

Needles in a haystack of protein diversity: Interrogation of complex biological samples through specialized strategies in bottom-up proteomics uncover peptides of interest for diverse applications.

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Manuel Ivan Villalobos Solis

August 2020

ACKNOWLEDGMENTS

I would like to thank everyone who made this PhD dissertation work possible. It has been a long road since I first decided to leave my home back in Mexico and travel abroad to continue my graduate education. Without a doubt I can say that this work started more than seven years ago when, while completing master's studies at McGill University, I was first introduced to the field of biological mass spectrometry. Along the road I have met exceptional people that have widened my academic and cultural horizons.

"Teaching should be such that what is offered is perceived as a valuable gift and not as a hard duty" – Albert Einstein. My first thanks go to Dr. Robert (Bob) L. Hettich. The work presented during this dissertation is also a reflection of the teachings of my advisor. I will be always grateful with Bob to allow me the opportunity to conduct graduate research with his group at Oak Ridge National Laboratory. Different than what I heard from other colleagues during my time in the PhD, Bob as an advisor, even with his busy schedule, always made time to talk about research updates and always encouraged us to find ways to bring our work to reality. He is a very positive guy and when you talk with him, that positivism gets into you too. During my time in his group, it always surprised me how Bob could enable interesting discussions amongst his students and how he would motivate us to think "outside the box". In the same way I would like to thank Dr. Paul E. Abraham and Dr. Richard J. Giannone, both staff scientists in my group, for all their support during these years. If we imagine my research group as an old Greek temple, while Bob is the foundation, those two guys are the pillars of it. Their open-door policies with students and all the creative thinking and energy that they bring to each research project is contagious. I am thankful for all the conversations that we had during these years but above all for the patience they had in teaching me about the nuts and bolts of mass spectrometry.

"Buying bread from a man in Brussels. He was six-foot-four and full of muscle I said, Do you speak-a my language? He just smiled and gave me a vegemite sandwich" – Lyrics of the song Land Down Under by Men at Work. My decision of studying abroad also brought with it the joy and opportunity to meet many extraordinary people from all over the world. I have friends from India, Nepal, China, U.S., Canada, and other places. My interactions with them helped me understand more of their ways of thinking and living while at the same time making me realize about the social bonds that we have as human beings. I am now a fan of mango lassi and dosas, momos, fried mashi, smoked meat and Knoxville-style barbeque meat. The Hettich's lab is made of very bright people and specially I would like to say to my PhD colleagues: धन्यवाद, धन्यवाद, 谢谢, Thank you! For all the talks and laughter that we shared over the years.

"Education is the most powerful weapon which you can use to change the world." -Nelson Mandela. Behind of the person that I am today, is my family. Both my mom, Miriam Solis Dominguez. and my dad, Manuel Villalobos Garcia, gave me a foundation with which to live honestly, humbly, and responsibly. I will always say that the best gift that they could have given me was my education and the courage to go further. Both my parents came from humble families who believe that education was the way to open the door of having a good life. The PhD was a challenging journey but one that I could endure thanks to my mom, my dad, and my brother, Cesar David Villalobos Solis, who were always there for me. Thank you for all your advices, support, and unconditional love! Thanks for believing in me no matter what life could bring. Finally, this dissertation is dedicated to all my family but specially to two special persons who I lost during my studies, my maternal grandma, Luz Alicia Dominguez Aguilar, and my paternal granddad, Manuel Villalobos Cunjama. Both were great examples to me of hard work, humility and how to live with a smile on your face. The sacrifices and love that you both did and put until the end so your families and future generations could achieve better lives will never be forgotten... And today I proudly said: Let many more successes come and MUCHAS GRACIAS for paving the way!

ABSTRACT

Peptide identification is at the core of bottom-up proteomics measurements. However, even with state-of the-art mass spectrometric instrumentation, peptide level information is still lost or missing in these types of experiments. Reasons behind missing peptide identifications in bottom-up proteomics include variable peptide ionization efficiencies, ion suppression effects, as well as the occurrence of chimeric spectra that can lower the efficacy of database search strategies. Peptides derived from naturally abundant proteins in a biological system also have better chances of being identified in comparison to the ones produced from less abundant proteins, at least in regular discovery-based proteomics experiments. This dissertation focused on the recovery of the “missing or hidden proteome” information in complex biological matrices by approaching this challenge under a peptide-centric view and implementing different liquid chromatography tandem mass spectrometry (LC-MS/MS) experimental workflows. In particular, the projects presented here covered: (1) The feasibility of applying a liquid chromatography-multiple reaction monitoring MS methodology for the targeted identification of peptides serving as surrogates of protein biomarkers in environmental matrices with unknown microbial diversities; (2) the evaluation of selecting unique tryptic peptides *in-silico* that can distinguish groups of proteins, instead of individual proteins, for targeted proteomics workflows; (3) maximizing peptide identification in spectral data collected from different LC-MS/MS setups by applying a multi-peptide-spectrum-match algorithm, and (4) showing that LC-MS/MS combined with *de novo* assisted-database searches is a feasible strategy for the comprehensive identification of peptides derived from native proteolytic mechanisms in biological systems.

