sequencing technologies depends on the experimental goals and no one is found to be ideal for all the samples. Short read approaches are cheaper and more accurate; however, it is very challenging to assemble the contig from these short reads. In contrast, long read technologies have the advantage that it is much easier to assemble the contig from longer reads, but they are more expensive and associated with higher error rates. Hybrid technology which combines both is also employed for metagenome sequencing.[52] In chapter 3 of this

dissertation, one optimized protein extraction method was developed for soil metaproteomics study to eliminate the negative impacts caused by soil-borne humic interference. This work contributes to the environmental metaproteomics fields by addressing one significant challenge associated with the sample preparation of soil mate-proteomics study.

Coupling of cross-linking and affinity enrichments and mass spectrometry extends the use of mass spectrometry based proteomics from identification, quantification and characterization of protein complexes into resolving protein-protein interactions.[53] A similar MS-based proteomics approach is also employed to study the protein-protein interactions in this dissertation, termed as affinity proteomics. Although significant success has been achieved by affinity proteomics in revealing protein interactions, it still faces several challenges, including the difficulty in capturing and purifying the interacting complex of interest, distinguishing the real interacting proteins with background proteins, unambiguously identifying the interacting proteins and interacting sites by bioinformatics tools. The detailed introduction of these challenges and the methods to alleviate these challenges are introduced in the following section 1.3. In chapter 6, an affinity proteomics work is presented, by carefully evaluating all the aspects in the workflow, from optimization of cross-linking parameters to MS measurement to the selection of proper bioinformatics tools. The methodology can be extended to identify different protein interactions in different systems.

1.3 Applications of mass-spectrometry based proteomics on environmental microbial investigations

1.3.1 Environmental proteomics of single organisms and microbial communities

Microbes such as bacteria, fungi, and viruses have evolved to occupy almost any niche on the earth, ranging from human body (human gut, oral cavity, skin surface, etc) to environmental systems (water, soil, air, etc). It is well known that environmental microbes play important roles in biogeochemical cycles, as they contribute to global nutrient cycling, organic matter decomposition, eutrophization, thereby exerting a lasting effect on biosphere and climate.[54] Natural soil and water

environment are two systems that are of utmost interest for microbiologists since these habitats often contain a very diverse population of prokaryotic species and the abundance and size can vary over many orders of magnitude.[55] These microorganisms play very important roles for the global nutrient cycling and organic matter transformation. In particular, some specific metabolic capabilities of microorganisms, such as dehalogenation, methanogenesis, denitrification and sulfate reduction, are being studied in detail and are being used for biotechnology development. Additionally, the abilities of some microorganisms to survive at extreme environmental conditions, or to use many broad electron donors and acceptors, or to tolerate radiation and toxic chemicals are all of interest.[56]

Traditionally, laboratory-based approaches are widely adopted to study microbiology. Different microbial species are isolated from the specific environment and various aspects, such as physiological properties, metabolic activities, and the molecular mechanisms underlying specific functions, are explicitly explored. These studies really broaden our knowledge on the microbial diversity; however, more than 90% of the environmental microorganisms are uncultivated, hence hindering our understanding towards the microbial functional roles and interactions in their natural surroundings. Currently, it is well accepted that microorganisms in natural environments generally function in communities, thus the cell-cell communication and interactions are very important for understanding how naturally occurring microbial communities respond to perturbations such as desiccation, starvation or freeze-thaw cycles.[56] Therefore, techniques that can bypass the need to isolate and culture individual microbial species will definitely offer promising new insights into microbial ecology: 1) 16S rRNA sequencing approaches can provide information regarding to the species composition; 2) novel shotgun sequencing and pyrosequencing techniques reveal the whole metagenomes; 3) proteomics approaches access proteins composition and interactions qualitatively and quantitatively. Valuable information gained from single or combined global techniques might revolutionize our previous understanding of microbial diversity and physiology within complex consortia and up to entire ecosystems.[54]

With advancement of DNA sequencing technology, the whole genome of microbial isolates and communities are accessible, which further prompted the application of proteomics to microbial studies.[57] It is known that in their natural habitat, microorganisms can alter their protein expression profile as a common strategy to overcome expeditious and harsh environmental changes, such as temperature, humidity, nutrient availability, and predators. Therefore, only focusing on the gene regulation is not sufficient to fully understand microbial adaptation strategies. Proteomics analyses are required to investigate the physiology of complex microbial consortia at the protein level.[54] Mass spectrometry based proteomics which offers protein expression profiles, opens a new window for us to look deep into the environmental microbial world. In particular, metaproteomics is an emerging but very promising approach to tackle the high complexity of the microbial world as a whole, thus providing a way to directly assess how microbial communities function. Besides the protein abundance information, proteomics study also assists extensive characterization of proteins primary structure (maturation, post-translational modifications and degradation), interaction network, topology, function, regulation and turnover.[58]

There are three types of environmental proteomics studies: 1) microbial isolates in the laboratory; 2) artificial communities created in the laboratory; and 3) natural communities collected from the environment.[56] Currently, there is still a large portion of environmental proteomic investigations are based on model microorganisms cultured in the laboratory. These studied species usually have special traits such as sulfur/metal reducing, high tolerance to toxic compounds, versatile usage of electron donors/acceptors, nitrogen fixation, etc. These qualities make these organisms attractive for environmental biotechnology applications, and proteomics can lead to a better understanding of their functions in specific habitats. In particular, by investigating the proteome change under monitored laboratory conditions, overall microbial physiological activates could be revealed and molecular processes or proteins that are responding to specific stimulations can be determined. Comparatively, proteomics studies on single organism are much simpler than on communities. The genome is more easily sequenced and annotated for single, laboratory cultivable organisms, thus improving the proteome

identification according to this high quality database. When dealing with batch cultures of microbes, sufficient sample amount also allows for flexibility during method development for more complete proteome profiling. Therefore, it is technically feasible to achieve coverages in proteome analyses of around 80% for a number of microbial proteomes.[59] Some sequenced bacterial species are well known for their unique capabilities. For example, *Shewanella oneidensis strain* MR1 can use more than ten electron acceptors, thus reducing a wide range of organic compounds, metal ions and radionuclides;[60-61] several species within the *Bacillus* genus can survive under extreme environments;[62-63] and the genera *Halobacteria* and *Haloarchaea* can tolerate the highly saline environments, etc.[64-65] The proteomic analysis for numerous microbial isolates not only contributes substantially to our understanding of individual organisms at the cellular level under laboratory conditions, but also provides guidance for the experimental design of following biological studies (e.g. gene knockout experiment, environmental manipulation).

With the development of all the omics techniques, today's microbial proteomics are not only limited to basic *in vitro* characterization of single organisms to describe the proteome content of a cell at a given time or unravel the physiological implications of stress/starvation events, but are also used in sophisticated settings of high complexity such as host–pathogen interactions, mixed microbial communities and, ultimately, native microbial metaproteomics. Although laboratory-based investigations of microbial communities provide a way to investigate the microbial functions and interactions as a community under controllable conditions, they do not represent the real natural environmental communities. Therefore, the metaproteomics study of microbial communities in their native environments is highly desirable, given its capability to provide a realistic picture of microbial functions, interactions and adaptations to specific environment. Around 30 years ago, ecologists already started to realize that composition and interactions of the community members during nutrient competition, predation, and cellular signaling largely determine the microbial activity and physiology in a certain environment.[54] In 2005, the metaproteomics characterization of a natural acid mine drainage microbial biofilm community

carried out by Banfield group became the landmark study for environmental community proteomics.[66] This group further published another work on a strain-resolved community proteomics study, identifying proteins from dominant community members, with strain specificity. This study provided very valuable evidence of gene exchange during adaptation to specific ecological niches.[67] Since then, more and more environmental communities are investigated by proteomics: such as community proteomics study of marine symbionts of *R. pachyptila*;[68] proteome analyses of waste water treatment plants and activated sludge;[69-70] and metaproteome studies of highly complex groundwater and soil environments.[71] Among all the proteomics investigated environmental systems, the soil environment is regarded as one of the most complex habitats since a handful of soil can contain up to 10 billion bacteria and tens of thousands of species.[72] Our lab has examined various soil samples, and found significant experimental and bioinformatics challenges: 1) sequencing of soil samples genome is very difficult, leading to most of the sequenced soil metagenomes are horribly incomplete and woefully assembled/annotated; 2) protein extraction from some types of soil is extremely difficult due to the low biomass and abundant interferences; 3) the enormous genetic heterogeneity within microbial populations; 4) huge and unknown dynamic range of protein expression levels within microbial cells, etc. Regardless of these challenges, the enormous potential of soil metaproteomics study still drives scientists to improve every aspect including improvement of protein extraction methods, downstream MS technology and subsequent bioinformatics tools.

Generally, systems biology workflows, which collect different levels of biological information like genomics, transcriptomics, proteomics, and metabolomics, are proved to be well suited for studying native microbial systems to ultimately reveal the real physiological situations.[59]

1.3.2 Metabolic characterization of mercury methylating bacteria

The dissimilatory sulfate-reducing bacteria (DSRB) and dissimilatory iron-reducing bacteria (DIRB) are two intensively studied groups of bacteria that are of economic, environmental and biotechnological importance owing to their capability to use various electron donor or acceptor. However, one capability of these two groups of bacteria is less studied: their capability to generate methylmercury. Significant efforts have been devoted to elucidate the process of mercury methylation in natural environment, considering that the methylmercury is neurotoxic and can be bioaccumulated in the food chain. Our understanding of the mechanisms behind mercury methylation was greatly enhanced by the discovery of two genes *hgcA and hgcB* that are essential for mercury methylation in 2013.[73] Subsequent studies further established that microbial methylation is tightly coupled with Hg(II) uptake, Hg(0) oxidation and MeHg export processes.[74] However, fundamental mechanisms regarding to all these mercury associated processes are poorly understood. The status of MS-proteomics affords the ability to characterize these microbes, along with their *hgcAB* gene deficient strain (Δ*hgcAB*) [73] in the system level, by revealing all the possible proteins / pathways that are associated with Hg related processes and how these bacteria respond to Hg stress. This information is of utmost importance to understand the bacterial mercury methylation processes and further contributes to the controlling process of global mercury pollution.

Heavy metals, such as mercury (Hg) and uranium (U), are emitted into the environment by natural or anthropogenic sources, which pose risk to environment and human health. Uranium is generated from the milling processes for nuclear weapons and reactor fuel, and soluble U(VI) can leach from the bottom to the surface water and then magnify in the food chain.[75] Certain microorganisms, such as Clostridium sp, Fe(III) reducers (IRB) and sulfate-reducing bacteria (SRB), are capable of reducing the U(VI) to uranium [U(IV)] which causes much less harm.[76-78] While U is not used frequently nowadays, the widespread occurrence of Hg in the automotive, electronic, agricultural, dental, and health care industries causes global concern. In nature, Hg is a rare and toxic metal element that exists as both

inorganic and organic compounds.[79-80] Human activities, such as agricultural irrigation, industrial water wastes and coal burning or trash incineration are major factors that impact the global mercury levels in the environment. Once atmospheric Hg deposits into the water system, about 98% of the Hg is immobilized in sediments. While inorganic mercury is toxic, methylmercury is several times more carcinogenic and neurotoxic: it does not only bioaccumulate in the food chain, but also causes irreversible MeHg poisoning.[81-83]

Microbial activities play critical roles in modulating environmental Hg mobility and toxicity.[84] Many phyla, such as Firmicutes, Actinobacteria, and Proteobacteria have been reported to have a mercury resistance mechanism.[85] A significant amount of these organisms are found to be able to reduce Hg(II) to Hg(0), which is carried out by *merA*-encoded mercuric reductase protein. This resistance mechanism is one way of microbial adaptation to environmental mercury stress and is very critical for Hg detoxification and biogeochemical cycling in the environemnt.[85] However, some other bacteria, in particular, certain dissimilatory sulfate-reducing bacteria (DSRB) and dissimilatory iron-reducing bacteria (DIRB) living under anoxic environment, are found to be able to methylate inorganic mercury to neurotoxic methylmercury, CH_3Hg^+, and further export it to the environment.[86] Methylmercury in the environment does not only impact the mercury cycling, but also poses risk for wild life and human health. Although mercury methylation has been studied for several decades, relatively few methylating bacteria have been discovered and fundamental mechanisms regarding to how bacteria convert inorganic mercury into methylmercury is poorly understood.

G. sulfurreducens PCA, an iron-reducing bacterium, and *D. desulfuricans* ND132, a sulfate-reducing bacterium, are two known Hg(II) methylators and extensively used as the model organisms for understanding mercury associated activities, including mercury reduction, oxidation and methylation. In 1994, mercury methylation mechanism in *D. desulfuricans* LS was partially unraveled by Bartha and coworkers.[87] By tracing the origin of the C element on the generated methylmercury by isotopic labeling of C element of different sources, they discovered that the source of the methyl group in $^{14}CH_3$-Hg was

$^{14}CH_3$-THF and one 40 kDa corrinoid protein was identified, but was not characterized. Based on this result, the reductive acetyl-CoA (Wood- Ljungdahl) pathway was proposed to be involved. However, further studies could not establish the positive correlation between reductive acetyl-CoA pathway and the mercury methylation capability, since some bacteria that are lacking this pathway are found to methylate mercury. Additionally, no clear evolutionary relationship is found among the mercury methylating bacteria. In 2013, with the development of the sequencing technology and the computational homology modeling techniques, scientists discovered the genetic basis of the mercury methylation: two genes, termed *hgcA* and *hgcB* are found to be essential for the mercury methylation. Sequence analysis, homology modeling, *in vivo* knockout studies, and site-directed mutagenesis indicate that *hgcA* encodes a membrane-associated corrinoid protein that is possible to be responsible for transferring a methyl group to a Hg substrate, while *hgcB* encodes a putative 2[4Fe-4S] ferredoxin, which presumably enables turnover by delivering low-potential electrons to reduce the corrinoid cofactor on HgcA to the cobalt (I) [Co(I)] state.[73] Although HgcA and HgcB are not part of the acetyl-CoA pathway, the sequence of the HgcA cobalamin-binding domain (HgcA-CBD) has detectable similarity to the cobalamin-binding domain of the well-characterized corrinoid iron-sulfur protein (CFeSP), indicating a common evolutionary origin. CFeSP is a key component of the acetyl-CoA (Wood-Ljungdahl) pathway involved in the transfer of methylcarbocations from CH_3-H_4folate to acetyl-CoA synthase (ACS). The HgcA-CBD is unique with respect to the presence of a strictly conserved cysteine, which is essential for Hg methylation in vivo. A methyl group from CH_3-H_4folate can only be transferred to a corrinoid cofactor in the Co(I) state. Thus, a source of low-potential electrons is required to reduce the cofactor to Co(I). The roles of *hgcA* and *hgcB* are consistent with the proposed methylation pathway. The ferredoxin HgcB presumably serves as the donor of low-potential electrons to HgcA, which continuously activates the corrinoid cofactor to accept a methyl group from a methyl donor. Although it is likely that CH_3-H_4folate is the proximal methyl donor, the relationship between Hg methylation and C1 metabolism and cellular electron transport processes remains unclear.[88] Comparative proteomics is a powerful tool for characterizing dynamic molecular responses in cellular networks induced by environmental stress or the deletion of individual genes, and

thus is employed in this dissertation work to investigate mercury associated activates, including mercury transport, mercury reduction and mercury methylation in these model organisms.

1.4 Affinity proteomics for characterizing protein interactions

Affinity proteomics extends the use of mass spectrometry based proteomics from identification, quantification and characterization of proteins into resolving protein-protein interactions. In order to explore the various levels of proteomic characterizations, we sought to use a similar MS-based protein measurement approach, label-free shotgun proteomics, to examine protein-protein interactions. Protein–protein interactions are the fundamental components for understanding protein-mediated biological processes. Traditional methods used for studying protein-protein interactions are carried out by combination of co-immunoprecipitation of protein complexes with SDS-PAGE, and protein members are then identified by following western blotting. More recently, high throughput techniques such as affinity purification-mass spectrometry (AP-MS)[89-90] and the yeast two-hybrid (Y2H) approaches[91] have been introduced and widely applied for this purpose. In particular, AP-MS, that can identify functional interactors under near-physiological conditions, holds great potential and has been applied for interactome mapping in several organisms.[92-93] Two prominent improvements recently really drive the AP-MS field even further: 1) advanced mass spectrometers currently can offer high resolution, fast sequencing speed and good sensitivity at the same time; 2) quantitative MS strategies assist distinguishing truly interacting proteins from background proteins.

In the early days of affinity purification-mass spectrometry, tagged bait proteins were always overexpressed to enhance their recovery during the pull-down procedure. However, overexpression of the bait protein can obscure the true situation happening in the cell, thus leading to the detection of false interactions.[94] Nowadays, improved MS is able to detect the bait proteins and their interactors expressed at endogenous levels, increasing the chance to detect the real functional interactions.[95] In some simple organisms like yeast, genes of interest can be engineered and tagged directly in their genetic loci, thus

they can be expressed under their native promoter. The affinity purification step is also evolving over time. Previous, affinity purification is coupled with nonquantitative MS, indicating that all identified proteins by MS are potential interactors. Therefore, in order to reduce co-purifying "contaminants", stringent tandem affinity purification using dual affinity tags was employed. However, this multi-step of affinity purification always results in loss of weak or transient interactors, along with sample amount.[96] Although this stringent purification method eliminates a lot of non-specific interactions, we still cannot be sure that all the obtained proteins are true interactors. The introduction of quantitative MS strategies, including intensity-based relative label-free quantification and SILAC- or TMT-based quantification, to this field around ten years ago was a paradigm shift, as it offered a new way to differentiate the unspecific binding proteins and true interactors.[97] It also works for one step affinity purification scenario, thus bringing the single-step AP protocols back to the dominating position again. It is well acknowledged that single-step AP is much milder and faster, and hence increases the chance to detect the weak and transient interactors.

In order to capture the weak and transient interactors, another chemical method is also widely employed: using a crosslinking reagent to form the covalent bond between two spatially adjacent residues of the interacting proteins, thus preserving their interaction until the mass spectrometry analysis. These days, a large number of chemical cross-linking reagents, which have different properties, have been developed. Crosslinking reagents contain two reactive sites to specific functional groups, connected through a spacer or linker region. The reactive groups of cross-linkers can target different chemical groups in the proteins and peptides. For example, N-hydroxysuccinimidyl or sulfosuccinimidyl esters are used to target the primary amino group of lysine (or the protein N termini). In general, crosslinking reagents can be classified in several categories according to their reactivity: Amine-to-Amine crosslinker; Amine-to-Sulfhydryl crosslinker; Carboxyl-to-Amine crosslinker; Sulfhydryl-to- Carbohydrate crosslinker; Sulfhydryl-to- Sulfhydryl crosslinker; Photoreactive crosslinker and Chemoselective ligation

crosslinker. A photoactivated cross-linking by (photoactivatable p-benzoyl-L-phenylalanine (Bpa)) is adopted in this dissertation for protein interaction investigation.

Photoactivated cross-linking of amino acid residues using UV lamps has been used for investigating protein-protein and protein-peptide interactions for a long time. To induce the crosslinking reaction between Bpa with nearby amino acids, a UV light at a wavelength of ~360 nm is often used. When coupling with high resolution mass spectrometry, cross-linked peptides along with the specific interacting sites can be identified with the substitution of Bpa into specific sites in a peptide or protein. Relatively long irradiation times (30 min as the normal UV irradiation time) and the difficulty in incorporating Bpa into cellular proteins are regarded as two conventional limitations of this methodology. To address the limitation of a long irradiation time, UV lasers have been used extensively to cross-link proteins and nucleic acids.[98] By this method, the interaction between nucleic acids (proteins) and proteins will be frozen in real time, thus allowing a snapshot of the steps in the assembly of such complexes.[99] Therefore, the ability to capture interactions in the living cell in a very short time using a laser should allow "real-time" analysis of protein-protein interactions of intracellular pathways and processes.

This dissertation specifically aims at identifying the interaction partners of G protein-coupled receptors in model organism yeast by coupling photoactivated cross-linking and affinity proteomics. G protein-coupled receptors (GPCR) are the targets of approximately 40% of all modern medicinal drugs. Despite significant progress in terms of elucidating the structures of G protein-coupled receptors (GPCRs), as well as the contacts between GPCRs and their cognate ligands outside the cell, knowledge is very limited regarding what happens in the intracellular region when the signal is transduced across the cell membrane to activate intracellular signaling pathways. Since this process is mediated by the interaction of the GPCR with other intracellular protein partners, it is vital to be able to characterize the interaction types and partners for this signaling process. To this end, we have developed a rapid, time-resolved method to integrate laser-induced cross-linking with MS readout to investigate the protein interaction partners, with a focus on a G protein-coupled receptor *in vivo*.

Ste2 (431 amino acids), our target protein, is a yeast mating pheromone receptor that is also a model GPCR. It participates in virtually all aspects of cellular physiology, including hormonal responses, neuronal transmission, and mediation of taste, smell, and vision. In order to investigate the interactions between GPCRs and intracellular effectors, the photoactivatable unnatural amino acid Bpa was engineered into the different positions of Ste2p, to more effectively capture and identify proteins associated with the intracellular domains of Ste2p in resting state and in the ligand-activated state. Ste2p was engineered with a FLAG tag at the C–terminus to allow affinity purification. To differentiate between Ste2p expressed on the plasma membrane vs. inside the cell, MTSEA-biotin was used to label the T199C residue engineered into the second extracellular loop of Ste2p, which is only surface exposed in the plasma membrane associated receptor. Cells containing Bpa mutants were irradiated with a UV laser to induce cross-linking with intracellular protein(s) in close proximity to the Bpa residue. Total cellular membranes were extracted with either a single step NeutrAvidin bead enrichment of a biotin -labeled Ste2p or a double enrichment method (FLAG beads / NA beads) to obtain surface expressed Ste2p and any cross-linked interacting proteins. Instead of using in-gel digestion, on-beads digestion was carried out here to minimize the sample loss.

Due to the difficulty and time required to prepare the Ste2p interacting complex samples, the use of laser-activation to facilitate cross-linking of Bpa-containing molecules to proteins was validated in a model system: bovine serum albumin (BSA) and two biologically relevant peptides: the tridecapeptide budding-yeast mating pheromone α-factor, and the decapeptide human gonadotropin-releasing hormone (GnRH). While the α-factor used for laser-induced cross-linking reaction is using a biotinylated, photoactivatable p-benzoyl-L-phenylalanine (BPA) modified analog. The best laser irradiation energy and irradiation time was explored in this study and the irradiation condition is further used for the Ste2p interacting study. An unexpected surprise from this study is that we found both peptides can specifically bind with BSA and they even compete for the same binding site, we thus also fully explored this subject in this dissertation work.

1.5 Overview of this dissertation

This dissertation explores a wide range of applications for a mass spectrometry based proteomics approach in different biological systems, ranging from the metaproteomics of soil microbial community, to proteomics of single bacterium, and finally to the affinity proteomics in yeast. The goal of this dissertation was to refine environmental proteomics to deliver a high-performance mass-spectrometry based approach that is able to address a wide range of biological questions. Of course, the application explored in this dissertation is just the tip of the iceberg.

A major part of this dissertation work focuses on the application of mass spectrometry based proteomics approach in environmental microbiology, for either single organisms or environmental microbial communities. Chapter 2 provides all the detailed information pertinent to the workflow of mass spectrometry based proteomics research, including sample preparation, mass spectrometry measurement and data analysis. Chapter 3 provides the optimized sample preparation method designed for removing humic interference from soil samples, thus enhancing the mataproteome analysis of the soil microbial. Chapter 4 and 5 focus on the proteome analyses of mercury methylating bacteria, in an effort to reveal the proteins/metabolic processes that are involved in mercury associated activities, and how bacteria respond to mercury stress. Chapter 6 focuses on the application of cross-linking affinity proteomics for studying the protein-protein interactions and protein-peptide interactions. Chapter 7 is the conclusion of this dissertation work and further perspectives of this general research area.

CHAPTER 2

Experimental and informatics methodologies for MS-based microbial proteome investigations

2.1 General workflow of shotgun proteomics

A shotgun proteomics approach based on liquid chromatography coupled with nano-electrospray tandem mass spectrometry (LC-nEST-MS/MS) was employed for all the proteomics work discussed throughout this dissertation. A typical bottom-up proteomics workflow consists of three parts, including sample preparation, mass spectrometry measurement, and the data analysis, as illustrated in Figure 2.1. In the sample preparation part, the sample is first harvested from the organism, culture, patient or environmental community to obtain proteins from a biological specimen. The proteins in the sample are readily accessible via lysis and extraction from the cells in the sample. The extracted proteins are later denatured and digested by a sequence-specific enzyme such as trypsin. Protease digested peptides are loaded onto the high-performance liquid chromatography (HPLC) system for separation and the liquid chromatography can be directly interfaced with electrospray ionization sources of mass spectrometer. Once the ions enter into the mass spectrometer, they are separated by their m/z in mass analyzer and their intensities are recorded. For peptide sequence, top n abundant peptides are subjected for fragmentations, to generate tandem mass spectra (MS/MS). Experimental MS/MS spectra are compared with the theoretical tandem spectra, and the best match is found that meet the minimum threshold for statistical significance. In this manner, the experimental spectra are assigned with peptide sequence based on the best matched predicted spectra. Once the peptide sequence is determined, they are traced back to the originating proteins. Overall, this typical pipeline is powerful for proteomics analysis in various biological systems, especially for the proteome characterization of single microbes by identifying around 50% to 60% proteins compared with all predicted proteins. However, it still needs to be optimized given the special needs. For example, for soil metaproteomics study, due to the present of abundant soil-borne interferences, optimization is required to remove the interferences before MS measurement, as discussed

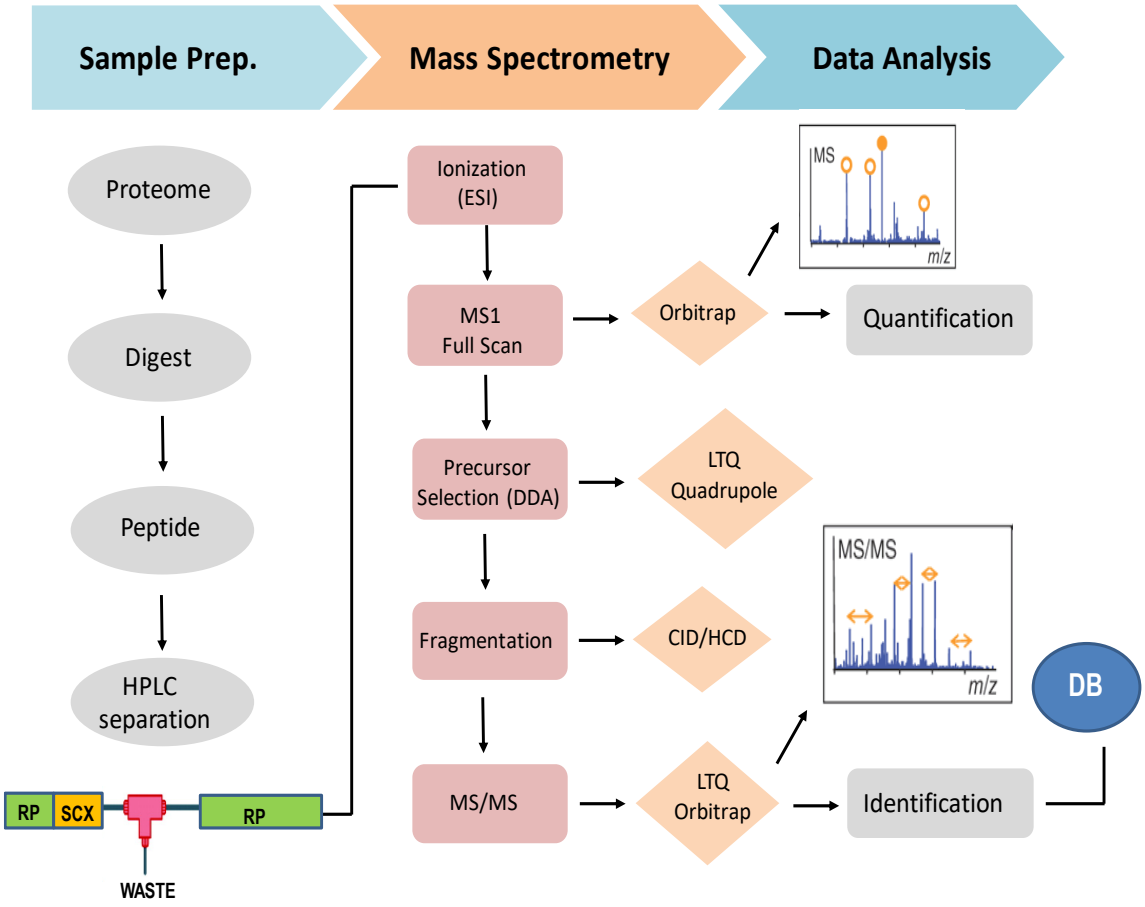


Figure 2.1. The schematic illustration of the workflow typically employed for the bottom-up proteomics analysis.

in detail in chapter 3. Additionally, for protein interaction research, affinity enrichments have to be incorporated into this pipeline to enrich the interacting complex of the target protein, as discussed in detail in chapter 6.

2.2 Sample preparation

2.2.1 Cell lysis

Sample preparation is a critical step for proteome analysis since its quality and reproducibility significantly impact downstream mass spectrometry measurement. Due to the complexity of the proteome samples, there is no standard protocol workable for extracting proteome from all different backgrounds. Different factors have to be considered before proteome extraction, including source, type, complexity, cellular localization, dynamic range, physical properties and interferences. Steps such as optimization of cellular lysis, depletion of abundant proteins, subcellular fractionation or enrichment of specific group of proteins are incorporated into the sample preparation method to resolve these issues.[100-102]

Cell lysis is normally the first step in proteome extraction and purification. In order to select the best cell lysis method, we need to first understand the structure of cells. All cells have a plasma membrane which consists of lipid and proteins that separate the intracellular parts from the extracellular environment. Membrane proteins are embedded in the lipid bilayer and always have several transmembrane regions. In different cell types and different organisms, the protein lipid interaction and content varies. Since my dissertation work is largely focusing on the bacterial proteome extraction, I will only focus on the bacterial cell structure. Besides the membrane, bacteria are always surrounded by a rigid cell wall outside of the membrane. The cell wall differs in Gram-negative and Gram-positive bacteria. The cell wall of Gram-negative bacteria consists of a thin layer of peptidoglycan between outer membrane and inner membrane. The outer membrane contains lipopolysaccharides, ribose and antigens. The Gram-positive bacteria have the cell wall consisting a thick layer of peptidoglycan and lipoteichoic acid surrounding the single membrane.[103] Therefore, it is notable that cell wall of Gram-positive bacteria

is more difficult to break than that of Gram-negative bacteria. Thus different lysis methods have to be considered for different bacteria cell lysis.

Generally, the lysis methods are divided into two basic categories, physical lysis and chemical lysis. Physical lysis approaches normally includes: bead beating, freeze-thaw, mortar mill grinding, grinding under liquid nitrogen, ultra-sonication, and thermal shock. While the chemical lysis methods normally employ different detergents to assist breaking the membrane structure.[104] Sometimes, protease and phosphatase inhibitors are also added to the lysis reagents due to their capability to degrade or artificially modify the extracted proteins. Considering the chemical property of detergents, which consist of a hydrophobic "tail" and a hydrophilic "head", they are useful for extraction and solubilization of membrane proteins by breaking the hydrophobic interactions, but are not powerful enough for disrupting the thick cell wall of Gram-positive bacteria. Therefore, in order to get the best lysis efficiency, it is usually best to combine the physical lysis method and chemical lysis method together for dealing the environmental microbial cell lysis. Note that critical micelle concentration (CMC) of detergent is important for the efficiency of lysis buffer and also impacts the subsequent removal steps. And for different detergent, different CMC is required to perform the proper function.[105] The detailed steps of the this part used for my dissertation are listed: cells that were in liquid medium were pelleted by centrifugation at 2850× g for 10 min, and the supernatant was discarded. 1 mL lysis buffer (100 mM Tris-HCl, pH 8.0, 4% w/v sodium dodecyl sulfate (SDS) and 10 mM dithiothreitol (DTT)) was added into the cell pellets and followed by boiling for 5 minutes and further disrupted by sonication (10 s on, 10 s off) for 2 min with an ultrasonic disruptor (Branson) at 20% amplitude. Crude lysates were subsequently centrifuged at 21,000×g for 10 min to remove any cell debris. Following the cell lysis, protein concentration was determined by the BCA assay (Pierce Biotechnology, Waltham, MA). It is a critical step to standardize the protein amount subject for the subsequent investigation and to assist determination of protease amount added into the solution based on the required protein/protease ratio.

2.2.2 Protein extraction and precipitation

Although detergents are shown to improve the lysis efficiency significantly, most of them are incompatible with the following liquid chromatography–tandem mass spectrometry (LC–MS/MS); in particular, detergent like SDS also interferes with protease activities. Therefore, removing detergents without impacting the proteins is a big challenge during the sample preparation. At the same time, removal of other unwanted cellular material, such as genomic DNA and lipids is also preferred to prevent any ion suppression or chromatographic interference from these molecules. To achieve this goal, precipitation is normally carried out to enrich the proteins and remove the detergents by using trichloroacetic acid (TCA), acetone, chloroform/methanol or ammonium sulfate and ultrafiltration. Previous study showed that precipitation with TCA and acetone yields higher concentration of proteins [106] and thus was routinely used in my lab. Besides the precipitation, molecular weight cutoff filtering is another option to remove SDS, which is used in filter-aided sample preparation (FASP) protocol.[107] Recently, another detergent sodium deoxycholate (SDC) has been evaluated to substitute for SDS, as the previous study showed that it yields better reproducibility by reducing all technical variations during sample preparation.[108] The detailed steps used for my dissertation is listed as follows: For each sample, 1 mg of protein was precipitated using 20% trichloroacetic acid (TCA) and stored at −80 °C overnight. TCA-precipitated proteins were centrifuged at 20,800×g (4ºC) for 15 min, washed twice with ice–cooled acetone and air dried.

2.2.3 Protein digestion and peptide clean up

Post-translational modification, truncation and metal ion adduction can all alter the molecule weight of the proteins, leading to the difficulty to identify intact proteins solely based on m/z values. Tandem MS of the intact proteins provides extensive sequence information, but not complete. Proteolytic digestion of proteins to characteristic peptides can provide more detailed sequence information and proteins can be further identified by peptide mapping. Additionally, analysis of peptides is more easily

achievable by mass spectrometry than directly measuring intact proteins. For this purpose, precipitated proteins from TCA/acetone step is further conducted for denaturation, reduction, alkylation, and digestion. Urea or Guanidine is commonly used as the denaturing reagent and dithiothreitol (DTT) and iodoacetamide (IAA) are used as reducing and alkylating reagent. Trypsin is the most commonly used protease for proteolytic digestion, but other proteases were also selected for better digestion or improved sequence coverage. Different digestion environment is required for different proteases and the detail is illustrated in Table2.1.[109] Since only trypsin was used for my dissertation work, I will discuss trypsin in more detail. After tryptic digestion, generated peptides have 10 ~15 amino acids on average, with the molecule masses ranging from 400 to 2000 m/z, well suited for the mass range of any mass analyzer. Since the cleavage site of trypsin is Arginine and Lysine that both have basic side chains at neutral pH, the C terminal of the tryptic peptides after protonation normally carry two charges. Denaturing reagent, such as urea or Guanidine, performs best denaturing effects at concentration around 8M or 6M, but this concentration is not compatible with trypsin digestion. Therefore, in order to get the complete digestion, two cycles of tryptic digestion were performed. The first was carried out in 4M urea solution, trypsin is not fully active, but the proteins can still be mostly denatured, so the first digestion is not fully complete. When the urea concentration dropped to 2M during the second digestion, trypsin is fully active. Even the denaturing power of urea dropped, previous uncomplete digestion prevents the proteins to fold back, thus allowing the complete digestion during the second time. After trypsin digestion, peptide clean-up is essential before mass spectrometry measurement. Undigested proteins can be removed by molecular weight cutoff spin column filter and salts and buffer can be removed by C-18 tips or columns. The detailed steps carried out in my dissertation are: Protein pellets were suspended in 8 M urea, 100 mM Tris-HCl, pH 8.0 buffer with 5 mM DTT and allowed to incubate at room temperature for 30 min. To facilitate re-solubilization of proteins, the pellets were sonicated again and adjusted to 20 mM iodoacetamide and allowed to incubate in the dark for 15 min. Samples were then diluted 1:100 (w/w) in 4 M urea with 100 mM Tris-HCl (pH 8.0) and 10 mM CaCl2 and pre-digested with sequencing grade trypsin (Promega, Madison, WI) overnight at room temperature on a nutator mixer (Barnstead

Table 2.1. Summary of the commonly used protein proteases

Protease	Cleavage Site(s)	Activity parameters
Trypsin	Arg-X,Lys-X (for X≠Pro)	< 0.1% SDS, 1M Guan. HCl, inactive at pH >11 or pH <4)
Chymotrypsin	F-X,Y-X,W-X,L-X	pH 7.5-8.5, 2M Guan. HCl, 0.1% SDS
Endoproteinase (Glu-C)	E-X and D-X (for X≠P)	0.2% SDS, 1M Guan HCl, 4M urea, up to 100 C, divalent cations and EDTA can be present
Pepsin	F-X,L-X,E-X (not at V ,A, G)	4M urea, 3M Guan.HCl, active at T < 10 C, pH <6
Endoproteinase (Asp-N)	X-D or X-cysteic acid (some X-E cleavage)	2M urea, active at pH ~7
Endoproteinase (Lys-C)	K-X (some N-X)	0.5% SDS and 5M urea
Endoproteinase (Arg-C)	R -X (some K-X)	0.1% SDS, urea < 4M
Elastase	not specific (C-term. of G,A,S,V,L,I)	0.1% SDS, 0.1mM Ca^{2+}, pH ~ 8.0 - 8.5
Pronase	not specific (any peptide bond)	pH 7-8, Ca^{2+}, 1% SDS, 1% Triton X

International). Samples were further diluted to 2 M urea and digested with a second aliquot of trypsin for an additional 3 hours. Following digestion, the peptide solution was adjusted to 200 mM NaCl, 0.1% formic acid (FA) and filtered through a pre-cleaned (with 100 mM Tris-HCl, pH 8.0 buffer) 10 kDa molecular weight cutoff spin column filter (Vivaspin 2, GE Healthcare Life Sciences, Pittsburgh, PA) for final clean up. The peptide enriched samples were quantified by the BCA assay (Pierce Biotechnology, Waltham, MA) and loading certain amounts of peptides onto the column that is packed with reverse-phase (C18) resin for desalting.

2.3 Peptide separation by liquid chromatography

Due to the complexity of the digested peptide mixture, fractionation is always required prior to the mass spectrometry measurement in order to better resolve all the peptides. At the beginning of the proteome research, two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) was the most widely used method to reduce the complexity of the proteome samples by separating proteins via two dimensions: isoelectric point (pI) and molecular weight (MW). [110] However, the shortcomings of this method, such as time consuming and significant bias on different groups of proteins, hindered its further application for deep unbiased proteome measurement.[110] An alternative separation method based on non-gel-based chromatography was later introduced by Yates group, termed multidimensional protein identification technology (MudPIT).[33] This method still separates proteins/peptides in a two dimensional manner in terms of charge and hydrophobicity, but with liquid chromatography consisting of a strong cation exchange (SCX) and a reversed phase (RP) resin to replace gel. Further study proved that this method is superior to be coupled with mass spectrometry compared with previous two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), in terms of easy operation, good reproducibility, unbiased identification and deep measurement.[33] This 2D-LC can be performed online or offline. Online method refers to the elute from first dimension (SCX) is directly eluted into the secondary dimension (RP); while offline method means that each fraction from first dimension is collected first, then subjected to secondary dimension separation. Both methods have advantages and disadvantages. For example, online approach

affords fast speed, minimized sample loss and good sensitivity. It is also highly automated and present high reproducibility. Comparatively speaking, offline approach is more flexible and each fractionation can be repeated separately. And it has much higher tolerance to the existence of a variety of salts and buffers that are MS incompatible.

Based on the principle of this liquid chromatography based multidimensional separation, addition to use the strong cation exchange as the first dimensional separation resin, other chromatography based separation methods were also proposed. For example, ZIC-HILIC-RP was developed by Mohammed group that used zwitterionic hydrophilic interaction chromatography coupled reversed phase and they got better chromatographic resolution compared with SCX-RP method.[111] Another alternative method is RP – IPRP, which was developed by Huber group that employs high-pH reversed phase chromatography and low-pH ion-pair reversed-phase chromatography as the two dimensions. Their results showed that although with lower degree of orthogonality, they gained better separation efficiency and more homogeneous distribution of eluted peptides when comparing this method to conventional SCX-RP method.[112] And recent study carried out in Mann's group also showed the potential of single run with only reversed phase chromatography by an EASY nLC ultra-high pressure system coupled to a Q Exactive HF.[113]

In this dissertation work, MudPIT with SCX and RP was used for all of the proteome measurement. In detail, a biphasic silica back-column (150 µm i.d., Polymicro Technology, Phonix, CA) was packed in-house, consisting of ~3 cm of reversed-phase (C18) resin (Aqua, 5 µm particle size, 125 Å pore size, Phenomenex, Torrance, CA) followed by ~3 cm strong cation exchange (SCX) resin (Luna, 5 µm particle size, 100 Å pore size, Phenomenex, Torrance, CA). Peptide samples (25µg ~ 50µg) were pressure bomb-loaded onto the back-column and washed offline with 15 minutes of solvent A first to remove salts, then with the 30 minutes gradient to 100% B for eluting peptides from reversed phase resin to strong cation exchange resin. The back-column was then coupled in-line with an in-house pulled analytical column, consisting of ~12 cm reverse-phase packed nanospray emitter (100 µm i.d., Polymicro

Technology, Phonix, CA). Further peptide separation was carried on by this MudPIT device interfaced with mass spectrometry via a nanospray source (Proxeon, Denmark). Since the SCX serves as the first dimensional separation, the sample loaded onto SCX is further eluted onto the RP column across a series of 11 different concentrations of ammonium acetate solution (25, 35, 50, 60, 75, 85, 100, 125, 175, 250 and 500mM). Once a fraction of peptdies are eluted onto the RP column, the SCX valve is closed and each fractionation (1 to 10) was followed by the second dimensional separation by reversed phase with 105 minutes of gradient from 0% B to 50% B, while last fractionation was separated with 135 minutes of gradient from 0% B to 100% B for complete eluting. LC solvents consisted of A: 95% H_2O, 5% acetonitrile (ACN), 0.1% formic acids (FA) and B: 30% H_2O, 70% ACN, 0.1% FA and D: 500mM ammonium acetate, 5% ACN, 0.1% FA. The working scheme is illustrated in Figure 2.2. The flow provided by our HPLC is in the several hundred μl/min scale, but given the nano-source interfaced with mass spectrometry, we require the flow entering into the front column to be within several hundred nL/min scale. Thus, the usage of two MircoTee here is a very important design to split the flow. The flow that enters into the first MicroTee is normally controlled at 100 ~ 200 μl/min, most of the liquid is directed to flow into the waste bottle, and provided the back pressure around 70 ~ 80 bar. After the flow splitting, the liquid entering into the front column is at a flow rate around 400 nL/min. Given this flow rate, optimized voltage (3~5KV) is provided to form the electrospray.

2.4 Mass spectrometry instrumentation

Mass spectrometry, based on its superior performance, has played a critical role in analytical chemistry for very long time.[30] During the last few decades, mass spectrometry has experienced a new revolution with the advent of new hybrid instrumentation and new applications in large-scale characterization of biomolecules such as proteome, lipidome, and metabolome. A mass spectrometer is composed of three typical components: ionization source, mass analyzer and the detector. The gas-phase ions are first produced in ionization source and later transmitted into the analyzer for separation based on their mass-to-charge ratio (m/z). Detector records the signal and the computer system transfers the signal

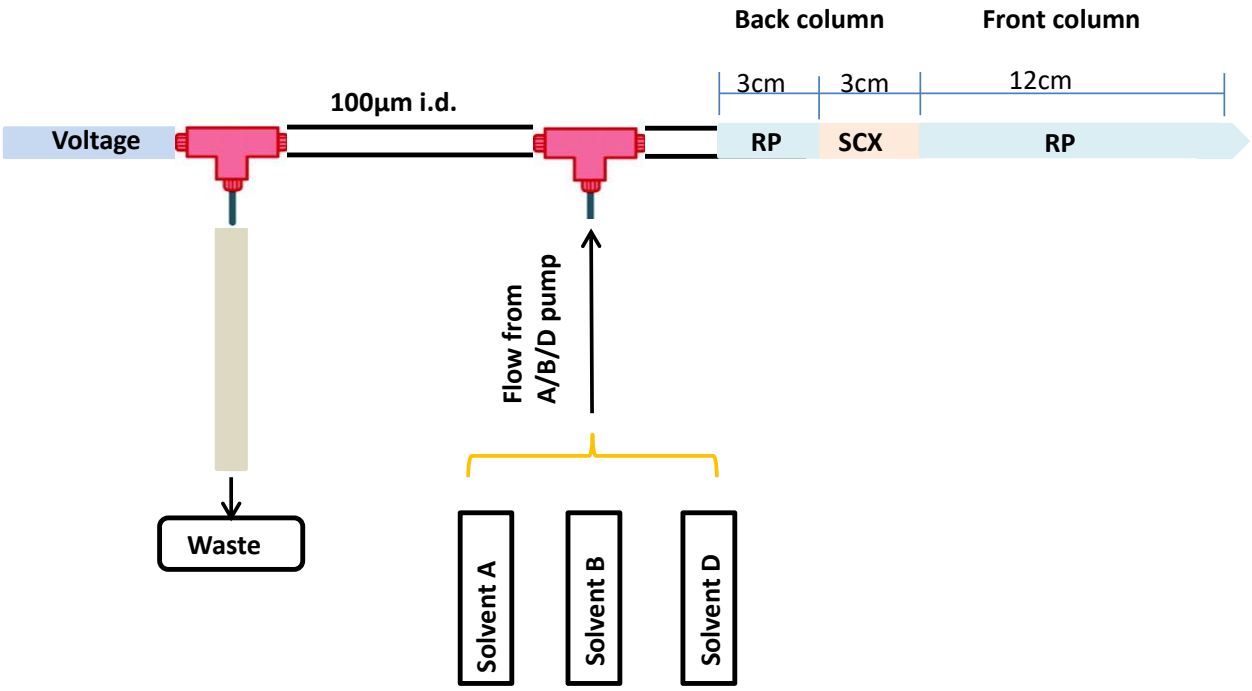


Figure 2.2. The working scheme of the MudPIT experimental setup.

to the spectra.

2.4.1 Ionization Methods

Mass spectrometry separates different analytes based on their mass-to-charge ratio. The first step in the mass spectrometry measurement is the generation of the gas phase ions of the molecules. A variety of ionization techniques are available for mass spectrometry, and the selection of the proper ionization technique depends mostly on the sample features. The ionization techniques can be divided into two groups: low pressure ionization method and atmospheric pressure ionization method. Low pressure ionization methods include electron ionization (EI) and chemical ionization (CI).[114] Atmospheric pressure ionization methods includes electrospray ionization (ESI), matrix-assisted laser desorption/ionization (MALDI),[115] atmospheric-pressure chemical ionization (APCI).[116] EI and CI are only suitable for vapor phase ionization and thus they are incompatible with nonvolatile and thermally labile molecules, such as most of the biomolecules. In addition, EI is also known as a hard ionization method, which induces large degrees of fragmentation along with the ionization. For atmospheric pressure ionization methods, APCI and ESI operate at atmospheric pressure. MALDI normally preformed in vacuum, but can be performed at atmospheric pressure as well.[117]APCI works for molecules with a molecular mass below 2000Da, the sensitivity, ruggedness and reliability makes it a good choice for pharmaceutical application.[116, 118]

MALDI and ESI are the two major ionization techniques used for large biomolecular characterization. In MALDI, the sample is typically mixed with a matrix. When the sample is bombarded with a laser, the matrix absorbs the laser radiation and transfers the proton to the sample. MALDI holds the advantage of easy sample preparation and large tolerance to contaminates such as salts, buffer and detergents. However, typical MALDI spectra contain mainly the singly charged ions, and the nature of the MALDI as a pulsed ionization technique, limits the mass analyzer that can be coupled to it. Time of flight (TOF) mass analyzer is the most common one that is coupled with MALDI since TOF analyzer has a wide mass range and well suited to the pulsed ionization.[119-120]

ESI can generate multiple charged ions, which facilitates the characterization of large biomolecules like proteins by allowing their molecular weight to be determined with lower mass range instruments. In addition, ESI also achieves high sensitivity and can be easily coupled with separation techniques such as high-performance liquid chromatography HPLC, μHPLC or capillary electrophoresis. The principle of the ESI has been extensively studied.[121] Generally, the charged gaseous species of ESI were generated by applying a strong electric field to the liquid which flow through a heated capillary tube with a slow flux. The electric field is produced by applying the potential difference between the capillary and the counter-electrode. This electric fields is strong enough to accumulate sufficient charges on the surface of the liquid at the end of the capillary and generates the "Taylor cone" – the drying process and the coulomb explosion further cause the breakdown of liquid to a lot of highly charged droplets. This process can be nicely explained by Rayleigh equation, which clearly illustrates that proper balancing of surface tension and charge is necessary to keep the stable shape of a charged liquid droplet. Once the charge gets sufficiently large compared to the surface tension, the shape of the droplet becomes unstable and fission takes place.[122] When the droplets enter the capillary, the remaining solvent molecules will be finally removed with the assistance of the heated capillary or heated inert gas (Figure 2.3).

Actually, the spray starts from the "initial voltage" given the specific source and the surface tension of solvent that is used. The voltage needs to be set high enough to conquer the surface tension of the solvents, thus be able to change the shape of the drop to "Taylor cone". Solvent which has higher surface tension always needs higher initial voltage, and vice versa.[123] Small charged droplets are released from the Taylor cone and they start shrinking due to the solvent evaporation. When Rayleigh stability limit is reached, charged droplets start to divide and explode. This kind of division continues for several generations until the electric field on their surface becomes large enough. Then desorption of the charged molecules starts from the surface. It explains that ions of higher concentration in the surface, which are normally more hydrophobic, are always having better sensitivity and can cause the ion suppression on

those ions sit inside. Ions will be further transferred to the high vacuum mass analyzer for m/z determination. The electrospray ionization process is presented in Figure 2.3.

Nano-ESI (nanoelectrospray ionization) has emerged ten years after conventional ESI, which requires very low amount of sample, but achieves higher sensitivity than ESI.[124] As mentioned above, The ESI process is largely depending on the electrochemical process happened in the probe tip, which is influenced by the physical and chemical properties of the sample liquid (surface tension, conductivity, ionizable sites) and external factors such as: voltage, liquid flow rate, length of the capillary. After the occurrence of the ESI, a number of sprayer modifications, such as pneumatically assisted electrospray, ultrasonic nebulizer electrospray, electrosonic spray and nanoelectrospray, have been developed to expand the range of ESI applications. The nanoelectrospray is the most popular one among all these techniques. [121]

Nano-ESI is a more efficient spraying technique than conventional ESI. By decreasing the flow rate to nL/min, lower sample amount and lower voltage are required for nanospraying. Conventional spray needle is replaced by borosilicate glass capillary with a fine tip (1–4 μm inner diameter) which is pulled with a micropipette puller. By using Nano-ESI, the size of first generation charged droplets are much smaller than that generated by conventional ESI, the desolvation process thus becomes much easier and quicker, and leading to the improvement of ion efficiency by 50~60%. Currently, nano-ESI is the most widely used ionization technique in peptide and protein analysis. All the Mass Spectrometry instruments used in our lab are coupled with nano-ESI.[124]

2.4.2 Mass Analyzers and Detectors

When gas-phase ions are transmitted into the analyzer, they are separated by their m/z. Different mass analyzers, which operate on different principles (Table 2.2), are available.

The most basic principle is that all the analyzers use static or dynamic electric and magnetic fields which can be combined or used alone. Each analyzer has its own advantages and limitations and

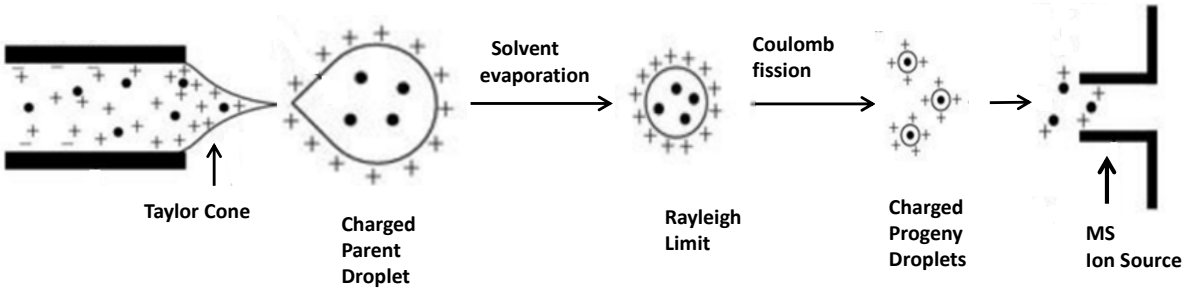


Figure 2.3. Schematic representation of the electrospray ionization process.

Table 2.2. Types of mass analyzers and principles of separation

Mass Analyzer	Ion Source	Ion Manipulation	Mass separation	Operating Pressure (Torr)
Sector(s)	Continuous	Electric and/or Magnetic Field	Kinetic Energy/Momentum	$\leq 10^{-6}$
Time-of-Flight (TOF)	Pulsed	Electric Field	Velocity (Flight Time)	$\leq 10^{-6}$
Fourier transform ion cyclotron resonance (FTICR)	Pulsed	Magnetic Field	Orbiting Frequency	$\leq 10^{-8}$
Quadrupole	Continuous	Electric Field	Stable Trajectory	$\leq 10^{-5}$
Ion trap (IT)	Continuous	Electric Field	Unstable Trajectory	$\leq 10^{-3}$

no one is ideal to all the applications. The selection of the mass analyzer relies on the sample type and experimental goal and the performance of a mass analyzer can be evaluated by the following figures of merit: mass range, scan speed, sensitivity, mass accuracy and resolution. Mass range measures the range of the lowest mass and highest mass that can be detected. Scan speed refers to the speed that the mass analyzer takes to measure over a particular mass/charge range. Sensitivity gives the signal response to different quantity of analyte, generally presents as the slope of the calibration curve. Mass accuracy indicates the difference between the theoretical m/z and the measured m/z and often expressed in parts per million (ppm). Resolution determines the capability to distinguish two m/z with very small difference and is obtained by calculating FWHM (full width at half maximum) of a peak. Ion trap and Orbitrap are the two types of analyzer used in this dissertation, thus their operational principles are introduced below. Additionally, the principles of quadrupole analyzers were also introduced, since it represents a very important separation scheme in MS analyzers, and it also helps better understand the principles of the ion trap analyzer.

Quadrupole Analyzers

Both the quadrupole and ion trap mass analyzer separate different m/z by oscillating electric fields, to manipulate the stable trajectories of different ions. A quadrupole is made up of four electrically isolated hyperbolic or cylindrical rods which are arranged in perfect parallel position. The principle of the quadrupole was described by previous paper.[125-126] The combination of an alternating field and a constant field creates a hyperbolic field, which is an electrical region with strong focusing and selectivity. The ions enter the space between the quadrupole rods along the z axis with certain speed, but they are also forced to move towards x and y axis due to the electric fields. [126] As long as the ion does not hit the rods ($x \neq r_0$, $y \neq r_0$), they can keep a stable trajectory along with z directory towards the detector. But for other ions which do not have the stable trajectory, they will strike the rods and neutralize upon contact, thus will not be detected. Ion motion in a quadruple follows the Mathieu equation, which can be further rearranged as follows:

$a_u = a_x = -a_y = \frac{8zeU}{m\omega^2 r_0^2}$ and $q_u = q_x = -q_y = \frac{4zeV}{m\omega^2 r_0^2}$

For a given quadrupole, r_0 (radius) is the constant, $\omega=2\pi v$ (ω the angular frequency (in radians per second=$2\pi v$, where v is the frequency of the RF field) is also maintained constant, leaving the U (direct potential) and V (amplitude of the RF voltage) as variables to determine the m/z. For an ion of any mass, the stability area can be represented in a V, U diagram. Separation of ions with different m/z can be achieved by ramping the RF and DC voltages at a nearly constant ratio. And different resolution can be achieved by adjusting the slope of V, U diagram.

Besides serving as the mass analyzer, a quadrupole can also be used as ion guide or collision cell. Ion guide, also called ion focusing device, is used to efficiently transmit ions from one point to another without mass separation. Ion loss, normally caused by their collision with residual gas molecules or space charge effects, can result in lower sensitivity. Therefore, an ion guide should transmit ions efficiently and reduce the ion loss. Quadrupole, when operating in RF-only mode (U=0), can keep ions of a large range of m/z in their stable trajectories under certain V. However, when V is set very low, although it can keep more ions stable, it has poor focusing on large m/z. In order to increase the transmission of ions with larger m/z, V has to be increased, with the sacrifice of losing lower m/z due to the unstable trajectories. There is thus a compromise to find the appropriate V.

Ion Trap Analyzers

The ion trap is a device that not only can trap ions by oscillating electric fields; but also can be used as mass filter and collision cell. The quadrupole analyzer manipulates the potentials to allow only the selected m/z go through the rods. In ion trap, all the ions of different masses are trapped in the trap, and are then expelled out of the trap according to their m/z to reach the detector.

There are two types of ion trap: 2D ion trap and 3D ion trap. The two types of ion trap operate on the similar principles; however, due to the better performance of the 2D ion trap on trapping more ions

before the space charge limit, the 2D ion trap thus has a better sensitivity and a wider dynamic range than 3D ion trap. The 2D ion trap is also known as linear ion trap (LIT), which is commonly combined with Orbitrap mass analyzer nowadays for proteomics measurements.

LIT is composed of an array of four rods and ending in lenses that can repel ions. Similar as in quadrupole analyzer, the direct potential (DC) and alternating potential (AC) are applied to the metal rods to control the ion motion. In the trap, ions are confined in the radical dimension by RF electric field and in the axial dimension by a static electric field using DC voltage at the ends of the trap. Once in the LIT, ions are cooled with an inert gas to reduce collision with neighboring ions and ion motion follows the Mathieu equations:

$$a_x = -a_y = \frac{8eU}{m\Omega^2 r_0^2} \; ; \quad q_x = -q_y = \frac{4eV_{RF}}{m\Omega^2 r_0^2} \; ; \quad \xi = \frac{\Omega t}{2}$$

$$\omega_n = (2n + \beta)\frac{\Omega}{2} \quad 0 \le \beta \le 1 \quad n = 0, \pm 1, \pm 2, \ldots.$$

U is the DC voltage, V is the RF amplitude, r_0 is the radius of the ring electrode, ω is the RF angular frequency. Three key parameters exist in the equation: a, q, β. β is an important parameter since the amplitude and the frequency of the oscillation performed by the ions in the trap are closely related to β. β is a function of a and q. For low value of q, and when a=0, β is given by $\beta \approx q/\sqrt{2}$.

Ion trap can have two different ways to expel ions out of the trap. The first scenario is to take advantage of the stability limit and let unstable ions at the boundaries of stability diagram to be scanned out first. In this case, DC voltage is set as 0, and only the fundamental RF voltage is applied to the rod. Its frequency is set as the constant but with the amplitude V varied. Then the additional RF voltage of selected frequency and amplitude is applied to the end cap. Since heavier ions have lower β value and lower secular frequency, when V is increased, β is increased and will hit the stability limit of lighter ions first, so lighter ions will be expelled from the trap.[127]

The second way to expel ions out of the trap is by resonant ejection. An ion oscillates in the trap at the secular frequency follows: $f_z = \beta_z v/2$. Since β_z is related to q_z, for an ion of given m/z oscillates at a selected frequency f_z, V can be finally calculated based on q_z. If an RF voltage at frequency f_z is applied on the end caps, the ion will resonate along with z axis, and the amplitude of its oscillation will increase. When the amplitude increased to a certain level, the oscillation of the ion will be so large and cause the ion become unstable, and thus the unstable ion will be expelled out of the trap. And a range of m/z can also be ejected by applying multiple frequencies.

Orbitrap analyzer

Orbitrap is an electrostatic ion trap, and the first commercial instrument was introduced into the market in 2005 and immediately attracted a great attention. The design of the Orbitrap analyzer is based on a new concept, described by Cooks in 2005.[128] The typical Orbitrap is composed of an external electrode shaped as a barrel, and a central electrode with a spindle shape. The central electrode was applied with electrostatic voltage, opposite to the charge of the ions; while the outer electrode is at ground potential. Ions are injected into the trap and start to oscillate around the central electrode along the z axis under the electrostatic field. Careful mathematical calculation is required to ensure a reasonably circular or oval trajectory of the ions around the central electrode. For example, for positive ions, while the inner electrode is set at about −3200V, the ions need to have a kinetic energy around 1600 eV when they enter the trap. Ions with different m/z have different frequencies while they oscillate around the central electrode, which is independent of the kinetic energy of the injected ions:

$$\omega = \sqrt{\left(\frac{q}{m}\right) k}$$

ω is frequency; m is mass; q is the total charge (q=ze); k is field curvature.

This is a very important property of the Orbitrap, indicating the separation of different ions is only depending on m/z value. The current induced by the oscillating ions is measured and converted by

Fourier transform to obtain mass spectra. In a complete stack of Orbitrap, including an atmospheric pressure source, an analytical quadrupole, a storage linear trap and an Orbitrap, the ions are transmitted from atmospheric pressure source to the vacuum area and accumulated in the C trap. The curved shape of C trap is allowing fast injection of fair amount of ions into the Orbitrap, thus improving the detection sensitivity. It is this ion trapping device that makes the coupling of continuous ion sources and Orbitrap mass analyzer possible since Orbitrap operates in a pulsed fashion. In addition, because of this trapping device, the amount of ions injected into the Obitrap can be controlled, further reducing the space-charge effects. Ions with the same m/z have very coherent oscillation along z axis, while the rotation around the central electrode does not have this kind of coherence, which reduces the background noise. Therefore, this perfect coherent oscillation achieves very high resolution and accuracy. The resolution is positively correlated to the detection time, increasing the detection time can obtain even higher resolution (>240,000 at m/z 400). High mass accuracy: <3ppm with external calibration and <1ppm with internal calibration can be achieved with Orbitrap.

Hybrid MS Instrumentation

Since each analyzer has its own strength and limitation, combination of different analyzers are engineered to obtain better performances. Different combinations can be found in the market, while the LTQ-Orbitrap is one of the favorite instrument platform in proteomics fields. In the hybrid configuration, LTQ is normally used for ion trapping, ion selection, ion fragmentation and low resolution ion detection; while the Orbitrap is used for high resolution, high mass accuracy ion detection. This hybrid instrument can operate on two different detection modes. One is FT-IT mode, which means the MS1 precursors ions are detected in Orbitrap while the MS2 fragment ions are detected in LTQ. The other mode is FT-FT mode, referring to the situation that both MS1 and MS2 are detected in Orbitrap. There are different generations of the LTQ as well as Orbitrap analyzer with continuously improved performance.[129] LTQ-Orbitrap Elite (Figure 2.4), combined the most recent generation of LTQ and ultra-high resolution Obitrap, is introduced in detail here. S-lens in the front side is an ion guide with spaced stacked ring shape, which

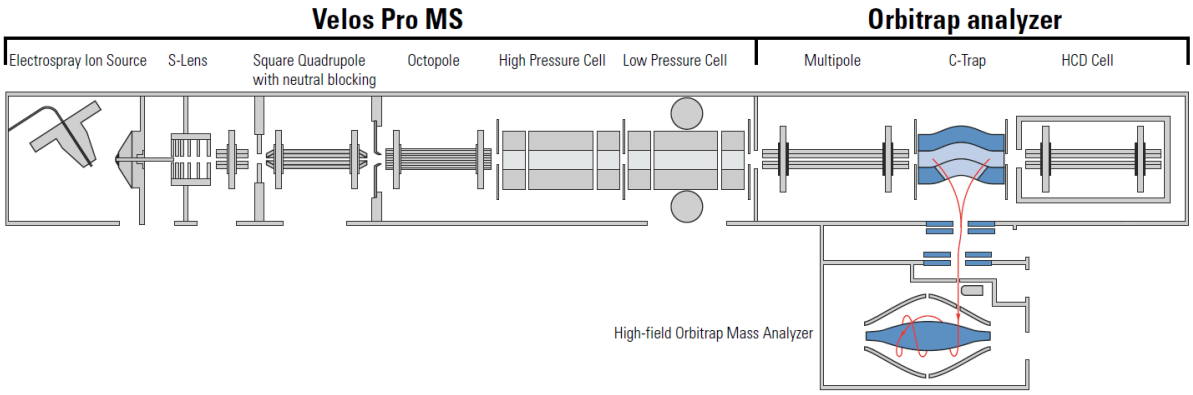


Figure 2.4. Schematic of the Orbitrap Elite Hybrid Mass spectrometer. Image source: Thermo Scientific LTQ Orbitrap Series Hardware Manual, http://www.thermoscientific.com/en/product/ltq-xl-linear-ion-trap-mass-spectrometer.html.

improves the pumping efficiency and ion transmission. Square Quadrupole with neutral blocking facilitates ion focusing. The dual-pressure linear ion trap is designed to achieve better whole-package performance since high pressure cell has higher efficiency on ion trapping, isolation and fragmentation; while the low-pressure cell has higher scan speed, resolving power and mass accuracy. Following gas-free multipole optics provides higher ion transmission and gas-filled C-Trap efficiently traps all the ions. The new collision cell (HCD) next to the C-Trap offers another way for fragmentation (Thermo Fisher Scientific Orbitrap Elite Hardware Manual). HCD refers to higher-energy collisional dissociation, which is also one CID technique specific for Orbitrap analyzer. The precursors ions trapped in C-Trap are transmitted into the HCD cell and fragmentation occurs, later on, fragmented ions are transferred back to the C-Trap and finally reach the Orbitrap for high resolution detection. Compared with traditional ion trap-based collision-induced dissociation, HCD has higher activation energy and shorter activation time. It also has no low mass cut-off restriction and can generate more high quality MS/MS spectra. One drawback is that the spectral acquisition time is up to two-fold longer.[130] The Orbitrap was developed with a decreasing gap between the inner and outer electrodes, thus providing even higher frequencies of ion oscillations and hence higher resolving power. Overall, the key performance characteristics of Orbitrap Elite are listed as follows: Mass Range: m/z 50 - 2,000 or m/z 200 - 4,000; Resolution (FWHM): > 240,000 at m/z 400; Mass Accuracy: < 3 ppm with external calibration and < 1 ppm using internal calibration; Dynamic Range: > 5,000 within a single scan; MS Scan Power: MSn, for n = 1 through 10.

Detectors

The ions passing through the mass analyzer are then detected and recorded by different detectors. Several different types of detectors exist, but only two detectors (Electron Multiplier and Image Current) that are commonly coupled with LTQ and Orbitrap are described here. The selection of the detectors relies on the types of the instruments and their analytical applications. Electron Multipliers are the most widely used detector in mass spectrometry. In this detector, the conversion dynode is held at a high

potential from ±3 to ±30 kV, opposite to the charges of the targeted ions. Due to this high potential, ions from the analyzers are accelerated to a high velocity to enhance the detection efficiency. Ions strike the conversion dynode and thus generate several secondary particles. These secondary particles then strike the first dynode, leading to the emission of secondary electrons. These electrons are amplified by a cascade effect to produce the final current. There are two types of electron multipliers, one is discrete dynode, and the other type is continuous dynode. The discrete dynode electron multipliers are composed of a series of 12 to 20 dynodes, which are held at decreasing negative potentials. The first dynode is held at the highest negative potential, while the output of the multiplier is held at ground potential. So when the secondary electrons generated after striking the first dynode by the secondary particles, these electrons are accelerated to the next dynode due to the lower potential. They strike the second dynode and generate more electrons, and this process is repeated for other dynodes until reaching the ground potential. And the amplified electric current is generated at the end of the electron multiplier. LTQ-Obitrap-Elite, which is introduced above, is using this type of electron multiplier. The other type of electron multiplier replaces the discrete dynodes by one continuous dynode, named continuous dynode electron multipliers. We will use channeltron as one example of this type to briefly illustrate its principles. The channeltron has a curved tube which has a voltage applied between the two extremities, thus producing a gradient accelerating field along tube walls. Secondary particles generated from the conversion dynode by one ion strike the curved inner wall and produce secondary electrons. These electrons pass further and strike the wall of the electron multiplier, thus generating more and more electrons. Thus the cascade of electrons is created and amplified electric current is measured. Continuous-dynode electron multiplier is also widely used in LIT detection system. The number of the secondary particles and the multiplying factor of the dynodes determine the amplifying power of the electron multiplier. Given the features of high amplification and fast response time, electron multipliers are able to be coupled with rapid scanning analyzers, such as quadrupole or linear ion trap analyzer. However, electron multipliers have limited lifetime (1 to 2 years) due to the surface contamination from ions or relatively poor vacuum.

The image current detection is commonly employed in Orbitrap and FTICR detection system. Unlike electron multiplier, the detector has a pair of metal plats within the mass analyzer region, thus is very close to the ion trajectories. To be detected, ions of a certain mass circulate as a tight packet in the same orbit. The ions of the same mass are excited to the same energy and thus have the same orbiting frequency. The movement of ions between rods induces an electric current which can be measured and then converted to mass/charge spectra by a fast Fourier Transform algorithm. This detection method can count multiple masses and detect all the ions that arrive at the same time. In addition, image current detection does not require the amplification of signals.

2.4.3 Tandem mass spectrometry

Tandem Mass Spectrometry, abbreviated as MS/MS or MS^n (n=2,3,…), normally referring to two or n stages of mass analysis associated with ion fragmentation process. Tandem mass spectrometry is employed to improve the specificity of the detection, in particular to facilitate structure elucidation by fragments information and enhance the signal-to-noise of the spectra. For MS/MS, a first mass analyzer is used for isolating the precursor ion of interest, which further undergoes the ion activation to produce product ions and neutral fragments.

Generally, tandem mass spectrometry can be operated in space or in time. In space tandem mass spectrometry requires coupling two distinct mass analyzers, such as triple quadrupoles (QqQ). The fragmentation happens in the second quadrupole (q), and the first and third quadrupole are two mass analyzers, which are used to measure MS1 and MS2 individually. Tandem mass spectrometry can also be achieved in time, which means that MS1 and MS2 are measured sequentially in the same mass analyzer assisted by an ion trap device. Mass analyzers suitable for in time separation are like ion trap, Orbitrap and FTICR.

Ion activation is achieved by increase of the ion internal energy to induce dissociation. There are three major methods which can supply the energy for ion dissociation: (1) Energetic ion collisions with

neutral collision gas, including collision-induced dissociation (CID) or collision-activated dissociation (CAD); (2) Photons, including photo dissociation (PD) and infrared multiphoton dissociation (IRMPD); (3) Electrons, including electron-capture dissociation (ECD).

Different ion activation processes fragment ions in different ways. By using CID or CAD, ion kinetic energy is converted to ion internal energy upon the collision with a neutral collision gas, bringing the ion to its excited stage and later leading to the unimolecular dissociation of the activated ions. Since the dissociation rate is slower than the rate of energy randomization, energy is redistributed among all of the vibrational modes before dissociation occurs. Under this condition, the weakest bonds are preferentially cleaved. Upon colliding with neutral collision gas, only a fraction of the ion kinetic energy can be converted to ion internal energy. This energy fraction can be expressed by the following equation:

$$E_{com}=E_{lab} * m_{collision\ gas}/ (m_{collision\ gas} + m_{ion})$$

E_{com} is the maximum kinetic energy that can be converted to internal energy during a single collision event; E_{lab} is the ion kinetic energy. From the equation, it is easy to notice that increasing the ion kinetic energy or the mass of the collision gas will increase the energy available for the conversion. In practice, there are two different collision regimes: high energy (several thousand electronvolts) and low energy (1-100eV). Normally, only TOF, electromagnetic or hybrid instrument can function at high energy, while the quadrupole, ICR or ion trap instruments can only perform at low energy. Generally speaking, high energy collision always generate simpler and more clear fragmentations, and low energy collision leads to more diverse fragmentation pathway, sometimes including rearrangements. As indicated by the equation, the mass of the collision gas also impacts the convertible energy. Helium is the commonly used collision gas at high energy collision, while the heavier gas such as argon or xenon are normally applied in low energy collision. As we mentioned above, the weakest bond is always broken first by CID, so for peptides, the peptide bonds are always preferentially cleaved and b, y ions are thus largely formed in CID (Figure 2.5).

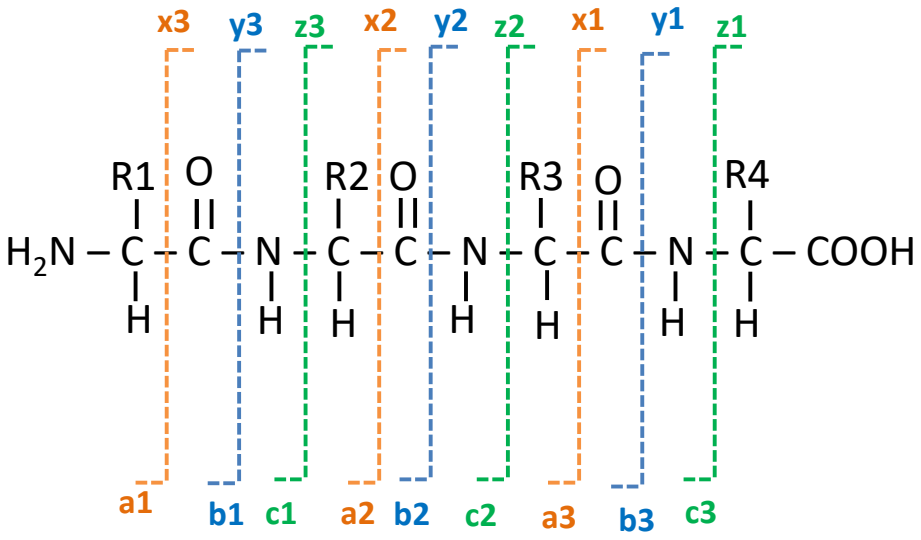


Figure 2.5. Ion fragmentation generated under different activation methods. b, y ions are two most abundant ions formed in CID. Image source: Book, Mass Spectrometry, Principles and Applications.

There is an inherent disadvantage associated with CID given the nature of this collision method: the converted energy is limited, thus the fragmentation is limited. When MS/MS is used to fragment large molecules, the energy needs to be re-distributed in a great number of bonds, resulting in a slower reaction rate of fragmentation. In addition, the time for collision gas introduced into the instrument further compromise the vacuum. Thus, other ion activation methods (SID, ECD, IRMPD) were developed to avoid these drawbacks. Since none of these methods was used in my dissertation work, they are not introduced here.

2.4.4 Data acquisition in proteome measurement

Advanced development in chromatography and MS instruments, offers significantly improved peptide separation and scan speed, leading to the increase of peptide identification in a shorter time manner. The selection of peptide ions for MS/MS analysis, which is controlled by data acquisition methods, is critical for the final results. There are two data acquisition methods available currently: data-dependent acquisition method (DDA) and data-independent acquisition method (DIA).

Data-dependent acquisition refers to the method that top n abundant precursor ions can be selected for MS/MS analysis in a user defined time. The parameters included in this process are: repeat count, repeat duration, minimum MS signal, and dynamic exclusion. Optimized setting of these parameters was shown to improve the proteome coverage. In particular, the number of MS/MS events/cycle and the dynamic exclusion setting are directly associated with the depth of identified proteome and protein quantification. [131-132]

Data-independent acquisition method (DIA) method is not looking for top n abundant precursor ions, instead, it divides the entire m/z into a series of consecutive windows and precursor ions appear in the same window are simultaneously fragmented regardless of the intensity. In this case, all peptides are systematically fragmented in the entire m/z range, which delivers more complete proteome information. The differences between DDA and DIA were examined by a previous paper. [131] However, it also brings in

the fragmentation of more interference and thus leads to more complicated data interpretation, and this challenge increases with the increased complexity of proteome samples. Therefore, data-dependent acquisition method is still the most widely used scheme in proteome analysis and used through this dissertation.

2.5 Bioinformatics for proteome interrogation

With the advent of the robust and high efficient LC/MS/MS platform, systemically analysis of all the proteome within a cell/community becomes feasible. However, manual assignment of this large amount of raw data to peptides is impossible, leading to the development of various bioinformatics tools to accomplish this task. Due to the complexity of the proteome data, it still represents a significant challenge for these bioinformatics tools. In principle, if a peptide can be fragmented completely, the mass difference between neighboring two tandem mass spectra represents the mass of one amino acid and the whole tandem mass spectra can deliver the information of the amino acid sequence. Based on this principle, different computational strategies were proposed to identify peptides/proteins from tandem mass spectrometry (MS/MS) data. One method is depending on *de novo* assignment, which is database independent. The deduction of amino acid sequences from MS/MS spectra by using this method largely rely on very high quality of MS/MS spectra and subsequent sophisticated mathematical models, sometimes also assisted by the chemical techniques such as isotopic labelling. This method is superior in terms of the potential to provide all the information collected by mass spectrometry, but the measured tandem mass spectra in real world is always much more complicated than it is predicted to be, thus adding more challenges for this method. There are several software available currently, such as Lutefisk, PEAKS [133] and PepNovo [134], which use this method to extract entire or partial amino acid sequence without the use of databases. Given the difficulty of the *de novo* sequencing, the other more widely adopted method is database based searching algorithm, which compares the experimental spectra with theoretical spectra that is generated from the known proteome database and further assigns the matched spectra to specific peptides by a series of statistical calculation.

2.5.1 Database searching algorithms for peptide assignment

The prerequisite to perform database searching algorithm for peptide identification is the availability of proteome database. Although various MS/MS database search algorithms have been developed, the basic principle is similar. Here, I am going to introduce the first developed database searching algorithm: SEQUEST, and followed by several others that are used in my dissertation. The user provided database goes through the in-silico digestion process to generate the theoretical peptide list. The program takes each experimental MS/MS spectrum of a precursor ion to compare with theoretical sequences by matching the molecular weight within a defined mass tolerance. The next step is to assign the score Sp to all the candidates by using several different criteria, including the number and intensity of the predicated fragment ions that matched to the measured fragment ions, the continuity of an ion series and the presence of immonium ions for the ammo acids His,Tyr, Trp, Met, or Phe. The Sp score is calculated by following formula:

$$Sp = (\sum i_m) n_i (1 + \beta)(1 + \rho)/n_t$$

i: intensity of matching ions; n: the number of matching ions within 1Da mass tolerance; β: the level of continuity of an ion series; ρ: the presence of immonium ions for the ammo acids His,Tyr, Trp, Met, or Phe containing peptides; n_t: total predicted sequence ions.

The next step is using cross-correlation analysis to compare the top 500 candidates that are ranked by Sp score with acquired tandem mass spectrum in terms of the spectra similarity. To compare the amino acids sequence from the database search with the measured tandem mass spectrum via cross-correlation analysis, theoretical fragmentation spectra is reconstructed for top 500 candidates which contains predicted b and y ions. The experimentally obtained spectra are also normalized and noise is reduced to create measured spectra barcode-ish. In this way, predicted tandem spectra can be compared with the measured spectrum in the same barcodes style and these spectra are further assigned with the XCorr score to indicate the level of similarity. If the score is above a given threshold and significantly

better than the next best score, it is reported in final output. At the same time, there is another score that is calculated, termed DeltCn, which is used to evaluate how well SEQUEST can distinguish the best hit and the second best match. The higher this score, more confident is the match.[29] Once the spectrum is assigned with the peptide sequence, the next question is how to distinguish if it is a true match. Therefore, FDR (false discovery rate) is employed to control the false positive identifications. The calculation of FDR is based on searching the MS/MS spectra against the target reference database as well as the decoy database such as reversed database, randomized or shuffled database. FDR is calculated by using the formula [(2*decoy IDs]/(total IDs)] * 100. The FDR value is impacted by the filter cutoffs, like the XCorr.[135]

Following the development of SEQUEST, other database searching algorithms are also developed. MyriMatch and Andromeda are introduced here since they are applied in my dissertation work. Based on the similar principle by comparing the observed tandem spectra to the theoretical spectra, MyriMatch further improved the statistical model by scoring peptide matches considering intensity. The first step is called “tunable preprocessing and scoring”. In this step, the sum of fragment ion intensities (TIC) for each MS/MS is calculated and ranked in descending order. The top 95% were retained and other 5% were removed as noise. All the retained peaks are divided into user defined numbers of intensity classes based on their intensity. The basic rule is to make the most intense class holding the fewest peak, and the adjacent class sizes are in a ratio of 2:1. The second step is to generate the theoretical spectra according to the provided database, and MyriMatch employs novel system for modeling fragments by considering the charge differences. The third step is the typical comparison between acquired spectra and predicated spectra. Measured m/z value is first matched to the predicted m/z based on the instrument determined mass tolerance, coupling with the intensity class information, the probability of this match occurring by random chance is evaluated by multivariate hypergeometric (MVH) distribution. For each spectrum, the negative of log (MVH) value is reported; indicating that lower is this value, less chance it is a random match; in other word, lower is this value, higher confidence is this match.[136]

Andromeda is the database searching engine that uses a probabilistic scoring model for the scoring of peptide-spectrum matches. This searching engine was developed to improve the searching space to enable analysis of complex data sets, such as data with different modifications and labeling, in a simple analysis workflow on a desktop computer. Probabilistic scoring method was first employed in Mascot by emphasizing that the match between experimental data set to each sequence is depending on the probability. The match which has the lowest probability is regarded as the best match, and the significance is also impacted by the size of the database.[137]

Generally, it first builds the theoretical fragment ions based on the provided database, coupling with the consideration of modification and specific molecule loss, such as H_2O, NH_3. At the meanwhile, the algorithm also processes the measured spectra by centroiding, de-isotoping and transferring all other charges to charge 1. Then it counts the number of the predicted spectra that are falling into the defined mass tolerance with measured spectra. Based on the number of matching ion in spectrum and total number of theoretical ions, the probability of getting this number of matches by chance is calculated. Other information, including the intensity, peptide length, number of modifications or missed cleavages further assist the specific assignment of peptide to spectra. [138]

2.5.2 Protein inference

After database searching, peptide sequences are given the specific score metrics which indicate their levels of match with spectra; however, due to the complexity of these results, automated handling is still required to further process this immature information to get more confident peptide identification and assemble peptides back to proteins. IDPicker is one of automated handling tools that filters peptide identifications to a desired FDR by using decoy database matches and further assembles peptides to proteins. Since IDPicker is only coupled with MyriMatch for my dissertation work, how IDPicker process the results generated by MyriMatch is introduced in details here.

Basically, IDPicker filters the peptide identification based on the user defined peptide spectrum match FDR (normally 2%). In IDPicker, FDR is reported in three levels, including peptide spectrum match FDR, peptide level FDR or the protein level FDR. The most widely used FDR threshold is the peptide level FDR $\leq$ 0.01. Several parameters are known to impact the FDR value. In the peptide-spectrum-match filter, the maximum q value, minimum spectra per peptide, minimum spectra per match, and maximum protein groups can be adjusted individually. IDPicker can be configured to use either one score or "Monte Carlo weighted multiple scores" for computing the q-values. In our setting, we normally applied Monte Carlo method to calculate the q value. This method figures out the best combination of multiple scores (XCorr, MVH, and mzFidelity output from MyriMatch) to compute the q-values. The score weights of combination are figured out using a Monte Carlo simulation. MVH score is used to evaluate the probability of the match occurring by random chance, mzFidelity score is used to access the proximity of the observed and expected fragmented m/z value, XCorr reflects the quality of the match. In the protein level filters, minimum distinct peptides, minimum additional peptides, minimum spectra can be optimized individually to adjust the desired FDR value. Normally, in order to confidently identify a protein, one unique peptide and one additional peptide is required. The most important information contained in IDPicker output is the spectra counts of identified protein and peptides, which is used for further protein quantification.

2.5.3 Protein quantification

Protein identification indicates the presence of the protein in the sample and mostly based on the tandem mass spectra; but the abundance or relative abundance information is provided via protein quantification which is normally based on full-scan mass spectra. Quantitative proteomics is the approach to determine the abundance ratios for thousands of proteins between different conditions. Mass spectrometry basically records the intensity for each spectrum, and the intensity is always used for quantification. MS intensity is influenced by analyte concentration, sample recovery during sample preparation and LC, ionization efficiency and instrument sensitivity. Therefore, the MS intensity of

different analyte does not necessarily reflect their abundance difference under natural condition. To do protein quantifications, there are several different methods depending on the research goals. The first one is the absolute quantification, which takes advantage of the internal standard with known concentration, to determine the analyte's actual abundance in a sample. This method is also applicable to compare different analytes and always requires the labeling technique like AQUA (absolute quantification of proteins) for constructing the corresponding internal standard. Second method is the relative quantification, which determines the abundance ratio of the same analyte between different samples. This relative quantification method is routinely used in proteomics analysis for protein quantification. Relative quantification can be achieved by label free method or labeled method. For label free method, spectral counting, NSAF (normalized spectral abundance factors), spectral indexing, MS1 intensity, matched ion intensity, etc can be used as the quantification matrices. For labeling-based methods, there are different labeling techniques. Chemical labeling includes iTRAQ (isobaric tag for relative and absolute quantification) and TMT (tandem mass tag), metabolic labeling includes SILAC (stable isotope labeling by amino acids in cell culture) and N^{15} (nitrogen-15 metabolic labeling).

For label free quantification method which is carried out in my dissertation, I used spectral counting output from IDPicker or MS1 intensity output from MaxQuant for relative protein quantification. Spectral counting of a protein is the sum of the spectral counting of all its peptides, which refers to the number of MS/MS spectra assigned to each peptide. In DDA analyses, more abundant peptides are sampled more often than lower abundance ones. Although it is an easy-to-apply method and also shows high correlation to the relative protein concentration when the protein has more than two fold changes, the drawbacks of this quantification method are obvious: 1) peptide ion properties are not the inherent characteristics of the peak recorded by mass spectrometry, thus potentially discarding other important information of a peak; 2) measurement can be easily affected by the level of competition with other ions for DDA selection within or across different samples and by the setting of dynamic exclusion methods; 3) saturation effects are observed for the spectral counting method; 4) the quantification power is weak for

low abundant proteins.[19, 41, 139] Therefore, MS1 intensity method is superior in terms of the quantification power to the spectral counting method by avoiding all the drawbacks listed above. With the development of high resolution instrument and advanced separation skills, peptides currently can be resolved well and the information of extracted ion currents (XICs) of peptides can be obtained, which is critical for accurate MS1 intensity based quantification. MaxLFQ is a method developed by Dr.Matthias Mann's group to achieve more accurate label free quantification for shotgun proteomics analysis. They employed a new method to normalize the differences originating from different fractions and developed a new method to extract maximum peptide ratio for protein quantification. Previously, there are always arguments on the selection of proper peptides intensity that can represent for the proteins. The simplest way is to sum up all peptide signals for each protein and compares the protein ratio. Alternative methods include only using the top n abundant peptides for protein quantification or using the average to represent the protein intensity. However, these methods discard the ratio information carried by individual peptides. Actually, the signal ratio of one peptide should be able to represent for the ratio of protein. To avoid this drawback, there is another method which selects peptide that is always identified across samples; however, with the increase of the sample numbers, the chance that one peptide is always identified decreased dramatically. Therefore, they proposed a new method to conquer these limitations. Generally, this LFQ method calculates each peptide ratio in pair-wise, and requires at least 2 pair-wise ratios to determine the final intensity of the proteins. In other words, across all the samples, the quantifiable proteins need to have at least two shared peptides that are identified in at least two samples. After the pair-wise ratio is calculated for each protein, the pair-wise protein ratio is defined as the median of the peptide ratios. Then a least-squares analysis is performed to reconstruct the abundance profile according to individual protein ratio and assign intensity to each protein. The method is proved to be much accurate than other label free quantification methods.[140]

2.5.4 Statistical methods for multiple comparisons in proteomics data analysis

High throughput proteomics experiment usually generates a large scale of dataset, which includes abundance information of thousands of proteins under different conditions. Given that the observed differences can come from sample variances, experimental variances and instrumental variances, appropriate statistical methods are required to evaluate high-dimensional data and prevent false conclusions. [141]

Before applying the statistical methods, data pretreatment is necessary due to two reasons: 1) technical variations introduced during sample preparation and MS measurement should be corrected 2) shotgun proteomics data is always skewed on the intensity distribution of all identified proteins, but statistical tests assume Gaussian distributions of the dataset. Date pretreatment generally includes two critical steps, one is normalization, and the other one is transformation. Normalization is required to adjust the variations coming from systematic biases which are caused by nonbiological reasons, thus making samples more comparable to each other. Different normalization methods for proteomics data analysis have been compared in a recently published paper.[142] Additionally, in order to meet data distribution assumptions of statistical tests, log transformation is always applied for proteomics data to adjust the date distribution.

One common goal of proteomics study is to simultaneously compare thousands of proteins under different conditions, thus searching for those differentially abundant proteins. In this case, a significance test is carried out for each single protein.[143] The statistical approach consists of a null hypothesis that this protein is not differentially expressed and an alternative hypothesis that this protein is differentially expressed. Based on the probability distribution of the dataset, we can decide each protein either for null hypothesis or alternative hypothesis. α is the significance level or probability level that is used for deciding to accept or reject the null hypothesis. Two levels of significance are commonly used in statistics: $\alpha=0.05$ or $\alpha=0.01$. There are two probability values: α is for controlling the false positive and β is for

controlling the false negative, these two probability values are dependent on each other. When α is set as 0.05 or 0.01, β is controlled by choosing adequate sample size.

However, the statistical power of significance test applied for proteomics data analysis suffers from one common feature of proteomics data which contains a large number of simultaneous measurement but with limited number of samples.[144] Thousands of proteins are tested simultaneously which cause the possibility that observed differences (found by setting the significant threshold at 0.05) can solely result from these multiple hypothesis testing, but not real causative correlations.[143] We know that the probability of false positive discoveries can be greatly increased by multiple testing, so the traditionally acknowledged significance level 5% is drastically magnified when we compare thousands of proteins at the same time. For example, if we conduct the significance analysis to 1000 identified proteins simultaneously by setting the significance level 0.05, 50 (1000*0.05) proteins can be found as significant even without any real biological changes. Therefore, appropriate multiple testing correction method has to be applied to avoid over-inflating the results.

There are several multiple testing methods available, and selection of the proper method is largely depending on the data and the research goal. [145] Some correction methods can be very stringent, while others can be comparatively loose. Generally, in any hypothesis test, we compare the null hypothesis with the alternative hypothesis and set the threshold to decide to accept or reject the null hypothesis. There are two types of errors associated with this hypothesis test: type I error denotes rejecting the null hypothesis when it is true (false positive), while type II error means accepting the null hypothesis when the alternative hypothesis is true (false negative). Type I error can be terribly increased in multiple tests without any correction.

Since thousands of proteins are tested simultaneously, instead of controlling α and β for each single test, the error rate has to be controlled considering the multiple testing procedures. Family-wise error rate (FWER) or the false discovery rate (FDR) are the two common error rates that are applied in

"omics" study. The FWER refers to the probability to have at least one false positive among all the tests. In order to control FWER ≤ α, we need to apply the multiple testing corrections to the significance level. For example, Bonferroni-adjusted significance level α/m can be used for each test of single protein. In this method, adjusted *p*-value, which is theoretical α/(number of comparison), is used to set the threshold of significance. That is to say, for 1000 identified proteins, the significantly changed protein has to have the *p*-value below than 0.05/1000. It is obviously too stringent for most of the proteomics data, especially for those experiments which aim at looking for proteome changes in response to environmental stimuli. Obviously, this correction method dramatically increases the type II error; but it is sometimes applied for establishing strong biomarkers.

FDR is another correction method proposed by Benjamini-Hochberg in 1995[146] that better controls the number of rejected hypotheses than FWER, and thus commonly applied in proteomics analysis. Unlike FWER, which gives out the specific corrected *p*-value as the significance level, FDR controls the ratio of numbers that are false positives among all numbers that are accepted as alternative hypothesis. In Benjamini-Hochberg FDR, unadjusted *p*-value is first ranked from smallest to largest: $pr_1 \leq pr_2 \leq \cdots \leq pr_m$. If controlling the FDR at the level of 0.05, it means that each p-value has to be compared with its corresponding Benjamini-Hochberg critical value (i/m)Q. The i denotes the rank number, m is the total number of test, Q is the false discovery rate we choose (0.05). For example, pr_1 needs to be compared with (1/m)*0.05, and all following p values are compared with their corresponding Benjamini-Hochberg critical value in the same way until one *p*-value (pr_j) is found to be larger than its corresponding Benjamini-Hochberg critical value. In this case, all the proteins that have the *p*-value ranked ahead of pr_j are regarded as the significantly changed proteins. [147] Note that false discovery rate is not necessary to be set at 0.05 or 0.01, especially dealing with a very large number of statistical tests. Basically, selection of a proper FDR is a choice between false positives and false negatives: smaller the FDR, better control the false positives, but increasing the number of false negatives. So the research goal has to be considered before choosing the FDR. For example, if the proteomics experiment is at the exploratory step of the

whole experimental design, aiming at providing more candidates for further experiments, and if the cost of the further experiments is not high, but the cost of false negative is very high, we should go for higher FDR value (0.1 or 0.2)[148] in order to capture as many candidates as we can. Of course, under this situation, we also allow more false positives in our final differentially expressed protein list.

This multiple correction has to be applied for proteomics data analysis vertically even under two different conditions, since large number of proteins is simultaneously compared. Horizontally, we need to apply significance analysis for each single protein to detect significant changes across the conditions of interest. If we only have two conditions, the two-sample t-test is commonly used for this binary comparison. The null hypothesis of this t-test is that the mean of two groups is the same. H_0: $\mu_1 - \mu_2 = 0$. We can define the significant level $\alpha = 0.05$ or $\alpha = 0.01$. Note that the two-sample t-test is only applied for data that follow the normal distribution; this is the reason that proteomics data is always log transformed. P-value is the output of the t-test which indicates the probability of observing the difference between two means is random. But if we have more than two conditions, different methods have to be considered depending on how the comparisons are carried out. As we discussed above, doing pairwise comparison over and over again for all possible pairs is not working since the type I error is significant increased. Therefore, Bonferroni correction or Benjamini-Hochberg FDR methods can be used if the multiple tests are independent of each other, but if the comparison of different conditions is dependent on each other, then ANOVA is the proper method to be used. Just like two-sample t-test, ANOVA also deals with the differences between sample means, but with more than two groups to be compared. ANOVA tests the null hypothesis that H_0: $\mu_1=\mu_2=\mu_3=\ldots\ldots=\mu_k$. , where μ is the mean of the group and k is the number of the groups. The alternative hypothesis is that the means are not all the same. In other words, if there is any two groups means are significantly different, then ANOVA returns a statistically significant result. ANOVA can only reveal that at least two groups that are significantly different, but will not tell us which group is different from which group. Therefore, ANOVA is regarded as a preliminary test that informs us if we should continue the investigation of the data. Subsequent analysis by post hoc test, such

as t-tests on all pairs, Bonferroni, Least significant difference (LSD) or Tukey's Honestly Significant Difference (HSD) test, etc. is employed after ANOVA to investigate which of the groups of samples were different from each other. The advantages and disadvantages are all compared in another paper. [149] Unlike the t-test, which uses means to evaluate H_0, ANOVA uses the F test that is based on the ratio of variances. There are two types of variances calculated in ANOVA: one is between group variability, and the other one is within group variability. So the basic idea is to calculate the means within each group, then compare the variance among these means to the average variance within each group. When the difference among group means is larger than variance within each group, the null hypothesis is rejected. It can be expressed by the formula: $F = \frac{Variances\ among\ group\ means}{Average\ variance\ within\ groups}$. In the normal output of ANOVA test, we also observed the values of two degrees of freedom (d.f.). One is for the degrees of freedom among groups, which is calculated by group number minus one. The other one is the degrees of freedom within groups, which is calculated by total number of observations minus the number of groups. Finally, the corresponding *p*-value is also reported.

ANOVA can be used in three ways: one-way ANOVA, two-way ANOVA and N-way Multivariate ANOVA. One-way ANOVA denotes only one independent variable, whereas two-way ANOVA involves two independent variables. In more detail, when we are interested in measuring the outcome/response on the same independent variable, one-way ANOVA can be applied. In other words, one-way ANOVA is carried out for more than 2 groups of interest, but aiming at investigating one variable. This independent variable should be normally distributed in each group that is being compared. For example, one-way ANOVA can be used to understand whether weight can lead to certain disease among a group of people. People thus are divided into three groups according to their weight. Therefore, one-way ANOVA can only be used to evaluate one variable to one outcome among three or more groups of interest. Two-way ANOVA is dealing with two independent variables and is used to investigate the interactions between these two independent variables on the outcome/response. For example, two-way ANOVA can be used to understand whether there is an interaction between age and gender on certain

disease amongst a group of people. In this case, age and gender are two independent variables, and certain disease is the outcome. N-way Multivariate ANOVA is using the same principle as the two-way ANOVA.

In conclusion, for significance analysis of proteomics data, proper multiple testing methods should be employed to avoid over-inflating the results. Proteomics data can be viewed in two dimensions: horizontal data (row) and vertical data (column). We and others normally apply the two-sample t-test [113, 150-153] or ANOVA[154-156] for significance analysis for each single protein under different conditions. As we discussed above, the selection of two-sample t-test or ANOVA is not simply relying on the number of experimental groups. The easiest situation is that only two groups of interest are compared, two-sample t-test is performed under this situation. If more than two groups of interest are present, we should first look at what is the independent variance and how many independent variances existing among all the groups. If there is only one independent variance among all the groups, one-way ANOVA can be performed. If there are two independent variances among all the groups, and the interest is to investigate the interaction of two variables on the outcome, two-way AVOVA can be applied. If there are two or more than two independent variances among all the groups, but no interest on the interaction of these variances, we should regard them as independent comparison, thus pairwise t-test with proper correction can be performed. Since thousands of the proteins are compared simultaneously, although independently, the error rate has to be controlled by either family-wise error rate (FWER) or the false discovery rate (FDR). Correction methods for controlling this type of error due to the multiple comparisons are necessary for any proteomics' data, no matter how many conditions are compared.

CHAPTER 3

Development of An Optimized Extraction Method to Remove Humic Acid Interferences from Soil Samples Prior to Microbial Proteome Measurement

Text and figures were adapted from:

Qian, C.; Hettich, R. L.; An Optimized Extraction Method To Remove Humic Acid Interferences from Soil Samples Prior to Microbial Proteome Measurement. J Proteome Res 2017(PMID: 28537741)

Chen Qian's contributions included: Experimental design, sample preparation, mass spectrometry measurement, data analysis and manuscript writing and revision

3.1 Introduction

A handful of soil can contain up to 10 billion bacteria and tens of thousands of species, which makes soil one of the most complicated ecosystems to study. [72] These micro-organisms are not passive members, but in fact play important roles in organic matter transformation and nutrient cycling. The complex interaction among these diverse organisms impact soil characteristics, and therefore can impact larger-scale aquatic and atmospheric environments. [157-160] More specifically, soil microbes are known to help remediate toxic metal contamination, as well as affect the carbon transformation and nitrogen-cycling. [161-162] Thus, characterization of microbial communities and their metabolic activities in soil is necessary for understanding soil ecosystems and their impact on local and global natural processes. Currently, a variety of experimental assays are available to measure specific soil microbial extracellular enzyme activities in the presence of the soil interference.[163] However, these assays are limited to enzymes and, more importantly, cannot provide large-scale, untargeted characterization of the range and nature of microbial functionality in natural environmental systems. This is the potential of soil metaproteomics, which aims to investigate the entire set of proteins belonging to soil microorganisms, and thus could become a key approach to facilitate the study of the biological diversity, activity, and functionality of the soil microorganisms at the system-interrogation level.[164]

Currently, shotgun proteomics via multi-dimensional liquid chromatography coupled with tandem mass spectrometry is the most common approach for large-scale characterization of metaproteomes.[164] However, the limited microbial biomass and the presence of abundant interferences in some soil samples present major challenges to the mass spectrometry measurement.[165] The advent of high resolution mass spectrometry and sophisticated data acquisition modes, such as DDA (data-dependent acquisition) or DIA (data-independent acquisition), have enabled moderately deep proteomic measurements. [166] Nevertheless, these techniques require relatively "clean" samples, in which there is detectable levels of proteolytic peptides with minimum interferences that impede consistent and high quality mass spectrometric results. Thus, even high-end MS instrumentation platforms are dependent on removing interferences from soil samples before mass spectrometry measurement.

Soil contains a vast array of interferences that can impact protein extraction and subsequent mass spectrometry measurement. One of the most common and problematic interferences in soil samples are humic substances (HS), which are heterogeneous organic matter produced by microbial metabolism and consist of a range of aromatic and aliphatic compounds in a complex array of molecular sizes and structures. [167] Humic substances are divided into three main fractions: humin, humic acids (HA) and fulvic acids (FA). These three fractions of humic substances can be separated by differential solubility: humin cannot be dissolved in either alkali or acid solvents, while the fulvic acids are soluble in a wide range of pH conditions; humic acids are insoluble in strong acids but are soluble in alkali conditions.[168]Given the insolubility of humin fraction, extractable humic acids and fulvic acids are the main humic interferences that impede protein characterization.

Humic substances are known to interfere with protein extraction, purification, and identification / quantification.[169] For instance, studies have shown that the hydrophobic interactions between humic substances and protein can lead to unwanted modifications, thus hampering the correct identification of certain proteins; [169-170] humic acids can also suppress the electrospray ionization of other analytes [171] and humic acids contain readily ionizable low molecular weight fractions (below ~ 2000Da) that can be easily

detected in ESI mass spectrometry.[172] Additionally, we and others have found that peptide samples containing a notable amount of humic acids are difficult to load onto LC trapping columns, often leading to flow obstruction. Considering the limited microbial biomass in some soil samples, it is often necessary to load the entire sample for a single MS to yield satisfactory signal intensity in the mass spectra, thus making the clogging issue a high risk for sample loss. Therefore, an unbiased separation of peptides from humic material is regarded as a key step during the soil microbial proteome extraction process.[71]

Phenol is the most well-known reagent used to clear humic substances from the protein samples by dispersing humic substances and proteins into two different phases (water phase / phenol phase); however, even without considering the toxicity of phenol, protein recovery has not been found to be very high [71, 173-174] and the use of phenol to remove humic acids is detrimental for highly complex samples.[175] Recently, cell lysis and protein extraction methods coupling sodium dodecyl sulfate (SDS) containing lysis buffer and trichloroacetic acid (TCA) precipitation has become a commonly used method applied to soil samples.[176-177] Although TCA precipitation is expected to remove a portion of the humic substances,[178] this is often not sufficient. Since acid precipitation is known to separate fulvic acids from humic acids based on their different solubility,[179] TCA precipitation does help remove fulvic acids from the protein pellet. However, humic acids that remain insoluble under acidic conditions will be co-precipitated with proteins (the pH of 20% TCA lysis buffer is around 0~1). Although subsequent C18 Sep-Pak solid phase extraction can help remove the remaining humic acids materials, we have found this step to result in a significant decrease in final peptide concentrations, likely due to competitive binding of the humic acids vs. peptides for the reversed-phase resin, thus limiting the effective enrichment of the peptides. This loss would have a severe impact especially on some soil samples, many of which have limited microbial biomass. Therefore, an optimized method that efficiently removes humic substances yet has minimal effect on the concentration and content of extracted proteins is highly desirable.

In this study, given the lack of proper metaproteome database of the soil, and in order to avoid the confound results brought in by the uncontrolled and undesirable variables of soil background; we are not

aiming to investigate any changes regarding to soil endogenous proteins which are previously existing in the soil. Therefore, the objectives of this study were: (i) to investigate how humic acids impact large-scale peptide identification in shotgun microbial proteomics measurements, (ii) to evaluate a novel approach to remove humic acids, and (iii) to examine whether this new approach impacts protein identification or abundance profiles. Our modified approach employed a pH adjustment to precipitate out the remaining humic acids followed by subsequent removal via a 10kDa molecular weight spin column filter to remove the interfering humic acids from the final proteolytic peptide sample. Our results indicated that this humic acids removal step did not negatively affect protein identification or abundance profile, thus successfully providing an unbiased soil proteome characterization after removing humic acids.

3.2 Materials and Methods

Microbial Strains and Culture Conditions

Simple microbial isolate: *Pseudomonas putida* F1 was grown overnight aerobically at 32 °C with vigorous shaking (200 rpm) in 50 mL Luria−Bertani (LB) medium to the stationery phase (OD ≈ 1). The cell culture was then vortexed and 5 mL aliquots from liquid medium was transferred into separate, new 50 mL Falcon tubes (VWR). Three subsequent samples were prepared: 1) control sample (microbial solution only), 2) humic acids sample (microbial solution plus addition of 1mg commercial humic acids), and 3) soil sample (microbial solution added to 5 g topsoil in a 50 mL Falcon tube). Both the inoculated soil and the liquid medium (LM) were further incubated for 3 hours at 32 °C, without shaking.

Four microbial mixture: All the four microbial mixtures related experiments were processed after the mass spectrometry measurement of simple microbial isolate. Thus, all the incubation processes were separated from the simple microbial isolate incubation described above. *Pseudomonas putida* F1, *Pseudomonas putida* KT2440, *Shewanella oneidensis* MR-1 and *Shewanella putrefaciens* CN-32 were grown separately under the same incubation conditions as described in simple microbial isolate. Three subsequent samples were prepared: 1) control sample (mixture of four organisms); 2) humic acids

samples (microbial mixture plus addition of 1mg commercial humic acids); and 3) soil samples (microbial mixture with 5 g topsoil). Both the inoculated soil and the liquid medium (LM) were further incubated for 3 hours at 32 °C, without shaking. To avoid unwanted competition for survival, these four organisms were not incubated together or added into a same soil falcon tube, they were incubated separated and later combined together. Given the easy accessibility of the gram-negative soil bacterial strains, only gram-negative bacteria, but no gram-positive bacteria, were used in this study. Although the different cell structure of gram-positive and gram-negative bacteria leads to differences on the cell lysis efficiency; there is no difference on protein precipitation by TCA method. Since our method did not optimize the cell lysis step, but focused on the peptide clean-up step after protein precipitation, this method should be widely applicable for all the protein samples originating from different biological sources.

Spiking of standard peptide mixture with humic acids

Proteolytic peptides of a six microorganism "artificial" mixture (*Ignicoccus hospitalis* KIN4/I, *Nanoarchaeum equitans* Kin4-M, *Escherichia coli* K12-W3110, *Clostridium thermocellum* LQ8-DSM 1313, *Clostridium thermocellum* LQ8-DSM 1313, *Sc strain* W303) were previously prepared by an SDS/TCA procedure [180] and stored in -80 °C. This peptide mixture is routinely used as a standard sample in our lab to evaluate mass spectrometry performance. Given the general availability and reasonable complexity, this peptide mixture was thus used to evaluate the impact of humic acids on large-scale peptide identification. This sample will be referred to as "standard peptide mixture" in this manuscript. Different ratios of humic acids to the standard peptide mixture were tested. 12.5 µg, 25 µg and 50 µg commercially available humic acid standard (Aldrich, Sigma) were added into 25 µg of the standard peptide mixture solution separately and then incubated overnight at room temperature.

Measurement of humic acids clearance by pH-induced precipitation and filtering

Three tubes were prepared, with each containing 1 mg humic acids dissolved in 1 mL 100 mM Tris-HCl buffer (pH=8). Formic acid was added to adjust the pH of the solution. Since humic acids cause

color interference on pH indicator strips, all the pH measurements were processed with the same volume of 100 mM Tris-HCl buffer added with same amount of formic acid without humic acids. After 10 minutes of equilibrium, all the tubes were filtered by 10 kDa cutoff spin column filter (Vivaspin 2, GE Health, Pittsburgh, PA) at 4500 x g for 15 minutes. The content of humic acids in the flow-through was measured by UV−vis spectrophotometer (Evolution 201, Thermo Scientific, Waltham, MA, USA). Samples were diluted 5 folds before measurement to reach the instrument sensitivity range. Due to the various structures of the humic acids, the 250 nm to 700 nm wavelength range was scanned and absorbance at each wavelength was recorded.

Protein Extraction and Digestion

After 3h inoculation, cells that were in liquid medium were pelleted by centrifugation at 2850× g for 10 min, and the supernatant was discarded. 1 mL lysis buffer (100 mM Tris-HCl, pH 8.0, 4% w/v sodium dodecyl sulfate (SDS) and 10 mM dithiothreitol (DTT)) was added into the control sample and humic acids sample individually. The two tubes were then boiled for 5 minutes and further disrupted by sonication (10 s on, 10 s off) for 2 min with an ultrasonic disruptor (Branson) at 20% amplitude. To each tube, chilled 100% tricholoroacetic acid (TCA) was added to achieve 20% TCA final concentration (to achieve a pH of ~1) and the sample was vortexed briefly and kept in -80 ˚C freezer overnight. The samples were then allowed to thaw and precipitated protein was pelleted by centrifuging the samples for 15 min at 21000 x g at 4˚C. Protein pellets were then washed twice by ice-cold acetone. Samples were allowed to air dry and the protein pellets were re-solubilized in 8 M urea 100 mM Tris-HCl pH 8.0 with 5 mM DTT by incubation at room temperature for 30 minutes. 20 mM iodoacetamide was then added into the protein solution and the sample was allowed to incubate in the dark for 15 minutes at room temperature. Two separate aliquots of 20 μg sequencing grade trypsin (Promega, Madison, WI) were added into each sample and then incubated at room temperature for 3 hours for the initial digestion period, and then overnight for the second time. For each time of digestion, trypsin was dissolved in the same volume of 10mM $CaCl_2$ tris buffer as the sample volume, thus urea was diluted to 4M for the first time

digestion and 2M for the second time digestion, Following digestion, the control sample was adjusted to 200 mM NaCl and 0.1% formic acid and filtered through a 10 kDa cutoff spin column filter (VWR brand). The sample with added humic acids was adjusted to 200 mM NaCl and 1% formic acid (~pH 3) and filtered through a 10 kDa cutoff spin column filter (Vivaspin 2, GE Health,Pittsburgh, PA) at 4500 x g for 30 minutes to remove the humic acids. The peptide enriched flow through was quantified by BCA assay and then stored at -80 ˚C freezer until analysis. For soil samples, 10 mL lysis buffer was added into the 5 g soil samples which contained the microbial inoculants.[176] The soil was vortexed and boiled in a water bath for 10 minutes. The samples were then centrifuged at 3200 x g for 10 minutes and the supernatants were collected. For each sample, TCA precipitation and sequential digestion steps were processed the same way as described above. In the final clean-up step, soil samples were adjusted to 1% formic acid and filtered the same way as humic acids samples. A schematic illustration of the workflow to analyze the soil metaproteome by incorporating the newly developed humic acids removal approach is present in the Figure 3.1.

Peptide Measurement by 2D-LC-MS/MS

25 μg of proteolytic peptides were loaded and washed as previously described.[50] LC-separated peptides were analyzed via 22-hr MudPIT [33] 2D-LC-MS/MS on different mass spectrometer according to the different sample complexity. Control samples, humic acids samples and soil samples containing simple microbial isolate were all run on a linear trapping quadrupole (LTQ XL) mass spectrometer (ThermoFisher Scientific, San Jose, CA). Control samples, humic acids samples and soil samples containing four soil microbial mixture were all run on a hybrid LTQ Orbitrap XL mass spectrometer (ThermoScientific). Standard peptide mixtures with/without addition of humic acids were run on a hybrid LTQ Orbitrap Elite mass spectrometer (ThermoScientific). All the data was acquired in the data-dependent mode: the top 5 abundant precursor ions were selected for MS/MS fragmentation by collision-induced dissociation (at 35% energy level) in the LTQ-MS; the top 10 were selected in LTQ Orbitrap XL-MS, and the top 20 were selected in LTQ Orbitrap Elite-MS. In the LTQ Orbitrap hybrid instrument,

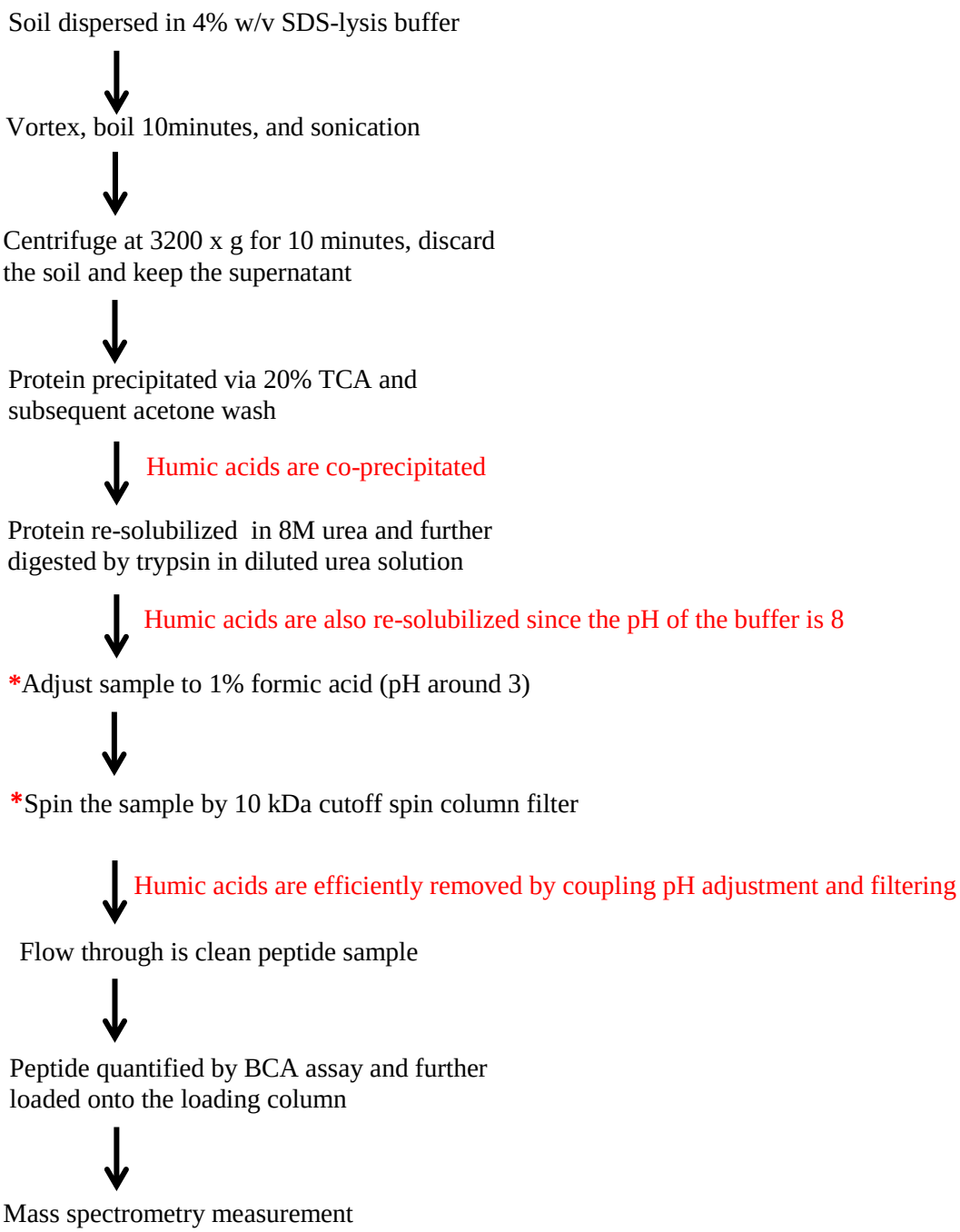


Figure 3.1. A schematic illustration of the workflow to analyze the soil metaproteome by incorporating the newly developed humic acids removal approach. The two steps marked with * are key steps to remove humic acids from the peptide solution.

precursor ions were measured in the Orbitrap analyzer with high resolution (15,000 at m/z 400 in Orbitrap XL, 30, 000 at m/z 400 in Orbitrap Elite).

Peptide and protein identification from LC-MS/MS

The predicted proteome databases for the specific organisms used here were all downloaded from JGI's Integrated Microbial Genomes server (the most current versions were downloaded). All the MS/MS spectra were searched by the Myrimatch v2.1 algorithm [136] against appropriate FASTA proteome database to be assigned to specific peptide sequences. For the simple microbial isolate, all the spectra from control, humic acids, and soil samples were searched against *Pseudomonas putida* F1 proteome database, with common contaminant proteins (e.g., keratin and trypsin) included. For the microbial mixture, all the spectra from control, humic acids, and soil samples were searched against the concatenated proteome database which included all the proteome information of four organisms and common contaminants. For standard peptide mixtures with/without addition of humic acids samples, all the spectra were searched against the concatenated proteome database which included all the proteome information of six organisms and common contaminants. The configuration parameters set for LTQ mass spectrometry measurement are: fully tryptic peptides with maximum 2 miscleavages, an average precursor mass tolerance of 1.5 m/z, and a fragment mass tolerance of 0.5 m/z. The modifications were set as: static carbamidomethylation modification (+57.0214) to cysteine, dynamic carbamylation modification (+43.0058) to N-terminal and oxidation modification (+15.9949) to methionine. The configuration parameters set for LTQ Orbitrap MS measurement were as follows: fully tryptic peptides with any number of miscleavages, an average precursor mass tolerance of 1.5 m/z for precursor ions with no charge state or a monoisotopic precursor mass tolerance of 10 ppm for precursor ions with a determined charge state, and a fragment mass tolerance of 0.5 m/z. Modification was set the same as in LTQ mass spectrometry. Peptide filtering and protein assembling was processed via IDPicker v3.0 [181] with peptide-level FDR controlled below 1%.

3.3 Small amounts of humic acids in proteome samples adversely affect the peptide measurements

After TCA precipitation, humic acids are the main humic interfering component that is co-precipitated with the proteins. Previous humic acids clean up approach based on sep-pack was found to cause significant sample loss for humic acids contained samples (Figure 3.2). Therefore, our optimized clean-up approach was designed to specifically remove humic acids from proteolytic peptide solutions and to avoid using the sep-pack approach. A comparison between a control sample (well-characterized microbial inoculants) and a humic acids sample (well-characterized microbial inoculants containing commercial humic acids) was conducted to evaluate how the humic acids removal step impacts protein identification and quantitation. Given the varied abundance of humic acids in different types of soil [182-183], a 1 mg humic acids addition was used in each sample as a reasonably high amount, considering that 5 g soils are normally used for soil microbial proteome extractions. In order to assess the feasibility and robustness of this modified method for diverse samples, the effects of humic acids removal on peptide identification were examined on two sample types: a simple microbial isolate and a four microbial mixture. In each sample, these microbial inoculants were also spiked into the normal topsoil to investigate the protein recovery from natural humic-enriched soil.

Although the effects of humic acids on the identification of targeted proteins via mass spectrometry have been examined before, there is limited information on how humic acids impact large-scale proteome identifications via a shotgun proteomics technique. In unpublished work in our lab on a variety of soil samples, it was observed that the normal protein extraction method works well for some types of soil,[176] but is quite ineffective for others, especially for soil samples containing high levels of humic substances. Thus, we aimed to design an optimized method to minimize the negative impact of humic substances interference while retaining the extractable peptides.

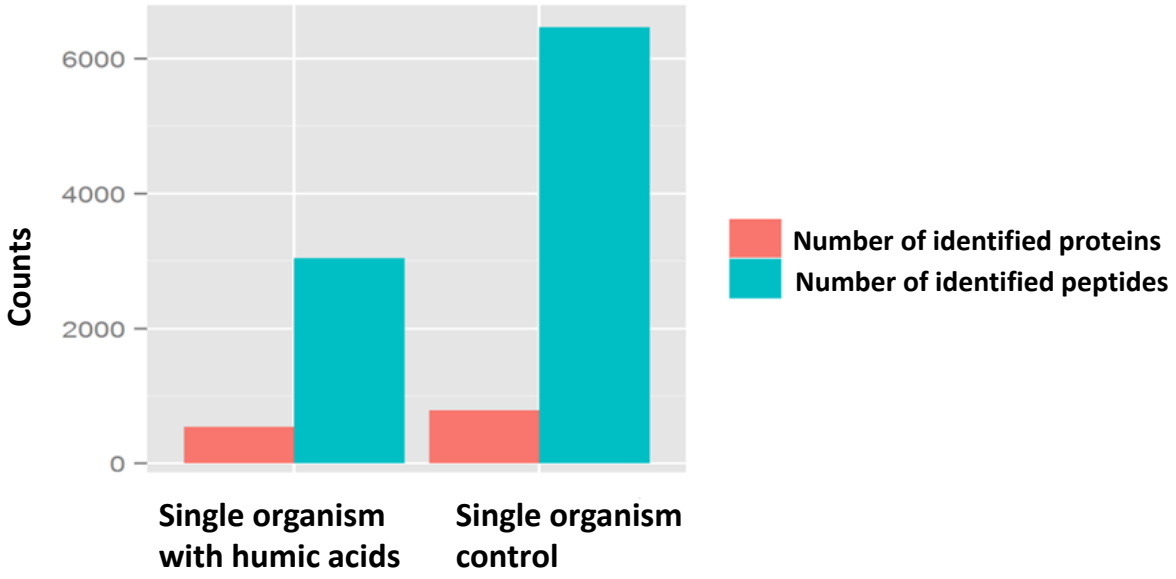


Figure 3.2. Depiction of how Sep-pack approach impacts protein/peptide identification. Control sample (microbial inoculant Pseudomonas putida F1), humic acids sample (microbial inoculant Pseudomonas putida F1 added with humic acids) were prepared by using previous protocol which used sep pack to remove humic acids from the peptide solution. All the samples were measured in technical replicates. The value presented in the figure is the averaged value of technical replicates in each condition. Results showed that the number of identified peptide/protein dropped significantly by applying this sep-pack approach.

Initial work was focused on an experiment designed to identify the minimum amount of humic acids that would cause significant impact on peptide identification; however, the task of loading humic acids containing peptide samples onto a chromatographic back column is very challenging, in that humic acids can easily cause the clogging in the column. Thus, the goal was modified to examine if a loadable amount of humic acids would impact peptide identification. Due to the complexity of humic acids, it is difficult to quantify their residual amounts remaining in the peptide samples after applying differentextraction and digestion methods. Therefore, different ratios of humic acids to the standard peptide mixture of the six microbes (1:2, 1:1, 2:1), were tested. In this case, the loading process for all humic acids contained samples was very slow and problematic, with several of LC columns clogging immediately after beginning the loading procedure, especially the sample containing 50 µg humic acids. We speculate that the clogging was probably due to the overloading of humic acids since they are very hydrophobic and will bind to reversed-phase (RP) resin as well, especially since the RP resin binding capacity is limited. 25 µg humic acids were eventually chosen as the maximum loadable amount with our standard sample loading procedure (pressure cell ~1000 PSI).

In general, the mass spectrometry measurement revealed that this small amount of humic acids did decrease the identified peptide number slightly (~ 8%)(Table 3.1). To further investigate the reason for this identification decrease, the charge state distribution of all the precursor ions that were recorded in MS1 scans were examined (Figure 3.3). Charge information of all precursor ions was retrieved by Raxport program.[184] The results showed that this small amount of humic acids suppressed the detection of both high charged (+5, +6) precursor ions and +1 charged ions, while the +2 and +3 charged precursor ions, which have the highest probability to be assigned as tryptic peptides, were not affected as much. In terms of the identification, there was around 22% decrease in identified peptides that were +1 charged, and around 38% decrease in identified peptides that were +5 or +6 charged in the humic acids containing samples (Table 3.1). However, since the majority of identified peptides consisted of +2 or +3 charge states, this small amount of humic acids did not significantly impact

Table 3.1. Overview of the mass spectrometry measurement results of standard peptide mixture & standard peptide mixture with 25 µg of humic acids added to each sample.

Summary	Proteins	Peptides	Assigned Spectrum
6_organisms_mixture_run1	3947	23099	156057
6_organisms_mixture_run1	3715	21620	151245
6_organisms_mixture_HA_run1	3619	20673	141320
6_organisms_mixture_HA_run2	3747	20750	141122

(a)

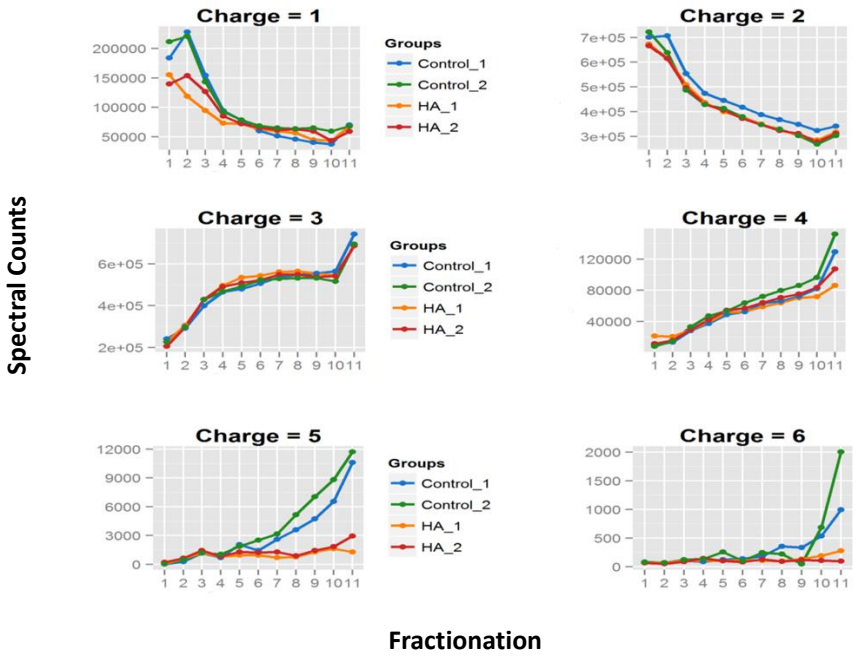


Figure 3.3. The impact of humic acids on standard peptide identification was investigated as a function of their charge and hydrophobicity. Control samples are 25µg standard peptide mixture. HA samples are 25µg standard peptide mixture supplemented with 25 µ g humic acids. (a) The charge information of all detectable precursor ions was retrieved by Raxport program and plotted in R. The number 1 to 11 in the x axis refers to the 11 fractionations generated by MudPIT based on different concentrations of ammonium acetate. The original concentration of stock ammonium acetate is 500mM, and the 11 fractionations were produced by eluting: 5%, 7%, 10%, 12%, 15%, 17%, 20%, 25%, 35%, 50% and 100% of stock ammonium acetate, individually.

(b)

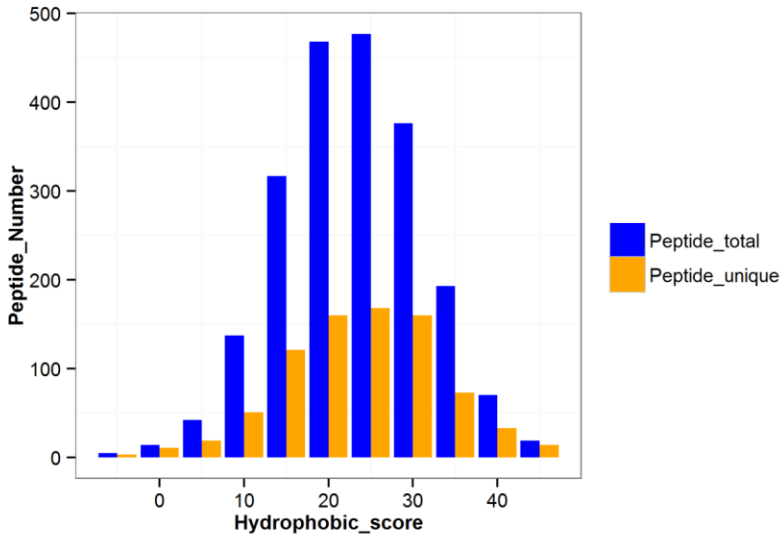


Figure 3.3. Continued (b) The hydrophobicity distribution of peptides (+1 charge) that were identified in control samples (peptide_total), and peptides (+1 charge) that were unique to the control samples (peptide_unique) was compared. Peptides that were unique to the control samples, representing the group of peptides which were impacted by the existence of humic acids since they were not identified in humic acids contained samples.

the whole proteome identification. This biased impact of humic acids on detection and identification of +1, +5 and +6 ions, was likely due to competitive binding of the humic acids with peptides to reversed-phase resin, impacting detection of the least abundant species (in this case, peptides observed with +1, +5 or +6 charges) most dramatically.

Considering the very low amount of identified peptides with +5 or +6 charge, we chose to examine whether humic acids would suppress singly-charged peptides as a function of their hydrophobicity. To this end, we compared the hydrophobicity distribution of *total* peptides identified in control samples relative to peptides *unique* to control samples (peptides unique to control samples were identified by comparing peptides in control samples and humic acids contained samples, whereas those peptides that were not identified in the humic acids samples were classified as peptides unique to control samples). Hydrophobic scores of identified peptides were calculated by in-house written peptide hydrophobicity calculator. The hydrophobic score is determined by the sum of hydrophobic values of all amino acids divided by the peptide length. Thus, the higher the score, the more hydrophobic is the peptide. The result revealed that humic acids suppress peptides with a wide range of hydrophobicities (Figure 3.3), as might be expected considering the wide range in complexity of the humic acids pool. Since "humic acids" is not a pure compound, different molecular weight humic acids have different hydrophobicities and thus elute at a range of different retention times in an LC run,[185] they present a range of ionization competition with peptides, in addition to the binding competition for reversed-phase resin. The wide ranging hydrophobicity of humic acids also likely impacts the biased effect of humic acids on the detection and identification of highly charged ions. More specifically, the fraction of humic acids with high molecular weight bound irreversibly to the RP resin,[185] will increase the hydrophobicity of the RP resin, thus making the elution of more hydrophobic peptides (higher charged peptides normally are more hydrophobic) challenging.

Overall, our results revealed that humic acids impact the peptide detection and identification, but are biased with respect to peptide charge state, suggesting that besides the ion suppression effects

previously reported,[171] it is likely that humic acids negatively impact the interaction between RP resin and peptides, either by competing for binding or hindering the elution. Additionally, considering that it is difficult to quantify peptides in the real soil samples containing humic acids (due to spectroscopic interferences),[186] and even small amounts of humic acids can easily clog loading column, removing humic acids from the final peptide solutions is highly desirable and enhances sample loading as well as mass spectrometry measurements, thus helping to enhance an unbiased characterization of the whole proteome.

3.4 Humic acids can be efficiently removed by a combination of pH adjustment and filtering

The complications outlined above prompted a need to implement and optimize a straight-forward approach to remove humic acids prior to proteome measurements. Humic acids share similar characteristics with proteins, which make them difficult to remove at the protein level; however, the size and solubility differences between humic acids and peptides make removal at the peptide level much more achievable. Since most of the humic acids have molecular weights above 10 kDa and are insoluble at low pH conditions, the combination of pH adjustment and filtering was postulated to provide a useful route for efficient separation of peptides and humic acids. To evaluate this, initial work was done with a pure humic acids solution, to examine the efficiency of clearing humic acids from the buffer solution. Three tubes each containing 1 mg humic acids dissolved in 1 mL 100 mM Tris-HCl buffer (pH=8) were prepared. To these tubes, formic acid was adjusted to 1% (~pH 3) in the first tube, 0.1% formic acid (~pH 6) in the second tube, and the last tube had no added formic acid (~ pH 8). The second tube was used to mimic real proteolytic peptide sample conditions, in which adjusting final sample to 0.1% formic acid is a normal procedure to protonate peptides prior to MS measurement.[180] After the three different treatments of pH adjustment and filtering, the color differences of the flow-through in three tubes are readily apparent (Figure 3.4a). At ~pH 3, the solution was very clear, while the other two tubes still exhibited brown coloration (Figure 3.4a). To more definitively quantify the concentration difference of humic acids after three different treatments, we measured the UV absorbance of the flow through at a wide range of wavelengths from 250 nm to 700 nm. The results are shown in the Figure 3.4b. In our original

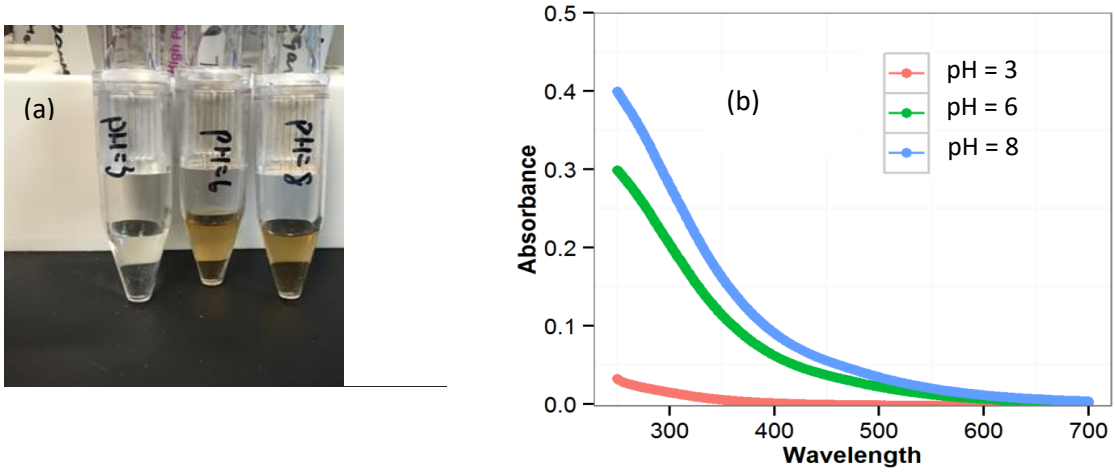


Figure 3.4. Humic acids can be efficiently removed by combination of pH adjustment and filtering by a molecular weight spin column. (a) The picture was taken immediately after filtering: pH was adjusted to 1% formic acid (pH=3), pH was adjusted to 0.1% formic acid and filtering (pH=6), filtering alone (pH=8). (b) UV absorbance of the flow through of three humic acids samples was measured by UV−vis spectrophotometer.

measurement of three flow-through samples without any dilution, the highest absorbance reading was larger than 1 in both of the two samples that contained observable amount of humic acids; we thus diluted all of three sample by 5 folds to reach the instrument sensitivity range (absorbance within 1). In the selected wavelength range, it is clear that the content of humic acids was reduced to an almost undetectable level by the combination of lowing pH to around 3 and filtering. Based on the total intensity across the wavelength range, more than 96% humic acids were found to be removed with this method, compared with filtering alone. Adjusting final solution to 0.1% formic acid (pH around 6), which is a common prep for peptide MS measurements, coincidentally removed around 30% of the humic acids, but not to an appropriate level.

3.5 Humic acids removal does not impact the peptide identification or abundance profiles

Sample clean-up is a key and usually necessary step in mass spectrometry measurements. This new pH precipitation/filtering procedure to remove humic acids is readily compatible with existing proteome extraction/preparation methods; however, a successful clean-up approach should be able to remove impurities but retain the target analytes. In order to demonstrate that this additional humic acids removal step did not cause sample loss nor impact the peptide/protein identification or abundance profiles, we evaluated this method by comparing the proteome identification results in three different samples: control samples containing only microbial cells; humic acids samples containing microbial cells with added commercially available humic acids, and natural topsoil samples inoculated with the same amount of microbial cells. We tested our method in different levels of sample complexity and biomass. In the first part, only a simple microbial isolate was used in three different samples: control samples, humic acids samples and soil samples. In the second part, a microbial mixture of four soil organisms was employed throughout the three different samples. Finally, a lower microbial biomass sample (1/8 of inoculated cells from the four soil organisms' mixture) was also tested in those three samples.

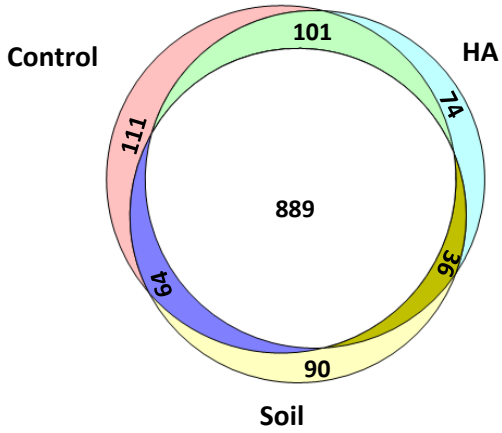
The results of protein identification and spectral assignments are summarized in Table 3.2. A very similar number of identified proteins and peptides in all samples indicated that there was no detectable amount of sample loss by removing humic acids prior to measurement. In the simple microbial isolate sample, the Venn diagram (Figure 3.5a) showed a large overlap of identified proteins in these three sample groups, which also revealed good recovery of proteins from soil samples. The correlation matrix (Figure 3.5b) demonstrated good correspondence between control and humic acids samples, which is the most compelling evidence for the unbiased characterization with this humic acids removal procedure. Furthermore, COG distributions are depicted in Figure 3.6, indicating that the functional characterization was not affected by humic acids removal. These same analyses were performed for samples contained microbial mixture as well, and gave very similar results, as shown in the supplemental section (Figure 3.7).

The lower biomass samples for both control samples and humic acids samples also gave very similar protein identifications and good correlations of the protein abundance (Table 3.3, Figure 3.8). These results also indicated that the humic acids removal step did not deleteriously impact protein identification even under this lower biomass condition. However, the number of identified proteins dropped more than 50% in the soil samples (Table 3.3), indicating that the protein recovery from soil samples decreased with the lower amount of spiked biomass. In order to more closely evaluate the reason for the poorer performance in soil samples, we plotted ScanRanker score distributions for soil samples which contained original four microbial mixture (Figure 3.9) and lower biomass of microbial mixture (Figure 3.9b). ScanRanker is a software developed to access the quality of tandem mass spectra via sequence tagging – the higher the ScanRanker score, the more “peptide-like” are the associated mass spectra. The red color denotes the distribution of ScanRanker scores for all collected spectra and the blue bars reflect the quality of the assigned spectra for the microbial isolate added to the sample. When comparing the soil samples with original microbial mixture, there was a significant amount of high quality spectra that were not assigned in this lower biomass sample. These high quality but unassigned

Table 3.2. Overview of the mass spectrometry measurement results of each sample in simple sample type in which only a microbial isolate was used, and in the moderately more complex sample type that a microbial mixture was employed.

	Simple microbial isolate						Microbial mixture					
Summary	Control_1	Control_2	HA_1	HA_2	Soil_1	Soil_2	Control_1	Control_2	HA_1	HA_2	Soil_1	Soil_2
protein counts	1317	1288	1263	1218	1207	1312	3447	3909	3358	3408	3601	3487
peptide counts	8528	8206	8724	8345	8075	9193	15033	16836	13069	13415	16292	15958
spectral counts	47143	45102	40842	43204	41511	48261	111444	115832	108331	109563	96915	98906

(a)



(b)

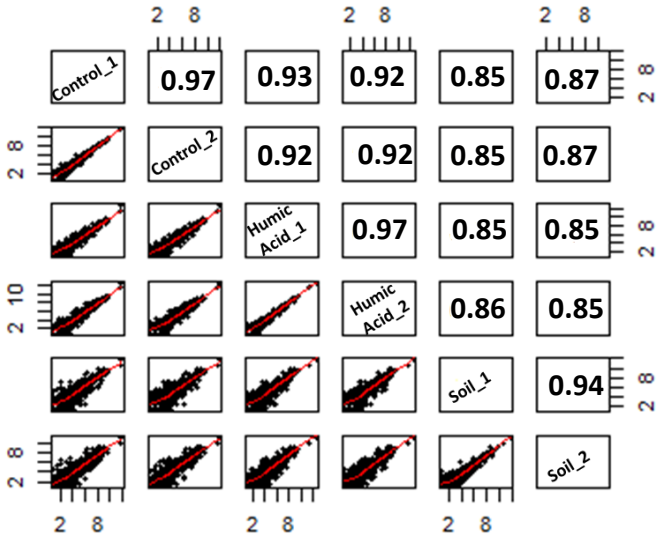


Figure 3.5. Control sample (microbial inoculant Pseudomonas putida F1), humic acids sample (microbial inoculant Pseudomonas putida F1 added with humic acids) and soil sample (natural topsoil spiked with microbial inoculant Pseudomonas putida F1) were prepared accordingly and measured in mass spectrometry. (a) Venn diagram illustrated the unique and shared proteins identified among these three groups. (b) Pearson correlation coefficients were calculated for each two samples under R environment.

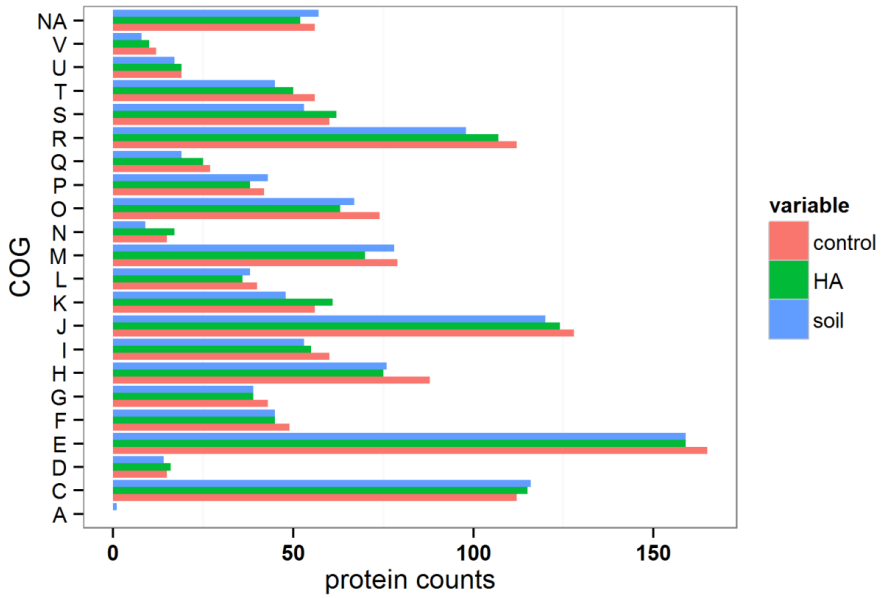
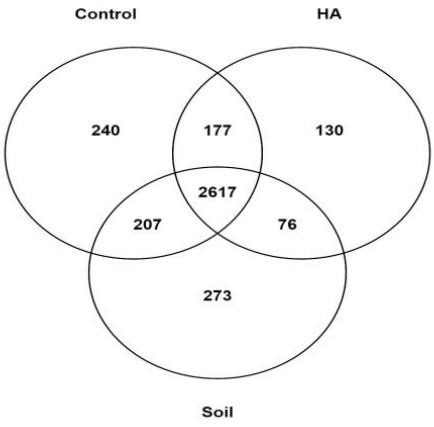


Figure 3.6. Numbers of proteins assigned to Clusters of Orthologous Groups (COG) among control runs, humic acids runs, and soil runs containing simple microbial isolate. COG categories A: RNA processing and modification; C: Energy production and conversion; D: Cell cycle control, cell division, chromosome partitioning; E: Amino-acid transport and metabolism; F: Nucleotide transport and metabolism; G: Carbohydrate transport and metabolism; H: Coenzyme transport and metabolism; I: Lipid transport and metabolism; J: Translation, ribosomal structure and biogenesis; K: Transcription; L: Replication, recombination and repair; M: Cell wall/membrane/envelope biogenesis; N: Cell motility; O: Posttranslational modification, protein turnover, chaperones; P: Inorganic ion transport and metabolism; Q: Secondary metabolites biosynthesis, transport and catabolism; R:General Functional Prediction only; S:Function unknown; T: Signal transduction mechanisms; U: Intracellular trafficking, secretion and vesicular transport; and Z: Cytoskeleton; NA: not assigned to any COG.

(a)



(b)

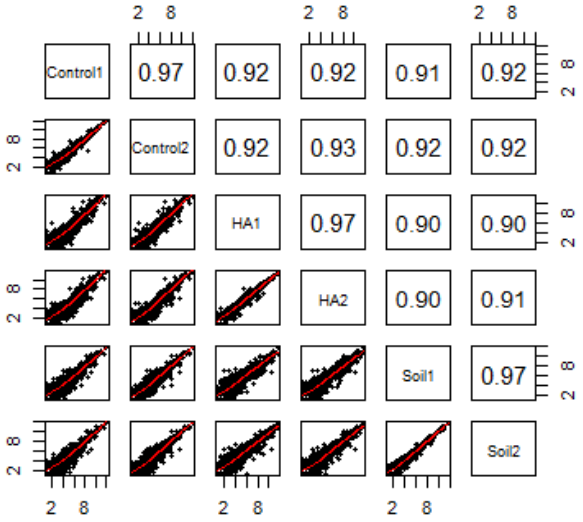


Figure 3.7. Control sample (four microbial mixture), humic acids sample (four microbial mixture added with humic acids) and soil sample (natural topsoil spiked with four microbial mixture) were prepared accordingly and measured in mass spectrometry. (a) Venn diagram illustrated the unique and shared proteins identified among these three groups. (b) Pearson correlation coefficients were calculated for each two samples under R environment.

(c)

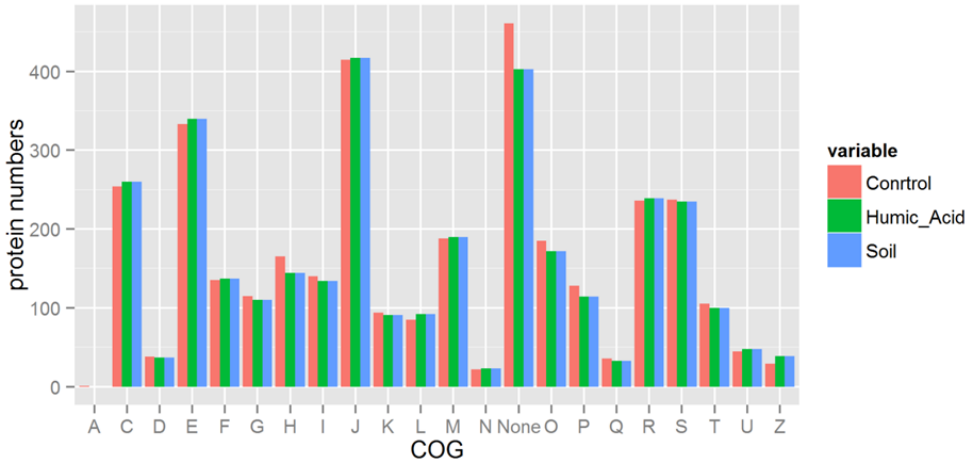


Figure 3.7. Continued. (c) Numbers of proteins assigned to Clusters of Orthologous Groups (COG) among control runs, Humic acids runs, and Soil runs that contained four microbial mixture samples. No obvious bias on specific functional group was observed.

Table 3.3. Overview of the mass spectrometry measurement results of each sample in lower biomass samples.

Summary	Proteins	Peptides	Assigned Spectrum
Lower_biomass_Control_Run1	3619	15479	122772
Lower_biomass_Control_Run2	3051	12672	122330
Lower_biomass _HA_Run1	3158	11578	118873
Lower_biomass _HA_Run2	3226	12185	122695
Lower_biomass _Soil_Run1	1211	3303	18879
Lower_biomass _Soil_Run2	1175	3153	19132

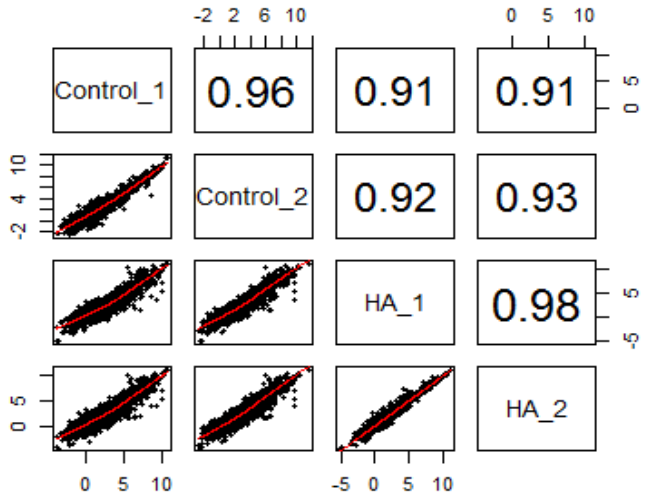


Figure 3.8. For lower biomass samples, pearson correlation coefficients were calculated between control samples and humic acids samples.

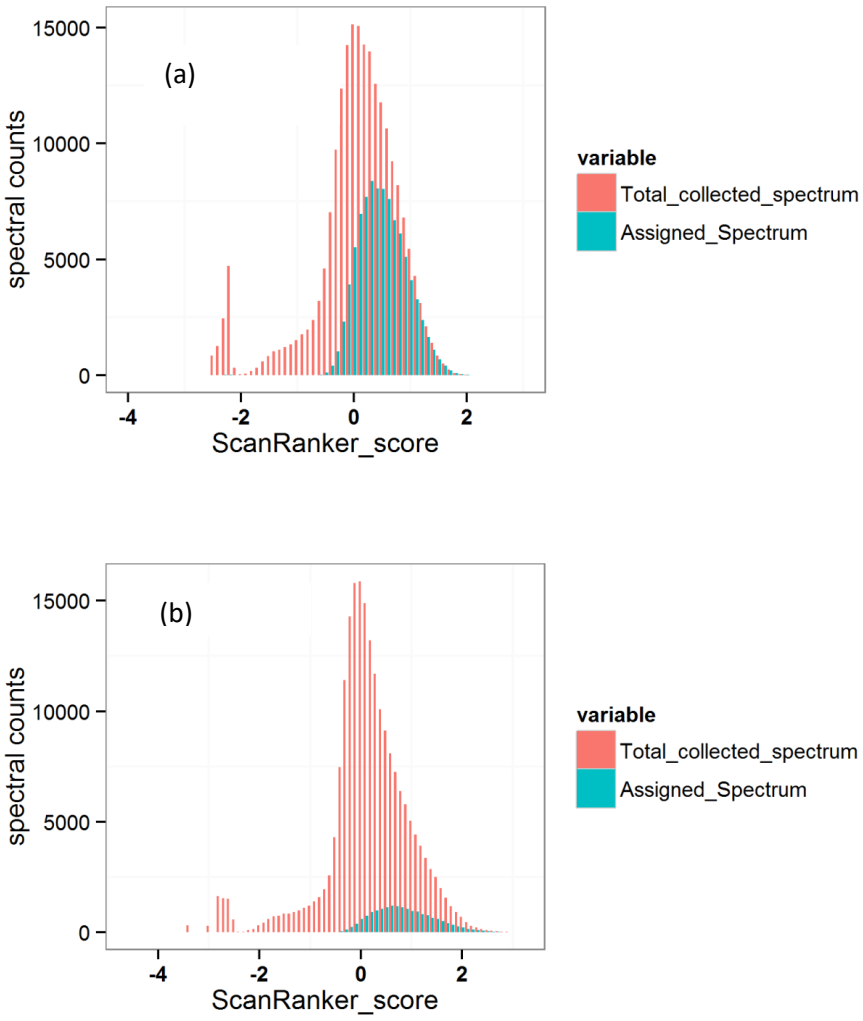


Figure 3.9. Distribution of ScanRanker scores of collected spectra gained from two soil samples which have been spiked with different biomass. The red color denotes the distribution of ScanRanker scores for all collected spectra and the blue color denotes the distribution of ScanRanker scores for assigned spectra. (a) Microbial mixture was added into the soil samples and proteome were extracted and measured by mass spectrometry. (b) Lower microbial biomass was added into the soil samples and proteome were extracted and measured by mass spectrometry.

spectra are very likely coming from the endogenous microbial peptides since these cannot be found in the database of spiked bacteria. More interestingly, the total number of high quality spectra (peptide like spectra) from this lower amount of spiked biomass is more or less very similar to that from original microbial mixture, indicating that our method successfully extracted a lot of endogenous proteins along with the proteins of spiked bacteria. Although we do not have the matched database to assign all of these high quality spectra to specific peptides, this ScanRanker figure serves as an evidence to suggest that our "humic acids removal step" did a great job in removing this interference and opened the door to a high level of reasonably high quality unidentified mass spectra that are originating from the endogenous proteins.

3.6 Conclusions

In this study, a novel and highly efficient humic acids removal step for soil metaproteomics study is presented. The optimized method was found to be highly efficient for the removal of endogenous soil-containing humic substances, without biased or interfering with the subsequent peptide identifications in the proteome MS measurements for the conditions used in this study. Additionally, the humic acids removal facilitates standard peptide colorimetric quantification assays (such as the BCA assay). This greatly assists in standardizing the HPLC loading amounts for each sample and thus provided a more reliable label-free quantification. For other major protocols employed in proteomics research, our humic acids removal step can also be easily incorporated by adding this step before final sample loading or offline fractionation. By applying the optimized methods on a range set of samples which differed by microbial composition and biomass, this clean-up approach consistently showed that the humic acids removal step did not induce sample loss or negatively impact on protein identification, thus successfully providing an unbiased soil proteome characterization. Furthermore, the application of this approach is not limited to soil metaproteomics measurements, but is likely to be more broadly applicable for samples in which complex compounds, such as humic acids mixtures, confound protein measurements.

CHAPTER 4

Global Proteome Response to Deletion of Genes Related to Mercury Methylation and Dissimilatory Metal Reduction Reveals Changes in Respiratory Metabolism in *Geobacter sulfurreducens* PCA

Text and figures were adapted from:

Qian, C.; Johs, A.; Chen, H.; Mann, B. F.; Lu, X.; Abraham, P. E.; Hettich, R. L.; Gu, B., Global Proteome Response to Deletion of Genes Related to Mercury Methylation and Dissimilatory Metal Reduction Reveals Changes in Respiratory Metabolism in Geobacter sulfurreducens PCA. J Proteome Res 2016, 15 (10), 3540-3549.

Chen Qian's contribution includes: equal contribution on experimental design with Chen,H, data acquisition, data analysis, manuscript writing and revision

4.1 Introduction

From technical perspective, we have optimized microbial proteome techniques, including for difficult environmental samples such as environmental soils. Subsequently, we sought to extend the general proteomics approach for exploring real biological questions regarding to the microbial physiology studies, such as the microbial response to metal exposure, in order to evaluate the ability of proteomics to characterize microbial metabolism at a systems-biology level. Since the samples used in this study did not contain any humic interferences, the same sample preparation protocol as described in chapter 3 was used, but without the humic acids removal step. The details of the sample preparation steps can be found in materials and methods section.

G. sulfurreducens PCA is a widely used model organism for iron [Fe(III)] reduction and exoelectrogenic reaction.[187-188]It also has been shown to concurrently reduce, sorb and methylate mercuric Hg [Hg(II)].[189-190] Methylmercury (MeHg) is a potent neurotoxin[191-193] that bioaccumulates up trophic levels and thus enters the food chain for human consumption. Although Hg methylation activity has been linked to the presence of the essential gene pair *hgcAB in vivo,*[73, 194] our understanding of how Hg methylation relates to other biochemical pathways is currently limited. Sequence analysis, homology

modeling, *in vivo* knockout studies, and site-directed mutagenesis indicate that *hgcA* encodes a corrinoid protein that is responsible for transfer of a methyl group to a Hg substrate, while *hgcB* encodes a 2[4Fe-4S] ferredoxin, which presumably enables turnover by delivering low-potential electrons to reduce the corrinoid cofactor on HgcA to the cobalt (I) [Co(I)] state.[73, 195] The sequence of the HgcA cobalamin-binding domain (HgcA-CBD) is similar to the cobalamin-binding domain of the well-characterized corrinoid iron-sulfur protein (CFeSP), indicating a common evolutionary origin. CFeSP is a key component of the acetyl-CoA (Wood-Ljungdahl) pathway involved in the transfer of methylcarbocations from methyltetrahydrofolate (CH_3-H_4folate) to a NiFeS cluster in acetyl-CoA synthase (ACS).[196] The HgcA-CBD is unique with respect to the presence of a strictly conserved cysteine located in proximity to the Co center of the cobalamin cofactor, which is essential for Hg methylation *in vivo*.[73, 194]

Early experimental studies with *Desulfovibrio desulfuricans* LS indicated that Hg methylation may be linked to cellular one-carbon (C1) metabolism, possibly as part of the acetyl-CoA pathway.[87, 197] Genes associated with this pathway appear in a diverse set of anaerobic bacteria and archaea, where it is used for both energy conservation and assimilation of C1 compounds.[194, 198] In the context of Hg methylation, it was found that a methyl group from CH_3-H_4folate is transferred to a corrinoid protein before it is further transferred to a Hg substrate resulting in the formation of MeHg.[87] A methyl group from CH_3-H_4folate can only be transferred to a corrinoid cofactor in the Co(I) state. Thus a source of low-potential electrons is required to reduce the cofactor to Co(I). The proposed roles of *hgcA* and *hgcB* are consistent with a Hg methylation pathway with the ferredoxin HgcB serving as a donor of low-potential electrons, which continuously activates the corrinoid cofactor of HgcA to accept a methyl group from CH_3-H_4folate.[73] However, the relationship between Hg methylation and other metabolic processes remains unclear. Previous experiments with *G. sulfurreducens* PCA also suggested that extracellular Hg reduction, sorption, Hg uptake and methylation are closely coupled.[190, 199] Deletion of the genes for five outer membrane *c*-type cytochromes (OmcB, OmcE, OmcS, OmcT, OmcZ) has been shown to impair electron transfer to extracellular electron acceptors.[200] More recently, the *omcBESTZ* deficient strain was

found to be less efficient in Hg reduction but showed increased Hg sorption and methylmercury production by the cells.[201] On the other hand, deletion of the gene pair *hgcAB* increased Hg reduction but did not prevent cells from taking up Hg.[199] We thus hypothesize that deletion of *hgcAB* not only abolishes Hg methylation but also impacts the proteome and specific metabolic pathways such as the abundance of proteins associated with C1 metabolism. In addition, due to the requirement for a continuous source of low-potential electrons, the loss of *hgcAB* may also affect relative abundances of proteins in the electron transport chain.

Comparative proteomics is a powerful tool for characterizing dynamic molecular responses in cellular networks induced by environmental stress or the deletion of individual genes. Currently, shotgun proteomics via two-dimensional liquid chromatography (LC) coupled with tandem mass spectrometry (MS/MS) is a major platform for large scale characterization of proteomes.[202] We therefore utilized this approach and examined differential protein abundances in response to the deletion of the *hgcAB* and *omcBESTZ* gene clusters in order to assess their impact on metabolic processes in *G. sulfurreducens* PCA.

4.2 Materials and methods

Bacterial strains and culture conditions

Details for the preparation of *G. sulfurreducens* PCA wild-type (WT) and its mutant strains Δ*hgcAB* and Δ*omcBESTZ* have been described previously.[73, 200-201] Briefly, cells were cultured anaerobically in nutrient broth Basal salts (NB) medium at 30°C.[189, 201] The NB medium contained 20 mM acetate as the electron donor and 40 mM fumarate as the electron acceptor at pH 6.8. All cell cultures were harvested at the same time at the late exponential growth phase with an optical density of 0.4-0.5 at 600 nm. Cell pellets were collected by centrifugation (1500×g, 10 min, 25°C) in an anoxic glove chamber (Coy) containing 98% N_2 and 2% H_2. Cells were then washed three times with a de-oxygenated phosphate-buffered saline (PBS), consisting of 0.14 M NaCl, 3 mM KCl, 10 mM Na_2HPO_4, and 2 mM KH_2PO_4 at pH 7.4. The buffer was first autoclaved and de-oxygenated by boiling and purging with ultra-

high purity N_2 gas and subsequently kept in the glove chamber for at least 24 h before use. All washing steps were conducted in the glove chamber and, after the final wash, the supernatant was decanted and the remaining cell pellets were collected and stored at –80 °C until protein extraction. Duplicate cultures were prepared for each strain of *G. sulfurreducens* and the mutants.

Protein extraction and proteolytic digestion

Protein extraction and subsequent digestion was performed as previously described.[180] Briefly, cell pellets (50-100 mg) were suspended in 0.5 mL sodium dodecyl sulfate (SDS) lysis buffer (4% w/v SDS, 100 mM Tris-HCl, pH 8.0) and boiled at 95°C for 5 min. The suspension was vortexed and pulse-sonicated (10 s on, 10 s off) for 2 min with an ultrasonic disruptor (Branson) at 20% amplitude. Crude lysates were subsequently centrifuged at 21,000×g for 10 min to remove any cell debris. A small amount of sample was removed to quantify the total protein concentration by the BCA assay (Pierce Biotechnology, Waltham, MA), and the remaining samples were adjusted to 20 mM dithiothreitol (DTT) and boiled again at 95°C for 5 min. For each sample, 1 mg of protein was precipitated using 20% trichloroacetic acid (TCA) and stored at −80 °C overnight. TCA-precipitated proteins were centrifuged at 20,800×g (4°C) for 15 min, washed twice with ice–cooled acetone and air dried. Protein pellets were suspended in 8 M urea, 100 mM Tris-HCl, pH 8.0 buffer with 5 mM DTT and allowed to incubate at room temperature for 30 min. To facilitate re-solubilization of proteins, the pellets were sonicated again and adjusted to 20 mM iodoacetamide and allowed to incubate in the dark for 15 min. Samples were then diluted 1:100 (w/w) in 4 M urea with 100 mM Tris-HCl (pH 8.0) and 10 mM $CaCl_2$ and pre-digested with sequencing grade trypsin (Promega, Madison, WI) overnight at room temperature on a nutator mixer (Barnstead International). Samples were further diluted to 2 M urea and digested with a second aliquot of trypsin for an additional 3 hours. Following digestion, the peptide solution was adjusted to 200 mM NaCl, 0.1% formic acid (FA) and filtered through a pre-cleaned (with 100 mM Tris-HCl, pH 8.0 buffer) 10 kDa molecular weight cutoff spin column filter (Vivaspin 2, GE Healthcare Life Sciences, Pittsburgh, PA) for

final clean up. The peptide enriched samples were quantified by the BCA assay (Pierce Biotechnology, Waltham, MA) and then stored at –80˚C until mass spectrometry analysis.

LC-MS/MS

LC-MS/MS data were collected in biological duplicates as previously described.[180] Briefly, 50 µg of proteolytic peptides were loaded onto a biphasic silica back-column packed in-house with ~3 cm of strong cation exchange (SCX) resin (Luna, Phenomenex) followed by ~3 cm of reverse-phase (C18) resin (Aqua, Phenomenex). Samples were pressure bomb-loaded onto the back-column and washed offline to remove salts. The back-column was then coupled in-line with an in-house pulled, ~12 cm reverse-phase packed nanospray emitter (100 µm ID) interfaced with a hybrid LTQ Orbitrap XL mass spectrometer (Thermo Scientific) operating in a data-dependent mode. LC-separated peptides were analyzed via 22-hr MudPIT 2D-LC-MS/MS. Precursor ions were measured in the Orbitrap analyzer with high resolution (15,000 at m/z 400). Dynamic exclusion was enabled with the following settings: repeat count was 1, the repeat duration was 30 s, and the exclusion duration was 15 s. The top 10 most abundant parent ions were selected for further fragmentation by collision-induced dissociation (at 35% energy level).

Peptide and protein identification from LC-MS/MS

MS/MS spectra were searched against the *G. sulfurreducens* PCA strain (ATCC 51573/DSM12127/PCA) proteome database (UniProt Proteome ID UP000000577) concatenated with common contaminant proteins by the Myrimatch v2.1 algorithm,[136] with the parameters set as follows: fully specific digestion with trypsin, automatic mode for precursor tolerance threshold (average precursor mass tolerance of 1.5 m/z for precursor ions with no charge state or a monoisotopic precursor mass tolerance of 10 ppm for precursor ions with a determined charge state), a fragment mass tolerance of 0.5 m/z, carbamidomethylation of cysteine (+57.02 Da) as a static modification and a maximum of two dynamic modifications per sequence, considering amine carbamylation (+43.00 Da) and methionine oxidation (+15.99 Da) as dynamic modifications.[203] The peptide results were compiled into protein

identifications via IDPicker v3.0 [181] with a final peptide-level false discovery rate (FDR) using the following parameters: peptide-spectrum-match filter at maximum FDR as 1%, a protein level filter minimum two distinct peptides per protein, and minimum three spectra per protein.[204]

Data analysis and differential comparison of identified proteins

Relative protein quantitation was achieved by comparing normalized spectral protein abundance values, which were calculated by $Ni = Ri * \frac{\overline{C}}{Ci}$, where N_i and R_i are the normalized and raw spectral counts of a protein in run *i*, Ci is the total spectral count of run *i*, and $\overline{C}$ is the average total spectral count of all runs under comparison.[205] Normalized spectral counts were further log2 transformed and filtered using Perseus software.[206] For quantitative comparisons, data were filtered to have two valid values in at least one group (one group refers to two biological replicates for each strain) and then missing values were imputed by generating random numbers from a Gaussian distribution which can best simulate the distribution of low abundant proteins (down-shift value is set to 1.8 and the width is set to 0.3). Since we are specifically interested in the comparison between wild-type and the two mutant strains, we performed two-sample *t*-tests in Perseus between WT and the Δ*hgcAB*, as well as WT and the Δ*omcBESTZ*, individually. Differential protein abundances between WT and Δ*hgcAB* and between WT and Δ*omcBESTZ* were considered significant at *p*-values ≤ 0.05.

Sequence analysis

The Basic Local Alignment Search Tool (BLAST)[207] was used to search for enzymes or proteins of known function in the acetyl-CoA and related C1 metabolic pathways using the translated NCBI reference genome NC_002939.5 for *G. sulfurreducens* PCA. The gene set includes essential genes required for the conversion of C1 precursors such as CO, CO_2 and formate and the biosynthesis of CH_3-H_4folate (Table 4.2).

Subcellular localization and functional classification

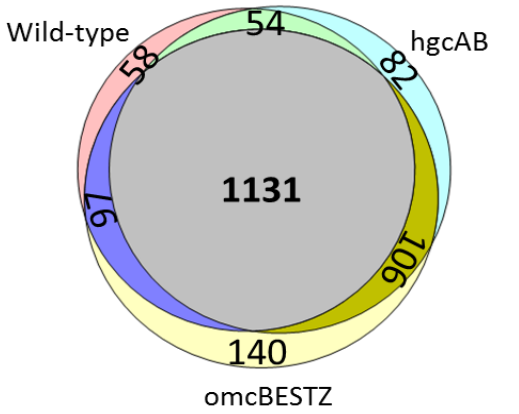
The subcellular localization of proteins was predicted using PSORTb v3.[208] A whole-genome dataset for *G. sulfurreducens* PCA was downloaded from PSORTdb.[209] Each protein identified by mass spectrometry-based proteomics was assigned a predicted localization. The Gene Ontology (GO) enrichment was carried out by BiNGO.[210] A hypergeometric test was chosen as the statistical test method and Benjamini & Hochberg False Discovery Rate (FDR) was selected for multiple testing corrections with the significance level set at $p = 0.05$. GO_Biological_Process, GO_Cellular_Component, and GO_Molecular_Function were then subjected to the enrichment analysis.

4.3 Overall comparison of proteomes

4.3.1 Differential protein abundances among the three strains

To determine the differences between protein abundance profiles on a broader scale among the three strains, a Venn diagram (Figure 4.1a) and correlation matrix were constructed (Figure 4.1b). In general, there is a large overlap in the number of proteins identified (1131) across all three strains. The Pearson correlation matrix showed that the correlation between different strains is ~ 0.84 and correlation between the biological replicates is ~ 0.94, indicating that the deletion of these genes caused detectable changes in the relative abundance of some proteins. This observation was further corroborated via a PCA plot (Figure 4.2), which shows biological replicates clustered together, while wild-type and mutant strains were clearly separated.

A two-sample *t*-test was conducted separately between wild-type and the two mutant strains. A total of 281 proteins were found to have significantly differential abundances between wild-type and the Δ*hgcAB* mutant, while 417 proteins were significantly different between wild-type and the Δ*omcBESTZ* mutant. These differentially abundant proteins were clustered in a heatmap for illustration purposes (Figure 4.3).



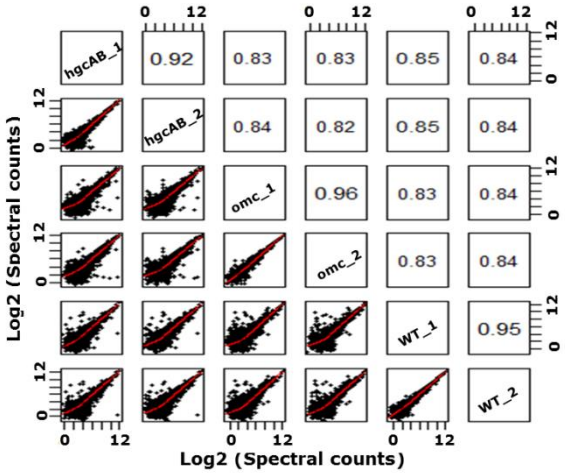


Figure 4.1. (a) Venn diagram of the numbers of identified proteins and their overlaps among different strains of Geobacter sulfurreducens PCA. Only proteins that were identified in both biological replicates were used to represent the number of identified proteins in each strain. (b) Pearson correlation coefficients were calculated for each two samples under R environment.

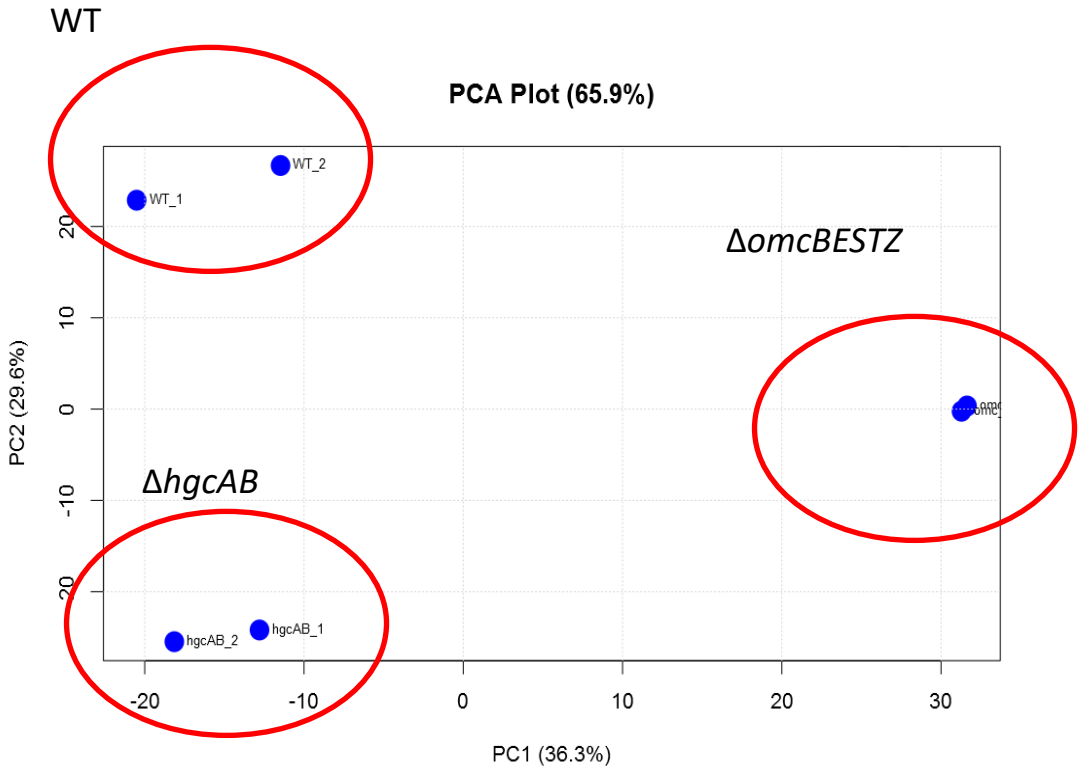


Figure 4.2. Principal component analysis (PCA) for biological replicates of wild-type and mutant Geobacter sulfurreducens PCA cultures. Proteomics results of biological replicates from the wild-type strain are shown on the top left , replicates for the Δ*omcBESTZ* mutant are shown on the right, and replicates for the ΔhgcAB mutant are shown on the bottom left. The plot was generated using the InfernoRND software.

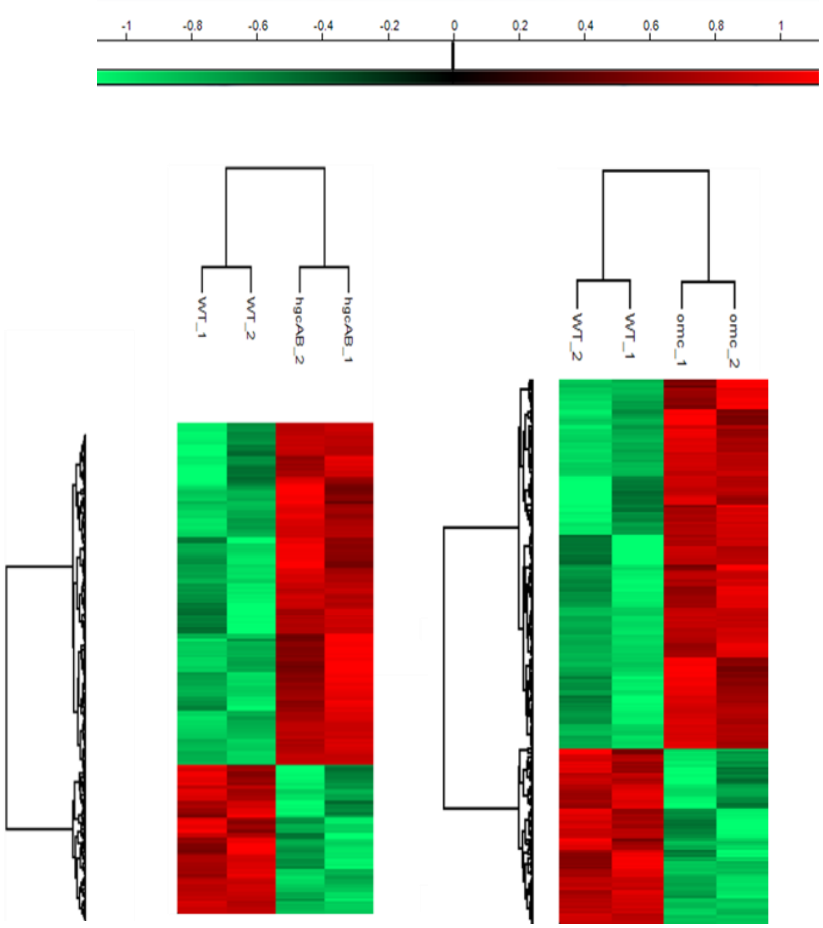


Figure 4.3. Spectral counts of all proteins were normalized by z-score and hierarchical clustering of differently expressed proteins was performed in Perseus. The data was divided into four sections: proteins of higher/lower abundance in wild-type relative to the Δ*hgcAB*, and proteins of higher/lower abundance in wild-type relative to the Δ*omcBESTZ*.

We subsequently examined the localization distribution of these differentially abundant proteins and compared them against the localization distribution of all identified proteins in each strain (Figure 4.4). The results show that the distribution of proteins among all three strains is similar with respect to all localizations. However, for differentially abundant proteins between wild-type and the Δ*hgcAB* mutant, most of the changes appeared in the cytoplasmic and outer membranes, suggesting that the deletion of *hgcAB* impacted membrane-associated biological processes. For differentially abundant proteins between wild-type and the Δ*omcBESTZ* mutant, we found significant changes in the periplasmic space, the outer membrane and cytoplasm, indicating that deletion of *omcBESTZ* encoding outer membrane cytochromes not only impacts the outer membrane region, but broadly affects protein abundance in the periplasmic space and the cytoplasm, which is consistent to the following functional enrichment analysis.

4.3.2 Functional classification of differentially abundant proteins

The GO enrichment analysis was performed to link differentially abundant proteins with major biological functions, and only the most significantly changed biological processes and/or molecular functions with the corrected *p*-value < 0.01 are shown in Table 4.1. The GO enrichments showed that proteins related to central metabolic processes, such as acetyl-CoA metabolism and tricarboxylic acid cycle were detected at higher levels after deletion of *omcBESTZ* (Table 4.1). In contrast, proteins related to other molecular function coupled to the transmembrane transport processes, such as amino acid transmembrane transporter activity andcarboxylic acid transmembrane transporter activity, were found to be down-regulated after deletion of *omcBESTZ* (Table 4.1), indicating that the deletion of these five genes had a significant impact on cellular transport processes. The GO enrichment results agree well with the analysis of subcellular localization shown above and thus offered additional insights into changes of proteins related to metabolic processes in wild-type and the Δ*omcBESTZ* mutant.

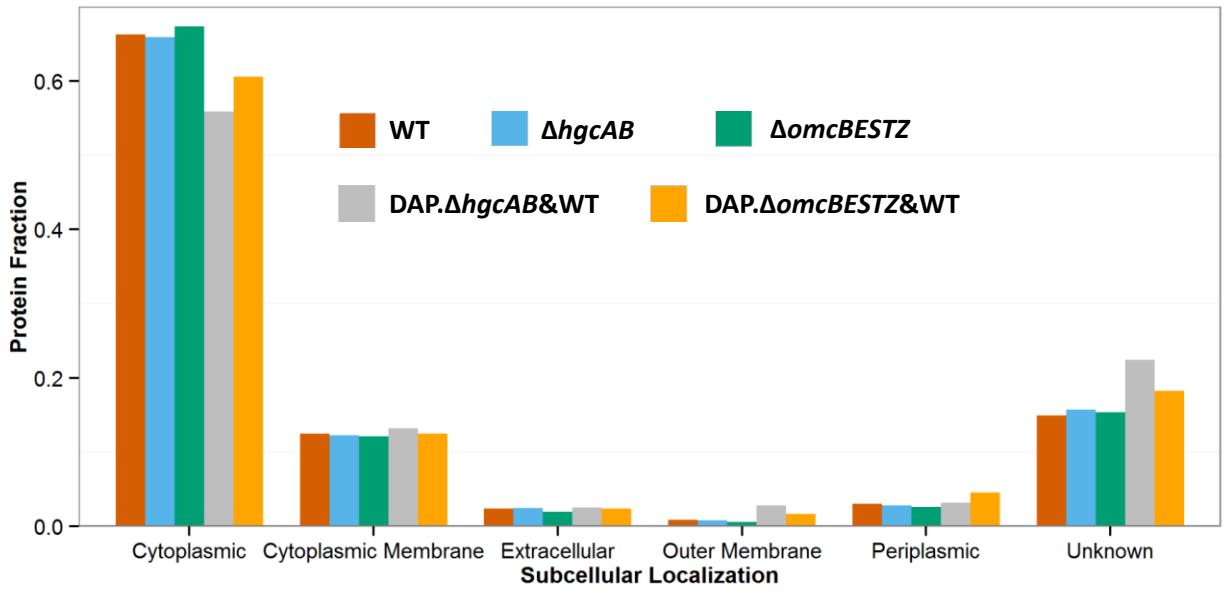


Figure 4.4. Differentially abundant proteins (DAP) between *Geobacter sulfurreducens* PCA wild-type (WT) and the two mutant strains were mapped to the predicted localization using their respective NCBI accession number. The number of proteins in each localization was counted. Protein fraction is the number of the proteins in each predicted localization divided by the total number of proteins in each case. DAP.Δ*hgcAB*&WT corresponds to DAP between *hgcAB*-deficient strain and wild-type; and DAP.Δ*omcBESTZ*&WT for DAP between the *omcBESTZ* deficient strain and wild-type.

Table 4.1. Gene ontology (GO) enrichment analysis of the biological processes (BP) and molecular functions (MF) that are over-represented or under-represented in Δ*omcBESTZ* strain compared to WT strain. Only the most significantly changed BP and MF are shown here with a corrected p-value well below 0.01.

Enriched Biological Process or Molecular Function	Corrected *p*-value	Properties
Metabolic process	1.29E-06	
Cellular metabolic process	1.36E-04	
Cellular biosynthetic process	1.36E-04	
Acetyl-coa metabolic process	1.66E-04	
Primary metabolic process	2.11E-04	
Biosynthetic process	2.73E-04	Biological Processes Over-Represented in Δ*omcBESTZ* strain than in wild type
Small molecule metabolic process	2.73E-04	
Acetyl-coa catabolic process	3.72E-04	
Aerobic respiration	3.72E-04	
Coenzyme catabolic process	3.72E-04	
Tricarboxylic acid cycle	3.72E-04	
Cofactor catabolic process	3.72E-04	
Cellular process	1.45E-03	
Heterocycle metabolic process	6.39E-03	
Catalytic activity	3.72E-04	Molecular Function Over-Represented in Δ*omcBESTZ* strain than in wild type
Transferase activity, transferring pentosyl groups	5.30E-03	
Aconitate hydratase activity	9.21E-03	
Amino acid transport	7.04E-03	Biological Processes Under-Represented in Δ*omcBESTZ* strain than in wild type
Amine transport	7.04E-03	
Carboxylic acid transport	9.60E-03	
Organic acid transport	9.60E-03	
Amine transmembrane transporter activity	7.04E-03	Molecular Function Under-Represented in Δ*omcBESTZ* strain than in wild type
Amino acid transmembrane transporter activity	7.04E-03	
Organic acid transmembrane transporter activity	9.60E-03	
Carboxylic acid transmembrane transporter activity	9.60E-03	

However, we did not observe any functional enrichment for differentially abundant proteins between wild-type and the Δ*hgcAB* mutant at the *p*-value cutoff of 0.05. This result suggests that the alteration to the proteome in Δ*hgcAB* is less global and more specialized as implicated by specialized roles of *hgcA* and *hgcB* in this organism.[73] Deletion of these two genes have also been shown to cause no impairment in the rate or extent of growth of the Δ*hgcAB* mutant.[73]

4.3.3 Impact on membrane protein abundance

Following the global proteome analysis, we examined the top 10 proteins which showed the largest changes in abundance between wild-type and the Δ*hgcAB* or the Δ*omcBESTZ* strain (Figure 4.5). Interestingly, we found that deletion of *hgcAB* and *omcBESTZ* genes both resulted in significantly lower abundances of a set of proteins annotated as metal ion efflux pumps, the resistance-nodulation-cell division (RND) superfamily (locus tag: GSU 1330, GSU1331, GSU1332), and ABC transporters (locus tag: GSU1340, GSU1341) (Figure 4.5a & c). The RND superfamily is widespread among Gram-negative bacteria and is used for catalyzing the efflux of metal ions and other substrates via an H^+ antiporter mechanism.[211] ABC transporters are involved in the active transport of metal ions and other substrates using an ATP-dependent mechanism.[212] These five genes (GSU1332, GSU1331, GSU1330, GSU1340, GSU1341) were also found to be upregulated when *G. sulfurreducens* PCA is grown in the presence Fe(III), instead of fumarate as the terminal electron acceptor.[213] These results highlight the potential roles of these RND proteins and ABC transporters in maintaining the metal homeostasis.

Based on the observation that the deletion of *hgcAB* and *omcBESTZ* had a significant impact on the abundance of RND family proteins, we further examined the abundance of all RND family proteins in the database. A total of 35 proteins annotated as RND family proteins were found in the *G. sulfurreducens* PCA genome. We compared the abundance of RND family proteins among three strains, and the result suggests that, among those proteins identified in the RND family, only one inner membrane protein (UniProt ID: Q74DI4), one periplasmic membrane fusion protein (UniProt ID: Q74DI5), and one

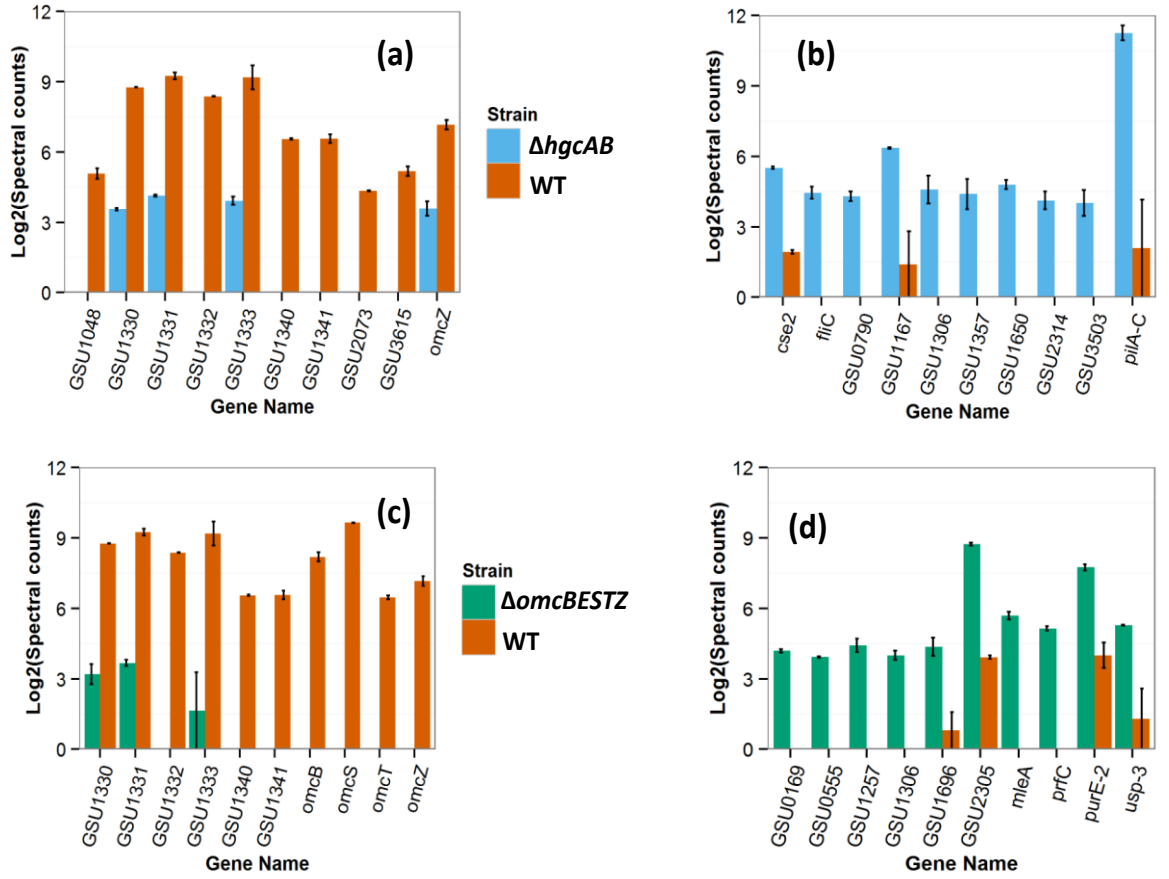


Figure 4.5. Top 10 significantly changed proteins in Δ*hgcAB* and Δ*omcBESTZ* compared to wild type (WT): (a) lower abundance in Δ*hgcAB*, (b) higher abundance in Δ*hgcAB*, (c) lower abundance in Δ*omcBESTZ*, and (d) higher abundance in Δ*omcBESTZ*. All differentially abundant proteins passing a t-test (p-value ≤ 0.05) were sorted by fold change. The calculation of fold change for proteins that were only identified in one strain was based on imputed values. After the list is generated, we replaced the imputed value back to 0 to reflect the true identification.

outer membrane protein (UniProt ID: Q74DI6) were found at very high abundance in the wild-type strain, but the abundance level decreased significantly in both mutant strains. The combination of a cytoplasmic membrane protein, a periplasmic accessory protein and an outer membrane-anchored channel can form a tripartite efflux pump which spans the inner and outer membranes for efficient transport of metal ions and other substrates.[212, 214] In addition, the sequences for Q74DI4 and Q74DI5 show significant homology (48% and 31% identity, respectively) to subunits of the tripartite efflux system CusCBA involved inmethionine-mediated metal ion efflux by *E. coli*. Therefore, we speculate that this RND efflux pump may also be involved in maintaining the Hg homeostasis since previous studies showed that deletion of *omcBESTZ* genes and *hgcAB* genes can impact Hg reduction and methylation.[199] Future experiments, such as RND protein knockout studies could potentially provide new insights into the role of efflux pumps in the transport of toxic metals such as Hg.

4.4 Biological processes change by deletion of *hgcAB* genes

4.4.1 Increased abundance of PilA-C and c-type cytochromes after deletion of *hgcAB*

We previously observed that Hg reduction rates followed the order of Δ*hgcAB* > PCA-WT > Δ*omcBESTZ,*[189-190, 199] and hypothesized that deletion of *hgcAB* may have an impact on the relative abundance of proteins involved in dissimilatory metal reduction. A previous study demonstrated that *c*-type cytochromes (*c*-Cyts), multi-copper proteins and pilin proteins, are involved in extracellular electron transfer by *G. sulfurreducens*.[215] Indeed, among the top 10 significantly changed proteins that were more abundant in the Δ*hgcAB* mutant compared to the wild-type strain (Figure 4.5b), PilA-C showed the most significant change. The *pilA-C* gene (GSU1497) encodes a structural geopilin domain essential for the assembly of pili, which are required for electron transfer to insoluble electron acceptors.[216] Furthermore, OmpB and OmpC are two multi-copper proteins required for Fe(III) oxide reduction by *G. sulfurreducens*.[217] OmpB was shown to have increased abundances in the Δ*hgcAB* mutant relative to the wild-type, while OmpC was not detected in any of the strains.

We subsequently examined how *hgcAB* gene deletion affects *c*-type cytochromes. The proteome database of *G. sulfurreducens* PCA was searched for all possible *c*-type cytochromes containing at least one heme-binding motif CXXCH using in-house scripts. A total of 132 *c*-type cytochromes were identified in the genome, suggesting that electron transport in *G. sulfurreducens* PCA is highly redundant and flexible.[218] The abundance of *c*-type cytochromes increased in the Δ*hgcAB* mutant relative to the wild-type strain (Figure 4.6a). An increase in protein abundance was observed for 42 proteins in the Δ*hgcAB* mutant out of a total of 65 *c*-type cytochromes identified in at least one strain (Δ*hgcAB* or wild-type). A comparison of only outer membrane cytochromes showed that besides OmcZ, all other identified outer-membrane cytochromes were found at higher levels in the Δ*hgcAB* mutant relative to the wild-type strain with OmcS showing the largest increase (Figure 4.6b). A previous study showed that OmcS is localized along pili to facilitate electron transfer to Fe(III) as terminal electron acceptor,[219] suggesting that OmcS and PilA together enhance extracellular electron transfer in the Δ*hgcAB* strain. Overall, our data show that deletion of *hgcAB* can impact proteins involved in the extracellular electron transfer of *G. sulfurreducens*, which is consistent with an increased rate of Hg reduction observed experimentally.[190, 199] Furthermore, the proposed function of HgcA and HgcB as a system to catalyze the transfer of methylcarbanions requires a continuous supply of low potential electrons.[73] Thus, deletion of *hgcAB* may result in diversion of excess reducing equivalents to exogenous electron acceptors.

As expected, none of the outer membrane cytochromes OmcB, OmcE, OmcS, OmcT, OmcZ were detected in the *omcBESTZ*-deletion mutant. Interestingly, however, other cytochromes showed either a decreased abundance (e.g., OmcC, OmcN, OmcQ) or an increased abundance (e.g., VorA, KorD, GSU0274) in the Δ*omcBESTZ* mutant relative to wild-type (Supplementary Data in Excel). These results are consistent with the observation that Hg reduction is not completely inhibited following the deletion of *omcBESTZ*, and suggest that other cytochromes may assume a role in the reduction of Hg.[190, 199]

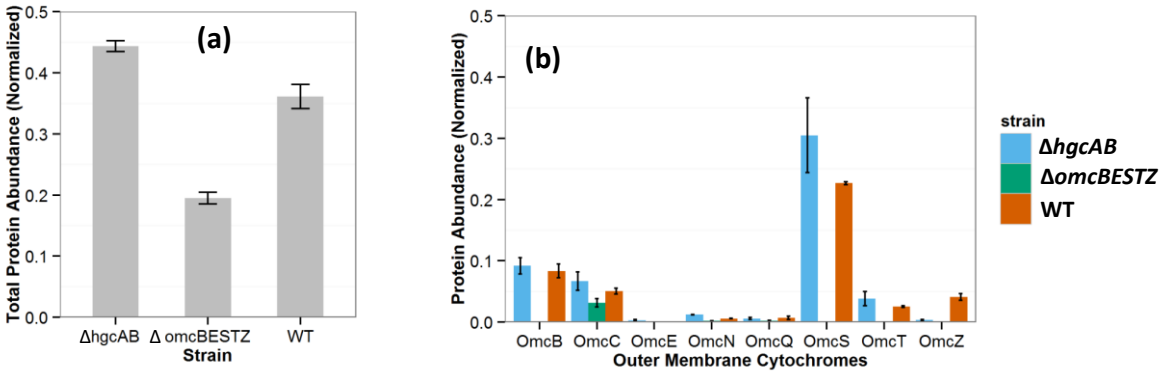


Figure 4.6. (a) Abundance levels of c-type cytochromes with at least one CXXCH motif were compared among the three strains. Data were normalized to unity. (b) The abundance level of each identified outer membrane cytochrome was compared among the three strains.

4.4.2 Increased abundance of proteins annotated as methyltransferases after deletion of *omcBESTZ*

The deletion of *omcBESTZ* increases Hg methylation, while deletion of *hgcAB* abolishes Hg methylation.[73, 199, 201] To further evaluate the broader impact of gene deletions on C1 metabolism, we retrieved 78 genes annotated to encode methyltransferases from the genome of *G. sulfurreducens* PCA and compared the abundance of each protein between wild-type and both mutant strains. A subset of 46 methyltransferases was detected in at least one strain and the abundance of these methyltransferases in different strains was summed and compared (Figure 4.7a). Results show that deletion of *omcBESTZ* led to overall higher relative abundances of methyltransferases, while deletion of *hgcAB* slightly decreased or had no effects on methyltransferase abundance levels. Individual paired *t*-tests were also conducted comparing methyltransferases between the Δ*omcBESTZ* mutant and wild-type, as well as the Δ*hgcAB* mutant and wild-type. The overall abundance of methyltransferases in Δ*omcBESTZ* was significantly higher than that in the wild-type strain (p = 1.894e-05) (Figure 4.7b), whereas the differences between Δ*hgcAB* and the wild-type strain were statistically insignificant (p = 0.3537). These findings agree with experimental results [199] that show Hg methylation activity follows the order Δ*omcBESTZ* > wild-type > Δ*hgcAB* and are thus indicative of elevated C1 metabolic activity in the Δ*omcBESTZ* mutant regarding to the methyl group transferring process.

4.4.3 Impact of gene deletions on carbon metabolism in *G. sulfurreducens* PCA

As described above, Hg methylation is predicted to be linked to cellular C1 metabolism and possibly part of the acetyl-CoA pathway.[73] In this pathway carbon dioxide can be utilized as an electron acceptor and as a precursor for the synthesis of acetyl-CoA, which is a fundamental building block for cellular metabolism.[198] Analysis of homologous genes required for the C1 metabolism reveals that *G. sulfurreducens* PCA encodes an incomplete acetyl-CoA pathway (Table 4.2); only genes corresponding to

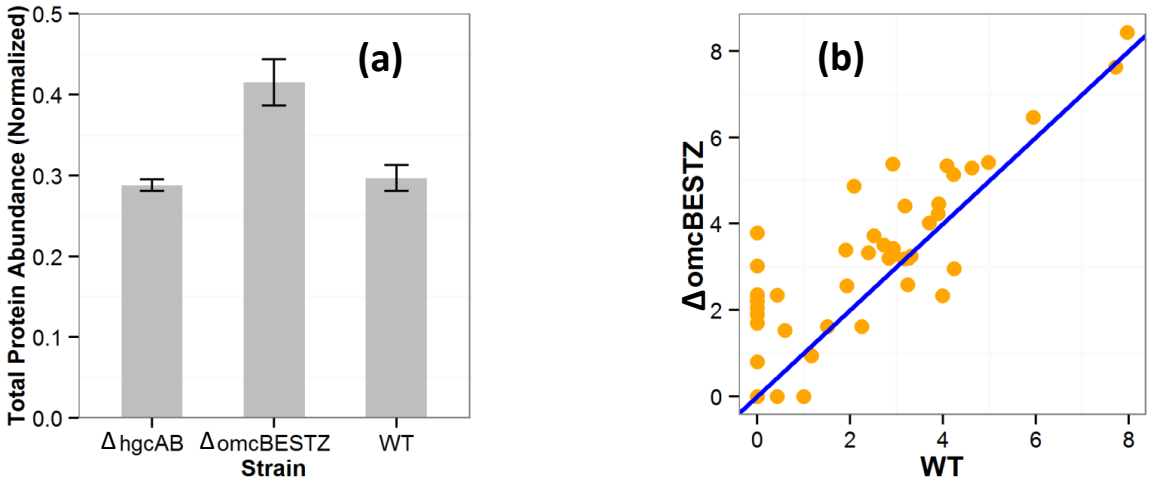


Figure 4.7. (a) Total abundance levels of 46 gene products annotated as methyltransferases identified in at least one strain. The abundance of these methyltransferases in different strains was averaged over all biological replicates and then normalized to unity. (b) Overall abundance of methyltransferases in Δ*omcBESTZ* was significantly higher than that in the wild-type strain (statistically significant at p = 1.894e-05 by a paired t-test).

Table 4.2. Homologs of proteins with known function in the acetyl-CoA and related C1 metabolic pathways identified in *Geobacter sulfurreducens* PCA. The table lists enzyme classification (EC), accession numbers, and protein abundance expressed as normalized spectral counts obtained from mass spectrometry.

Protein (function)	EC	Present in PCA	NCBI accession (locus tag)	Protein abundance (Normalized spectral counts)		
				wild-type	Δ*hgcAB*	Δ*omcBESTZ*
5,10-methylenetetrahydrofolate dehydrogenase and methenyltetrahydrofolate cyclohydrolase (bifunctional)	1.5.1.5 3.5.4.9	+	WP_010941526 (GSU0862)	130.05	128.77	107.85
5,10-methylenetetrahydrofolate reductase and homocysteine S-methyltransferase (bifunctional)	1.5.1.20	+	WP_010943603 (GSU2974)	5.23	6.50	10.07
5-methyltetrahydrofolate CFeSP methyltransferase (MeTr)	2.1.1.258	-	n/a	n/a	n/a	n/a
glycine/serine hydroxymethyltransferase	2.1.2.1	+	WP_010942252 (GSU1607)	250.26	213.60	347.34
formate dehydrogenase	1.2.1.2	+	WP_010941441 (GSU0777)	374.69	231.28	25.14
formate-tetrahydrofolate ligase	6.3.4.3	-	n/a	n/a	n/a	n/a
corrinoid iron-sulfur protein (CfsA)	2.3.1.169	-	n/a	n/a	n/a	n/a
corrinoid iron-sulfur protein (CfsB)	2.3.1.169	-	n/a	n/a	n/a	n/a
acetyl-CoA synthase (ACS)	2.3.1.169	-	n/a	n/a	n/a	n/a
carbon monoxide dehydrogenase (CODH)	1.2.99.2	+	WP_010942742 (GSU2098)	111.59	64.55	35.22
5-methyltetrahydrofolate-homocysteine S-methyltransferase, cobalamin-dependent (MetH)	2.1.1.13	+	WP_010943552 (GSU2921)	61.29	43.26	88.89
pyruvate-flavodoxin/ferredoxin oxidoreductase (PFOR)	1.2.7.1	+	WP_010940774 (GSU0097)	1274.65	1397.03	1140.39

the folate branch of the reductive acetyl-CoA pathway are present in the genome. We further examined the relative abundance levels of the corresponding proteins and found that all proteins associated with the folate pathway were identified. Moreover, these proteins were all shown to have decreased protein abundances in the Δ*hgcAB* mutant compared to wild-type, indicating that enzymes in the folate pathway required for the synthesis of CH_3-H_4folate were downregulated in the Δ*hgcAB* mutant. This result supports our hypothesis that the physiological function of HgcA and HgcB may be linked to a specialized role in the utilization of C1 substrates derived from the folate pathway, which is responsible for the formation of MeHg. However, we did not observe the same trend for the Δo*mcBESTZ* mutant when compared to wild-type. In fact, a subset of enzymes in the folate branch such as serine/glycine hydroxymethyltransferase, which catalyzes the transfer of the C3 of serine to tetrahydrofolate,[87] was shown to have increased protein abundance in the Δo*mcBESTZ* mutant relative to the wild type. These results support our hypothesis that the function of HgcA and HgcB may be linked to cellular electrontransport and C1 metabolism and the folate pathway in *G. sulfurreducens* PCA, as illustrated in Figure 4.8. Previous studies also suggested that HgcB shuttle electrons from various electron donors to HgcA, resulting in the reduction of its corrinoid cofactor.[73] Enzymes related to the folate pathway may thus play a role in the reduction of various C1 precursors resulting in the generation of CH_3-H_4folate. The methylcarbocation from CH_3-H_4folate is then transferred to the reduced Co(I) corrinoid cofactor of HgcA, followed by the transfer of a methylcarbanion to Hg to form methylmercury.[73, 220]

4.5 Conclusions

Comparative proteomics was conducted to delineate the global response of *G. sulfurreducens* PCA after deletion of *hgcAB* and *omcBESTZ* genes, which are associated with Hg methylation and dissimilatory metal reduction, respectively. To our knowledge, this is the first study comparing differences in the proteome of deletion mutants and evaluating the impact of gene deletion on fundamental metabolic processes in *G. sulfurreducens* PCA. Our results suggest that deletion of *hgcAB* increased the abundance of proteins involved in the extracellular electron transfer of *G. sulfurreducens*

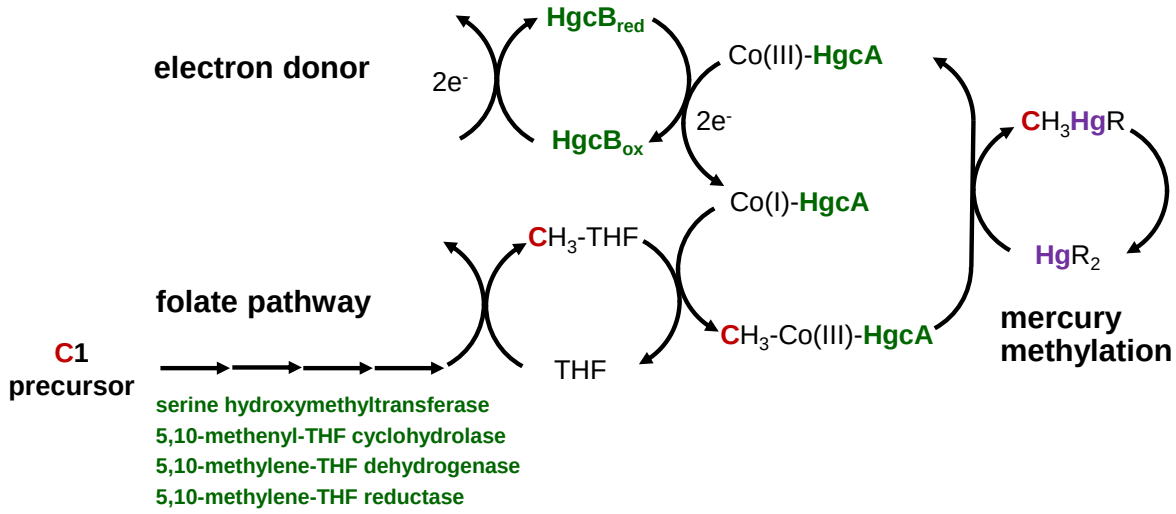


Figure 4.8. The proposed function of HgcA and HgcB and linkages to the folate branch of the reductive acetyl-CoA (Wood-Ljungdahl) pathway.[73] Some enzymes in the folate pathway were found to be upregulated after deletion of *omcBESTZ* compared to the wild-type strain. This figure is provided by co-author Dr.Alexander Johs.

PCA, including PilA-C, OmpB and many outer membrane cytochromes, consistent with an increased Hg reduction rate in the Δ*hgcAB* mutant. In contrast, deletion of *omcBESTZ* results in increased central metabolic processes, particularly an increased abundance of methytransferases, leading to increased Hg methylation by the Δ*omcBESTZ* mutant. In addition, the abundance of a set of RND-family transporters were significant impacted by the deletion of these gene-clusters, pointing to the possibility that the function of RND-family proteins may be linked to Hg methylation and metal ion homeostasis. Finally, we show that certain pathways associated with carbon metabolism are also affected by loss of these genes. A set of enzymes related to the folate branch of the acetyl-CoA pathway are under-represented in the Δ*hgcAB* mutant strain relative to the wild-type, supporting the hypothesis that the function of HgcA and HgcB may be linked to one carbon metabolism through the folate pathway. Future studies are warranted and, in particular, knockout studies would help to confirm the role of specific genes encoding proteins related to C1 metabolism and metal ion efflux pumps.

CHAPTER 5

Quantitative Proteome Analysis of *Desulfovibrio desulfuricans* ND132 reveals an additional function of *hgcAB* genes

Text and figures were adapted from:

Qian, C.; Lu, X; Chen, H; An, J; Johs, A.; Pierce, E; Hettich, R. L.; Gu, B., Quantitative Proteome Analysis of Desulfovibrio desulfuricans ND132 reveals an additional function of hgcAB genes.(This manuscript is currently under preparation, and thus the title and author lists are not finalized)

Chen Qian's contributions include: equal contribution on experimental design with Lu,X and Chen,H, sample preparation, MS measurement, data analysis, manuscript writing

5.1 Introduction

In chapter 4, comparative proteomics was successfully conducted to delineate the global response of *G. sulfurreducens* PCA after deletion of mercury methylation essential genes *hgcAB*; however, deletion of *hgcAB* was found to cause very mild perturbation to the whole proteome. In this study, we ought to investigate if deletion of this gene pair would cause same impacts to the proteome in another mercury methylating bacterium *Desulfovibrio desulfuricans* ND132. Mercury has been treated as an environmental pollutant for decades. However, certain anaerobic bacteria that thrive in anoxic zones (such as anaerobic sediments, soil, bottom water, periphyton biofilms), can convert its inorganic form to an even more toxic form known as methylmercury (MeHg).[221] Methylmercury can be bio-accumulated in organisms and biomagnified in food webs, thus posing risk for wildlife and human health. To date, the most well-known and most dominant Hg-methylating microorganisms are sulfate- (SRB) and iron- (FeRB) reducing bacteria, but some new species were also identified to be able to methylate mercury, suggesting more diverse MeHg distribution on the earth.[222]

Sulfate-reducing bacterium *Desulfovibrio desulfuricans* ND132 is commonly used as a model organism for understanding bacterial mercury methylation since it is a strong producer of MeHg.[223] Although it was found that Hg methylation is associated with the presence of the essential gene pair

hgcAB in vivo,[73] our understanding of the mercury transport and mercury methylation mechanisms in this model organism is still very limited. Additionally, *hgcAB* genes were found to be expressed at very low level in another *Desulfovibrio* organism,[224] resulting in the fully understanding and characterization of HgcA and HgcB proteins very challenging.

Following the discovery of the two essential genes required for mercury methylation: *hgcA*(which encodes a putative corrinoid protein) and *hgcB* (which encodes a 2[4Fe-4S] ferredoxin), a possible mercury methylation mechanism was proposed. The methyl group is transferred from methylated HgcA protein to inorganic Hg(II), and HgcB protein is involved in HgcA protein turnover by delivering low-potential electrons to reduce the corrinoid cofactor on HgcA.[225] Previous study found that the mercury methylation is likely to be involved in one carbon metabolism, and CH_3-THF probably serves as the methyl group donor.[87, 197] However, no further experimental evidence is available to support this mechanism.

Besides the biochemical mechanism behind the mercury methylation, our understanding of mercury uptake and methylmercury export is also very limited. Different metals pose different roles to bacterial life: some of them are beneficial to the cellular metabolism, while some others are toxic. Bacteria have its own discriminatory mechanism for different metals, so that the cellular response is specific to specific metal.[226] However, bacteria may lack the discriminatory mechanism for mercury, thus allowing mercury accidently uptake into the cells by other metals' transporters. Converting toxic Hg(II) to even more toxic methylated mercury is likely to be a detoxification mechanism since it is observed that methylmercury can be immediately exported out of the cells. The recent finding that certain divalent metals can inhibit Hg(II) methylation and uptake provided more support for this accidental uptake hypothesis.[227]

Proteomics is a powerful tool to investigate the dynamic molecular responses in cellular networks induced by environmental stress or the deletion of individual genes. Liquid chromatography (LC) coupled

with tandem mass spectrometry (MS/MS) is currently acknowledged as one of the most effective methods to study proteomes. Therefore, quantitative proteomics analysis was utilized to achieve two goals in this study: (i) to examine the impact of deletion of *hgcAB* genes on the proteome; (ii) to investigate how wild type strain respond to the mercury exposure. Our proteome results revealed another function of *hgcAB* genes in this model organism related to acetate formation, besides the possible involvement in one carbon metabolism previously proposed. Additionally, no differentially abundant protein was found between the proteome of wild type strain with or without exposure to the mercury, indicating that the mercury concentration we added is far less than the toxic level, and thus did not trigger specific proteome response to it.

5.2 Materials and methods

Bacterial strains and culture conditions

Desulfovibrio desulfuricans ND132 (wild type) and its mutant strain *hgcAB* were cultured in a modified MOY medium containing 40 mM fumarate and 40 mM pyruvate at 30°C.[228-229] Cells were harvested during the early-, mid- and late-exponential phase and washed three times by repeated centrifugation (at 1200g, 10 min, 25°C) and resuspension in a deoxygenated phosphate-buffered saline (PBS) at pH 7.4. PBS consisted of 0.14 M NaCl, 3 mM KCl, 10 mM Na_2HPO_4, and 2 mM KH_2PO_4. The buffer was first autoclaved and deoxygenated by boiling and subsequently kept in an anaerobic glove chamber (Coy) with ~98% N_2 and 2% H_2 for at least 24 h before use. All the washing steps and Hg(II) methylation assays were conducted in the glove chamber as previously described.[228, 230]

Protein extraction and proteolytic digestion

Protein extraction and proteolytic digestion was carried out as previously described.[50] Generally, cell pellets (around 1E+11 cells) were suspended in 1mL sodium dodecyl sulfate (SDS) lysis buffer (4% w/v SDS, 100 mM Tris-HCl, pH 8.0) and vortexed. The suspension was further boiled at 95°C for 5 min and sonicated by (10 s on, 10 s off) for 2 min with an ultrasonic disruptor (Branson) at 20% amplitude.

High speed centrifugation at 21,000×g was conducted for 10 minutes to remove cell debris. Protein concentration was measured by BCA assay (Pierce Biotechnology, Waltham, MA), and crude lysates were adjusted to 20 mM dithiothreitol (DTT) and boiled again at 95°C for 5 min. For each sample, 0.5mg proteins were taken out and precipitated by 20% trichloroacetic acid (TCA) and further incubated at −80 °C overnight. TCA-precipitated protein pellets were gained by centrifugation at 20,800×g (4°C) for 15 min, followed by two times of acetone wash. Clean precipitated protein pellets were re-solubilized in 8 M urea containing 5mM DTT and disulfide bonds reformation was blocked by adding 20 mM iodoacetamide. Samples were then digested with sequencing grade trypsin (Promega, Madison, WI). Following digestion, the peptide solution was adjusted to 200 mM NaCl, 0.1% formic acid (FA) and filtered through a 10 kDa molecular weight cutoff spin column filter (Vivaspin 2, GE Healthcare Life Sciences, Pittsburgh, PA) for final clean up. The peptide concentration was measured by the BCA assay (Pierce Biotechnology, Waltham, MA) and samples were stored at –80˚C until mass spectrometry analysis.

LC-MS/MS

LC-MS/MS analysis was performed in biological triplicates. 50 µg of proteolytic peptides were loaded onto a biphasic silica back-column which is packed in-house and consisting of ~3 cm of reverse-phase (C18) resin (Aqua, Phenomenex) and ~3 cm of strong cation exchange (SCX) resin (Luna, Phenomenex). Samples were pressure bomb-loaded onto the back-column and washed offline with water-organic solvent gradient to remove salts and push all the peptides onto the SCX resin. The back-column was then coupled in-line with the analytic column which is in-house pulled, ~12 cm reverse-phase packed nanospray emitter (100 µm ID) and interfaced with a hybrid LTQ Orbitrap XL mass spectrometer (Thermo Scientific). LC-separated peptides were further analyzed via 22-hr MudPIT 2D-LC-MS/MS [33]. The data acquisition was operated in a data-dependent mode with top 10 most abundant parent ions were selected for further fragmentation by collision-induced dissociation (at 35% energy level). Precursor ions were measured in the Orbitrap analyzer with high resolution (15,000 at m/z 400) and monoisotopic

selection is enabled. Dynamic exclusion was enabled with the following settings: repeat count was 1, the repeat duration was 30 s, and the exclusion duration was 15 s.

Peptide and protein identification from LC-MS/MS

MS/MS spectra were searched against the UniProt downloaded *Desulfovibrio desulfuricans* ND132 strain proteome database, concatenating with common contaminant proteins by andromeda search engine, [138] which is incorporated in MaxQuant (version 1.5.5.1). [231] The key parameters set in MaxQuant is listed as follows: fully specific digestion with trypsin and allow 5 maximum missed cleavages; mass tolerance was set to 10 ppm at the MS level and 0.5 Da at the MS/MS level; The search included carbamidomethylation of cysteine as a static modification and amine carbamylation and methionine oxidation as dynamic modifications, with 5 maximum number of modification per peptide. A false discovery rate (FDR) for peptide-spectrum matches (PSMs) and protein identification was both controlled at 1% using a target–decoy approach. To achieve more accurate quantification, MaxLFQ algorithm was used with the default parameters.[140]

Data anaylsis

The results produced by MaxQuant were further analyzed in Perseus (version 1.5.5.3). [232] Categorical annotation was added in form of Gene Ontology (GO) biological process (BP), molecular function (MF), and cellular component (CC). All annotations were extracted from UniProt database. Data from the same condition are grouped in the same group and the LFQ intensities of the proteins were log2 transformed. For more accurate quantitative comparisons, data were filtered to have at least three valid values (greater than 0) in at least one group and missing values were imputed by generating random numbers from a Gaussian distribution which can best simulate the distribution of low abundant proteins (down-shift value is set to 1.8 and the width is set to 0.3). Two sample t-test were performed on different groups based on different questions by controlling FDR 0.05 with permutation-based FDR method. 2 D Annotation Enrichment analysis does not directly compare the intensity of the same proteins between

these two strains; instead, it plots the distribution of proteins in each annotation category and compares it with the global distribution (the distribution of all of the identified proteins). Significance is calculated based on the averaged abundance of each category and the background, significant deviation from the global distribution indicates that the specific category is upregulated or downregulated in the whole system (Benjamini-Hochberg FDR ≤ 0.05).

Pathway analysis

The Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology (KO) identifiers, or the K numbers for all the protein in the *Desulfovibrio desulfuricans* ND132 proteome database were assigned by KAAS (KEGG Automatic Annotation Server).[233] Each KO identifier represents an ortholog group of genes. Then KO identifiers of proteins with significant abundance change were mapped to KEGG pathway to search for the pathways that are impacted by the deletion of *hgcAB* genes. [234] BioCyc is another pathway library used to assist the pathway analysis. [235]

5.3 Global proteome comparison between wild type strain and Δ*hgcAB* strain

Wild type strain and mutant strain were both grown anaerobically to the late log phase and then conducted to the comparative proteomics analysis. Proteomics analysis reveals that very similar number of proteins was identified in these two different strains (around 1800 proteins), but with quite obvious differences of abundance profile, which was revealed by the Pearson correlation coefficients and Principal component analysis (Figure 5.1).

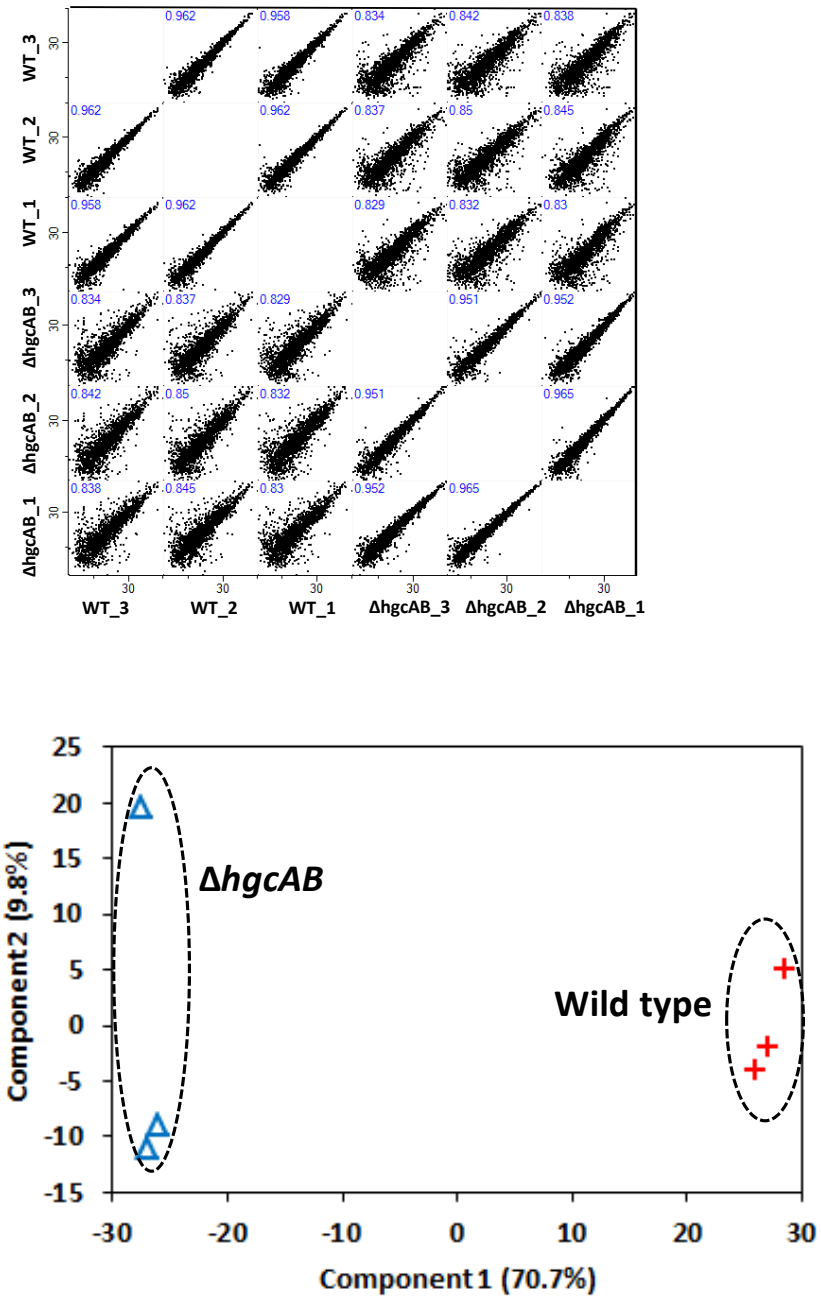


Figure 5.1. (a): Pearson correlation coefficients were calculated for each two samples in Perseus. Log2(FLQ intensity) was used for coefficients calculation; (b): Principal component analysis shows that mutant strain and wild type strain can be well separated. The figure is provided by co-author Dr. Hongmei Chen.

In order to ensure that different time points in log phase did not cause significant impact on the proteome, the wild type strain was grown and cells that are in three different time points of log phase (early time (ET), middle time (MT), and late time (MT) of log phase) were harvested separately and later used for proteome measurement and comparison. The results showed that different time points of log phase did not cause significant perturbation to the proteome abundance profile, as the neighboring two time points were very similar (Figure 5.2). Further, AVOVA testing was also performed on these three time points (Benjamini-Hochberg FDR $\leq$ 0.05) and only one protein was shown to have significant change, suggesting that harvesting cells at different time points of log phase did not bring in significant changes on proteome abundance profile. Therefore, even these two strains were not harvested at exactly same time points of log phase, this small growth difference would not bring in significant changes to the whole proteome, eliminating any concern that significant changes we observed coming from the different growth.

5.4 The expression level of HgcA and HgcB proteins are directly associated with mercury methylation

The *hgcAB* genes were discovered to be essential for the mercury methylation; however, no protein level characterization of these two genes is available. In our previous proteome analysis of *Geobacter sulfurreducens* PCA, we were not able to detect the HgcA and HgcB proteins by global proteome technique ,[50] leading to the speculation that these two proteins are either expressed in very low level or their expression need to be trigged by mercury. However, for the *Desulfovibrio desulfuricans* ND132 organism used for this study, we were able to detect the HgcA protein that was consistently recorded in three biological triplicates of wild type, but not in any of mutant strains. For the wild type strain at three different time points, HgcA protein can be accurately quantified by LFQ intensity (Figure 5.3b), but the abundance was low, as evidenced by its intensity rank compared to all the identified proteins (Figure 5.3a). HgcB cannot be accurately quantified, probably due to even lower expression

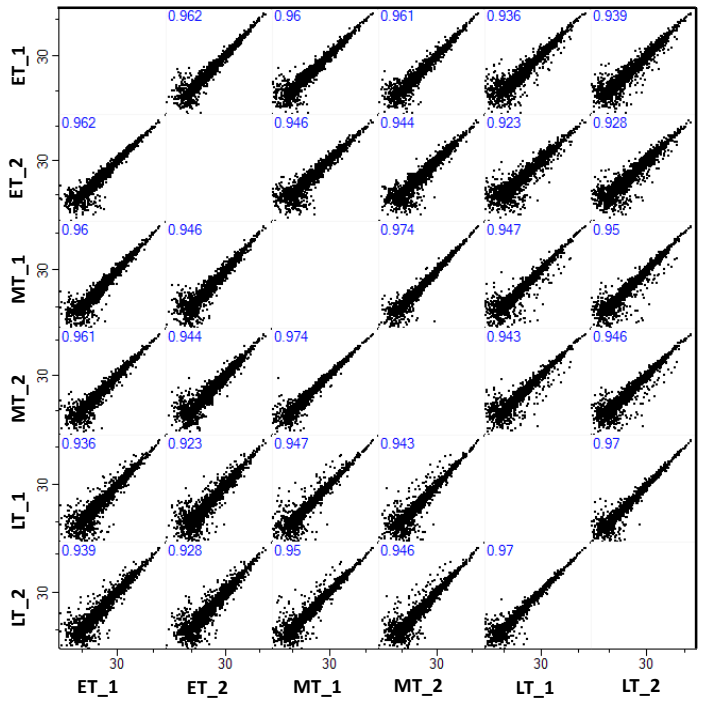


Figure 5.2. Pearson correlation coefficients were calculated for three different time points in log phase of the growth of *Desulfovibrio desulfuricans* ND132 wild type. Log2(FLQ intensity) was used for coefficients calculation.

levels or very limited number of peptides it can generate since it is a very short protein. Interestingly, we found that the intensity of HgcA decreased from early time point to late time point in both raw intensity and LFQ intensity (Figure 5.3b) and HgcB also follow this trend based on the raw intensity, although the change is very small. This result prompted us to hypothesis that if the expression level of HgcA and HgcB proteins is directly associated with mercury methylation, mercury methylation rate at these three time points should follow the same trend as the abundance of HgcA and HgcB proteins. We thus measured the mercury methylation rate at these three time points and found that the result supported our hypothesis (Figure 5.3c). This results indicated that the expression level of HgcA and HgcB proteins is probably another factor that determines the mercury methylation capability, addition to the two other factors (mercury bioavailability and thiols) described before. [236]

5.5 Functional Differences between wild type and mutant strains

To further investigate the differences between these two strains, a two-sample t-test was conducted between wild-type and the mutant strains and 522 proteins were found to have significantly differential abundances between these two strains (Permutation-based FDR $\leq$ 0.05). 304 proteins were found to have higher abundance in mutant strains than in wild type strain, while 218 proteins were found in the opposite direction. These differentially abundant proteins were clustered in a heatmap (Figure 5.4a). For functional classification of differentially abundant proteins, Fisher exact test (Benjamini-Hochberg FDR $\leq$0.01) was performed to get the enrichment of different GO categories for proteins in each cluster (Figure 5.4a). Biological processes, such as glucose metabolic process, generation of precursor metabolites and energy were found to be significantly upregulated in Δ*hgcAB* mutant strain, indicating the central carbon metabolism was impacted by deletion of *hgcAB* genes. The cellular component: proton-

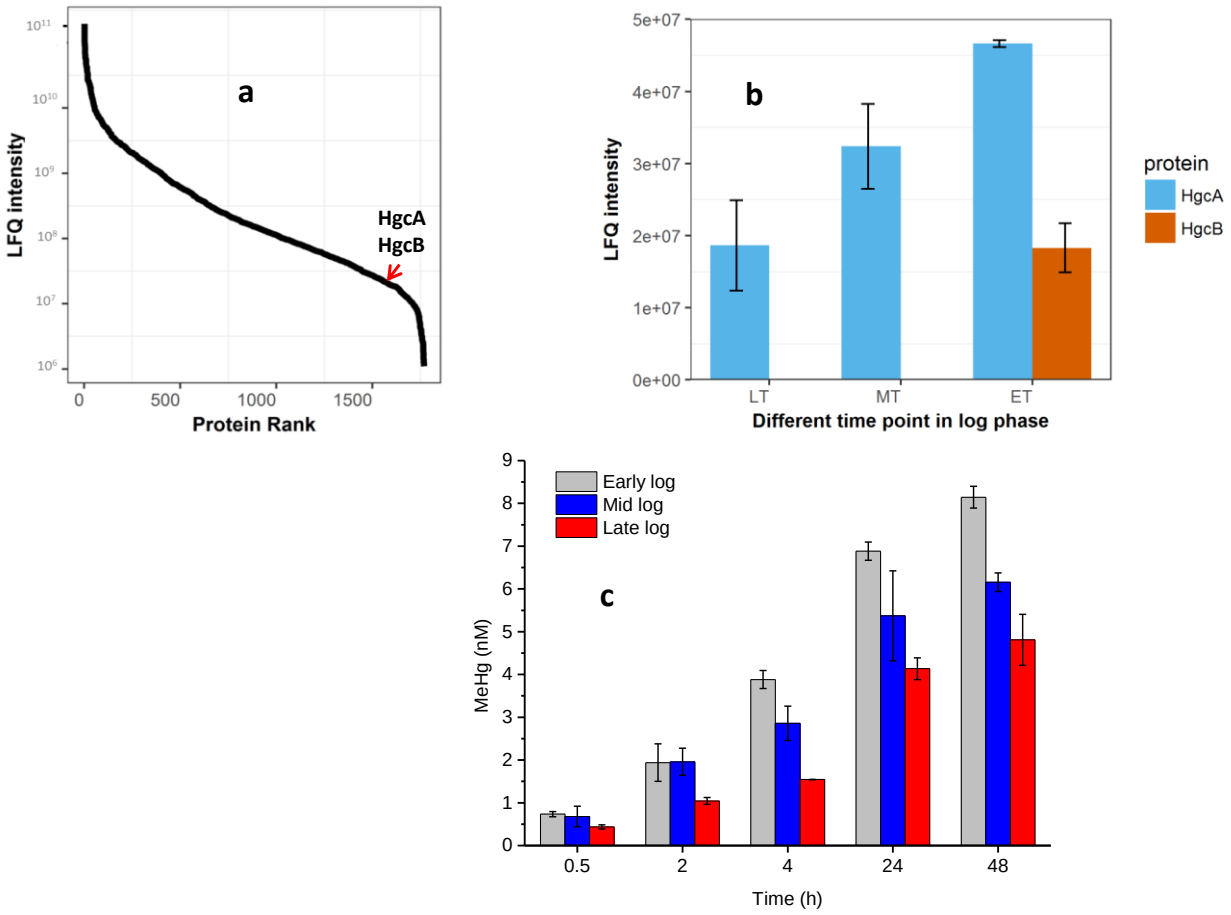


Figure 5.3. (a) Protein rank of all the proteins based on their LFQ intensity. (b) LFQ intensity of the HgcA and HgcB proteins in different time points of log phase. One imputed intensity value is used for HgcA protein in late log phase and HgcB protein in early log phase since only one of the replicates had the LFQ intensity value. (c) Mercury methylation rate was measured in the same three different times points of log phase.

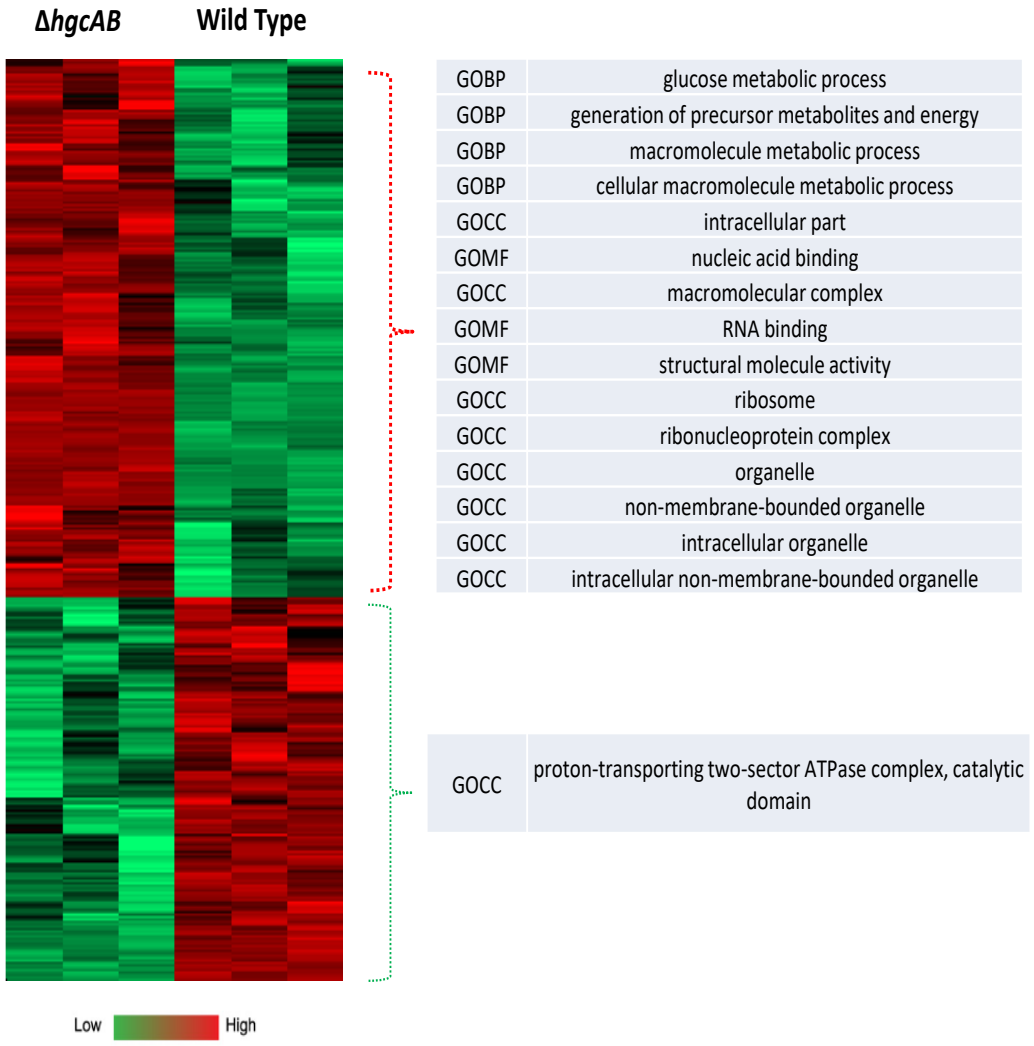


Figure 5.4. (a) Differentially abundant proteins were obtained by conducting a two-sample t-test at Permutation-based FDR ≤ 0.05. Fisher exact test was performed on differentially abundant proteins in each cluster to get the enrichment of GO function upregulated in wild type or in mutant strain (Benjamini-Hochberg FDR ≤ 0.01).

transporting two-sector ATPase complex, catalytic domain was found to be upregulated in wild type. Although the FDR value for this cellular component was 0.03 (higher than the FDR threshold we set), we still decided to keep it as a significant enrichment since there are only 5 proteins identified in this category with 4 was found to be significantly more abundant in wild type than in mutant strain. So the comparatively higher score of FDR is due to the very small size of this category. Proton-transporting two-sector ATPase complex are actively involved in ATP hydrolysis or synthesis, suggesting that several energy associated metabolic pathways could be impacted by the deletion of the genes. The detailed analysis regarding to these functional differences is illustrated in following pathway analysis part.

Comparative analysis of these two different strains highlights their significant differences but not their commonalities. For this purpose, we applied 2 D Annotation Enrichment[237], which calculated enrichments in each of the strain compared with the abundance distribution of all identified proteins and thereby revealed protein categories with significant regulation in the combined space of the two strains. The result (Figure 5.4b) suggests that all of enriched GO categories show simultaneously upregulated or downregulated trend in both strains and were very close to the diagonal line, indicating that these two strains are system-wide very similar. This 2 D Annotation analysis suggests that deletion of *hgcAB* genes did not cause the significant changes to the whole system, which is consistent with the observation that both strains grew very similar.

5.6 Deletion of *hgcAB* genes significantly impacts metal transporters and electron transporter chain

5.6.1 Deletion of *hgcAB* genes significantly impacts metal transporters

From above functional difference analysis, we did not see enrichment related to transport processes. However, from the examination on the top 15 proteins that showed the largest changes in abundance between wild type and mutant strain, several metal ion transport proteins were included. For example, among the 15 proteins which have higher abundance in mutant strain than in wild type,

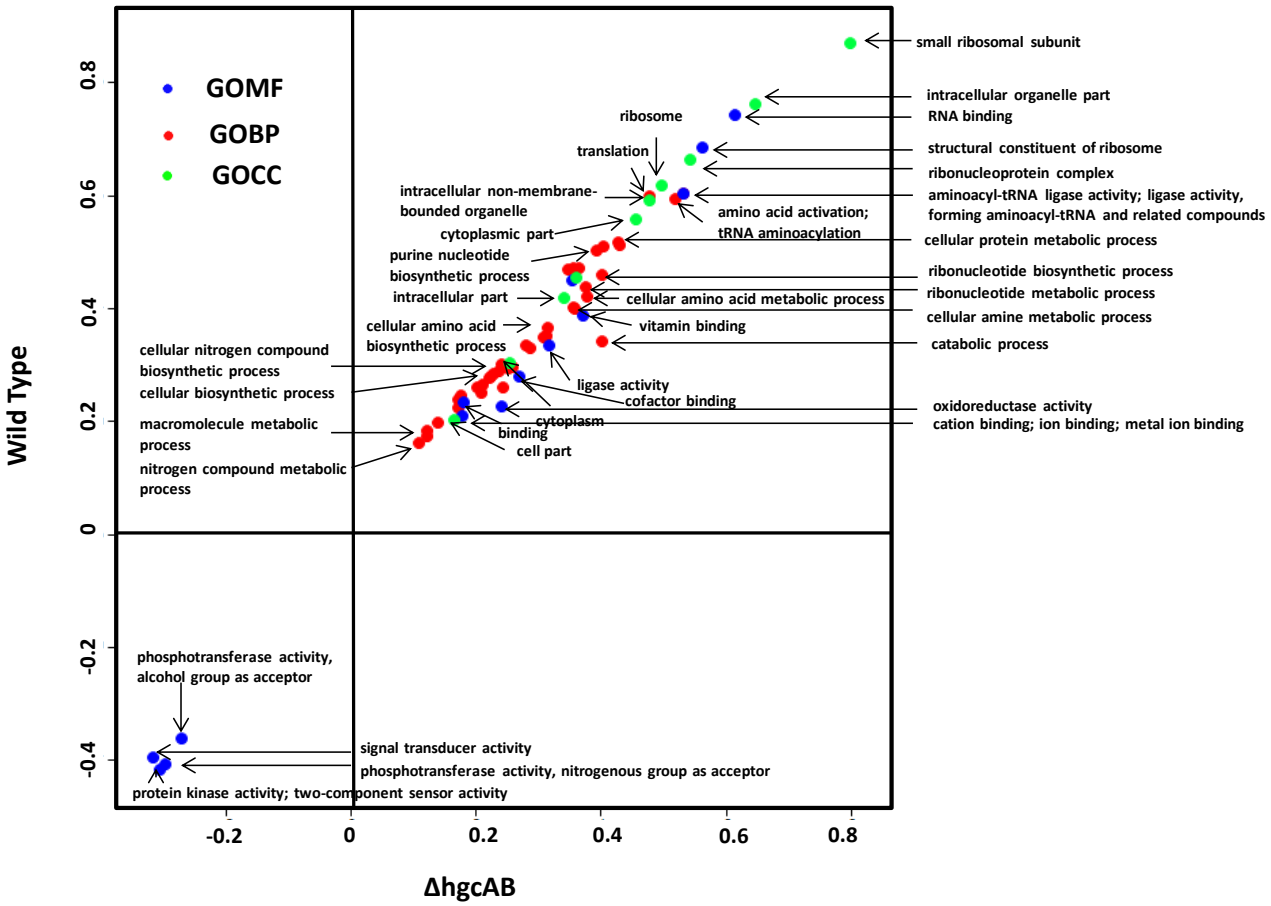


Figure 5.4. Continued (b) 2D annotation distribution. Scatter plot of normalized GO annotation changes between wild type strain and *hgcAB* deficient strain. The GO annotation analyzed were: GOMF(blue), GOBP(red), and GOCC(green).

five of them are associated with metal ion transport (Uniprot ID: F0JGV2 , F0JGH8, F0JGV5, F0JGV3, F0JCY2). F0JGH8 is annotated as the NikM subunit which is involved in nickel transport; F0JCY2 is one ABC transporter; F0JGV2, F0JGV3 and F0JGV5 belong to RND family proteins.. Additionally, other three RND family proteins are also very interesting since we also identified a tripartite RND efflux pump (Q74DI4, Q74DI5 and Q74DI6) differentially expressed between wild type strain and Δ*hgcAB* strain of *Geobacter sulfurreducens* PCA in our previous experiment.[50] More importantly, F0JGV5, F0JGV3 and F0JGV2 showed significant homology to the tripartite RND efflux pump that we identified in *Geobacter sulfurreducens* PCA (F0JGV5 is 30% identity to Q74DI6, F0JGV3 is 35% identity to Q74DI5, F0JGV2 is 46% identity to Q74DI4). However, the abundance of three RND family proteins had significantly higher level in wild type strain than in Δ*hgcAB* mutant strain in *Geobacter sulfurreducens* PCA, which is opposite to what we observed in *Desulfovibrio desulfuricans* ND132. We further searched for the KO number of these three proteins to look for possible linkage to specific metals, results showed that KO number is K07787 (Cu(I)/Ag(I) efflux system membrane protein CusA/SilA) for F0JGV2 and KO number is K07798 (membrane fusion protein, Cu(I)/Ag(I) efflux system) for F0JGV3, but no KO number was found for F0JGV5. This result indicated that the deletion of *hgcAB* genes probably significantly impact the Cu(I)/Ag(I) transport. And the interesting connection between deletion of *hgcAB* genes and the significant abundance change of this tripartite RND efflux pump in both organisms highlights possible connections between mercury transport / methylation and copper/silver transport.

Since we observed that several metal transporters showed significant changes after *hgcAB* genes deletion, we further examined the impact of deletion of *hgcAB* genes on all possible transport related proteins. All the proteins which are associated with metal uptake or export were collected together according to the annotation of different types of metal transporters described in a previous paper.[238] There are 135 proteins in total (100 associated with ABC transporter, 15 belong to the RND family, 11 in Major facilitator superfamily, 2 in Cation diffusion facilitator family, 2 are P-TYPE ATPase, and 5 are metalloregulatory protein). At the same time, we also added other transport associated proteins that have

their annotation specific for certain metals. In total, we have 146 proteins included in this list and after data filtering, we identified 70 proteins in this collected list, of which 30 of them showed significant change between wild type and mutant strain, suggesting that the deletion of *hgcAB* genes significantly impact the metal transport processes. The assignment of ABC transporters to specific substrates is determined by their KO (KEGG Orthology) numbers. All these differentially abundant transporters are presented in the volcano plot (Figure 5.5). Our results showed that besides the three Cu(I)/Ag(I) related RND family proteins and the nickel transporter discussed above, various other metal transporters were also found to be impacted, such as iron, zinc, molybdenum. Additionally, we noticed that some ABC transporters that are associated with polar amino acids transport were also impacted by the deletion of *hgcAB* genes. Previous paper found that Zn(II) and Cd(II) inhibited Hg(II) uptake and methylation, but Ni(II), Co(II), or Fe(II) did not.[227] Our data provided other candidates (Cu(I), Cu(II), Fe(III), Mo(II),. etc), that can be tested for this purpose. However, the impact of metals (under different concentration) on mercury uptake/methylation is probably not only due to the transport process but also result from certain enzymes' activities alteration given a lot of enzymes require proper concentration of metals as cofactors.

5.6.2 Deletion of *hgcAB* genes impacts extracellular solute-binding protein family

For the top 15 proteins which showed the largest changes in abundance between wild type and mutant strain, two protein families also attracted our attention, which were annotated as Methyl-accepting chemotaxis sensory transducer and extracellular solute-binding protein family. We subsequently examined all the proteins which fell into these two families. There are 20 proteins belonging to Methyl-accepting chemotaxis sensory transducer family, and 12 of them were identified in our data, with 7 of them showed significant change. Three of these 7 proteins were more abundant in mutant strain than wild type, 4 of them showed an opposite trend. There are 35 proteins belong to extracellular solute-binding protein family, and 23 of them were identified in our data, with 11 of them showed significant change.

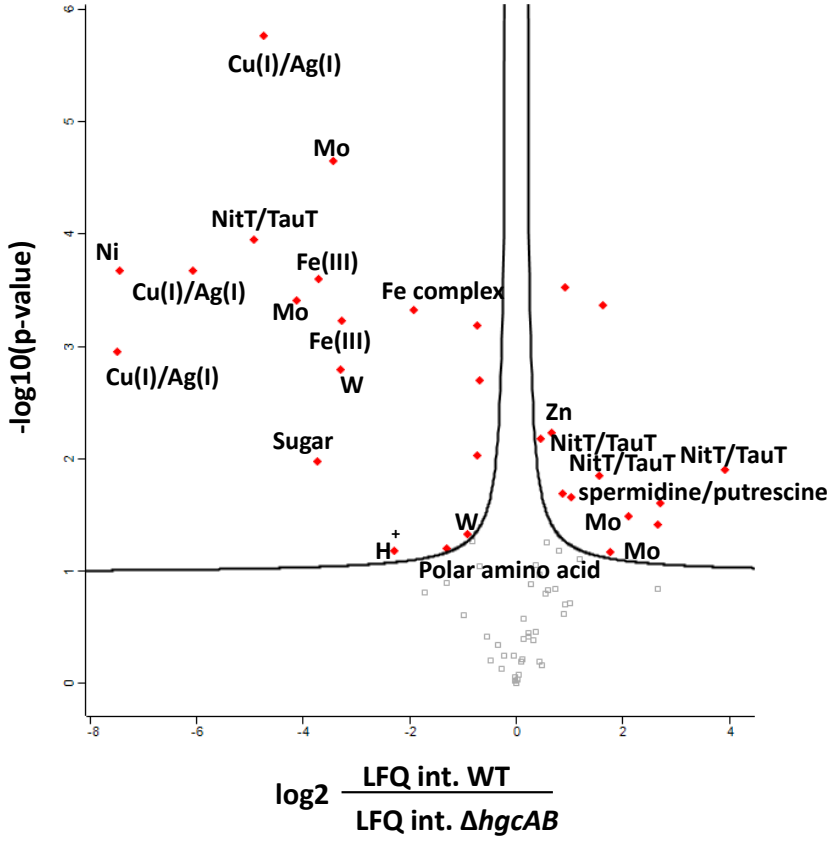


Figure 5.5. Comparison of the predicted metal transporters. There are 146 proteins annotated to be possibly associated with mental transport. After data filtering, we confidently identified 70 of them, and with 30 proteins showed significant abundance change between wild type strains and mutant strain (permutation-based FDR ≤5%, S0=0.1).

It has been known for long time that extracellular solute-binding proteins of bacteria serve as chemoreceptors, recognition constituents of transport systems, and initiators of signal transduction pathways.[239] For the 11 significantly changed proteins, only one belongs to extracellular solute-binding protein family 7, whereas all others belong to family 1 or family 3. The broad function of these different families of extracellular solute-binding protein is known; for example, cluster 1 solute-binding proteins are specific for malto-oligosaccharides, multiple oligosaccharides, glycerol 3-phosphate, and iron; while cluster 3 proteins are specific for histidine, lysine-arginine-ornithine, glutamine, octopine, nopaline, and basic amino acids.[239] The assignment of specific substrate to these differentially abundant extracellular solute-binding proteins was depending on the KO number of these proteins. The assignment results suggested that most of these solute-binding proteins are related to polar amino acid transport.

5.6.3 Deletion of *hgcAB* genes significantly impacts electron transport chain

As mentioned above, comparative analysis of these two different strains suggests that glucose metabolic process, generation of precursor metabolites and energy process were significantly upregulated in Δ*hgcAB* mutant strain. Since generation of precursor metabolites and energy process is a very broad category which contains pyruvate metabolism, glycolysis, TCA cycle and electron transport, the enrichment of this large category may come from certain specific processes. We thus examined the electron transport chain first. It is known that electron transport in microbial processes is used for energy conservation, and electron carriers, such as cytochromes, ferredoxin, quinones, flavins as key components in this process.[240] We initially investigated all the proteins that had the GO terms containing the biological process as electron transport chain and we found that 24 proteins in the database are involved in this biological process, with 7 of them identified, and 5 of them significantly changed after deletion of *hgcAB* genes. Then we looked at the specific electron carriers (cytochromes, ferredoxin, quinones, flavins) in the same way and listed the results in Table 5.1. Our results suggest that most of these electron carriers were impacted by deletion of *hgcAB* genes, with the ferredoxin altered most, which is as expected since HgcB is predicted to be a ferredoxin protein. A previous paper also reported that ferredoxin is involved in the

energy metabolism of sulfate reducing bacteria, by serving as a major electron donor in a lot of redox reactions. [241] Therefore, the disturbance to ferredoxins can further lead to the perturbation on energy associated processes, as we mentioned above in the functional difference analysis. Additionally, we also examined dehydrogenases in the same way (Table 5.1) since it is the primary enzyme to catalyze the redox reactions. All of these proteins were collected together and presented in the volcano plot for better observing the differences (Figure 5.6). Given the changes observed for electron carriers and primary redox enzymes, we concluded that deletion of *hgcAB* genes significantly impacted the electron transport chain of *Desulfovibrio desulfuricans* ND132.

5.7 *hgcAB* genes were involved in acetate formation pathways

Since glucose metabolic process, pyruvate metabolism, glycolysis and TCA cycle are included in the central carbon metabolism, we thus mapped our proteomics data on the central carbon metabolism pathways, including Glycolysis/Gluconeogenesis, TCA cycle, cysteine and methionine metabolism, and reductive acetyl CoA pathway. By comparing the abundance of proteins that are involved in these pathways, we found that all the pathways leading to the generation of the final product acetate was significantly upregulated in the mutant strain (Figure 5.7a).This result indicated a possible linkage between *hgcAB* genes deletion and acetate generation.

The upregulated acetate formation pathway in mutant strain also raised another question about why mutant strain needs to generate more acetate. It could simply be the requirement to regenerate NAD^+ and to recycle the coenzyme A (CoASH) for the central metabolism, or the requirement to use acetate as the carbon source when pyruvate is consumed[242]. In order to investigate the relationship between *hgcAB* genes and acetate formation, we monitored the acetate accumulation in the growth media after growing wild type strain and mutant strain with different electron donors and electron acceptors (e.g., pyruvate and fumarate; provide acetate and fumarate). Results showed that this organism is not able to use the acetate as the electron donor, evidenced by the middle figures that acetate is accumulating all the time in both

Table 5.1. Summary of identification of electron carriers and the primary enzyme (dehydrogenase) that are involved in electron transport chain.

Electron carriers / dehydrogenase	Proteins in database	Porteins identified	Proteins significantly changed
cytochromes	56	26	9
ferredoxin	40	24	14
quinones	15	5	2
flavins	14	12	2
dehydrogenase	67	49	22

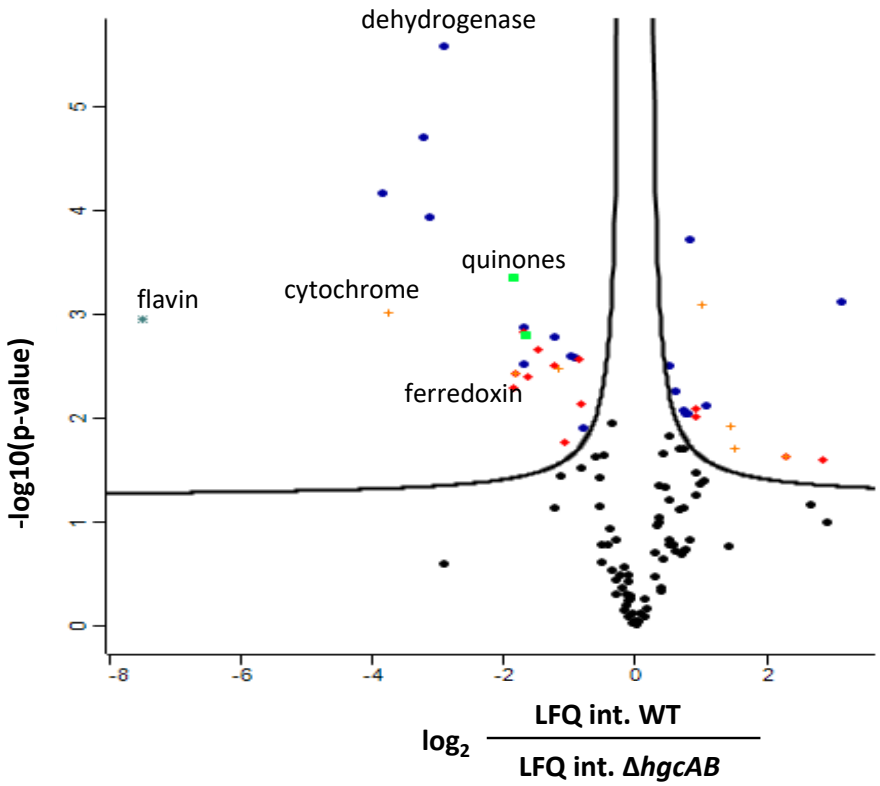


Figure 5.6. Comparison of all the electron carriers and dehydrogenase between wild type strain and mutant strain. Different color and different symbols represents for different group proteins. Blue filled circle (dehydrogenase); green filled square (quinones); aqua filled star (flavin); orange cross (cytochrome).

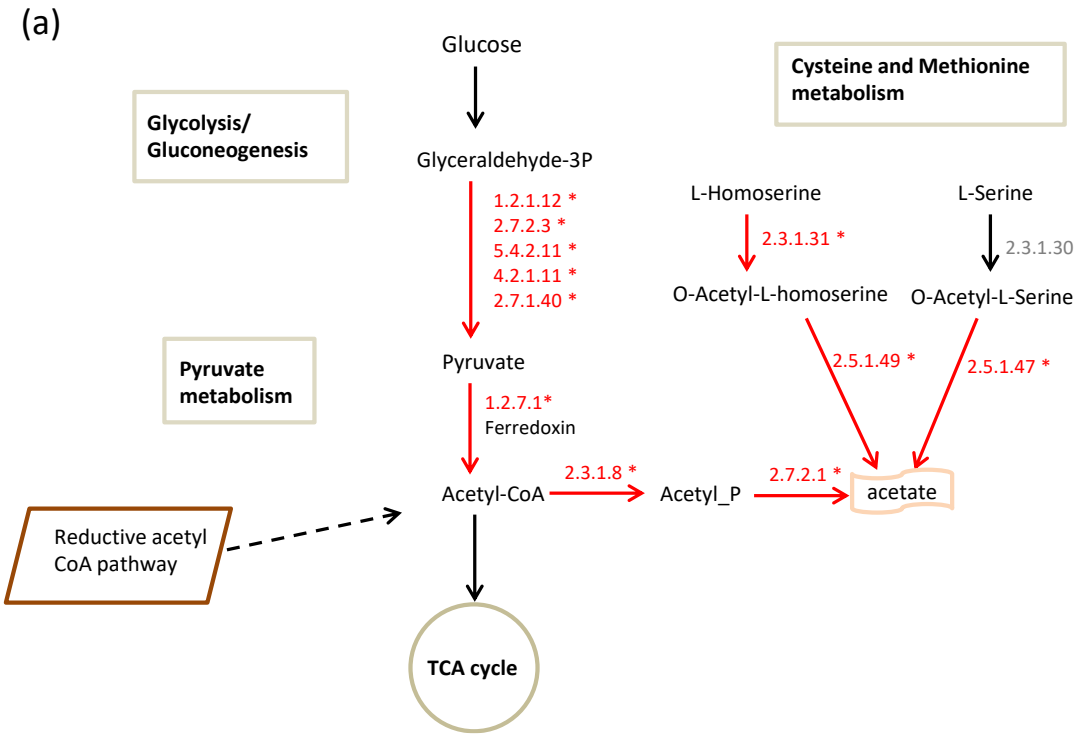


Figure 5.7. (a) Overview of the central carbon metabolism. Number showed in the figure refers to the EC number (Enzyme Commission number). The red color indicated that proteins under this EC number showed higher abundance in mutant strain than in wild type strain, with the star * means the change is statistically significant. Grey indicated the protein was not identified in our dataset.

strains (Figure 5.7b). However, it is interesting to find that wild type strain can use fumarate for growth, but mutant strain cannot (Figure 5.7b). Acetate was not utilized by either the wild type or the mutant strain (bottom figures), indicating that acetate was not used as electron donor for this organism under provided growth condition. Under fumarate fermentation condition, it was found that fumarate was not quantitatively converted to succinate (not 1:1 ratio), but was converted to succinate plus acetate based on the observation that the molar ratio of fumarate to succinate plus acetate is 1:1. The observation that mutant strain was not able to grow via fumarate fermentation, but wild type can, provided direct evidence that deletion of *hgcAB* genes disabled the electron transport chain from fumarate to acetate, indicating that *hgcAB* genes are playing roles in central carbon metabolism that leads to the acetate formation.

As the major precursor of the acetate, acetyl-CoA sits at the crossroads position of central metabolism. Besides the glycolysis and gluconeogenesis, certain amino acids metabolism and fatty acid metabolism, reductive acetyl coenzyme A pathway can also result in the generation of acetyl-CoA. We would like to highlight this pathway since early experimental studies with *Desulfovibrio desulfuricans* LS and further homology modeling analysis both indicated that Hg methylation may be linked to cellular one-carbon (C1) metabolism, possibly as part of the acetyl-CoA pathway.[73, 87] Thus, the identified proteins were mapped to this pathway and presented in Figure 5.7c. All the known proteins involved in this pathway were identified in our proteomics data and the protein abundance were all more abundant in mutant strain than in wild type strain. Since the homology of the 5-methyltetrahydrofolate CFeSP methyltransferase and acetyl-CoA synthase were not found in *Desulfovibrio desulfuricans* ND132 genome, it was unclear if the complete pathway exists in this bacterium. However, the significantly increased abundance of enzymes involved in both the branches of this pathway indicates two possibilities: either this pathway existing and was thus upregulated in the mutant strain to generate more acetate (*hgcAB* genes may or may not involve), or only part of this pathway is existing (no acetyl-CoA generated) and *hgcAB* genes are involved in this part, resulting in disturbance of the pathway after deletion of *hgcAB*

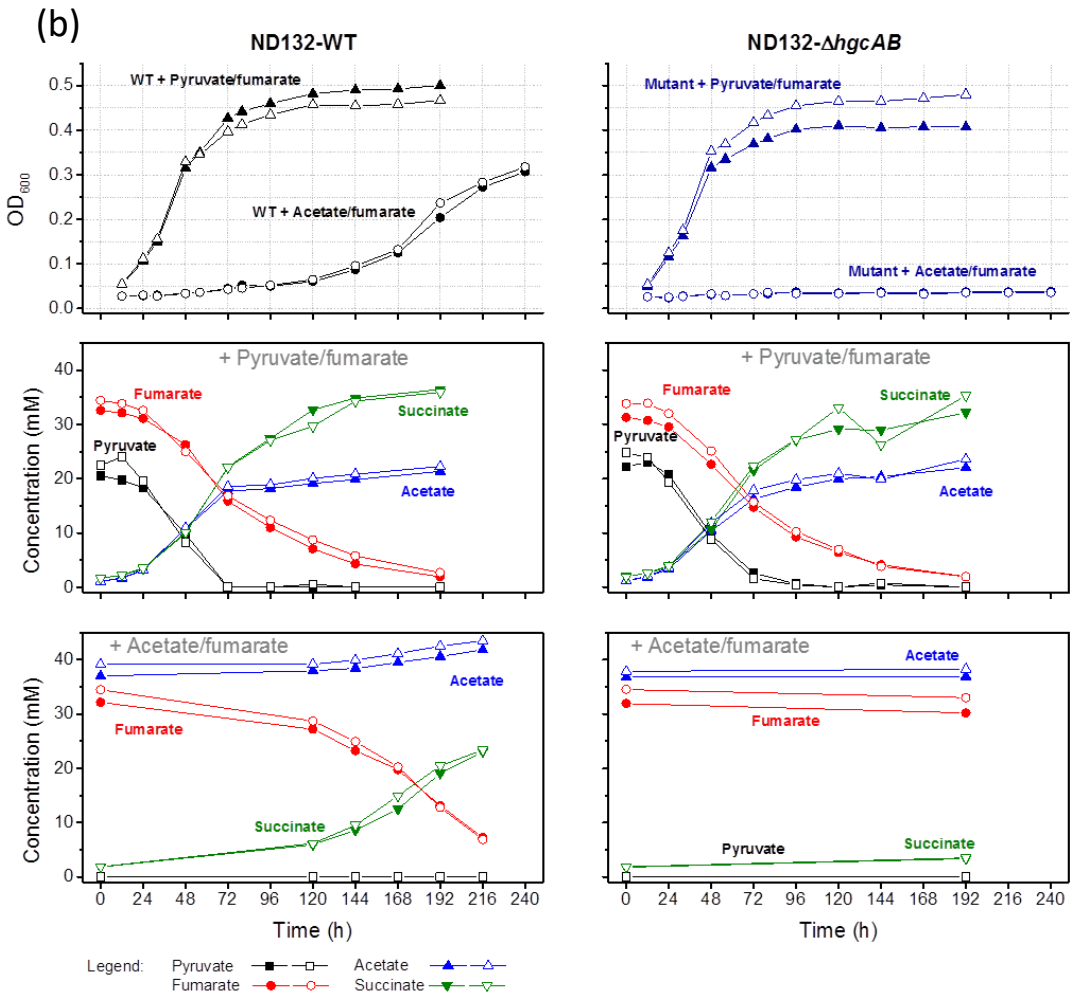


Figure 5.7. Continued (b) Growth curves of the wild type and *hgcAB* mutant strains and analyses of the production or consumption of added electron donors and acceptors (e.g., pyruvate, acetate, fumarate, succinate). This figure is provided by co-author Dr. Xia Lu.

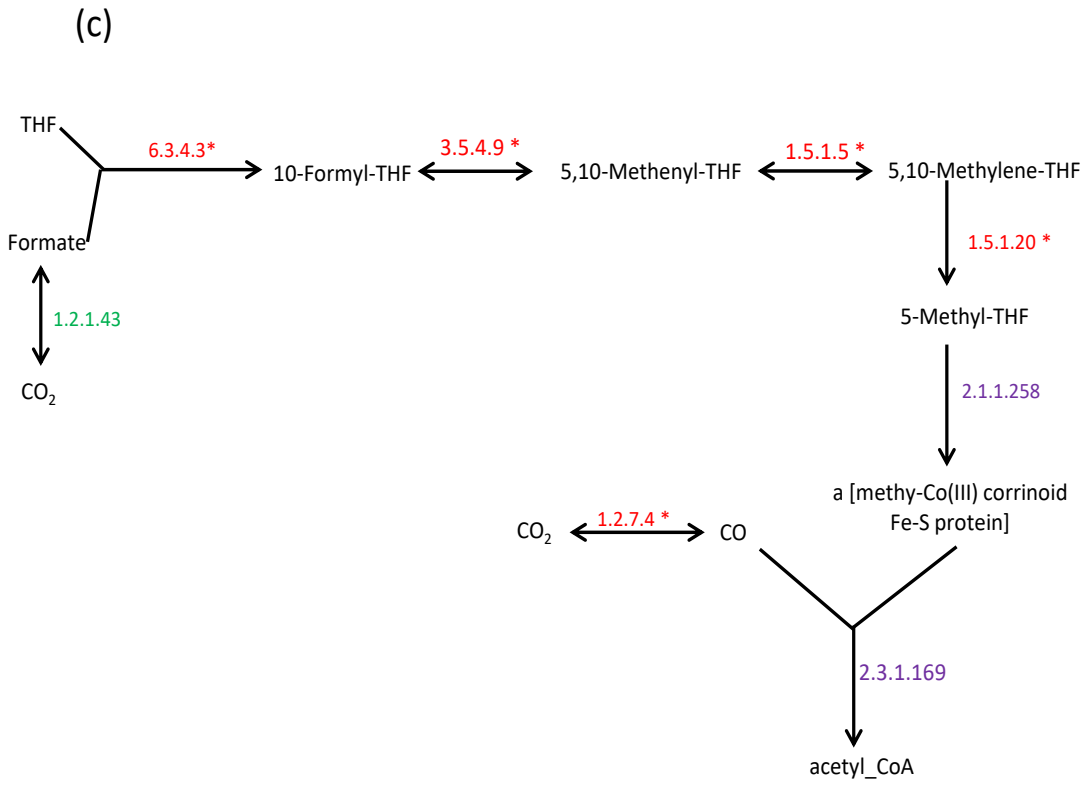


Figure 5.7. Continued (c) Overview of the reductive acetyl-CoA pathway. Number showed in the figure refers to the EC number (Enzyme Commission number). The red color indicated that proteins under this EC number showed higher abundance in mutant strain than in wild type strain, with the star * means that the change is statistically significant. Purple means that the existence of this enzyme is very uncertain. Grey indicated the protein was not identified in our dataset. Green indicates the proteins under this EC number showed higher abundance in wild type strain than mutant strain.

genes. Detailed information of all the proteins involved in Figure 5.7 can be found in the table 5.2.

5.8 Differential Protein Abundance during Hg Exposure

After examination of the impact of deletion of *hgcAB* genes, we further investigated how wild type strain responds to the mercury exposure. In order to confidently capture the response to the mercury addition, cells of wild type strain were harvested at three different time points of log phase individually, and they were then exposed to Hg(II) for 40 h (as $HgCl_2$), as described previously.[201, 243-244] Control cells without mercury exposure were performed in the same manner (incubated for 40 h) before proteome analysis. The concentration of $HgCl_2$ is very low and not expected to impact cell growth, based on previous knowledge.[245-246] The incubation time was selected based on previous results that maximum amount of methylmercury production usually peak around 24 hours.[236] In order to give the cell enough time to respond to mercury exposure, we selected 40 hours as the incubation time.

Correlation matrix showed that mercury addition caused very small fluctuations to the whole proteome in all three time points of log phase; this is within our expectation given the limited amount of mercury added. A two-sample t-test was carried out between wild-type strains with or without mercury in three different time points. We found that 69 proteins showed significantly differential abundances between wild type strains with and without mercury exposure in the middle time point of log phase (Permutation-based FDR ≤ 0.05), but most of the changes are within 2 fold. The seven proteins which had the fold change larger than 2 were listed in Table 5.3 for reference. Only 3 proteins of significantly differential abundances were found in early log phase, and none was found in late log phase. Overall, this result showed that mercury at this low concentration does not induce significant mercury specific responses that can be captured by proteome analysis.

Table 5.2. Detailed information of proteins that are involved in Figure 5.7.

UniProt	Gene	EC	Annotation
F0JG64	DND132_0596	2.7.1.2	Glucokinase
F0JIX0	DND132_2666	5.3.1.9	Glucose-6-phosphate isomerase
F0JIT4	DND132_2630	2.7.1.11	Phosphofructokinase
F0JIN6	DND132_2266		Phosphofructokinase
F0JEQ3	DND132_1575	4.1.2.13	Ketose-bisphosphate aldolase class-II
F0JFB0	DND132_0449	5.3.1.1	Triosephosphate isomerase
F0JET3	DND132_0351	1.2.1.12	Glyceraldehyde-3-phosphate dehydrogenase
F0JFA9	DND132_0448	2.7.2.3	Phosphoglycerate kinase
F0JIQ8	DND132_2288		Phosphoglycerate mutase
F0JH11	DND132_0735	5.4.2.11	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase
F0JDJ9	DND132_0150		Phosphoglycerate mutase
F0JC83	DND132_2453	4.2.1.11	Enolase
F0JF41	DND132_1635	2.7.1.40	Pyruvate kinase
F0JDH3	DND132_0124		Pyruvate flavodoxin/ferredoxin oxidoreductase domain protein
F0JDH5	DND132_0126		Pyruvate/ketoisovalerate oxidoreductase, catalytic domain containing protein
F0JG41	DND132_0573	1.2.7.1	Pyruvate ferredoxin/flavodoxin oxidoreductase, beta subunit
F0JJH0	DND132_2866		Pyruvate synthase
F0JIQ0	DND132_2280		Pyruvate/ketoisovalerate oxidoreductase, catalytic domain-containing protein

Table 5.2. (Continued) Detailed information of proteins that are involved in Figure 5.7.

UniProt	Gene	EC	Annotation
F0JIP8	DND132_2278	1.2.7.1	Pyruvate flavodoxin/ferredoxin oxidoreductase domain protein
F0JCG2	DND132_1251		Pyruvate/ketoisovalerate oxidoreductase, gamma subunit
F0JCG0	DND132_1249		Pyruvate flavodoxin/ferredoxin oxidoreductase domain protein
F0JIF3	DND132_0992		Pyruvate synthase
F0JGA1	DND132_0633		Pyruvate synthase
F0JG42	DND132_0574		Pyruvate flavodoxin/ferredoxin oxidoreductase domain protein
F0JIF7	DND132_0996	2.3.1.8	Phosphate acetyltransferase
F0JJH1	DND132_2867		Phosphate acetyltransferase
F0JIF8	DND132_0997	2.7.2.1	Acetate kinase
F0JCM4	DND132_2501	2.3.1.31	Homoserine O-acetyltransferase
F0JES2	DND132_0340	2.5.1.49	O-acetylhomoserine/O-acetylserine sulfhydrylase
F0JEX2	DND132_0390	2.3.1.30	Serine O-acetyltransferase
F0JEX3	DND132_0391	2.5.1.47	Cysteine synthase
F0JES2	DND132_0340		O-acetylhomoserine/O-acetylserine sulfhydrylase
F0JKP5	DND132_3291		Methionine-R-sulfoxide reductase
F0JKP1	DND132_3287		Cysteine synthase
F0JEH6	DND132_0322	1.2.1.43	Formate dehydrogenase, alpha subunit
F0JEH7	DND132_0323		Formate dehydrogenase subunit beta
F0JDR7	DND132_1390		Formate dehydrogenase
F0JEK4	DND132_1526		Formate dehydrogenase, alpha subunit

Table 5.2. (Continued) Detailed information of proteins that are involved in Figure 5.7.

UniProt	Gene	EC	Annotation
F0JEK5	DND132_1527		Formate dehydrogenase beta subunit
F0JK07	DND132_3053		Formate dehydrogenase, beta subunit, putative
F0JK08	DND132_3054	1.2.1.43	Formate dehydrogenase, alpha subunit
F0JDC8	DND132_2599		Formate dehydrogenase gamma subunit
F0JBP6	DND132_2345		Formate dehydrogenase subunit gamma
F0JC81	DND132_2451	6.3.4.3	Bifunctional protein FolD
F0JC81	DND132_2451	3.5.4.9	Bifunctional protein FolD
F0JC81	DND132_2451	1.5.1.5	Bifunctional protein FolD
F0JGR2	DND132_0716	1.5.1.20	Methylenetetrahydrofolate reductase
F0JCD8	DND132_1224		Methylenetetrahydrofolate reductase
F0JCP3	DND132_2520	1.2.7.4	Carbon monoxide dehydrogenase

Table 5.3. Mass spectrometry measurement of differentially abundant proteins between wild type strain and wild type strain exposed to mercury. The wild type strains were harvested in the middle time point (MT) of log phase.

Protein	Annotation	Log2 (LFQ intensity MT_1)	Log2 (LFQ intensity MT_2)	Log2 (LFQ intensity MT_3)	Log2 (LFQ intensity MT_Hg_1)	Log2 (LFQ intensity MT_Hg_2)	Log2 (LFQ intensity MT_Hg_3)
F0JDM1	Cytochrome c family protein	23.45	23.31	24.56	27.08	26.89	26.88
F0JBG0	Adenylate cyclase	21.96	21.92	21.96	24.62	25.12	25.26
F0JFR8	Spermidine/putrescine ABC transporter ATPase subunit	22.80	21.87	22.56	24.98	24.64	24.83
F0JCZ8	Type I secretion membrane fusion protein, HlyD family	24.64	23.70	24.19	25.73	25.91	26.17
F0JG13	Periplasmic binding protein	27.30	27.67	27.32	26.23	26.66	26.39
F0JF39	CBS domain containing protein	25.51	25.55	25.57	24.18	24.74	24.64
F0JD06	Nitrogenase	25.83	25.23	25.31	24.02	23.56	24.22

5.9 Conclusions

Quantitative proteomics were carried out to delineate the global response of *Desulfovibrio desulfuricans* ND132 to the deletion of *hgcAB* genes and to the exposure of mercury. For the first time, we provide the evidence to show the positive correlation between the expression level of HgcA protein and the capability of mercury methylation. Deletion of *hgcAB* genes significantly impacted a range of proteins that are associated with different subtracts of transport processes and electron transport chains. In particular, three RND family proteins that are associated with copper transport were found to be dramatically altered after deletion of *hgcAB* genes in two mercury methylation organisms: *D. desulfuricans* ND132 and *G. sulfurreducens* PCA, suggesting possible linkages between copper transport and mercury transport/methylation processes. More importantly, we found that *hgcAB* genes, besides the previously hypothesized involvement in one carbon metabolism, are involved in the acetate formation pathway. Additionally, the connection between *hgcAB* genes and acetate formation is only observed in *D. desulfuricans* ND132, but not in another mercury methylation organism *G. sulfurreducens* PCA, suggesting that *hgcAB* genes are probably associated not only with mercury methylation, but other functions as well. However, the specific role of *hgcAB* genes involved in the acetate metabolism is still not clear, and how this function can be linked to mercury methylation is yet to be determined. Future experiments have to be carried out to further validate the results presented in this dissertation: 1) although two independent growth experiments were carried out for each growth condition, these experiments need to be repeated at least one more time to confirm the growth differences observed; 2) complemented strains need to be measured and compared in the same manner to make sure the differences observed are only due to the deletion of *hgcAB* genes. The alternative way is to re-perform the sequencing on the genome of the wild type strain and mutant strain used in this study, to make sure that there is no other mutations occurred.

CHAPTER 6

Development of an MS approach for characterizing intracellular proteins interacting with a G protein-coupled receptor captured using laser-induced photo-activation

Part of the text and figures were adapted from:

Qian, C.; Hauser, M; King, S; Kaufman, S; Naider, F; Hettich, R; Becker, J. Identification of peptide binding sites within BSA by application of rapid, laser-induced covalent cross-linking combined with high-performance mass spectrometry.(J. Mol. Recognition, submitted and under review, 2017)

Chen Qian's contributions include: help experimental design, help part of the sample preparation, MS measurement, data analysis, part of the manuscript writing

6.1 Introduction

After using global proteomics in the previous chapters to examine the entire suite of protein machinery that drive microbial functional activities, we then sought to focus more sharply into a more specific study of protein-protein interactions, to ascertain how well the MS-based protein measurement approach would be to investigate specific interacting partners. To this end, we chose to initially focus on membrane proteins, since they are critical components of cellular communication and recognition. For example, G protein-coupled receptors (GPCR) are the largest and most diverse group of membrane receptors in eukaryotes and they are the targets of approximately 40% of all modern medicinal drugs. Ste2 (431 amino acids), a yeast mating pheromone receptor, is a model GPCR (G Protein-Coupled Receptor). Ste2p is participating in virtually all aspects of cellular physiology, including hormonal responses, neuronal transmission, and mediation of taste, smell, and vision.[247] Despite of the significant amount of progress that has been made in terms of the structure of G protein-coupled receptors (GPCRs) and the contacts between GPCRs and their cognate ligands outside the cell, knowledge is very limited regarding to what happened in the intracellular part when the signal was transduced across the cell membrane to activate intracellular signaling pathways, which is mediated by interaction of the GPCR with other intracellular protein partners. To this end, we developed a rapid, time-resolved method by

integrating laser-induced cross-linking/MS approach to investigate protein interaction partners with a G protein-coupled receptor *in vivo*. The photoactivatable unnatural amino acid Bpa was engineered into the different positions of Ste2p,[248] in order to capture and identify proteins associated with the intracellular domains of Ste2p in resting state and in the ligand-activated state. The photoactivatable p-benzoyl-L-phenylalanine (Bpa) is commonly used for photoactivated cross-linking of amino acid residues to form the covalent bond between two spatially adjacent residues when irradiated by UV light at a wavelength of ~360 nm. When coupling with high resolution mass spectrometry, cross-linked peptides along with the specific interacting sites can be identified with the substitution of Bpa into specific sites in a peptide or protein. Ste2p was engineered with FLAG tag at the C–terminus to allow affinity purification. To differentiate between Ste2p expressed on the plasma membrane vs. inside the cell, MTSEA-biotin was used to label the T199C residue engineered into the second extracellular loop of Ste2p, which is only surface exposed in the plasma membrane associated receptor.

Due to the difficulty of the purification of Ste2p associated complex, the use of laser-activation to facilitate cross-linking of Bpa-containing molecules to proteins was validated in the model system: bovine serum albumin (BSA) and two peptides, the tridecapeptide budding-yeast mating pheromone α-factor and the decapeptide human gonadotropin-releasing hormone (GnRH) (Figure 6.1). In contrast to chemical cross-linking agents, photoactivation can be temporally controlled, and photoreactive cross-linkers can be used to specifically target interactions between the proteins of interest. The benzophenone derivative of phenylalanine p-Benzoyl-L-phenylalanine (Bpa), has been widely used as a photoreactive cross-linker.[249] This phenylalanine analog can be chemically incorporated into peptides by solid-state synthesis[250] or genetically inserted into proteins in living cells using an orthogonal tRNA-tRNA synthetase pair which incorporates Bpa in response to a specific nonsense codon, such as the UAG amber engineered into specific sites of the target protein.[251-252] A Bpa-labeled protein engages in non-covalent interactions within a cell lysate or the interior of a living cell until the introduction of UV light energy into the system

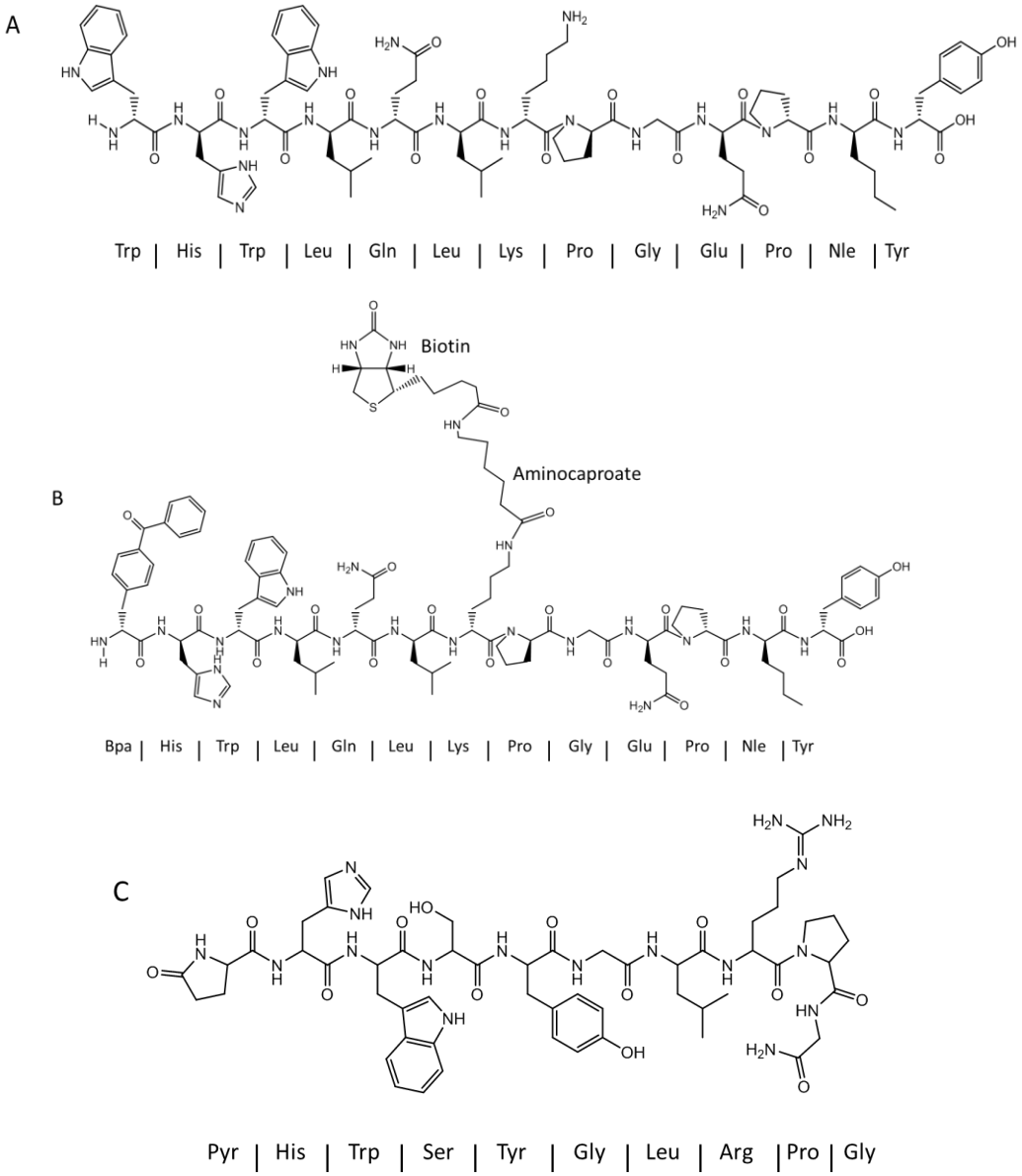


Figure 6.1. Peptide structures. (a) α-factor, with Nle12 substituted for Met12 (b) [bioBpa]α-factor, with Bpa1, biotin conjugated to Lys7 via aminocaproate, and Nle12. Biotinylated α-factor has the native Trp1 in place of Bpa1. (c) GnRH, pyroglutamic acid (Pyr) is the native residue at position 1. For biotinylated activates the photoreactive group, which then forms a bond to nearby interacting proteins. The resulting protein complex GnRH, Pyr1 is substituted with Biotinyl-(Gln[1]). The figure is provided by co-author Dr. Melinda Hauser.

can then be isolated and analyzed by biochemical or mass spectrometric approaches. This methodology has been used to identify active sites of enzymes, map protein channels, study interactions in protein complexes, find protein interacting partners with antimicrobial peptides, and examine receptor-ligand binding.[253-255]

A typical source in a UV chamber used to initiate cross-linking reactions is a 350 nm UV bulb; the recommended exposure time to activate photoreactive amino acids, such as Bpa, ranges from 30 to 120 minutes.[256] The exact cross-linking time varies, depending on the specific UV source, but is generally on the order of tens of minutes. Such long irradiation periods increase the chance of side reaction and non-specific crosslinking.[257] To interrogate transient protein-protein interactions in living cells, it is essential to complete the cross-linking reaction on a time scale of seconds rather than minutes. One obvious means towards this end is to deliver the necessary energy on a shorter time scale. A pulsed UV laser can generate a very short burst of irradiation with extremely high peak power, much higher than would be produced if it were operated in continuous wave mode. For example, a 355 nm, 1W laser operating at 50 kHz can generate peak power in the range of 3000 W per pulse. Unlike chamber-based irradiation where the energy is not directed onto the sample, the beam of the laser can be focused into a vessel holding the proteins of interest, ensuring the delivery of the UV energy directly to the sample.

In this report, we assessed the ability of a pulsed UV laser to stimulate the cross-linking using the model system of a Bpa-containing peptide, the tridecapeptide yeast pheromone α-factor, and bovine serum albumin (BSA). Serum albumin is a blood-borne protein which serves to bind, transport and deliver substances, including hormones and drugs, via the circulatory system.[258] The analysis of serum albumins, including crystal structures, solution structures and biochemical approaches reveals the presence of seven evolutionarily conserved high-affinity binding pockets for medium and long chain fatty acids, four additional sites for short chain fatty acids, as well as numerous sites which bind drugs or other ligands.[259-260] The capacity of BSA to bind hydrophobic ligands with high affinity, coupled with its well documented use in LC-MS as a model protein prompted its use as a substrate for Bpa-α-factor cross-

linking. In addition, in the study reported herein we investigated cross-linking of BSA to the decapeptide gonadotropin releasing hormone. GnRH (also known as luteinizing hormone releasing hormone, LHRH, or follicle stimulating releasing hormone, FSHRH), is a neurohormone produced in the hypothalamus and carried via the bloodstream bound to serum proteins to the pituitary gland where it binds its cognate G protein-coupled receptor. This peptide has sequence homology to α-factor and it has been demonstrated that α-factor can bind to and activate the rat GnRH receptor. The goals of our study were two-fold: (1) to determine the parameters under which UV laser energy would effectively stimulate photoactivation of [bioBpa]α-factor and GnRH cross-linking to BSA and (2) to utilize mass spectrometry to ascertain the specificity of the interaction. An outcome of this study was the determination of the site(s) of the peptide-BSA interactions.

6.2 Materials and methods

Laser Cross-Linking

Nle12α-factor (α-factor, Fig. 1), Bpa1K7(bio ACA)Nle12α-factor ([bioBpa]α-factor, Fig. 1) and K7 (bio ACA)Nle12α-factor ([bio]α-factor) were prepared as previously described. [261] L-norleucine (Nle), which is isosteric with L-methionine, was incorporated at position 12 in all synthetic α-factors to replace the naturally occurring L-methionine. This prevents oxidation of the sulfur atom of methionine during peptide synthesis and purification. Replacement of Met with Nle was shown previously to result in an analogue with activity and receptor affinity equal to those of the native pheromone.[262] BSA (Fraction V) was obtained from Research Products International (Mt. Prospect, IL). GnRH (also known as LHRH, Fig. 1) peptides (LHRH Acetate Salt and Biotinyl-(Gln[1])-LHRH trifluoroacetate salt) were purchased from Bachem Americas Inc. (Torrance, CA). Peptides were prepared as aqueous stocks and stored frozen in small volumes to minimize repeated freeze-thaw events. Phosphate buffered saline (PBS: 10 mM Na2HPO4, 1.5 mM KH2PO4, 3 mM KCl, 150 mM NaCl, pH 7.4) was prepared from analytical grade reagents.

One hundred microliters of BSA (0.1 µg/µl in PBS) was combined with the various peptides in the molar ratios specified for each experiment and dispensed into a 2.0 ml Eppendorf tube kept chilled on ice. The Explore One XP 355-1 UV laser (Spectra-Physics, Santa Clara CA) controlled by L-Win, a LabView-based graphical user interface, was used to irradiate samples. Unless otherwise specified, pulse repetition frequency was 50 kHz, the diode current was set to 30% of the maximum output (5.49 A), and the irradiation time was 30s. The beam was directed through a plano-concave lens (f 75.0 mm, Ø1" UV Fused Silica, Thorlabs, Newton, NJ) to expand the diameter to approximate that of the diameter of the sample tube (10 mm), and reflected off a fused silica UV-laser mirror (Thorlabs, Newton, NJ) down into the tube. To begin the irradiation, the tube containing the sample was placed into a benchtop freezer block, chilled to -20C, the block positioned such that the aperture of the tube was in alignment with the beam, and the samples were irradiated. Upon completion of the irradiation interval, samples were mixed with 50 µl 3X SDS sample buffer (30% glycerol, 15% β-mercaptoethanol, 3% sodium dodecyl sulfate or SDS, 0.01% bromophenol blue, 0.1875 M Tris, pH 6.8) and stored at -20C until used.

SDS PAGE and Protein Blotting

Samples (1.0 µg BSA per lane) were fractionated by discontinuous SDS PAGE (5% stacking gel, 10% separating gel). For MS analysis, gels were stained (50% methanol, 10% acetic acid, 0.2% Coomassie Brilliant Blue R-250) for at least one hour and destained (10% acetic acid, 10% methanol, 2% glycerol) overnight. The bands were excised and place into Eppendorf tubes, washed 3 times with distilled water and stored without liquid at -20C until used for MS analysis. Multiple lanes were run in parallel and bands were pooled to provide sufficient material for MS analysis.

For detection of BSA cross-linked to a biotinylated peptide the laser cross-linked BSA samples were transferred to Immobilon-FL PVDF membrane (Millipore, Billerica, MA). The membranes were allowed to air dry at room temperature, re-wetted in Odyssey Blocking Buffer and processed for detection (IRDye 800CW Streptavidin conjugate) using the Odyssey CLx Imaging System (Buffer, Dye conjugate

and Imaging System, LI-COR Biosciences, Lincoln, NE) as specified by the manufacturer. Band intensity was measured using the Image Studio 4.0 software (LI-COR Biosciences).

Double enrichments of Ste2p interacting protein complex

Cells expressing Ste2p variants (T199C, K151Bpa and T329Bpa) cultured overnight (50 ml) in MLTU medium and used to start a 500 ml overnight culture in the presence of 1 mM Bpa in MLTU. The cells were harvested, washed with PBS (2X) and finally resuspended to a final volume of 12 ml. The cell suspensions (T199C, K151Bpa and T329Bpa) were divided into 24-0.5 ml aliquots and held on ice. Each aliquot was irradiated (30s/50% power) and then re-pooled back together into a 50 ml conical tube. Upon completing the irradiation of the three sets of samples, the pooled cells (12 mls) were biotinylated with MTSEA biotin. The tubes were incubated in water at room temperature for 5 minutes to raise the temperature. To initiate the reaction, 60 µl MTSEA biotin (7.6 µg/µl) was added to each tube. The tubes were incubated at room temperature for 5 minutes with end-over-end mixing. The reaction was terminated by the addition of 3 mls ice cold sodium citrate/potassium phosphate buffer (0.25 M, pH 4) and the tubes chilled on ice for 5 minutes. The cells were pelleted at 4ºC and washed 3 times with 10 volumes of PBS. The cell pellet was ultimately resuspended to a final volume of 8 mls with ice cold PBS and divided into 16 x 0.5 ml aliquots for homogenization with glass beads and membrane isolation according to the normal protocol. Upon completion of the membrane isolation, protein concentration was determined. An aliquot of the total membrane protein was diluted in sample buffer and stored for immunoblot analysis. The remaining protein was solubilized in New RIPA buffer (13 mls final volume – see volumes above) overnight with end-over-end mixing at 4ºC. The solubilized material was cleared by centrifugation (4ºC, 5’ at 4300 RPM). A sample of the solubilized material was retained for immunoblot analysis. It was diluted to 0.5 mg/ml (T199C) or 2 mg/ml (K151, T329) with sample buffer and stored for immunoblot analysis. The cleared supernatant was supplemented with washed FLAG beads (250 µl, 50% slurry) and incubated O/N at 4ºC with mixing. The FLAG bead supernatant (100 µl) was collected, supplemented with sample buffer and stored for immunoblot analysis. The FLAG beads were washed and

bound protein eluted with Sample Buffer (200 µl). An aliquot of the eluate (20 µl) was retained and supplemented with sample buffer for immunoblot analysis. The remainder of the eluted material was diluted into New RIPA buffer (15 ml final volume) and supplemented with 150 µl (50% slurry) NeutrAvidin beads. The mixture was incubated O/N at 4ºC with mixing. An aliquot of the bead supernatant was retained and supplemented with sample buffer for immunoblot analysis. The beads were harvested and washed according to the standard protocol. A small aliquot of the beads was removed and bound material was extracted from the beads with sample buffer. The remaining beads were split into two equal aliquots, all supernatant was removed and the beads stored at -20ºC for MS analysis. The samples collected during the enrichment procedure were analyzed by SDS-PAGE followed by FLAG immunoblot.

Protein extraction and proteolytic digestion

Protein extraction from the gel bands and subsequent digestion was performed by following standard in-gel digestion protocol for mass spectrometric characterization of proteins.[32] Briefly, the bands were rinsed with two successive additions of 1 mL water before being crushed with a pestle in 100 µL of 50 mM acetonitrile/ 50 mM ammonium bicarbonate (50:50, v/v). An additional 500 µL of 100% acetonitrile was added to dehydrate and shrink the pulverized gel fragments, then as much liquid as possible was removed without disturbing any of the gel fragments. Twenty µg of porcine trypsin (Sigma Aldrich, St.Louis, MO) was dissolved in 1.5 mL of a buffer solution consisting of 10 mM ammonium bicarbonate (pH = 7.5) and 10% acetonitrile (50:50, v/v). This solution was added to the crushed, dehydrated gel pieces until the fragments were entirely submerged (≈50 µL). The gel fragments were rehydrated at 4ºC for 2 hours before being placed in a 37ºC water bath to digest overnight. The following morning, an extraction buffer containing 100% acetonitrile + 5% (vol/vol) formic acid was prepared, and was added to the digested peptide samples in a ratio of 1:2 sample volume: extraction buffer. After a 15 min incubation period at 37ºC, the solution was filtered through a 10 kDa molecular weight cutoff spin column filter (Vivaspin 2, GE Healthcare Life Sciences, Pittsburgh, PA) to recover the peptides for final clean up.

Beads-bounded Ste2p interacting complex was directly digested by trypsin following the on-beads digestion protocol.[263] After two times wash of the beads with 1 ml of cold 100 mM ammonium hydrogen carbonate (AMBIC) solution, which is freshly made and kept at 4 °C, around 100 ng of trypsin in 100 mM AMBIC was directly added onto the washed beads. After the overnight digest at 37 °C in a water bath without agitation, another 100 ng of trypsin was added and digest it for 4 h at 37 °C. After digestion, beads were precipitated by centrifugation at 6000 x g for 5 minutes and the supernatant, that contains the digested peptides, was collected.

Nano-2D-LC-MS/MS Analyses

After trypsin digestion, the peptide fragments were separated by HPLC. Samples were pressure-bomb loaded onto a trap column (150 µm i.d.), consisting of 2cm reversed phase (C18) resin (Aqua, Phenomenex) and washed offline with 100% solvent A for 20 minutes. The separation of peptides was later carried out on an in-house pulled analytical column (100 µm i.d.), packed with 12cm reversed phase (C18) resin (Aqua, Phenomenex) with gradients from 0 to 50% B (45 min); 50% to 100% B (5min); 100% B (7min). LC solvents consisted of A: 95% H_2O, 5% ACN, 0.1% FA, and B: 30% H_2O, 70% ACN, 0.1% FA.

The analytical column was interfaced with a hybrid LTQ Orbitrap Elite mass spectrometer (ThermoScientific, Waltham, MA) operating in a data-dependent mode. Each full scan acquired in the Orbitrap over the mass range of m/z 400-1700 at R=30,000 (m/z=400), was followed by HCD (Higher-energy collisional dissociation) fragmentation of the 10 most abundant ions at normalized collision energy of 30. HCD was used to achieve higher energy dissociation for the higher charge state ions. Fragment ions were detected in the Orbitrap at a resolution of 15,000 (at m/z 400), which was essential to identify the higher charge state products. Precursor ions with unassigned charge or charge state 1 were rejected for further fragmentation. Dynamic exclusion was enabled to deepen the measurement, with the following settings: repeat count was 1, the repeat duration was 30 s, and the exclusion duration was 60s.

Identification of Cross-Linked Products

MS/MS spectra were searched against the database which contains the amino acid sequence of BSA and [bioBpa]α-factor peptide or biotinylated GnRH peptide. StavroX (version 3.6.0) was used for an automated comparison of MS and MS/MS data and identification of cross-linked products. Precursor ions were compared with theoretical masses of cross-linked products with mass accuracy precision set at 10 ppm and signal/noise ratio at 2:1. Fragment ions were identified with precision at 0.1 Da. Bpa was set as the cross linker with no mass loss and no preference of cross linking to any amino acid. For GnRH cross linking experiments, biotin was set as the non-specific cross linker. Oxidation of methionine was set as a variable modification. Up to six missed cleavages were allowed for Arg and Lys.

Cross-finder (Version 1.0, OMICtools, Rouen, France) was also utilized for searching the cross-linked products, to increase the confidence of the results.[264] Precursor ions were compared to theoretical cross-linked product masses with precision set to 10 ppm. Fragment ions were identified and scored with precision at 50 ppm. Bpa was chosen as the non-specific cross linker with no mass loss. Oxidation of methionine was allowed as a variable modification. Up to 6 missed cleavages were allowed for arginine and lysine. The top 30 ions in each MS2 were considered for scoring, only spectra in which 3 residues or more could be identified were considered for scoring and only MS2 spectra whose total ion chromatogram constituted at least 5% of the total TIC of their associated MS1 were considered for scoring. Only candidates with scores greater than 100 were retained and presented in the final summary.

Computational prediction of BSA and alpha-factor binding

UCSF DOCK (version 6.7) and UCSF Chimera (version 1.11.2) were employed to both provide a visual model of alpha-factor bound to BSA and to aid in predicting the most likely binding site of the ligand. The docking procedure followed previously reported protocols and parameters. In brief, an x-ray crystallographic model of dimeric BSA (PDB accession number: 4F5S) with bound triethylene glycol was downloaded using Chimera's built-in structure retrieval tool.[265] The preexisting ligand and all hydrogens

were removed, and the model was made monomeric by removing one of the pair of BSA molecules. A model of α-factor was separately created with Chimera's peptide builder tool. Due to software constraints regarding unnatural amino acids, Trp was subsituted for Bpa in the virtual peptide used for binding-site prediction. The van der Waal's surface of the protein was generated by Chimera and used by DOCK to determine potential binding sites. The α-factor peptide was designated as a flexible ligand, allowing DOCK to alter bond angles to find the lowest-energy orientation for binding. A scoring grid was calculated for the largest candidate site before being passed to the dock6 program for final energy minimization.

6.3 Optimization of laser activated cross-linking of [bioBpa]α-factor to BSA with respect to irradiation time and energy

BSA incubated in the presence or absence of a 100-fold molar excess of [bioBpa]α-factor was irradiated for 0, 10, 30 or 60 seconds using the Explorer One XP laser at a pulse repetition frequency (PRF) of 50 kHz. The laser output at 30% and 50% of the maximum diode laser current delivering 3.9 x 10^{-6} J per pulse and 17.5 x 10^{-6} J per pulse, respectively, was assessed for the ability to promote cross-linking. Irradiated BSA samples were transferred to a membrane and probed with an IR dye conjugated to streptavidin to detect [bioBpa]α-factor cross-linked to the protein. Band density was normalized to the most dense band (60s/50% laser current =100%). In the absence of laser energy, no signal was visually observed in the presence of a 100-fold molar excess of [bioBpa]α-factor [bioBpa]α-factor (Figure 6.2), although the band quantitation indicated a 5% signal at the molecular weight for the cross-linked product indicating that there was a very limited non-specific association between the peptide and BSA. Biotinylation increased as a function of time and laser output, with the highest signal intensity associated with sample irradiated for 60 s at 50% laser output. At the lowest irradiation time (10s), the band density doubled (27% to 64%) as the laser output increased from 30% to 50%. At the longest irradiation time (60s) increasing laser current from 30% to 50% resulted in a 1.4-fold increase in band density. However, as laser time and intensity increase, there was also an increase in high molecular weight forms of the

BSA-[bioBpa]α-factor complex as well, suggesting an increased tendency towards BSA aggregation. An increase in low molecular weight smearing of signal also increased as a function of time and laser output, likely due to cross-linked concatamers of [bioBpa]α-factor or degradation of BSA cross-linked to [bioBpa]α-factor (Figure 6.2).

Irradiating samples at 30% laser output for times as short as 1 s promoted cross-linking of the [bioBpa]α-factor to BSA, although at only 14% percent of that observed for the 30s irradiation (Figure 6.3A), so cross-linking is possible on much shorter time scales, but at the expense of yield. The software indicated that the energy per pulse at a pulse repetition frequency (PRF) of 50 kHz and laser output set at 30% was 3.9×10^{-6}J/pulse. Irradiating a sample with this energy for 30 s at 50,000 pulse/sec results in the delivery of approximately 6 J. Modulating irradiation time, laser current and PRF, a constant amount of energy, 6 J in this case, can be delivered to the sample in different ways. Samples were irradiated at either 50 kHz or 200 kHz adjusting the diode current and time such that 6J of energy was delivered per sample. The intent was to determine if shorter irradiation times at higher laser fluence was more effective or less effective than longer times at reduced laser fluence. It was determined that the method by which the 6 J was delivered did not influence significantly the amount of cross-linked product generated in all conditions except for 90% current for 5 sec that gave a higher amount of cross-linking (Figure 6.3B). The parameter which is important is the amount of energy delivered. The 30s/30% laser output at a PRF of 50 kHz was therefore used as the standard parameters for cross-linking to maximize cross-linked product formation while minimizing the formation of intermolecular cross-linking of BSA. In comparison to the laser-induced cross-linking, no detectable cross-linking was observed using a standard UV bulb (365nm) after one minute irradiation. Increasing cross-linking time to 10 minutes was required for a barely detectable α-factor-BSA cross-linking (data not shown).

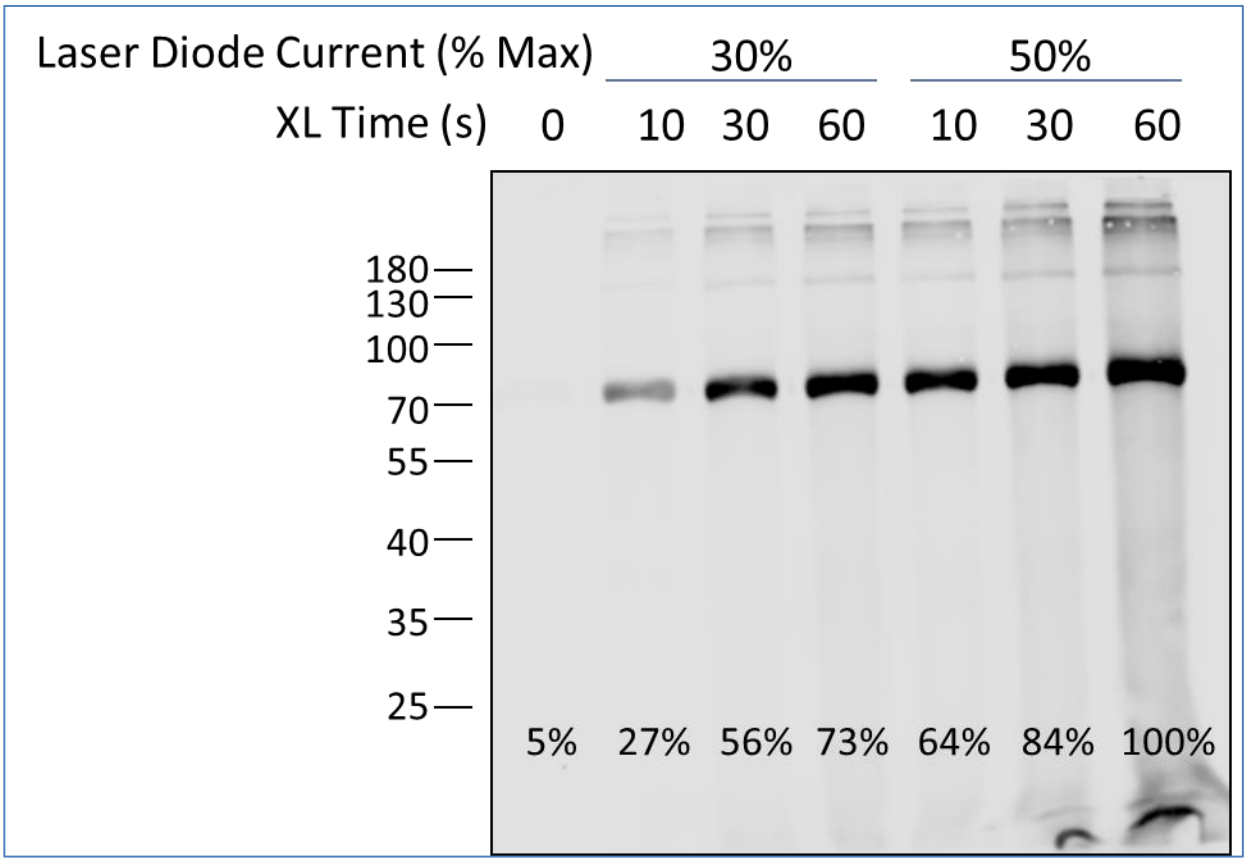


Figure 6.2. BSA – [bioBpa]α-factor cross-linking as a function of time and laser output. BSA solution was irradiated in presence of a 100-fold molar excess of [bioBpa]α-factor for the times indicated. The irradiation was conducted at either 30% or 50% laser output and the cross-linked product detected on immunoblots using an avidin-conjugated IR dye. Values under each lane are the band density normalized to the maximal signal at 50%/60 s (=100%). The figure is provided by co-author Dr. Melinda Hauser.

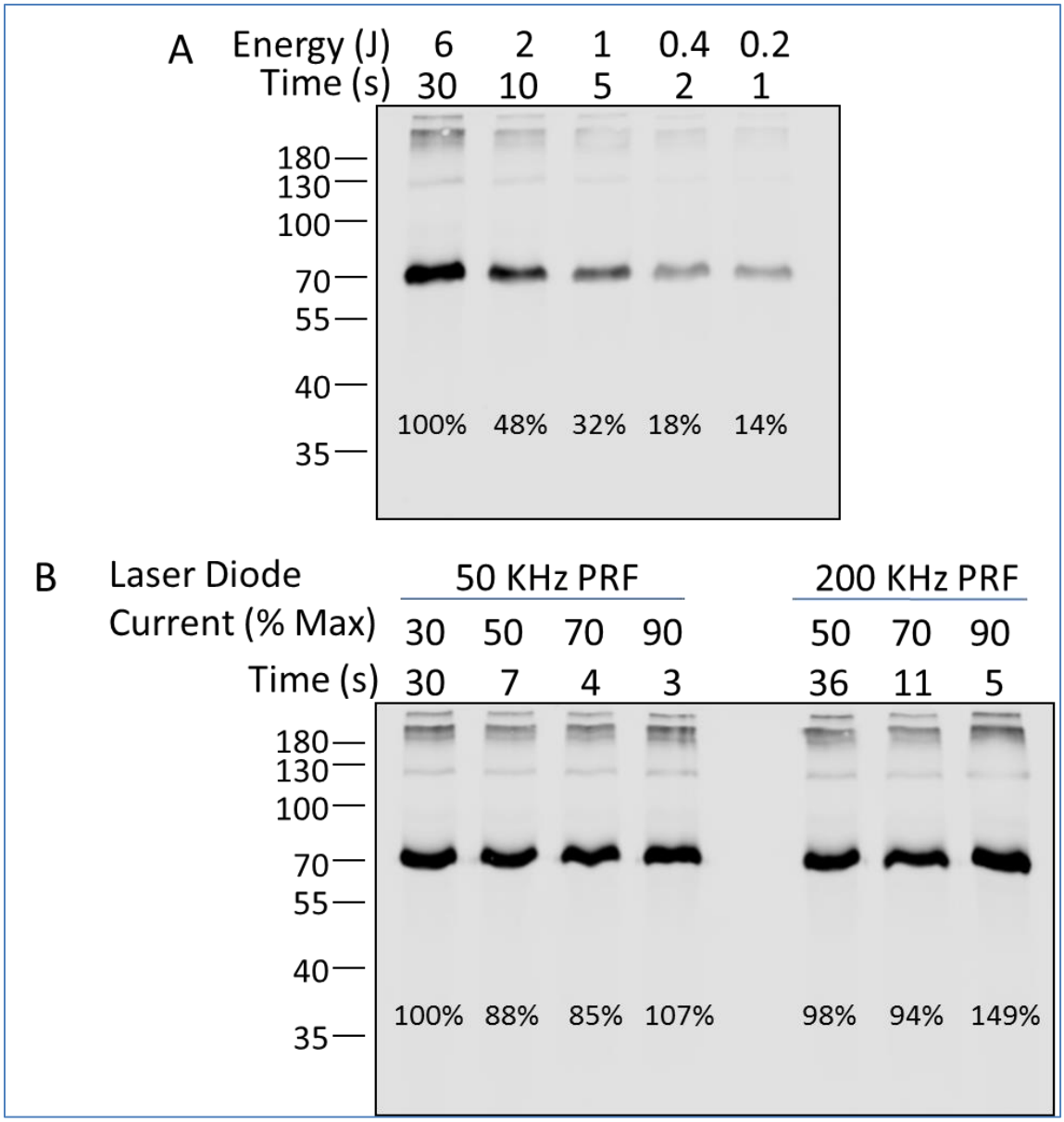


Figure 6.3. Cross-linking as a function of time and energy delivered. A.) The laser pulse repetition frequency was held at 50 kHz and samples irradiated for the times (s) indicated at a laser output of 30%. The energy delivered during this time (J) is also reported. B.) The laser diode current was varied and the samples irradiated to deliver 6 J of energy at pulse repetition frequencies of 50 kHz or 200 kHz. Values under each lane are the band density normalized to 30%/30 s (=100%) on each panel. The figure is provided by co-author Dr. Melinda Hauser.

6.4 Identification of Cross-Linked Products by HPLC-MS/MS

High resolution MS measurements were carried out for the analysis of the cross-linked peptides. Bands containing BSA-[bioBpa]α-factor cross-linked material were excised from stained SDS gels and processed for HPLC-MS/MS analysis. A "high-high" MS approach with a LTQ-Orbitrap-Elite was employed to measure both the parent peptide ions and fragment ions at high mass resolution. HCD fragmentation yields higher energy MS/MS spectra and also provides the charge information of fragment ions, which makes the manual interpretation of specific MS/MS spectra much easier. Note that we are able to simultaneously monitor all digested peptides, including the majority of abundant BSA peptides that do not contain a cross-linked component. While this data is not shown, it is essential to measure and identify all "normal" tryptic peptides to eliminate them appropriately and to observe and identify "unusual" higher charged peptides that may represent cross-linked candidates. Subsequently, cross-linked products were identified by StavroX. This software is designed for identifying cross links of peptides measured by mass spectrometry, and it has a Bpa setting as one default cross linker; we thus chose this software for our mass spectrometry data analysis.

Initially, BSA incubated in the presence or absence of a 100-fold molar excess of [bioBpa]α-factor under various laser conditions (irradiation times, power density) were evaluated by mass spectrometry. In order to investigate the generation of cross-linked products as a function of different laser current, four samples were examined by mass spectrometry, including controls (single BSA) and mixtures ([bioBpa]α-factor bound to BSA) under 30s/30% maximum diode current or 30s/60% maximum diode current. After cross linking, one good indicator of the presence of cross-linked peptides is high charge state ions.[53] Under various laser energy conditions, we observed that even without Bpa, intramolecular cross-linking within BSA was present, as evidenced by the high charge state ions in single BSA samples (Figure 6.4(a) and (b)).The charge state information of all the precursor ions that were recorded in MS1 scans were retrieved by Raxport program, to explicitly illustrate the impact of different laser conditions on the formation of high charge state ions in single BSA samples. For example, precursor

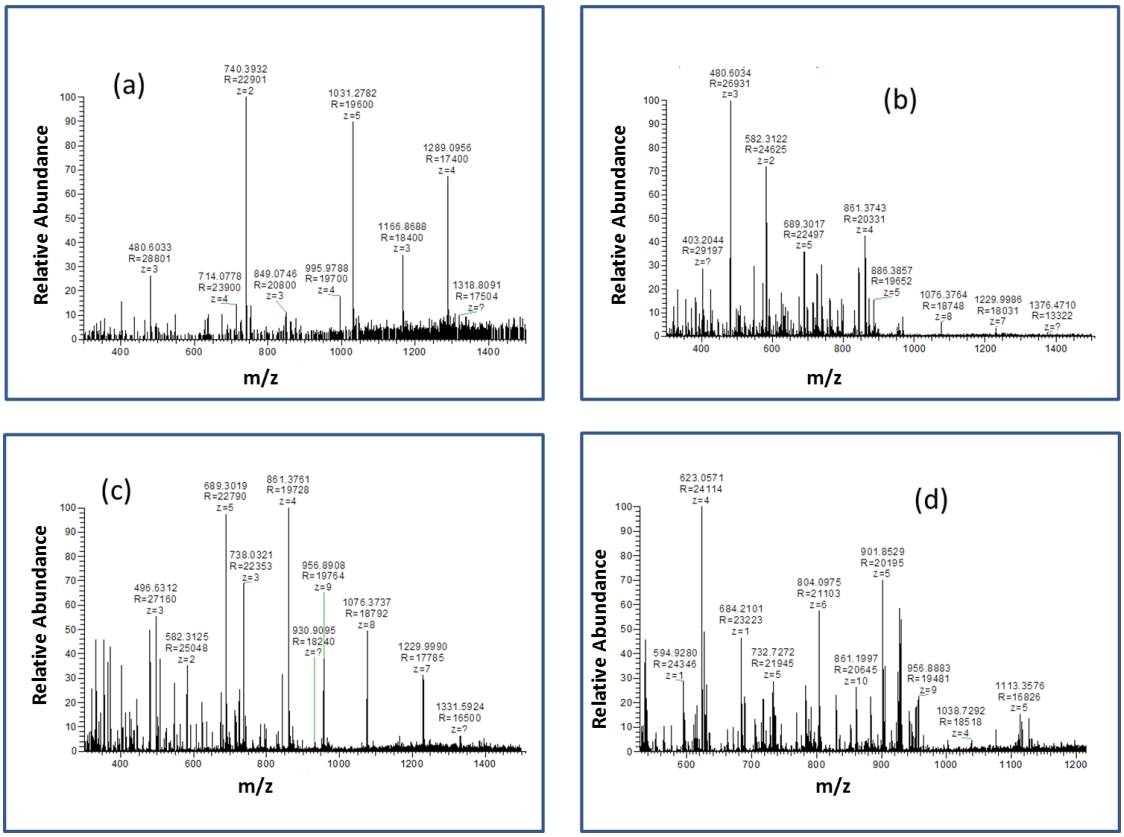


Figure 6.4. Various laser conditions (power density, irradiation times) were evaluated, both with controls (single BSA) and mixtures ([bioBpa]α-factor bound to BSA). (a) BSA, No [bioBpa]α-factor (30% laser power); (b) BSA, No [bioBpa]α-factor (60% laser power); (c) BSA with [bioBpa]α-factor (30% laser power); (d) BSA with [bioBpa]α-factor (60% laser power).

ions with charge state up to +4 and +5 were recorded in control BSA sample under 30s/30% conditions (Fig. 6.4a), while even higher charge state ions (up to +8) were recorded in the control BSA sample under 30s/60% conditions. This reveals the presence of much larger peptides (neutral molecular masses exceeding 13 KDa) in the higher irradiation samples suggesting that internal aromatic amino acids within BSA are susceptible to laser induced cross-linking and higher laser current initiated intramolecular cross-linking within BSA. This result showed also that besides the BSA aggregation observed in the gel bands (Figure 6.2) at increased laser output, formation of intramolecular cross-linking of BSA was also increased at higher laser output. Therefore, the protein gel data along with mass spectrometry data informed us that reduced laser output (30s/30%) would reduce the amount of undesirable inter-molecule and intra-molecule cross-linking, thus promoting only the Bpa-induced cross-linking.

Samples which have [bioBpa]α-factor bound to BSA under these two different laser outputs were also measured by mass spectrometry. Distinctly different high charge state ions were observed in [bioBpa]α-factor-BSA cross-linked samples (Figure 6.4c and d) relative to their corresponding control BSA samples (Figure 6.4a and b), revealing the presence of cross-linked BSA and [bioBpa]α-factor constituents. When searched with StavroX, multiple BSA peptides were found to be cross-linked to [bioBpa]α-factor, which suggests two possibilities. One is that BSA has multiple binding sites for [bioBpa]α-factor. The other possibility is that 100-fold molar excess of [bioBpa]α-factor allowed for random cross-links to residues of BSA. In order to minimize random cross-linking, the molar ratio of BSA to [bioBpa]α-factor was reduced sequentially to 1:50, 1:25, 1:12.5, 1:6.25 in order to identify the binding site with highest affinity. Protein blot results showed that signal intensity increased as a function of the amount of [bioBpa] peptide present, although there was no difference between 6.25 and 12.5-fold excess. Considering that as low as 6.25-fold molar excess of [bioBpa] pheromone still yielded significant cross-linked product and it is easier to identify the unambiguous cross-linked BSA peptides and [bioBpa]α-factor at lower molar ratio, the two lowest molar ratio samples (where BSA: [bioBpa]α-factor is 1:12.5, 1:6.25) were selected and further measured in mass spectrometry. Additionally, cross-linking

with a 10-fold excess of [bioBpa]α-factor under different laser conditions was also measured. We anticipated that the BSA peptide which contains the binding site of highest binding affinity would be consistently identified in these different molar ratio samples and under several different laser conditions.

Mass spectrometry results revealed that two BSA tryptic peptides were identified (Table6.1): Peptide 1 - HPYFYAPELLYYANKYNGVFQECCQAEDKGACLLPKIETMREK and Peptide 2 - DDSPDLPKLKPDPNTLCDEFKADEKKFWGK. Note that this table reveals distinct versions (i.e., different lengths and charge states) of the same basic two peptide regions. The 3D structure of BSA revealed that these two regions are spatially very close. When reducing the molar ratio of BSA to [bioBpa]α-factor to 1:6.25, peptide 1 was still identified, but not peptide 2. This suggested that the Peptide 1 binding region, previously identified as fatty acid binding site 1, may have the highest binding affinity for [bioBpa]α-factor. However, the signal decreased dramatically at this molar ratio, making the assignment of fragment ions more difficult, resulting in a somewhat lower confidence score in StavroX.

A second analysis was conducted using Cross-finder to independently validate the search results from StavroX. Cross-finder identified 6 different binding regions for BSA using the same set of samples under different conditions. The region which had the highest score overlaps the same Peptide 1: (RHPYFYAPELLYYANKYNGVFQECCQAEDKGACLLPKIETMREKVLASSAR) that was identified by StavroX.

6.5 MS reveals that GnRH binds to BSA at the same region as [bioBpa]α-factor

To determine the site(s) of GnRH cross-linked to BSA, mass spectrometry was employed. Cross-linked results obtained via the StavroX showed that the BSA peptide "YNGVFQECCQAEDKGACLLPKIETMREKVLASSARQR" yielded a score as cross linked to the bio GNRH. This is the same region as that identified in the BSA-[bioBpa]α-factor experiment, which supports the result of the competition experiment above. The +6 charge state of the BSA/Bio-GnRH complex at m/z= 933.629 was selected and fragmented by HCD and the ms/ms spectrum is illustrated in

Table 6.1. Peptides of BSA that are cross linked to α-factor as identified by StavroX

Molar Ratio (BSA:ALPHA_FACTOR)	Laser Power	Possible cross_linked peptide of BSA	m/z	charge	StavroX score
1:6.25	30%/30s	HPYFYAPELLYYANKYNGVFQECCQAEDKGACLLPKIETMREKVLASSARQRLRCASIQK	901.958	10	11
1:10	30%/10s	DDSPDLPKLKPDPNTLCDEFKADEKKFWGK	1109.966	5	56
	50%/7s	DDSPDLPKLKPDPNTLCDEFKADEKKFWGK	1109.966	5	61
	90%/5s	FWGKYLYEIARRHPYFYAPELLYYANKYNGVFQECCQAEDKGACLLPKIETMREK	1249.63	7	59
1:12.5	30%/30s	HPYFYAPELLYYANKYNGVFQECCQAEDKGACLLPKIETMR	1381.074	5	60
		VASLRETYGDMADCCEKQEPERNECFLSHKDDSPDLPKLKPDPNTLCDEFKADEK	1198.715	7	54

Figure 6.5. Multiple b, y ions of α peptide and β peptide were found (α peptide and β peptide are the two cross linked peptides, the peptide which has more fragments found is regarded as the leading peptide, termed as α peptide). Although the existence of y35 ion indicated that either the N-terminal Y or N amino acid of the BSA can be the crosslinking site for bio-GNRH; given the low intensity of this fragment, we did not consider it as solid enough evidence to localize the specific binding site.

6.6 Computational DOCKing analysis of BSA to ligands

Geometry optimization using DOCK6 suggested that the bound α-factor adopted an orientation similar to its biologically active geometry when interacting with the Ste2p receptor in vivo, such that the N-terminus (W1) is bound deep in a fold of the van der Waals surface of the protein, with the rest of the ligand protruding out of the fold before making a β-turn at P8 so that the C-terminus projects back into the fold at roughly half the depth of the N-terminus (Figure 6.6). An animated version of Figure 9 is submitted as Figure S1. Bovine serum albumin provides a scaffold to reasonably replicate this geometry, with the deep cleft between domains I and III of BSA serving as the analog to the opening formed by the transmembrane domains of Ste2p. DOCK6 places Trp1 in proximity to His145 (3.411 Å), Ser192 (3.308 Å), and Pro146 (3.635 Å). From Trp1, α-factor extends out of the cleft between domains I and III, staying close to the α-helix connecting domains I and II, until it makes a β-turn at Pro8 at the “top” of the cleft. At this point, Pro8 is nearest to Lys116 (3.246 Å). From here, α-factor extends back into the cleft so that at the C-terminus (Tyr13) is located 3.398 Å from Thr434 of BSA. Finally, the distances between the α-carbons of Trp1-Pro8, Pro8-Tyr13, and Trp1-Tyr13 are 20.640 Å, 15.004 Å, and 16.025 Å, respectively.

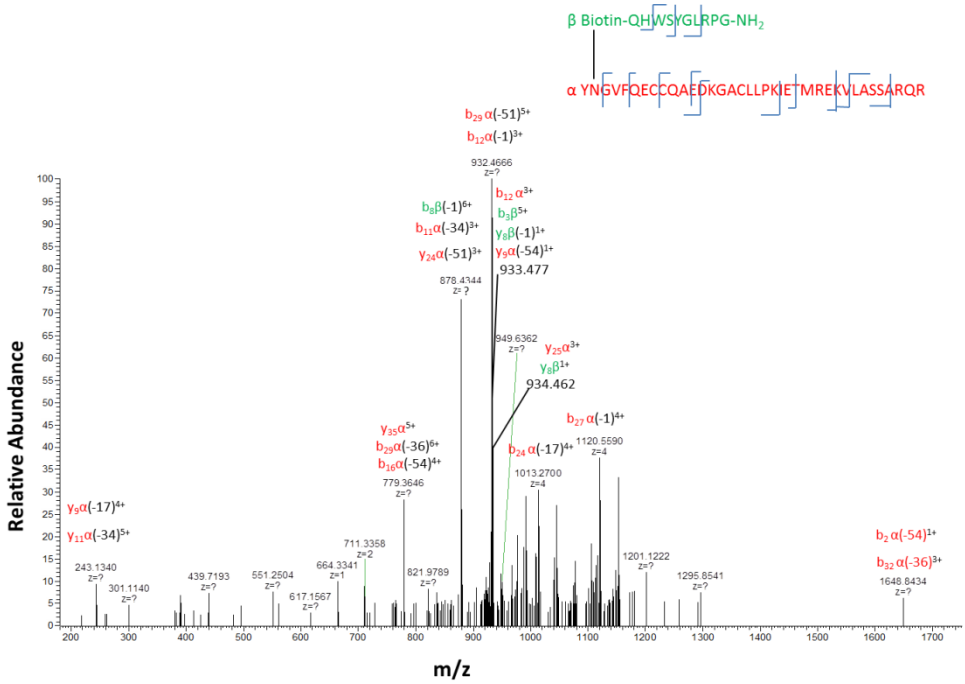


Figure 6.5. Identification of cross-linked products between biotinylated GnRH and BSA, irradiated with 30% power at 50s. Biotin was set as the cross linker, fragment ions of α-peptide are shown in red; fragment ions of β-peptide are shown in green. Note that all "b-type ions" contain the other cross-linked partner, verifying that GnRH is linked to this BSA tryptic peptide. The nomenclature of fragment ions is based on the StavroX results. In some cases, there was more than one possible fragment ion assignment, but all of these were still consistent with the same BSA peptide linked to GnRH.

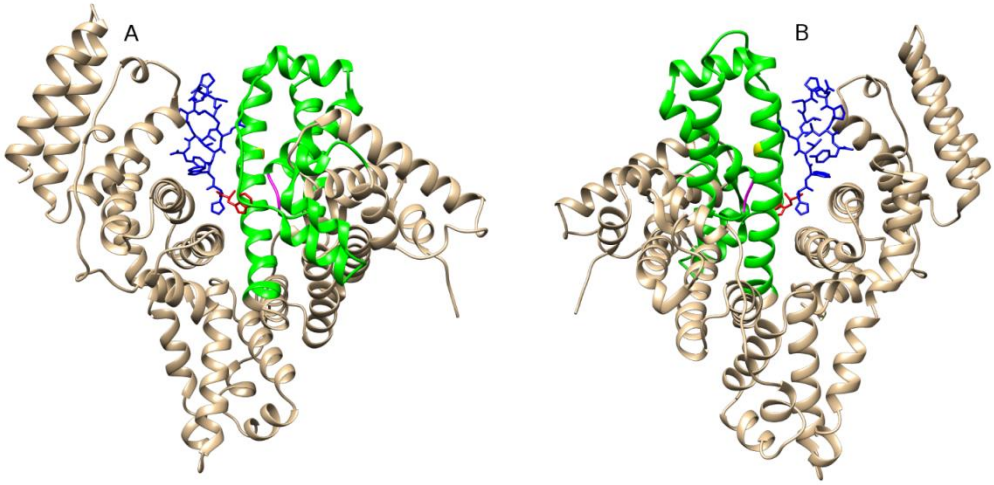


Figure 6.6. (a) DOCK-predicted binding mode of α-factor to BSA. Figure is color-coded for clarity: green represents the domain found experimentally (with StavroX) to be the primary binding region. Sequences corresponding to the green-colored domain can be found in Table 1. Red represents the crosslinking Bpa residue. In the figure, Trp is substituted for Bpa due to software constraints. Blue represents the remainder of the α-factor peptide. Magenta represents positions of His145 and Pro146 (just to the right of Bpa). Yellow represents position of Met184, although this is difficult to see due to its position in the helix. (b) The same model, rotated 180˚ about the vertical axis. The yellow Met184 residue is slightly more visible, as opposed to Figure 9-A, just to the left of the α-factor peptide. The figure is provided by co-author Tyler King.

6.7 Initial results on *in vivo* Ste2p crosslinking

Based on the encouraging methodology optimization for BSA, we proceeded to extend this approach for actual *in-vivo* crosslinking for the GPCR Ste2p. The workflow for identifying Ste2p interacting complex by coupling crosslinking, immunoprecipitation and MS measurement is illustrated in Figure 6.7. Acetyl-CoA carboxylase (YNR016C) is a biotin containing enzyme that is very abundant in yeast. This protein was identified in all samples, regardless of single or double enrichment preparations. Since this protein is known to not bind to Ste2p, it was used for normalization of spectral counts across different samples. Based on initial crosslinking results, YAL038W and YGL009C are two possible interacting proteins for Ste2p (YFL026W) (Table 6.2). While these results are interesting and encouraging, they need to be further verified with additional controls to sort out quantification issues based on spectral counts intensities. Since the initial results did not reveal many interacting protein partners for Ste2p, we surmised that Bpa at 329 position (original residue is T) in Ste2p did not capture any specific interactions, indicating that interaction sites are likely in other regions. We thus also investigated the Bpa in position 151 (original residue is K) in Ste2p for protein interaction binding. However, for laser irradiated K151 mutant strains, we only identified target protein Ste2p and acetyl-CoA carboxylase, likely due to low sample yields for just not yet optimized protein system. Thus, work so far revealed several challenges that need to be taken into consideration in further study: 1) After substitution of Bpa into specific site in Ste2p, biological activities can be lost due to the mutation in certain specific sites; thus the number of mutant strains which can be used for studying the protein-protein interaction is limited. 2) After two rounds of enrichment, the amount of retained Ste2 protein complex is very limited, especially for the mutant strain. It is possible that increasing the sample size of yeast incubation could solve this problem. 3) Since the position of the Bpa is set in the target protein, one position of Bpa probably will not capture a wide range of interactions since the intracellular region has around 100 amino acids. 4) How many Ste2p molecules are actually involved in forming the cross linking products is

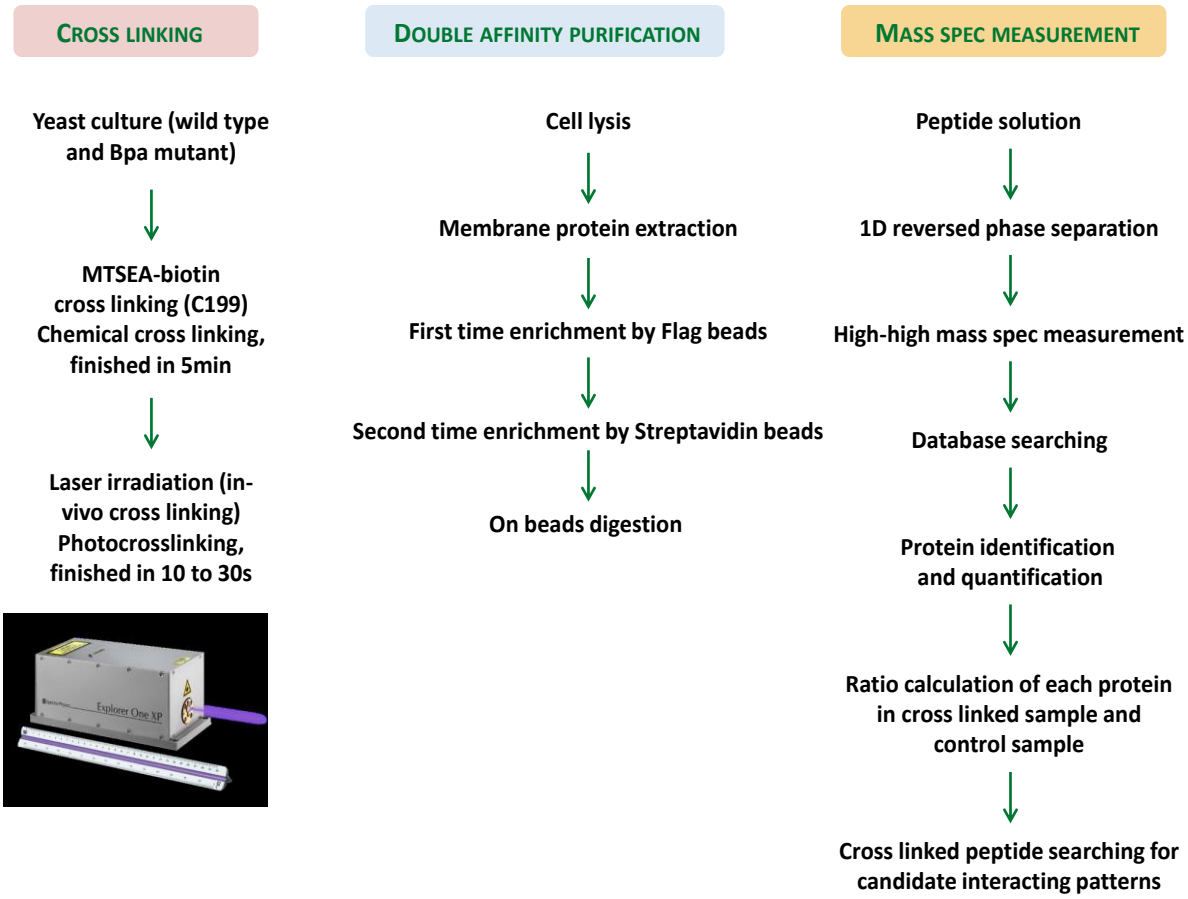


Figure 6.7. The schematically illustration of the workflow employed for Ste2p interaction analysis

Table 6.2. MS measurement results of the Ste2p interacting complex which is captured by the assistance of crosslinking and further enriched by tandem affinity enrichments.

Accession	BPA	Bpa_30	Bpa_60	Control	Control_30	Control_60
YAL005C	4.72	7.69	7.69	5.71	5.81	8.96
YAL038W	2.36	12.09	15.38	6.43	5.81	7.46
YBL075C	2.36	4.39	NA	3.57	3.73	4.48
YBR118W	18.11	13.19	NA	7.85	7.46	NA
YBR218C	7.08	13.18	10.25	5	7.05	NA
YDR502C	2.36	NA	NA	NA	1.65	5.97
YER103W	2.36	4.39	NA	3.57	3.73	NA
YFL026W	109.44	65.93	271.79	37.85	34.43	32.83
YGL008C	7.08	9.89	12.82	7.85	7.46	10.44
YGL009C	5.51	12.08	23.07	6.42	6.63	5.97
YGL062W	7.87	14.28	10.25	6.42	7.88	NA
YHR020W	2.36	NA	NA	NA	NA	NA
YLL024C	4.72	7.69	10.25	5	5.39	8.95
YLR044C	3.14	18.68	10.25	13.57	14.93	4.47
YNL178W	3.14	NA	NA	NA	NA	NA
YNR016C	100	100	100	100	100	100

unknown. 5) Spectral counting based label-free quantification method may not provide the necessary measurement dynamic range that is afforded by intensity based quantification approach or labeling strategies, which should be considered in the future study.

6.8 Conclusions

In this study, we have demonstrated the utility of laser-activated cross-linking of Bpa-containing peptides and, to a lesser extent, the ability of laser to cross-link native amino acids. Coupled with subsequent MS measurement, cross linked peptides can be successfully identified. Further, we extended our study to capture Ste2p-protein interactions in a live cell using unnatural amino acid incorporation of Bpa into Ste2p in the native state by employing this methodology. After double enrichment, surface Ste2p and its interacting proteins were successfully identified by MS; however, this work is too preliminary to permit verification of specific interacting proteins by incorporating Bpa in 329 and151 positions of Ste2p.

CHAPTER 7

Conclusions and Future Perspectives

7.1 Overview

Mass spectrometry based proteomics has emerged as a powerful and indispensable method for in-depth characterization of protein components and protein interactions of biological systems, continuously advancing the understanding of complex biological processes. In many pioneering studies, this approach was used to map the proteins to specific subcellular structures and organelles, thus building valuable repositories of information about the localization of proteins in cells and tissues. This methodology is also emerging as a powerful way to discern higher-order structural features of protein complexes, such as subunit orientation and stoichiometry.[266] More specifically, this dissertation focuses on two main applications: 1) to identify the protein 'signatures' that depict/determine the higher-order biological processes under investigation, specific to environmental microbial systems; 2) the functional characterization of protein complexes by investigating protein interactions. For the first application, a major goal of this type of proteomics analysis was the identification of proteins involved in dynamic biochemical processes, e.g. changes due to external stimuli, pH, and temperature changes, and changes by chemical treatment; however, a major shortcoming is that no information about specific protein interactions or about molecular recognition structures is obtained due to the denaturing measurement conditions. Therefore, this dissertation also explored the second application of proteomics analysis by coupling cross-linking and affinity proteomics. In this way, this dissertation is able to deliver a more comprehensive study of proteomics based applications. Each application faces its own challenges and how this dissertation work contributes to alleviate certain aspects of these challenges is fully discussed in the introduction and in each chapter. The remaining challenges and future perspective are discussed below. The methodology introduced in this dissertation can also be extended to study similar biological questions in different systems.

7.2 Conclusions and future perspectives on environmental proteomics

Conclusions

Microorganisms play critical roles to almost all biological systems on our planet. The structure and composition of the microbial communities are diverse in a wide range of environment, such as sediment and soil, marine and freshwater, activated sludge and acid mine drainage biofilms, etc.[267] Additionally, metabolic activities and cellular physiology of the microbial community frequently fluctuate with environmental changes.[268] The uniqueness and complexity of microbial communities in different environments have been extensively studied for decades; however, their response to various environmental changes and their ecological functions in certain environment has gained increasing attention in recent years.

The advent and rapid advances in "omics" technologies, including genomics, transcriptomics, proteomics and metabolomics, afford new ways to study the structure and function of environmental microbial systems that might revolutionize our inherent understanding of microbial diversity and physiology in natural communities. Decades ago, global molecular approaches were restricted to isolated organisms that can be cultured in the lab; however, currently it is possible to perform the global analyses on the whole complex microbial communities or even the entire ecosystems.[54] Among these global molecular approaches, this dissertation focus on one specific technique: proteomics, which characterizes all the proteins expressed at a given time for a specific biological sample. Proteomics is the key tool to interrogate the microbial functionality in the system level as proteins are directly associated with virtually all of the biological processes.

Therefore, environmental proteomics study is largely focusing on two types of research: 1) the proteome investigation of single organisms isolated from environmental systems, which have special functions. By deep understanding the metabolic mechanisms behind these functions, biotechnology can be developed to address different environmental problems. 2) the metaproteome investigation of microbial communities in natural ecosystems. Investigation of the whole proteome expression of a

microbial community offers an unprecedented way to understand the adaptive responses of microbial community to environmental stimuli or their interactions with each other or host cells.

The work presented in this dissertation contributes to both areas of environmental proteomics study: 1) optimizing a sample preparation method for enhancing the metaproteome measurement of soil microbial community, 2) investigating one group of single isolates that have the special capability of generating methylmercury.

Efficiently extracting the microbial proteins from soil samples is one big challenge for the soil metaproteomics experiments, as the low biomass and abundant soil-borne interferences present in most of the soil samples confound the work. To this end, we have optimized the sample preparation method by removing one well-known soil interference, humic acids, as illustrated in chapter 3. In this study, we presented an efficient way to remove humic acids from proteome samples, without impacting the protein identification and expression profile. More specifically, we first demonstrated that even small amount of humic acids can cause problems on sample loading and proteome identification. Then we showed that removing humic acids from peptides is much easier than removing them at the protein level since humic acids have two obvious different characterizations with peptides. One is the size differences, the other one is the solubility differences in acid solvent. However, humic acids share very similar characterizations with protein complex: 1) humic acids are not pure compounds, but mixtures and thus have a wide range of mass, which overlap with the mass range of proteins; 2) humic acids are also hydrophobic according to their basic structures, and thus show diverse range of hydrophobicities depending on size; 3) humic acids are insoluble in acid environments. By taking advantage of the differences between humic acids and peptides in a mixture, we demonstrated a new approach that combines pH adjustment and filtering together to enhance collection of the humic acid precipitate (containing both large and small humic acids) on the membrane, but passing soluble peptides through into the collection vial. Our data revealed that by adjusting the pH around 3 and coupling with subsequent filtering, it is possible to efficiently remove the majority of the humic acids. Further, we also demonstrated that this approach can be easily incorporated

in any currently available protein extraction method, and it did not induce any sample loss or negatively impact on protein identifications in the proteome MS measurements.

While optimization of sample preparation improves the power of proteomics approach from the technical perspective, more details about how a proteomics approach can be used to address real biological questions is presented in chapters 4 and 5. Mercury pollution is a pervasive global environmental problem, with methylmercury causing the biggest concerns since it is neurotoxic and can be bioaccumulated in the food chain. The question about how mercury methylation occurs in nature has been studied for decades, the most important information we obtained until now includes: 1) methylmercury is generated in the environment predominantly by anaerobic microorganisms, while the sulfate-reducing bacteria are the main producers of CH_3Hg^+, iron-reducing bacteria and methanogens can also be involved; 2) production of CH_3Hg^+ by the model methylating bacteria *Desulfovibrio desulfuricans* ND132 and *Geobacter sulfurreducens* PCA involves cellular uptake of Hg(II) by active transport, methylation of Hg(II) in the cytosol, and export of CH_3Hg^+ from the cell; 3) Hg methylation is an enzyme-catalyzed process and is proposed to be associated with the reductive acetyl–coenzyme A (CoA) pathway and likely to be linked to corrinoid proteins involved in this pathway; 4) two genes: *hgcA* that encodes a putative corrinoid protein, and *hgcB* that is a 2[4Fe-4S] ferredoxin, are found to be essential for mercury methylation. Following the discovery of these two genes, a mutant strain which has both genes deleted is available for two model organism *Desulfovibrio desulfuricans* ND132 and *Geobacter sulfurreducens* PCA. Chemical properties related to mercury uptake, mercury methylation and methylmercury export have been extensively compared between the wild type strain and its corresponding mutant strain; however, the knowledge of the underlying mechanisms to all these mercury associated processes is very limited. Therefore, a proteomics approach was employed to address several questions: 1) to reveal the global impacts from the deletion of *hgcAB* genes in both model organism; 2) to elucidate the pathways in which *hgcAB* genes are involved; 3) to demonstrate the relationship between *hgcAB* genes expression level and mercury methylation capability on the protein level; 4) to come up with

candidate proteins that are involved in mercury transport; 5) to examine the bacterial response to the mercury stress.

Our results showed that, deletion of these two genes caused different impacts on these two strains, with very mild perturbation on the *G. sulfurreducens* PCA proteome, but comparatively much more significant impact on *D. desulfuricans* ND132 proteome, evidenced by the number of significantly differential abundant proteins between wild type strain and Δ*hgcAB* mutant strain. The abundance levels of the corresponding proteins HgcA and HgcB are also different in these two strains. In *G. sulfurreducens* PCA, both proteins were not identified by the global proteome analysis. In *D. desulfuricans* ND132, both proteins were identified by the same MS instrument platform and same searching algorithm, although with very low abundance. We also found that in *D. desulfuricans* ND132, the abundance level of HgcA and HgcB protein is consistent with the capability of mercury methylation, by investigating cells harvested in three different time points in the log phase (early log, middle log and late log). We thus concluded that the expression of HgcA and HgcB does not need the trigger of mercury, indicating that these two genes have additional functions. Additionally, the expression level of HgcA and HgcB is also associated with the capability to generate methylmercury. Our results also revealed one same set of homology transporters that showed dramatic change after the gene deletion in both strains, and this set of transporters are annotated to be responsible for the Cu(I)/Ag(I) export, thus building the interesting linkage between Cu(I)/Ag(I) and mercury transport as well as mercury methylation. Further, in *D. desulfuricans* ND132, pathway analysis revealed that all the metabolic pathways leading to the acetate generation were significantly enhanced in Δ*hgcAB* mutant strain, indicating that *hgcAB* genes are involved in central carbon metabolic pathways. We also have the experimental evidence that Δ*hgcAB* cannot grow by fumarate fermentation, but wild type can grow, which further fortified our conclusions. Finally, we also investigated how wild type strain responds to the mercury exposure and our results showed that mercury at this low concentration does not induce significant mercury specific responses that can be captured by proteome analysis. This result is also consistent to the previous observation that low level of mercury is not toxic for the bacteria.

Remaining challenges and future perspective

The proteomics approach has emerged as an indispensable tool to investigate environmental microbes, no matter for lab cultured isolates or naturally collected microbial communities. From a technical perspective, this proteomics approach is very mature for single isolate studies in terms of the protein extraction and in-depth characterization of the whole proteome. However, the expense and time used for each sample's analysis still limits the number of samples that can be analyzed. This further causes two problems. First, as it is impossible to obtain 100% sequence coverage under one growth condition, some proteins of interest may not be identified. Second, the power of various statistical methods suffers from this undersampling issues, further impacting the biological conclusions drawing from the datasets. With development of the MS techniques, the scanning speed of MS has been dramatically improved recently. Therefore, this undersampling issues can be largely resolved in the near future.

In contrast, proteomics applications for natural microbial community, termed metaproteomics, is still at the beginning stage. It faces all the challenges from sample preparation, to in-depth measurement and comprehensive data interpretation. To understand the roles of microbial communities in their surroundings, a wide range of proteins has to be identified, which includes the low abundance proteins. Therefore, the first challenge is associated with the sample collection and preservation. For most of the environmental samples, the biomass is quite low and majority of microorganisms in the environment cannot be cultured. Ultrafiltration and flow cytometry serve as two major methods to collect environmental samples, but none of them can separate different microorganisms in the populations and obtain sufficient cell biomass for each different species. The protein information gained from the metaproteomics analysis reflects most of the abundant microbes in the populations, but hardly covers the information regarding to the rare or sparse species. This actually hinders our understanding of their roles of different microbial species in the environment. On the other hand, maintaining microbial communities at an *in situ* status is also a challenge. It is well acknowledged that the *in situ* environmental samples should not be altered too much in order to reflect the real microbial world since microbial communities

are very sensitive to environmental alterations and will respond very quickly. For some samples, it is convenient use quick preservation method at a low temperature, such as liquid nitrogen, immediately after sample collection. However, for some other samples which are collected from extreme environments, for example samples from habitats with very limited oxygen, extremely high or low temperatures or high pressure, it is difficult to maintain the natural conditions. Another obstacle for environmental metaproteomics analysis is the efficient protein extraction from complex environmental samples, which normally contains various organic and inorganic interferences, such as humic acids, lignin and various unknown degradation products. Currently, different protein extraction methods have been developed to deal with different environmental samples; however, due to several issues, such as heterogeneous species distribution, the huge dynamic range of protein abundance levels, and the unextractable proteins that are binding to the soil matrix, there is no ideal protocol that is suitable for extracting proteins from all different environmental samples. As the last but most important step, the quality of the metaproteome database influence significantly on the resulting peptide sequence matches. First, searching with a large database is known to suffer from the increased potential for false-positive identifications, which further lower the number of highly confident true matches. Second, too many shared peptides among homologous proteins make the unambiguous protein inference more difficult.[269] Although challenges still exist in environmental metaproteomics study, the improvements in protein extraction and downstream MS measurement along with the advanced bioinformatics tools and various databases should overcome these limitations and speed up microbial metaproteomic research in the near future.

7.3 Conclusions and future perspectives on affinity proteomics

After more than a decade of method developments and proof-of-concept studies, affinity proteomics has emerged as a powerful tool for studying protein interactions, especially when coupling with cross linking. In this dissertation, Bpa was engineered into the different position of target protein Ste2p, and we hypothesize that proteins associated with the intracellular domains of Ste2p in resting state and in the ligand-activated state can be captured and identified by coupling cross-linking, affinity enrichment and mass spectrometry.

Considering the difficulty of preparing the Ste2p interacting complex, the use of laser-activation to facilitate cross-linking of Bpa-containing molecules to proteins was validated in the model system: bovine serum albumin (BSA) and two biologically relevant peptides: the tridecapeptide budding-yeast mating pheromone α-factor, and the decapeptide human gonadotropin-releasing hormone (GnRH). This topic was also fully explored in this dissertation since we found that both peptides can bind with BSA and even compete for the same binding site; therefore, cross-linking coupled with mass spectrometry and computational modeling were employed to identify the binding site. In this study, we demonstrated the successful application of this workflow (Bpa based cross-inking, mass spectrometry and computational model) to explore the protein-peptide interaction. We found that cross-linking of α-factor, using a biotinylated, photoactivatable p-benzoyl-L-phenylalanine (Bpa) modified analog, was time and energy dependent and cross-linking was achieved within seconds of irradiation. By comparing the immunoblot results of the cross-linking products with/without Bpa substitution on α-factor, we found that the use of BPA facilitates cross-linking. Mass spectrometry along with the bioinformatics tool StavroX finally identified the same interacting region of BSA to both peptides, and computational modeling further verified this results.

We have extended this optimized laser-induced cross-linking approach to study the Ste2p-protein interactions. While several initial experiments have been undertaken, initial results has suggested some possible protein interactions, but the sample quantities and interferences have not yet been sorted out, and thus we presented preliminary results for this component of the work. For this study, we evaluated a single step enrichment and double enrichment in terms of sample complexity and sample amount. Single step refers to the enrichment over just the NA beads to pull down all cell surface biotinylated proteins, as well as endogenous biotinylated material. Tandem affinity enrichment method refers to using FLAG and NA Beads to enrich the clean surface Ste2p interacting complex. We found that although single enrichment method yielded much more sample material, there are too many background proteins present, thus decreasing the chance to unambiguously find the truly interacting proteins. We finally chose to using tandem affinity enrichment method, and to insure production of enough material for the MS analysis of

Ste2p, larger yeast cultures were grown. Based on initial results, YAL038W and YGL009C are two possible interacting proteins for Ste2p, but these results need to be further verified with additional controls to sort out quantification issues based on spectral counts.

Despite of the success that affinity proteomics has achieved for protein interaction analysis, the technology itself may still need further consolidation. The biggest challenge for this analysis is the development of a user-friendly and powerful database searching engine, which can directly identify the interacting proteins from the database with high confidence. Additionally, crosslinking and affinity enrichment methods need to be optimized according to the specific systems and the features of the target proteins.

7.4 Overall concluding perspectives

Mass spectrometry has become the cornerstone of analytical chemistry owing to its power to identify and quantify unknown/known compounds, and to elucidate the structure and chemical properties of simple molecules as well as complex macromolecules. At the intersection of mass spectrometry and biology, mass spectrometry based proteomics has emerged as an indispensable field of study that plays an essential role for deciphering complex biological processes, by providing in-depth characterization of the protein components and protein interactions of biological systems. This dissertation contributes to the understanding of microbial functions and responses to the environment stress/signaling from different perspectives and reveals the power of proteomics to address different biological questions. Continued improvements in sample preparation methods, high performance mass spectrometers and bioinformatics platforms will allow even more comprehensive insights into the biological functions and interactions within microbial world. Additionally, with the decrease of the cost for proteomics analysis in the near future, we believe that proteomics study will be more widely applied in biology related fields and continuously contribute to deepen our understanding for the biological world.

REFERENCES

1. Avery, O. T.; Macleod, C. M.; McCarty, M., Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med* **1944,** *79* (2), 137-58.

2. Watson, J. D.; Crick, F. H., Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **1953,** *171* (4356), 737-8.

3. Yoshikawa, H.; Haas, M., On the regulation of the initiation of DNA replication in bacteria. *Cold Spring Harb Symp Quant Biol* **1968,** *33*, 843-55.

4. Crick, F. H., The origin of the genetic code. *J Mol Biol* **1968,** *38* (3), 367-79.

5. Li, G. W.; Xie, X. S., Central dogma at the single-molecule level in living cells. *Nature* **2011,** *475* (7356), 308-315.

6. Crick, F., Central dogma of molecular biology. *Nature* **1970,** *227* (5258), 561-3.

7. Clark, W. T.; Radivojac, P., Analysis of protein function and its prediction from amino acid sequence. *Proteins-Structure Function and Bioinformatics* **2011,** *79* (7), 2086-2096.

8. Westerhoff, H. V.; Winder, C.; Messiha, H.; Simeonidis, E.; Adamczyk, M.; Verma, M.; Bruggeman, F. J.; Dunn, W., Systems biology: the elements and principles of life. *FEBS Lett* **2009,** *583* (24), 3882-90.

9. Sanger, F.; Coulson, A. R., A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **1975,** *94* (3), 441-8.

10. Sanger, F., The Croonian Lecture, 1975. Nucleotide sequences in DNA. *Proc R Soc Lond B Biol Sci* **1975,** *191* (1104), 317-33.

11. Fleischmann, R. D.; Adams, M. D.; White, O.; Clayton, R. A.; Kirkness, E. F.; Kerlavage, A. R.; Bult, C. J.; Tomb, J. F.; Dougherty, B. A.; Merrick, J. M.; et al., Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **1995,** *269* (5223), 496-512.

12. Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, Y.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J. C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, E. R.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J. F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R. A.; Muzny, D. M.; Scherer, S. E.; Bouck, J. B.; Sodergren,

E. J.; Worley, K. C.; Rives, C. M.; Gorrell, J. H.; Metzker, M. L.; Naylor, S. L.; Kucherlapati, R. S.; Nelson, D. L.; Weinstock, G. M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D. R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H. M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R. W.; Federspiel, N. A.; Abola, A. P.; Proctor, M. J.; Myers, R. M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D. R.; Olson, M. V.; Kaul, R.; Raymond, C.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G. A.; Athanasiou, M.; Schultz, R.; Roe, B. A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W. R.; de la Bastide, M.; Dedhia, N.; Blocker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J. A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D. G.; Burge, C. B.; Cerutti, L.; Chen, H. C.; Church, D.; Clamp, M.; Copley, R. R.; Doerks, T.; Eddy, S. R.; Eichler, E. E.; Furey, T. S.; Galagan, J.; Gilbert, J. G.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L. S.; Jones, T. A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W. J.; Kitts, P.; Koonin, E. V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T. M.; McLysaght, A.; Mikkelsen, T.; Moran, J. V.; Mulder, N.; Pollara, V. J.; Ponting, C. P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A. F.; Stupka, E.; Szustakowki, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y. I.; Wolfe, K. H.; Yang, S. P.; Yeh, R. F.; Collins, F.; Guyer, M. S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K. A.; Patrinos, A.; Morgan, M. J.; de Jong, P.; Catanese, J. J.; Osoegawa, K.; Shizuya, H.; Choi, S.; Chen, Y. J.; Szustakowki, J.; International Human Genome Sequencing, C., Initial sequencing and analysis of the human genome. *Nature* **2001,** *409* (6822), 860-921.

13. Reddy, T. B.; Thomas, A. D.; Stamatis, D.; Bertsch, J.; Isbandi, M.; Jansson, J.; Mallajosyula, J.; Pagani, I.; Lobos, E. A.; Kyrpides, N. C., The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res* **2015,** *43* (Database issue), D1099-106.

14. Hieter, P.; Boguski, M., Functional genomics: It's all how you read it. *Science* **1997,** *278* (5338), 601-602.

15. DeRisi, J.; Penland, L.; Brown, P. O.; Bittner, M. L.; Meltzer, P. S.; Ray, M.; Chen, Y. D.; Su, Y. A.; Trent, J. M., Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* **1996,** *14* (4), 457-460.

16. Zhou, J. H., Microarrays for bacterial detection and microbial community analysis. *Curr Opin Microbiol* **2003,** *6* (3), 288-294.

17. Wang, Z.; Gerstein, M.; Snyder, M., RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **2009,** *10* (1), 57-63.

18. Mesri, M., Advances in Proteomic Technologies and Its Contribution to the Field of Cancer. *Adv Med* **2014,** *2014*, 238045.

19. Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B., Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem* **2012,** *404* (4), 939-65.

20. Marguerat, S.; Schmidt, A.; Codlin, S.; Chen, W.; Aebersold, R.; Bahler, J., Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* **2012,** *151* (3), 671-83.

21. Aebersold, R.; Mann, M., Mass-spectrometric exploration of proteome structure and function. *Nature* **2016,** *537* (7620), 347-355.

22. Johnson, C. H.; Ivanisevic, J.; Siuzdak, G., Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol* **2016,** *17* (7), 451-9.

23. Rochfort, S., Metabolomics reviewed: a new "omics" platform technology for systems biology and implications for natural products research. *J Nat Prod* **2005,** *68* (12), 1813-20.

24. Wishart, D. S., Emerging applications of metabolomics in drug discovery and precision medicine. *Nat Rev Drug Discov* **2016,** *15* (7), 473-84.

25. Beger, R. D., A review of applications of metabolomics in cancer. *Metabolites* **2013,** *3* (3), 552-74.

26. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., Electrospray Ionization for Mass-Spectrometry of Large Biomolecules. *Science* **1989,** *246* (4926), 64-71.

27. Hunt, D. F.; Michel, H.; Dickinson, T. A.; Shabanowitz, J.; Cox, A. L.; Sakaguchi, K.; Appella, E.; Grey, H. M.; Sette, A., Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. *Science* **1992,** *256* (5065), 1817-20.

28. Mann, M., The Rise of Mass Spectrometry and the Fall of Edman Degradation. *Clin Chem* **2016,** *62* (1), 293-294.

29. Eng, J. K.; McCormack, A. L.; Yates, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **1994,** *5* (11), 976-89.

30. Mann, M., Origins of mass spectrometry-based proteomics. *Nat Rev Mol Cell Biol* **2016,** *17* (11), 678.

31. Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M., Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **1996,** *379* (6564), 466-9.

32. Shevchenko, A.; Tomas, H.; Havlis, J.; Olsen, J. V.; Mann, M., In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc* **2006,** *1* (6), 2856-60.

33. Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **2001,** *19* (3), 242-7.

34. Resing, K. A.; Ahn, N. G., Proteomics strategies for protein identification. *FEBS Lett* **2005,** *579* (4), 885-9.

35. Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M. X.; Wu, C.; Sweet, S. M. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L., Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011,** *480* (7376), 254-U141.

36. Catherman, A. D.; Skinner, O. S.; Kelleher, N. L., Top Down proteomics: Facts and perspectives. *Biochem Bioph Res Co* **2014,** *445* (4), 683-693.

37. Zhang, H.; Ge, Y., Comprehensive analysis of protein modifications by top-down mass spectrometry. *Circ Cardiovasc Genet* **2011,** *4* (6), 711.

38. Wehr, T., Top-down versus bottom-up approaches in proteomics. *Lc Gc N Am* **2006,** *24* (9), 1004-+.

39. Doerr, A., Targeted proteomics. *Nature Methods* **2011,** *8* (1), 43-43.

40. Ahrens, C. H.; Brunner, E.; Qeli, E.; Basler, K.; Aebersold, R., Generating and navigating proteome maps using mass spectrometry. *Nat Rev Mol Cell Bio* **2010,** *11* (11), 789-801.

41. Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry* **2007,** *389* (4), 1017-1031.

42. Yeh, C. C.; Hsu, C. H.; Shao, Y. Y.; Ho, W. C.; Tsai, M. H.; Feng, W. C.; Chow, L. P., Integrated Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) and Isobaric Tags for Relative and Absolute Quantitation (iTRAQ) Quantitative Proteomic Analysis Identifies Galectin-1 as a Potential Biomarker for Predicting Sorafenib Resistance in Liver Cancer. *Mol Cell Proteomics* **2015,** *14* (6), 1527-1545.

43. Bronstrup, M., Absolute quantification strategies in proteomics based on mass spectrometry. *Expert Rev Proteomic* **2004,** *1* (4), 503-512.

44. Patel, V. J.; Thalassinos, K.; Slade, S. E.; Connolly, J. B.; Crombie, A.; Murrell, J. C.; Scrivens, J. H., A Comparison of Labeling and Label-Free Mass Spectrometry-Based Proteomics Approaches. *J Proteome Res* **2009,** *8* (7), 3752-3759.

45. Prakash, A.; Piening, B.; Whiteaker, J.; Zhang, H.; Shaffer, S. A.; Martin, D.; Hohmann, L.; Cooke, K.; Olson, J. M.; Hansen, S.; Flory, M. R.; Lee, H.; Watts, J.; Goodlett, D. R.; Aebersold, R.; Paulovich, A.; Schwikowski, B., Assessing bias in experiment design for large scale mass spectrometry-based quantitative proteomics. *Mol Cell Proteomics* **2007,** *6* (10), 1741-1748.

46. Dowd, W. W., Challenges for Biological Interpretation of Environmental Proteomics Data in Non-model Organisms. *Integr Comp Biol* **2012,** *52* (5), 705-720.

47. Martyniuk, C. J.; Kroll, K. J.; Doperalski, N. J.; Barber, D. S.; Denslow, N. D., Environmentally relevant exposure to 17 alpha-ethinylestradiol affects the telencephalic proteome of male fathead minnows. *Aquat Toxicol* **2010,** *98* (4), 344-353.

48. Domon, B.; Aebersold, R., Challenges and opportunities in proteomics data analysis. *Mol Cell Proteomics* **2006,** *5* (10), 1921-1926.

49. Wilkins, M. R.; Appel, R. D.; Van Eyk, J. E.; Chung, M. C. M.; Gorg, A.; Hecker, M.; Huber, L. A.; Langen, H.; Link, A. J.; Paik, Y. K.; Patterson, S. D.; Pennington, S. R.; Rabilloud, T.; Simpson, R. J.; Weiss, W.; Dunn, M. J., Guidelines for the next 10 years of proteomics. *Proteomics* **2006,** *6* (1), 4-8.

50. Qian, C.; Johs, A.; Chen, H.; Mann, B. F.; Lu, X.; Abraham, P. E.; Hettich, R. L.; Gu, B., Global Proteome Response to Deletion of Genes Related to Mercury Methylation and Dissimilatory Metal Reduction Reveals Changes in Respiratory Metabolism in Geobacter sulfurreducens PCA. *J Proteome Res* **2016,** *15* (10), 3540-3549.

51. Abraham, P. E.; Giannone, R. J.; Xiong, W.; Hettich, R. L., Metaproteomics: extracting and mining proteome information to characterize metabolic activities in microbial communities. *Curr Protoc Bioinformatics* **2014,** *46*, 13 26 1-14.

52. Goodwin, S.; McPherson, J. D.; McCombie, W. R., Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **2016,** *17* (6), 333-51.

53. Giese, S. H.; Fischer, L.; Rappsilber, J., A Study into the Collision-induced Dissociation (CID) Behavior of Cross-Linked Peptides. *Mol Cell Proteomics* **2016,** *15* (3), 1094-1104.

54. Schneider, T.; Riedel, K., Environmental proteomics: Analysis of structure and function of microbial communities. *Proteomics* **2010,** *10* (4), 785-798.

55. Torsvik, V.; Ovreas, L.; Thingstad, T. F., Prokaryotic diversity--magnitude, dynamics, and controlling factors. *Science* **2002,** *296* (5570), 1064-6.

56. Lacerda, C. M.; Reardon, K. F., Environmental proteomics: applications of proteome profiling in environmental microbiology and biotechnology. *Brief Funct Genomic Proteomic* **2009,** *8* (1), 75-87.

57. Dworzanski, J. P.; Snyder, A. P., Classification and identification of bacteria using mass spectrometry-based proteomics. *Expert Rev Proteomic* **2005,** *2* (6), 863-878.

58. Armengaud, J., Microbiology and proteomics, getting the best of both worlds! *Environmental Microbiology* **2013,** *15* (1), 12-23.

59. Otto, A.; Becher, D.; Schmidt, F., Quantitative proteomics in the field of microbiology. *Proteomics* **2014,** *14* (4-5), 547-565.

60. Elias, D. A.; Monroe, M. E.; Marshall, M. J.; Romine, M. F.; Belieav, A. S.; Fredrickson, J. K.; Anderson, G. A.; Smith, R. D.; Lipton, M. S., Global detection and characterization of hypothetical proteins in Shewanella oneidensis MR-1 using LC-MS based proteomics. *Proteomics* **2005,** *5* (12), 3120-3130.

61. Fang, R. H.; Elias, D. A.; Monroe, M. E.; Shen, Y. F.; Mcintosh, M.; Wang, P.; Goddard, C. D.; Callister, S. J.; Moore, R. J.; Gorby, Y. A.; Adkins, J. N.; Fredrickson, J. K.; Lipton, M. S.; Smith,

R. D., Differential label-free quantitative proteomic analysis of Shewanella oneidensis cultured under aerobic and suboxic conditions by accurate mass and time tag approach. *Mol Cell Proteomics* **2006,** *5* (4), 714-725.

62. Hecker, M.; Volker, U., Towards a comprehensive understanding of Bacillus subtilis cell physiology by physiological proteomics. *Proteomics* **2004,** *4* (12), 3727-50.

63. Oosthuizen, M. C.; Steyn, B.; Theron, J.; Cosette, P.; Lindsay, D.; von Holy, A.; Brozel, V. S., Proteomic analysis reveals differential protein expression by Bacillus cereus during biofilm formation. *Appl Environ Microb* **2002,** *68* (6), 2770-2780.

64. Joo, W. A.; Kim, C. W., Proteomics of Halophilic archaea. *J Chromatogr B* **2005,** *815* (1-2), 237-250.

65. Ukani, H.; Purohit, M. K.; Georrge, J. J.; Paul, S.; Singh, S. P., HaloBase: development of database system for halophilic bacteria and archaea with respect to proteomics, genomics & other molecular traits. *J Sci Ind Res India* **2011,** *70* (11), 976-981.

66. Ram, R. J.; VerBerkmoes, N. C.; Thelen, M. P.; Tyson, G. W.; Baker, B. J.; Blake, R. C.; Shah, M.; Hettich, R. L.; Banfield, J. F., Community proteomics of a natural microbial biofilm. *Science* **2005,** *308* (5730), 1915-1920.

67. Lo, I.; Denef, V. J.; VerBerkmoes, N. C.; Shah, M. B.; Goltsman, D.; DiBartolo, G.; Tyson, G. W.; Allen, E. E.; Ram, R. J.; Detter, J. C.; Richardson, P.; Thelen, M. P.; Hettich, R. L.; Banfield, J. F., Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **2007,** *446* (7135), 537-541.

68. Robidart, J. C.; Bench, S. R.; Feldman, R. A.; Novoradovsky, A.; Podell, S. B.; Gaasterland, T.; Allen, E. E.; Felbeck, H., Metabolic versatility of the Riftia pachyptila endosymbiont revealed through metagenomics. *Environmental Microbiology* **2008,** *10* (3), 727-737.

69. Lacerda, C. M.; Choe, L. H.; Reardon, K. F., Metaproteomic analysis of a bacterial community response to cadmium exposure. *J Proteome Res* **2007,** *6* (3), 1145-52.

70. Park, C.; Helm, R. F.; Novak, J. T., Investigating the Fate of Activated Sludge Extracellular Proteins in Sludge Digestion Using Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis. *Water Environ Res* **2008,** *80* (12), 2219-2227.

71. Benndorf, D.; Balcke, G. U.; Harms, H.; von Bergen, M., Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. *Isme J* **2007,** *1* (3), 224-234.

72. Young, I. M.; Crawford, J. W., Interactions and self-organization in the soil-microbe complex. *Science* **2004,** *304* (5677), 1634-7.

73. Parks, J. M.; Johs, A.; Podar, M.; Bridou, R.; Hurt, R. A., Jr.; Smith, S. D.; Tomanicek, S. J.; Qian, Y.; Brown, S. D.; Brandt, C. C.; Palumbo, A. V.; Smith, J. C.; Wall, J. D.; Elias, D. A.; Liang, L., The genetic basis for bacterial mercury methylation. *Science* **2013,** *339* (6125), 1332-5.

74. Lin, H.; Lu, X.; Liang, L. Y.; Gu, B. H., Thiol-Facilitated Cell Export and Desorption of Methylmercury by Anaerobic Bacteria. *Environmental Science & Technology Letters* **2015,** *2* (10), 292-296.

75. Thomas, P. A.; Gates, T. E., Radionuclides in the lichen-caribou-human food chain near uranium mining operations in northern Saskatchewan, Canada. *Environ Health Persp* **1999,** *107* (7), 527-537.

76. Anderson, R. T.; Vrionis, H. A.; Ortiz-Bernad, I.; Resch, C. T.; Long, P. E.; Dayvault, R.; Karp, K.; Marutzky, S.; Metzler, D. R.; Peacock, A.; White, D. C.; Lowe, M.; Lovley, D. R., Stimulating the in situ activity of Geobacter species to remove uranium from the groundwater of a uranium-contaminated aquifer. *Appl Environ Microb* **2003,** *69* (10), 5884-5891.

77. Chang, Y. J.; Peacock, A. D.; Long, P. E.; Stephen, J. R.; McKinley, J. P.; Macnaughton, S. J.; Hussain, A. K. M. A.; Saxton, A. M.; White, D. C., Diversity and characterization of sulfate-reducing bacteria in groundwater at a uranium mill tailings site. *Appl Environ Microb* **2001,** *67* (7), 3149-3160.

78. Elias, D. A.; Krumholz, L. R.; Wong, D.; Long, P. E.; Suflita, J. M., Characterization of microbial activities and U reduction in a shallow aquifer contaminated by uranium mill tailings. *Microb Ecol* **2003,** *46* (1), 83-91.

79. Batten, K. M.; Scow, K. M., Sediment microbial community composition and methylmercury pollution at four mercury mine-impacted sites. *Microb Ecol* **2003,** *46* (4), 429-441.

80. Baralkiewicz, D.; Gramowska, H.; Goldyn, R., Distribution of total mercury and methyl mercury in water, sediment and fish from Swarzedzkie lake. *Chem Ecol* **2006,** *22* (1), 59-64.

81. Desrosiers, M.; Planas, D.; Mucci, A., Total mercury and methylmercury accumulation in periphyton of Boreal Shield Lakes: Influence of watershed physiographic characteristics. *Science of the Total Environment* **2006,** *355* (1-3), 247-258.

82. Zhang, M. Q.; Zhu, Y. C.; Deng, R. W., Evaluation of mercury emissions to the atmosphere from coal combustion, China. *Ambio* **2002,** *31* (6), 482-484.

83. Vishnivetskaya, T. A.; Mosher, J. J.; Palumbo, A. V.; Yang, Z. K.; Podar, M.; Brown, S. D.; Brooks, S. C.; Gu, B. H.; Southworth, G. R.; Drake, M. M.; Brandt, C. C.; Elias, D. A., Mercury and Other Heavy Metals Influence Bacterial Community Structure in Contaminated Tennessee Streams. *Appl Environ Microb* **2011,** *77* (1), 302-311.

84. Freedman, Z.; Zhu, C. S.; Barkay, T., Mercury Resistance and Mercuric Reductase Activities and Expression among Chemotrophic Thermophilic Aquificae. *Appl Environ Microb* **2012,** *78* (18), 6568-6575.

85. Oregaard, G.; Sorensen, S. J., High diversity of bacterial mercuric reductase genes from surface and sub-surface floodplain soil (Oak Ridge, USA). *Isme J* **2007,** *1* (5), 453-67.

86. Hu, H. Y.; Lin, H.; Zheng, W.; Tomanicek, S. J.; Johs, A.; Feng, X. B.; Elias, D. A.; Liang, L. Y.; Gu, B. H., Oxidation and methylation of dissolved elemental mercury by anaerobic bacteria. *Nat Geosci* **2013,** *6* (9), 751-754.

87. Choi, S. C.; Chase, T.; Bartha, R., Metabolic Pathways Leading to Mercury Methylation in Desulfovibrio desulfuricans LS. *Appl Environ Microbiol* **1994,** *60* (11), 4072-7.

88. Qian, C.; Johs, A.; Chen, H. M.; Mann, B. F.; Lu, X.; Abraham, P. E.; Hettich, R. L.; Gu, B. H., Global Proteome Response to Deletion of Genes Related to Mercury Methylation and Dissimilatory Metal Reduction Reveals Changes in Respiratory Metabolism in Geobacter sulfurreducens PCA. *J Proteome Res* **2016,** *15* (10), 3540-3549.

89. Gingras, A. C.; Gstaiger, M.; Raught, B.; Aebersold, R., Analysis of protein complexes using mass spectrometry. *Nat Rev Mol Cell Bio* **2007,** *8* (8), 645-654.

90. Oeffinger, M., Two steps forward--one step back: advances in affinity purification mass spectrometry of macromolecular complexes. *Proteomics* **2012,** *12* (10), 1591-608.

91. Rajagopala, S. V.; Sikorski, P.; Caufield, J. H.; Tovchigrechko, A.; Uetz, P., Studying protein complexes by the yeast two-hybrid system. *Methods* **2012,** *58* (4), 392-9.

92. Gavin, A. C.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A. M.; Cruciat, C. M.; Remor, M.; Hofert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M. A.; Copley, R. R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B.; Neubauer, G.; Superti-Furga, G., Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **2002,** *415* (6868), 141-7.

93. Ho, Y.; Gruhler, A.; Heilbut, A.; Bader, G. D.; Moore, L.; Adams, S. L.; Millar, A.; Taylor, P.; Bennett, K.; Boutilier, K.; Yang, L.; Wolting, C.; Donaldson, I.; Schandorff, S.; Shewnarane, J.; Vo, M.; Taggart, J.; Goudreault, M.; Muskat, B.; Alfarano, C.; Dewar, D.; Lin, Z.; Michalickova, K.; Willems, A. R.; Sassi, H.; Nielsen, P. A.; Rasmussen, K. J.; Andersen, J. R.; Johansen, L. E.; Hansen, L. H.; Jespersen, H.; Podtelejnikov, A.; Nielsen, E.; Crawford, J.; Poulsen, V.; Sorensen, B. D.; Matthiesen, J.; Hendrickson, R. C.; Gleeson, F.; Pawson, T.; Moran, M. F.; Durocher, D.; Mann, M.; Hogue, C. W.; Figeys, D.; Tyers, M., Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **2002,** *415* (6868), 180-3.

94. Gibson, T. J.; Seiler, M.; Veitia, R. A., The transience of transient overexpression. *Nat Methods* **2013,** *10* (8), 715-21.

95. Glatter, T.; Wepf, A.; Aebersold, R.; Gstaiger, M., An integrated workflow for charting the human interaction proteome: insights into the PP2A system. *Molecular Systems Biology* **2009,** *5*.

96. Gavin, A. C.; Maeda, K.; Kuhner, S., Recent advances in charting protein-protein interaction: mass spectrometry-based approaches. *Curr Opin Biotech* **2011,** *22* (1), 42-49.

97. Ranish, J. A.; Yi, E. C.; Leslie, D. M.; Purvine, S. O.; Goodlett, D. R.; Eng, J.; Aebersold, R., The study of macromolecular complexes by quantitative proteomics. *Nat Genet* **2003,** *33* (3), 349-355.

98. Benedetti, R.; Conte, M.; Carafa, V.; Della Ventura, B.; Altucci, C.; Velotta, R.; Stunnenberg, H. G.; Altucci, L.; Nebbioso, A., Analysis of chromatin-nuclear receptor interactions by laser-chromatin immunoprecipitation. *Methods Mol Biol* **2014,** *1204*, 25-34.

99. Itri, F.; Monti, D. M.; Della Ventura, B.; Vinciguerra, R.; Chino, M.; Gesuele, F.; Lombardi, A.; Velotta, R.; Altucci, C.; Birolo, L.; Piccoli, R.; Arciello, A., Femtosecond UV-laser pulses to unveil protein-protein interactions in living cells. *Cell Mol Life Sci* **2016,** *73* (3), 637-648.

100. Zolotarjova, N.; Mrozinski, P.; Chen, H.; Martosella, J., Combination of affinity depletion of abundant proteins and reversed-phase fractionation in proteomic analysis of human plasma/serum. *J Chromatogr A* **2008,** *1189* (1-2), 332-8.

101. Fonslow, B. R.; Stein, B. D.; Webb, K. J.; Xu, T.; Choi, J.; Park, S. K.; Yates, J. R., 3rd, Digestion and depletion of abundant proteins improves proteomic coverage. *Nat Methods* **2013,** *10* (1), 54-6.

102. Zhao, Y. M.; Jensen, O. N., Modification-specific proteomics: Strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* **2009,** *9* (20), 4632-4641.

103. Brown, L.; Wolf, J. M.; Prados-Rosales, R.; Casadevall, A., Through the wall: extracellular vesicles in Gram-positive bacteria, mycobacteria and fungi. *Nat Rev Microbiol* **2015,** *13* (10), 620-630.

104. Paul, T.; Basu, S.; Sarkar, K., SPION-mediated soil DNA extraction and comparative analysis with conventional and commercial kit-based protocol. *3 Biotech* **2014,** *4* (6), 669-677.

105. Feist, P.; Hummon, A. B., Proteomic Challenges: Sample Preparation Techniques for Microgram-Quantity Protein Analysis from Biological Samples. *Int J Mol Sci* **2015,** *16* (2), 3537-3563.

106. Jiang, L.; He, L.; Fountoulakis, M., Comparison of protein precipitation methods for sample preparation prior to proteomic analysis. *J Chromatogr A* **2004,** *1023* (2), 317-20.

107. Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M., Universal sample preparation method for proteome analysis. *Nat Methods* **2009,** *6* (5), 359-62.

108. Scheerlinck, E.; Dhaenens, M.; Van Soom, A.; Peelman, L.; De Sutter, P.; Van Steendam, K.; Deforce, D., Minimizing technical variation during sample preparation prior to label-free quantitative mass spectrometry. *Anal Biochem* **2015,** *490*, 14-9.

109. Giansanti, P.; Tsiatsiani, L.; Low, T. Y.; Heck, A. J., Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat Protoc* **2016,** *11* (5), 993-1006.

110. Rosen, R.; Ron, E. Z., Proteome analysis in the study of the bacterial heat-shock response. *Mass Spectrom Rev* **2002,** *21* (4), 244-65.

111. Boersema, P. J.; Divecha, N.; Heck, A. J. R.; Mohammed, S., Evaluation and optimization of ZIC-HILIC-RP as an alternative MudPIT strategy. *J Proteome Res* **2007,** *6* (3), 937-946.

112. Delmotte, N.; Lasaosa, M.; Tholey, A.; Heinzle, E.; Huber, C. G., Two-dimensional reversed-phase x ion-pair reversed-phase HPLC: An alternative approach to high-resolution peptide separation for shotgun proteome analysis. *J Proteome Res* **2007,** *6* (11), 4363-4373.

113. Grassl, N.; Kulak, N. A.; Pichler, G.; Geyer, P. E.; Jung, J.; Schubert, S.; Sinitcyn, P.; Cox, J.; Mann, M., Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome. *Genome Med* **2016,** *8*.

114. Vermeulen, N. P. E.; Burn, A. G.; Onkenhout, W., Electron-Impact and Chemical Ionization Gas-Liquid Chromatography-Mass Spectrometry - Principles and Applications in Pharmacokinetic and Biotransformation Studies. *Pharm Weekblad* **1984,** *6* (6), 251-251.

115. El-Aneed, A.; Cohen, A.; Banoub, J., Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers. *Appl Spectrosc Rev* **2009,** *44* (3), 210-230.

116. Song, M.; Zhao, H.; Wang, L.; Yang, L.; Hang, T.; Wen, A., Development and validation of a robust LC-MS-MS with atmospheric pressure chemical ionization for quantitation of carbazochrome sodium sulfonate in human plasma: application to a pharmacokinetic study. *Biomed Chromatogr* **2010,** *24* (9), 990-9.

117. Powers, T. W.; Jones, E. E.; Betesh, L. R.; Romano, P. R.; Gao, P.; Copland, J. A.; Mehta, A. S.; Drake, R. R., Matrix Assisted Laser Desorption Ionization Imaging Mass Spectrometry Workflow for Spatial Profiling Analysis of N-Linked Glycan Expression in Tissues. *Analytical Chemistry* **2013,** *85* (20), 9799-9806.

118. Wachsmuth, C. J.; Hahn, T. A.; Oefner, P. J.; Dettmer, K., Enhanced metabolite profiling using a redesigned atmospheric pressure chemical ionization source for gas chromatography coupled to high-resolution time-of-flight mass spectrometry. *Analytical and Bioanalytical Chemistry* **2015,** *407* (22), 6669-6680.

119. Santos, I. C.; Hildenbrand, Z. L.; Schug, K. A., Applications of MALDI-TOF MS in environmental microbiology. *Analyst* **2016,** *141* (10), 2827-2837.

120. Kasprzyk, J.; Stepien, E.; Piekoszewski, W., Application of nano-LC-MALDI-TOF/TOF-MS for proteomic analysis of microvesicles. *Clin Biochem* **2017,** *50* (4-5), 241-243.

121. Banerjee, S.; Mazumdar, S., Electrospray ionization mass spectrometry: a technique to access the information beyond the molecular weight of the analyte. *Int J Anal Chem* **2012,** *2012*, 282574.

122. Yarin, A. L.; Koombhongse, S.; Reneker, D. H., Taylor cone and jetting from liquid droplets in electrospinning of nanofibers. *J Appl Phys* **2001,** *90* (9), 4836-4846.

123. Kebarle, P.; Tang, L., From Ions in Solution to Ions in the Gas-Phase - the Mechanism of Electrospray Mass-Spectrometry. *Analytical Chemistry* **1993,** *65* (22), A972-A986.

124. Karas, M.; Bahr, U.; Dulcks, T., Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine. *Fresenius J Anal Chem* **2000,** *366* (6-7), 669-76.

125. Sheretov, E. P., Principles of the Theory for 3-Dimensional Quadrupole Mass-Spectrometry .2. *Zh Tekh Fiz+* **1979,** *49* (1), 41-46.

126. Sheretov, E. P., Principles of the Theory for Quadrupole Mass-Spectrometry .1. *Zh Tekh Fiz+* **1979,** *49* (1), 34-40.

127. Douglas, D. J.; Frank, A. J.; Mao, D. M., Linear ion traps in mass spectrometry. *Mass Spectrometry Reviews* **2005,** *24* (1), 1-29.

128. Hu, Q. Z.; Noll, R. J.; Li, H. Y.; Makarov, A.; Hardman, M.; Cooks, R. G., The Orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry* **2005,** *40* (4), 430-443.

129. Kalli, A.; Smith, G. T.; Sweredoski, M. J.; Hess, S., Evaluation and Optimization of Mass Spectrometric Settings during Data-dependent Acquisition Mode: Focus on LTQ-Orbitrap Mass Analyzers. *J Proteome Res* **2013,** *12* (7), 3071-3086.

130. Jedrychowski, M. P.; Huttlin, E. L.; Haas, W.; Sowa, M. E.; Rad, R.; Gygi, S. P., Evaluation of HCD- and CID-type Fragmentation Within Their Respective Detection Platforms For Murine Phosphoproteomics. *Mol Cell Proteomics* **2011,** *10* (12).

131. Johnson, D.; Boyes, B.; Fields, T.; Kopkin, R.; Orlando, R., Optimization of data-dependent acquisition parameters for coupling high-speed separations with LC-MS/MS for protein identifications. *J Biomol Tech* **2013,** *24* (2), 62-72.

132. Doerr, A., DIA mass spectrometry. *Nature Methods* **2015,** *12* (1), 35-35.

133. Ma, B.; Zhang, K. Z.; Hendrie, C.; Liang, C. Z.; Li, M.; Doherty-Kirby, A.; Lajoie, G., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Sp* **2003,** *17* (20), 2337-2342.

134. Frank, A.; Pevzner, P., PepNovo: De novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry* **2005,** *77* (4), 964-973.

135. Tang, W. H.; Shilov, I. V.; Seymour, S. L., Nonlinear fitting method for determining local false discovery rates from decoy database searches. *J Proteome Res* **2008,** *7* (9), 3661-7.

136. Tabb, D. L.; Fernando, C. G.; Chambers, M. C., MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* **2007,** *6* (2), 654-61.

137. Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999,** *20* (18), 3551-3567.

138. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J Proteome Res* **2011,** *10* (4), 1794-1805.

139. Old, W. M.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K. G.; Mendoza, A.; Sevinsky, J. R.; Resing, K. A.; Ahn, N. G., Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* **2005,** *4* (10), 1487-1502.

140. Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M., Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **2014,** *13* (9), 2513-26.

141. Urfer, W.; Grzegorczyk, M.; Jung, K., Statistics for proteomics: a review of tools for analyzing experimental data. *Proteomics* **2006,** *6 Suppl 2*, 48-55.

142. Valikangas, T.; Suomi, T.; Elo, L. L., A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform* **2016**.

143. Diz, A. P.; Carvajal-Rodriguez, A.; Skibinski, D. O. F., Multiple Hypothesis Testing in Proteomics: A Strategy for Experimental Work. *Mol Cell Proteomics* **2011,** *10* (3).

144. Levin, Y., The role of statistical power analysis in quantitative proteomics. *Proteomics* **2011,** *11* (12), 2565-2567.

145. Gyorffy, B.; Gyorffy, A.; Tulassay, Z., The problem of multiple testing and solutions for genome-wide studies. *Orv Hetil* **2005,** *146* (12), 559-63.

146. Benjamini, Y.; Hochberg, Y., Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **1995,** *57* (1), 289-300.

147. Bertolino, F.; Cabras, S.; Castellanos, M. E.; Racugno, W., Unscaled Bayes factors for multiple hypothesis testing in microarray experiments. *Stat Methods Med Res* **2015,** *24* (6), 1030-43.

148. Cabezas-Wallscheid, N.; Klimmeck, D.; Hansson, J.; Lipka, D. B.; Reyes, A.; Wang, Q.; Weichenhan, D.; Lier, A.; von Paleske, L.; Renders, S.; Wunsche, P.; Zeisberger, P.; Brocks, D.; Gu, L.; Herrmann, C.; Haas, S.; Essers, M. A. G.; Brors, B.; Eils, R.; Huber, W.; Milsom, M. D.; Plass, C.; Krijgsveld, J.; Trumpp, A., Identification of Regulatory Networks in HSCs and Their Immediate Progeny via Integrated Proteome, Transcriptome, and DNA Methylome Analysis. *Cell Stem Cell* **2014,** *15* (4), 507-522.

149. McHugh, M. L., Multiple comparison analysis testing in ANOVA. *Biochem Medica* **2011,** *21* (3), 203-209.

150. Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J., The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **2016,** *13* (9), 731-740.

151. Deshmukh, A. S.; Murgia, M.; Nagaraj, N.; Treebak, J. T.; Cox, J.; Mann, M., Deep Proteomics of Mouse Skeletal Muscle Enables Quantitation of Protein Isoforms, Metabolic Pathways, and Transcription Factors. *Mol Cell Proteomics* **2015,** *14* (4), 841-853.

152. Abraham, P. E.; Yin, H. F.; Borland, A. M.; Weighill, D.; Lim, S. D.; De Paoli, H. C.; Engle, N.; Jones, P. C.; Agh, R.; Weston, D. J.; Wullschleger, S. D.; Tschaplinski, T.; Jacobson, D.;

Cushman, J. C.; Hettich, R. L.; Tuskan, G. A.; Yang, X. H., Transcript, protein and metabolite temporal dynamics in the CAM plant Agave. *Nat Plants* **2016,** *2* (12).

153. Wippler, J.; Kleiner, M.; Lott, C.; Gruhl, A.; Abraham, P. E.; Giannone, R. J.; Young, J. C.; Hettich, R. L.; Dubilier, N., Transcriptomic and proteomic insights into innate immunity and adaptations to a symbiotic lifestyle in the gutless marine worm Olavius algarvensis. *Bmc Genomics* **2016,** *17*.

154. Bell, L. N.; Theodorakis, J. L.; Vuppalanchi, R.; Saxena, R.; Bemis, K. G.; Wang, M.; Chalasani, N., Serum Proteomics and Biomarker Discovery Across the Spectrum of Nonalcoholic Fatty Liver Disease. *Hepatology* **2010,** *51* (1), 111-120.

155. Barros, E.; Lezar, S.; Anttonen, M. J.; van Dijk, J. P.; Rohlig, R. M.; Kok, E. J.; Engel, K. H., Comparison of two GM maize varieties with a near-isogenic non-GM variety using transcriptomics, proteomics and metabolomics. *Plant Biotechnol J* **2010,** *8* (4), 436-451.

156. Oberg, A. L.; Mahoney, D. W.; Eckel-Passow, J. E.; Malone, C. J.; Wolfinger, R. D.; Hill, E. G.; Cooper, L. T.; Onuma, O. K.; Spiro, C.; Therneau, T. A.; Bergen, H. R., Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *J Proteome Res* **2008,** *7* (1), 225-233.

157. Hug, L. A.; Thomas, B. C.; Sharon, I.; Brown, C. T.; Sharma, R.; Hettich, R. L.; Wilkins, M. J.; Williams, K. H.; Singh, A.; Banfield, J. F., Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environ Microbiol* **2015,** *18*, 159-173.

158. Graham, D. E.; Wallenstein, M. D.; Vishnivetskaya, T. A.; Waldrop, M. P.; Phelps, T. J.; Pfiffner, S. M.; Onstott, T. C.; Whyte, L. G.; Rivkina, E. M.; Gilichinsky, D. A.; Elias, D. A.; Mackelprang, R.; VerBerkmoes, N. C.; Hettich, R. L.; Wagner, D.; Wullschleger, S. D.; Jansson, J. K., Microbes in thawing permafrost: the unknown variable in the climate change equation. *Isme J* **2012,** *6* (4), 709-712.

159. Hultman, J.; Waldrop, M. P.; Mackelprang, R.; David, M. M.; McFarland, J.; Blazewicz, S. J.; Harden, J.; Turetsky, M. R.; McGuire, A. D.; Shah, M. B.; VerBerkmoes, N. C.; Lee, L. H.; Mavrommatis, K.; Jansson, J. K., Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* **2015,** *521* (7551), 208-12.

160. Arsene-Ploetze, F.; Bertin, P. N.; Carapito, C., Proteomic tools to decipher microbial community structure and functioning. *Environ Sci Pollut Res Int* **2015,** *22* (18), 13599-612.

161. Tveit, A.; Schwacke, R.; Svenning, M. M.; Urich, T., Organic carbon transformations in high-Arctic peat soils: key functions and microorganisms. *Isme J* **2013,** *7* (2), 299-311.

162. Khan, A. G., Role of soil microbes in the rhizospheres of plants growing on trace metal contaminated soils in phytoremediation. *J Trace Elem Med Biol* **2005,** *18* (4), 355-64.

163. Jackson, C. R.; Tyler, H. L.; Millar, J. J., Determination of Microbial Extracellular Enzyme Activity in Waters, Soils, and Sediments using High Throughput Microplate Assays. *Jove-J Vis Exp* **2013,** (80), e50399.

164. 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* **2013,** *85* (9), 4203-14.

165. Gillespie, A. W.; Farrell, R. E.; Walley, F. L.; Ross, A. R. S.; Leinweber, P.; Eckhardt, K. U.; Regier, T. Z.; Blyth, R. I. R., Glomalin-related soil protein contains non-mycorrhizal-related heat-stable proteins, lipids and humic materials. *Soil Biology & Biochemistry* **2011,** *43* (4), 766-777.

166. Hoopmann, M. R.; Finney, G. L.; MacCoss, M. J., High-speed data reduction, feature detection and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass Spectrometry. *Analytical Chemistry* **2007,** *79* (15), 5620-5632.

167. Gao, T. G.; Xu, Y. Y.; Jiang, F.; Li, B. Z.; Yang, J. S.; Wang, E. T.; Yuan, H. L., Nodulation Characterization and Proteomic Profiling of Bradyrhizobium liaoningense CCBAU05525 in Response to Water-Soluble Humic Materials. *Sci Rep* **2015,** *5*, 10836.

168. Sutton, R.; Sposito, G., Molecular structure in soil humic substances: the new view. *Environ Sci Technol* **2005,** *39* (23), 9009-15.

169. Arenella, M.; Giagnoni, L.; Masciandaro, G.; Ceccanti, B.; Nannipieri, P.; Renella, G., Interactions between proteins and humic substances affect protein identification by mass spectrometry. *Biology and Fertility of Soils* **2014,** *50* (3), 447-454.

170. Arenella, M.; D'Acqui, L. P.; Pucci, A.; Giagnoni, L.; Nannipieri, P.; Renella, G., Contact with soil-borne humic substances interfere with the prion identification by mass spectrometry. *Biology and Fertility of Soils* **2014,** *50* (6), 1009-1013.

171. Piccolo, A.; Spiteller, M., Electrospray ionization mass spectrometry of terrestrial humic substances and their size fractions. *Analytical and Bioanalytical Chemistry* **2003,** *377* (6), 1047-1059.

172. Alexandra C. Stenson, W. M. L., Alan G. Marshall, and William T. Cooper, Ionization and Fragmentation of Humic substances in ESI trasnform ion cyclotron resonance mass spectrometry. *Anal. Chem* **2002,** *74*, 4397-4409.

173. Keiblinger, K. M.; Wilhartitz, I. C.; Schneider, T.; Roschitzki, B.; Schmid, E.; Eberl, L.; Riedel, K.; Zechmeister-Boltenstern, S., Soil metaproteomics - Comparative evaluation of protein extraction protocols. *Soil Biology & Biochemistry* **2012,** *54*, 14-24.

174. Giagnoni, L.; Magherini, F.; Landi, L.; Taghavi, S.; Modesti, A.; Bini, L.; Nannipieri, P.; Van der Lelie, D.; Renella, G., Extraction of microbial proteome from soil: potential and limitations assessed through a model study. *European Journal of Soil Science* **2011,** *62* (1), 74-81.

175. Khatabi, B.; Tabrizi, N. M.; Salekdeh, G. H., Holistic Sequencing: Moving Forward from Plant Microbial Proteomics to Metaproteomics. 2016; pp 87-103.

176. Chourey, K.; Jansson, J.; VerBerkmoes, N.; Shah, M.; Chavarria, K. L.; Tom, L. M.; Brodie, E. L.; Hettich, R. L., Direct Cellular Lysis/Protein Extraction Protocol for Soil Metaproteomics. *J Proteome Res* **2010,** *9* (12), 6615-6622.

177. Bastida, F.; Hernandez, T.; Garcia, C., Metaproteomics of soils from semiarid environment: Functional and phylogenetic information obtained with different protein extraction methods. *Journal of Proteomics* **2014,** *101*, 31-42.

178. Nicora, C. D.; Anderson, B. J.; Callister, S. J.; Norbeck, A. D.; Purvine, S. O.; Jansson, J. K.; Mason, O. U.; David, M. M.; Jurelevicius, D.; Smith, R. D.; Lipton, M. S., Amino acid treatment enhances protein recovery from sediment and soils for metaproteomic studies. *Proteomics* **2013,** *13* (18-19), 2776-85.

179. Kim, J. I.; Buckau, G.; Li, G. H.; Duschner, H.; Psarros, N., Characterization of Humic and Fulvic-Acids from Gorleben Groundwater. *Fresen J Anal Chem* **1990,** *338* (3), 245-252.

180. Giannone, R. J.; Huber, H.; Karpinets, T.; Heimerl, T.; Kuper, U.; Rachel, R.; Keller, M.; Hettich, R. L.; Podar, M., Proteomic Characterization of Cellular and Molecular Processes that Enable the Nanoarchaeum equitans-Ignicoccus hospitalis Relationship. *Plos One* **2011,** *6* (8), e22942.

181. Ma, Z. Q.; Dasari, S.; Chambers, M. C.; Litton, M. D.; Sobecki, S. M.; Zimmerman, L. J.; Halvey, P. J.; Schilling, B.; Drake, P. M.; Gibson, B. W.; Tabb, D. L., IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res* **2009,** *8* (8), 3872-81.

182. Elham A. Ghabbour, G. D.; John L. Daggett, J., Christopher A.; Worgul, G. A. W. a.; Sayedbagheri, M.-M., Measuring the humic acids content of commercial lignites ad argiculture top soils in the national soil project. *Annals of Environmental Science* **2012,** *(6)*, 1-12.

183. Kang, S. H.; Xing, B. S., Phenanthrene sorption to sequentially extracted soil humic acids and humins. *Environmental Science & Technology* **2005,** *39* (1), 134-140.

184. Pan, C.; Fischer, C. R.; Hyatt, D.; Bowen, B. P.; Hettich, R. L.; Banfield, J. F., Quantitative tracking of isotope flows in proteomes of microbial communities. *Mol Cell Proteomics* **2011,** *10* (4), M110 006049.

185. Trubetskoi, O. A.; Trubetskaya, O. E., Reversed-phase high-performance liquid chromatography of the stable electrophoretic fractions of soil humic acids. *Eurasian Soil Sci+* **2015,** *48* (2), 148-156.

186. Redmile-Gordon, M. A.; Armenise, E.; White, R. P.; Hirsch, P. R.; Goulding, K. W. T., A comparison of two colorimetric assays, based upon Lowry and Bradford techniques, to estimate total protein in soil extracts. *Soil Biology & Biochemistry* **2013,** *67*, 166-173.

187. Caccavo, F., Jr.; Lonergan, D. J.; Lovley, D. R.; Davis, M.; Stolz, J. F.; McInerney, M. J., Geobacter sulfurreducens sp. nov., a hydrogen- and acetate-oxidizing dissimilatory metal-reducing microorganism. *Appl Environ Microbiol* **1994,** *60* (10), 3752-9.

188. Jung, S.; Regan, J. M., Influence of external resistance on electrogenesis, methanogenesis, and anode prokaryotic communities in microbial fuel cells. *Appl Environ Microbiol* **2011,** *77* (2), 564-71.

189. Hu, H.; Lin, H.; Zheng, W.; Rao, B.; Feng, X.; Liang, L.; Elias, D. A.; Gu, B., Mercury reduction and cell-surface adsorption by Geobacter sulfurreducens PCA. *Environ Sci Technol* **2013,** *47* (19), 10922-30.

190. Lin, H.; Morrell-Falvey, J. L.; Rao, B.; Liang, L.; Gu, B., Coupled mercury-cell sorption, reduction, and oxidation on methylmercury production by Geobacter sulfurreducens PCA. *Environ Sci Technol* **2014,** *48* (20), 11969-76.

191. Morel, F. M. M.; Kraepiel, A. M. L.; Amyot, M., The chemical cycle and bioaccumulation of mercury. *Annu Rev Ecol Syst* **1998,** *29*, 543-566.

192. Krabbenhoft, D. P.; Sunderland, E. M., Environmental science. Global change and mercury. *Science* **2013,** *341* (6153), 1457-8.

193. Barkay, T.; Wagner-Dobler, I., Microbial transformations of mercury: potentials, challenges, and achievements in controlling mercury toxicity in the environment. *Adv Appl Microbiol* **2005,** *57*, 1-52.

194. Podar, M.; Gilmour, C. C.; Brandt, C. C.; Soren, A.; Brown, S. D.; Crable, B. R.; Palumbo, A. V.; Somenahally, A. C.; Elias, D. A., Global prevalence and distribution of genes and microorganisms involved in mercury methylation. *Sci Adv* **2015,** *1* (9), e1500675.

195. Smith, S. D.; Bridou, R.; Johs, A.; Parks, J. M.; Elias, D. A.; Hurt, R. A., Jr.; Brown, S. D.; Podar, M.; Wall, J. D., Site-directed mutagenesis of HgcA and HgcB reveals amino acid residues important for mercury methylation. *Appl Environ Microbiol* **2015,** *81* (9), 3205-17.

196. Doukov, T. I.; Iverson, T. M.; Seravalli, J.; Ragsdale, S. W.; Drennan, C. L., A Ni-Fe-Cu center in a bifunctional carbon monoxide dehydrogenase/acetyl-CoA synthase. *Science* **2002,** *298* (5593), 567-72.

197. Choi, S. C.; Chase, T., Jr.; Bartha, R., Enzymatic catalysis of mercury methylation by Desulfovibrio desulfuricans LS. *Appl Environ Microbiol* **1994,** *60* (4), 1342-6.

198. Ragsdale, S. W., Enzymology of the wood-Ljungdahl pathway of acetogenesis. *Ann N Y Acad Sci* **2008,** *1125*, 129-36.

199. Lin, H.; Hurt, R. A.; Johs, A.; Parks, J. M.; Morrell-Falvey, J. L.; Liang, L.; Elias, D. A.; Gu, B., Unexpected Effects of Gene Deletion on Interactions of Mercury with the Methylation-Deficient Mutant ΔhgcAB. *Environmental Science & Technology Letters* **2014,** *1* (5), 271-276.

200. Voordeckers, J. W.; Kim, B. C.; Izallalen, M.; Lovley, D. R., Role of Geobacter sulfurreducens outer surface c-type cytochromes in reduction of soil humic acid and anthraquinone-2,6-disulfonate. *Appl Environ Microbiol* **2010,** *76* (7), 2371-5.

201. Lin, H.; Lu, X.; Liang, L.; Gu, B., Cysteine Inhibits Mercury Methylation byGeobacter sulfurreducensPCA Mutant ΔomcBESTZ. *Environmental Science & Technology Letters* **2015,** *2* (5), 144-148.

202. Zhang, B.; VerBerkmoes, N. C.; Langston, M. A.; Uberbacher, E.; Hettich, R. L.; Samatova, N. F., Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res* **2006,** *5* (11), 2909-18.

203. Abraham, P.; Adams, R. M.; Tuskan, G. A.; Hettich, R. L., Moving away from the reference genome: evaluating a peptide sequencing tagging approach for single amino acid polymorphism identifications in the genus Populus. *J Proteome Res* **2013,** *12* (8), 3642-51.

204. Xiong, W.; Giannone, R. J.; Morowitz, M. J.; Banfield, J. F.; Hettich, R. L., Development of an enhanced metaproteomic approach for deepening the microbiome characterization of the human infant gut. *J Proteome Res* **2015,** *14* (1), 133-41.

205. Li, Z.; Adams, R. M.; Chourey, K.; Hurst, G. B.; Hettich, R. L.; Pan, C., Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on LTQ Orbitrap Velos. *J Proteome Res* **2012,** *11* (3), 1582-90.

206. Deshmukh, A. S.; Murgia, M.; Nagaraj, N.; Treebak, J. T.; Cox, J.; Mann, M., Deep proteomics of mouse skeletal muscle enables quantitation of protein isoforms, metabolic pathways, and transcription factors. *Mol Cell Proteomics* **2015,** *14* (4), 841-53.

207. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997,** *25* (17), 3389-402.

208. Yu, N. Y.; Wagner, J. R.; Laird, M. R.; Melli, G.; Rey, S.; Lo, R.; Dao, P.; Sahinalp, S. C.; Ester, M.; Foster, L. J.; Brinkman, F. S. L., PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **2010,** *26* (13), 1608-1615.

209. Gardy, J. L.; Laird, M. R.; Chen, F.; Rey, S.; Walsh, C. J.; Ester, M.; Brinkman, F. S. L., PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **2005,** *21* (5), 617-623.

210. Maere, S.; Heymans, K.; Kuiper, M., BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* **2005,** *21* (16), 3448-3449.

211. Nies, D. H., Efflux-mediated heavy metal resistance in prokaryotes. *FEMS microbiology reviews* **2003,** *27* (2-3), 313-339.

212. Valencia, E. Y.; Braz, V. S.; Guzzo, C.; Marques, M. V., Two RND proteins involved in heavy metal efflux in Caulobacter crescentus belong to separate clusters within proteobacteria. *Bmc Microbiol* **2013,** *13*.

213. Methe, B. A.; Webster, J.; Nevin, K.; Butler, J.; Lovley, D. R., DNA microarray analysis of nitrogen fixation and Fe(III) reduction in Geobacter sulfurreducens. *Appl Environ Microbiol* **2005,** *71* (5), 2530-8.

214. Long, F.; Su, C. C.; Lei, H. T.; Bolla, J. R.; Do, S. V.; Yu, E. W., Structure and mechanism of the tripartite CusCBA heavy-metal efflux complex. *Philos Trans R Soc Lond B Biol Sci* **2012,** *367* (1592), 1047-58.

215. Liu, Y.; Wang, Z.; Liu, J.; Levar, C.; Edwards, M. J.; Babauta, J. T.; Kennedy, D. W.; Shi, Z.; Beyenal, H.; Bond, D. R.; Clarke, T. A.; Butt, J. N.; Richardson, D. J.; Rosso, K. M.; Zachara, J. M.; Fredrickson, J. K.; Shi, L., A trans-outer membrane porin-cytochrome protein complex for extracellular electron transfer by Geobacter sulfurreducens PCA. *Environ Microbiol Rep* **2014,** *6* (6), 776-85.

216. Aklujkar, M.; Coppi, M. V.; Leang, C.; Kim, B. C.; Chavan, M. A.; Perpetua, L. A.; Giloteaux, L.; Liu, A.; Holmes, D. E., Proteins involved in electron transfer to Fe(III) and Mn(IV) oxides by Geobacter sulfurreducens and Geobacter uraniireducens. *Microbiology* **2013,** *159* (Pt 3), 515-35.

217. Holmes, D. E.; Mester, T.; O'Neil, R. A.; Perpetua, L. A.; Larrahondo, M. J.; Glaven, R.; Sharma, M. L.; Ward, J. E.; Nevin, K. P.; Lovley, D. R., Genes for two multicopper proteins required for Fe(III) oxide reduction in Geobacter sulfurreducens have different expression patterns both in the subsurface and on energy-harvesting electrodes. *Microbiol-Sgm* **2008,** *154*, 1422-1435.

218. Methe, B. A.; Nelson, K. E.; Eisen, J. A.; Paulsen, I. T.; Nelson, W.; Heidelberg, J. F.; Wu, D.; Wu, M.; Ward, N.; Beanan, M. J.; Dodson, R. J.; Madupu, R.; Brinkac, L. M.; Daugherty, S. C.; DeBoy, R. T.; Durkin, A. S.; Gwinn, M.; Kolonay, J. F.; Sullivan, S. A.; Haft, D. H.; Selengut, J.; Davidsen, T. M.; Zafar, N.; White, O.; Tran, B.; Romero, C.; Forberger, H. A.; Weidman, J.; Khouri, H.; Feldblyum, T. V.; Utterback, T. R.; Van Aken, S. E.; Lovley, D. R.; Fraser, C. M., Genome of Geobacter sulfurreducens: metal reduction in subsurface environments. *Science* **2003,** *302* (5652), 1967-9.

219. Leang, C.; Qian, X. L.; Mester, T.; Lovley, D. R., Alignment of the c-Type Cytochrome OmcS along Pili of Geobacter sulfurreducens. *Appl Environ Microb* **2010,** *76* (12), 4080-4084.

220. Shi, Z. F.; Rui, W.; Feng, Y. F.; Guo, J., Bioinformatics Study on Relationship of Hyperlipidemia and Phospholipids Based on Cytoscape Software. *Aer Adv Eng Res* **2014,** *5*, 357-360.

221. Hsu-Kim, H.; Kucharzyk, K. H.; Zhang, T.; Deshusses, M. A., Mechanisms regulating mercury bioavailability for methylating microorganisms in the aquatic environment: a critical review. *Environ Sci Technol* **2013,** *47* (6), 2441-56.

222. Gilmour, C. C.; Podar, M.; Bullock, A. L.; Graham, A. M.; Brown, S. D.; Somenahally, A. C.; Johs, A.; Hurt, R. A., Jr.; Bailey, K. L.; Elias, D. A., Mercury methylation by novel microorganisms from new environments. *Environ Sci Technol* **2013,** *47* (20), 11810-20.

223. Gilmour, C. C.; Elias, D. A.; Kucken, A. M.; Brown, S. D.; Palumbo, A. V.; Schadt, C. W.; Wall, J. D., Sulfate-Reducing Bacterium Desulfovibrio desulfuricans ND132 as a Model for Understanding Bacterial Mercury Methylation. *Appl Environ Microb* **2011,** *77* (12), 3938-3951.

224. Goni-Urriza, M.; Corsellis, Y.; Lanceleur, L.; Tessier, E.; Gury, J.; Monperrus, M.; Guyoneaud, R., Relationships between bacterial energetic metabolism, mercury methylation potential, and hgcA/hgcB gene expression in Desulfovibrio dechloroacetivorans BerOc1. *Environ Sci Pollut R* **2015,** *22* (18), 13764-13771.

225. Poulain, A. J.; Barkay, T., Environmental science. Cracking the mercury methylation code. *Science* **2013,** *339* (6125), 1280-1.

226. Hobman, J. L.; Wilkie, J.; Brown, N. L., A design for life: prokaryotic metal-binding MerR family regulators. *Biometals* **2005,** *18* (4), 429-36.

227. Schaefer, J. K.; Szczuka, A.; Morel, F. M., Effect of divalent metals on Hg(II) uptake and methylation by bacteria. *Environ Sci Technol* **2014,** *48* (5), 3007-13.

228. Hu, H.; Lin, H.; Zheng, W.; Tomanicek, S. J.; Johs, A.; Feng, X.; Elias, D. A.; Liang, L.; Gu, B., Oxidation and methylation of dissolved elemental mercury by anaerobic bacteria. *Nature Geosci.* **2013,** *6* (9), 751-754.

229. Lin, H.; Lu, X.; Liang, L.; Gu, B., Thiol-facilitated cell export and desorption of methylmercury by anaerobic bacteria. *Environ. Sci. Technol. Lett.* **2015,** *2* (10), 292-296.

230. Lin, H.; Morrell-Falvey, J. L.; Rao, B.; Liang, L.; Gu, B., Coupled mercury-cell sorption, reduction, and oxidation affecting methylmercury production by *Geobacter sulfurreducens* PCA. *Environ. Sci. Technol.* **2014,** *48* (20), 11969-11976.

231. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **2008,** *26* (12), 1367-72.

232. Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J., The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* **2016,** *13* (9), 731-40.

233. Moriya, Y.; Itoh, M.; Okuda, S.; Yoshizawa, A. C.; Kanehisa, M., KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* **2007,** *35* (Web Server issue), W182-5.

234. Kanehisa, M.; Goto, S., KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **2000,** *28* (1), 27-30.

235. Caspi, R.; Billington, R.; Ferrer, L.; Foerster, H.; Fulcher, C. A.; Keseler, I. M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Mueller, L. A.; Ong, Q.; Paley, S.; Subhraveti, P.; Weaver, D. S.; Karp, P. D., The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* **2016,** *44* (D1), D471-80.

236. Liu, Y. R.; Lu, X.; Zhao, L.; An, J.; He, J. Z.; Pierce, E. M.; Johs, A.; Gu, B., Effects of Cellular Sorption on Mercury Bioavailability and Methylmercury Production by Desulfovibrio desulfuricans ND132. *Environ Sci Technol* **2016,** *50* (24), 13335-13341.

237. Cox, J.; Mann, M., 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **2012,** *13 Suppl 16*, S12.

238. Ma, Z.; Jacobsen, F. E.; Giedroc, D. P., Coordination chemistry of bacterial metal transport and sensing. *Chem Rev* **2009,** *109* (10), 4644-81.

239. Tam, R.; Saier, M. H., Structural, Functional, and Evolutionary Relationships among Extracellular Solute-Binding Receptors of Bacteria. *Microbiol Rev* **1993,** *57* (2), 320-346.

240. Kracke, F.; Vassilev, I.; Kromer, J. O., Microbial electron transport and energy conservation - the foundation for optimizing bioelectrochemical systems. *Front Microbiol* **2015,** *6*, 575.

241. Buckel, W.; Thauer, R. K., Energy conservation via electron bifurcating ferredoxin reduction and proton/Na(+) translocating ferredoxin oxidation. *Biochim Biophys Acta* **2013,** *1827* (2), 94-113.

242. Wolfe, A. J., The acetate switch. *Microbiol Mol Biol Rev* **2005,** *69* (1), 12-50.

243. Lu, X.; Liu, Y.; Johs, A.; Zhao, L.; Wang, T.; Yang, Z.; Lin, H.; Elias, D. A.; Pierce, E. M.; Liang, L.; Barkay, T.; Gu, B., Anaerobic Mercury Methylation and Demethylation by Geobacter bemidjiensis Bem. *Environ Sci Technol* **2016,** *50* (8), 4366-73.

244. Lin, H.; Hurt, R. A.; Johs, A.; Parks, J. M.; Morrell-Falvey, J. L.; Liang, L. Y.; Elias, D. A.; Gu, B. H., Unexpected Effects of Gene Deletion on Interactions of Mercury with the Methylation-Deficient Mutant Delta hgcAB. *Environmental Science & Technology Letters* **2014,** *1* (5), 271-276.

245. Amara, I. E.; Anwar-Mohamed, A.; El-Kadi, A. O., Posttranslational mechanisms modulating the expression of the cytochrome P450 1A1 gene by methylmercury in HepG2 cells: a role of heme oxygenase-1. *Toxicol Lett* **2013,** *219* (3), 239-47.

246. Moberly, J. G.; Miller, C. L.; Brown, S. D.; Biswas, A.; Brandt, C. C.; Palumbo, A. V.; Elias, D. A., Role of morphological growth state and gene expression in Desulfovibrio africanus strain Walvis Bay mercury methylation. *Environ Sci Technol* **2012,** *46* (9), 4926-32.

247. Lee, B. K.; Jung, K. S.; Son, C.; Kim, H.; VerBerkmoes, N. C.; Arshava, B.; Naider, F.; Becker, J. M., Affinity purification and characterization of a G-protein coupled receptor, Saccharomyces cerevisiae Ste2p. *Protein Expr Purif* **2007,** *56* (1), 62-71.

248. Umanah, G.; Huang, L. Y.; Schultz, P. G.; Naider, F.; Becker, J. M., Incorporation of the Unnatural Amino Acid p-benzoyl-L-phenylalanine (Bpa) into a G Protein-coupled Receptor in its Native Context. *Peptides for Youth* **2009,** *611*, 333-335.

249. Kauer, J. C.; Erickson-Viitanen, S.; Wolfe, H. R., Jr.; DeGrado, W. F., p-Benzoyl-L-phenylalanine, a new photoreactive amino acid. Photolabeling of calmodulin with a synthetic calmodulin-binding peptide. *J Biol Chem* **1986,** *261* (23), 10695-700.

250. Becker, J. M.; Naider, F., Cross-linking strategies to study peptide ligand-receptor interactions. *Methods Enzymol* **2015,** *556*, 527-47.

251. Wang, L.; Brock, A.; Herberich, B.; Schultz, P. G., Expanding the genetic code of Escherichia coli. *Science* **2001,** *292* (5516), 498-500.

252. Chin, J. W.; Cropp, T. A.; Anderson, J. C.; Mukherji, M.; Zhang, Z.; Schultz, P. G., An expanded eukaryotic genetic code. *Science* **2003,** *301* (5635), 964-7.

253. Smith, E.; Collins, I., Photoaffinity labeling in target- and binding-site identification. *Future Med Chem* **2015,** *7* (2), 159-183.

254. Das, S.; Oliver, D. B., Mapping of the SecA.SecY and SecA.SecG Interfaces by Site-directed in Vivo Photocross-linking. *Journal of Biological Chemistry* **2011,** *286* (14).

255. Volke, D.; Krizsan, A.; Berthold, N.; Knappe, D.; Hoffmann, R., Identification of Api88 Binding Partners in Escherichia coli Using a Photoaffinity-Cross-Link Strategy and Label-Free Quantification. *J Proteome Res* **2015,** *14* (8), 3274-3283.

256. Farrelli, I. S.; Toroney, R.; Hazen, J. L.; Mehl, R. A.; Chin, J. W., Photo-cross-linking interacting proteins with a genetically encoded benzophenone. *Nature Methods* **2005,** *2* (5), 377-384.

257. Son, C. D.; Sargsyan, H.; Naider, F.; Becker, J. M., Identification of ligand binding regions of the Saccharomyces cerevisiae alpha-factor pheromone receptor by photoaffinity cross-linking. *Biochemistry* **2004,** *43* (41), 13193-13203.

258. Bertucci, C.; Domenici, E., Reversible and covalent binding of drugs to human serum albumin: Methodological approaches and physiological relevance. *Curr Med Chem* **2002,** *9* (15), 1463-1481.

259. Bujacz, A.; Zielinski, K.; Sekula, B., Structural studies of bovine, equine, and leporine serum albumin complexes with naproxen. *Proteins* **2014,** *82* (9), 2199-208.

260. Hamilton, J. A., NMR reveals molecular interactions and dynamics of fatty acid binding to albumin. *Biochim Biophys Acta* **2013,** *1830* (12), 5418-26.

261. Son, C. D.; Sargsyan, H.; Naider, F.; Becker, J. M., Identification of ligand binding regions of the Saccharomyces cerevisiae alpha-factor pheromone receptor by photoaffinity cross-linking. *Biochemistry* **2004,** *43* (41), 13193-203.

262. Raths, S. K.; Naider, F.; Becker, J. M., Peptide analogues compete with the binding of alpha-factor to its receptor in Saccharomyces cerevisiae. *J Biol Chem* **1988,** *263* (33), 17333-41.

263. Mohammed, H.; Taylor, C.; Brown, G. D.; Papachristou, E. K.; Carroll, J. S.; D'Santos, C. S., Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) for analysis of chromatin complexes. *Nat Protoc* **2016,** *11* (2), 316-26.

264. Mueller-Planitz, F., Crossfinder-assisted mapping of protein crosslinks formed by site-specifically incorporated crosslinkers. *Bioinformatics* **2015,** *31* (12), 2043-5.

265. Bujacz, A., Structures of bovine, equine and leporine serum albumin. *Acta Crystallogr D Biol Crystallogr* **2012,** *68* (Pt 10), 1278-89.

266. Cravatt, B. F.; Simon, G. M.; Yates, J. R., 3rd, The biological impact of mass-spectrometry-based proteomics. *Nature* **2007,** *450* (7172), 991-1000.

267. Williams, M. A.; Taylor, E. B.; Mula, H. P., Metaproteomic characterization of a soil microbial community following carbon amendment. *Soil Biology & Biochemistry* **2010,** *42* (7), 1148-1156.

268. Wilkins, M. J.; Wrighton, K. C.; Nicora, C. D.; Williams, K. H.; McCue, L. A.; Handley, K. M.; Miller, C. S.; Giloteaux, L.; Montgomery, A. P.; Lovley, D. R.; Banfield, J. F.; Long, P. E.; Lipton, M. S., Fluctuations in Species-Level Protein Expression Occur during Element and Nutrient Cycling in the Subsurface. *Plos One* **2013,** *8* (3).

269. Wang, D. Z.; Kong, L. F.; Li, Y. Y.; Xie, Z. X., Environmental Microbial Community Proteomics: Status, Challenges and Perspectives. *Int J Mol Sci* **2016,** *17* (8).

VITA

Chen Qian was born and raised in Yixing, Jiangsu, China. She received her Bachelor degree of Biological Engineering and Master degree of Microbiology in Nanjing Tech University, Nanjing, China. She began her Ph.D. study in UT-ORNL Graduate School of Genome Science and Technology in August 2012. She expects to receive her Ph.D. degree at August 2017.