TABLE OF CONTENTS

CHAPTER 1 Introduction to mass spectrometry-based proteomics and overview of research topics covered in this dissertation	1
1.1 The advent of MS-based proteomics in the biological sciences	1
1.1.1 Functional genomics and the wholistic view of biology	1
1.1.2 Proteomics gives meaning to genomics and transcriptomics information	2
1.1.3 The practical implications of proteomics in biological research	4
1.2 Overview of MS-based proteomics for biological research	6
1.2.1 Mass spectrometry bottom-up proteomics as the powerhouse for large-scale/high-throughput analysis of proteomes	6
1.2.2 Global proteomics by data-dependent acquisition-mass spectrometry (DDA-MS).....	7
1.2.3 Targeted proteomics by multiple reaction monitoring-mass spectrometry (MRM-MS).	8
1.2.4 Relative label free and absolute quantification of proteins in bottom-up proteomics..	11
1.3 Bottom-up proteomics in environmental microbiology applications	13
1.3.1 The role of bottom-up proteomics in microbial environmental bioremediation	13
1.3.2 From the clinic to the river: Environmental monitoring of protein biomarkers with targeted mass spectrometry.....	14
1.4 Isolation interference in bottom-up proteomics experiments	16
1.4.1 The occurrence of chimeric spectra in DDA bottom-up proteomics experiments.....	16
1.4.2 Exploiting chimeric spectra in database-based peptide searches to increase proteome depth	18
1.5 LC-MS/MS based peptidomics	19
1.5.1 Small and mighty: Proteolytic cleavage products (PCPs) across different domains of life.....	19
1.5.2 Challenges in the identification of PCPs by LC-MS/MS: Natural protein dynamic ranges, databases and PTMs	22
1.6 Dissertation summary	23
CHAPTER 2 Experimental methodologies, mass spectrometry instrumentation and bioinformatic approaches used for proteomics analyses in this dissertation	25
2.1 Overview of a bottom-up proteomics workflow	25
2.2 Protein extraction and digestion.....	28
2.2.1 Being mindful about the sample: Considerations before starting a bottom-up proteomics experiment.	28
2.2.2 Cell lysis and protein extraction via the SDS-TCA and CME methods	29
2.2.3 Protein solubilization with urea or SDC buffers and DTT and IAA.	31
2.2.4 Protein digestion in MS bottom-up proteomics with trypsin.....	32

2.2.5 Peptide clean up before loading into a LC column.	33
2.3 Peptide separation by nanoLC	34
2.3.1 Basics of peptide chromatographic separation by nanoLC.	34
2.3.2 HPLC vs. UHPLC	35
2.3.3 Reverse phase liquid chromatography (RP-LC)	37
2.3.4 Two-dimensional liquid chromatography (2D-LC)	37
2.4 Fundamentals of electrospray ionization (ESI)	39
2.5 Fundamentals of mass spectrometric detection	42
2.5.1 Figures of merit in mass spectrometric detection	42
2.5.2 Overview of tandem mass analyzers and detectors used in bottom-up proteomics applications	44
2.5.2.1 Targeted proteomics via multiple reaction monitoring MS on a triple quadrupole (MRM-MS).....	46
2.5.2.2 Global proteomics through spectral data-dependent acquisition on a Q Exactive hybrid quadrupole-Orbitrap	48
2.6 Fundamentals of tandem MS (MS/MS)	50
2.7 Downstream processing of MS/MS data in global and targeted proteomics	51
2.7.1 Introduction to database search driven algorithms in bottom-up global proteomics research	51
2.7.1.1 Peptide identification by database searching.....	53
2.7.1.2 Protein inference.....	55
2.7.2 Overview of database search engines used in this dissertation	57
2.7.2.1 Myrimatch-IDPicker	57
2.7.2.2 CharmeRT	58
2.7.2.3 <i>De novo</i> -assisted peptide database searches with PEAKS	59
2.7.3 Targeted proteomics data analyses with Skyline.	60
2.8 Post-processing of bottom-up proteomics data	62
CHAPTER 3 Evaluation of the application feasibility of a LC-MRM-MS targeted proteomics approach for the detection of <i>Dehalococcoides mccartyi</i> protein biomarkers in chloroethene contaminated groundwater	66
3.1 Introduction	66
3.2 Experimental procedures.....	67
3.2.1 <i>Dehalococcoides mccartyi</i> (Dhc) cultures and growth conditions.	67
3.2.2 Biomass recovery from groundwater samples.....	68
3.2.3 Sample preparation for global and targeted proteomics analyses.	68
3.2.4 Global proteomics of axenic cultures and groundwater samples.	69

3.2.5 RDase phylogenetic tree construction	70
3.2.6 LC-MRM-MS method development.....	71
3.2.7 Analyses of the commercial bioaugmentation culture BDI and groundwater samples by LC-MRM-MS.....	72
3.3 Results & Discussions	73
3.3.1 Global proteomics exploration of axenic <i>Dhc</i> cultures enables selection of candidate peptides for targeted proteomics.....	73
3.3.2 Selection of <i>Dhc</i> MRM-MS observable peptides and <i>in-silico</i> evaluation of their biological specificities	78
3.4 Conclusions.....	86
CHAPTER 4 <i>In-silico</i> assessment of the selection of a minimal set of unique peptides for glycoside-hydrolases families for the development of a targeted proteomics methodology: demonstrations with a global non-system specific and a system-specific approach	88
4.1 Introduction	88
4.2 Materials & Methods.....	91
4.2.3 CAZyDB.07312019 from dbCAN2 meta site	91
4.2.4 Unique tryptic peptides from GH proteins compared to backgrounds of other CAZymes	92
4.2.5 Generation of a minimum list of unique tryptic peptides for GH families	92
4.2.6 Super kingdom taxonomy of groups of proteins having unique peptides selected from the CAZy database	93
4.2.7 Re-processing of 1401 MAGs in dbCAN2	94
4.2.8 Selection of unique tryptic peptides from GH families in the biogas microbiome MAGS and generation of a minimum list of peptides	94
4.3 Results & Discussion	95
4.3.1 A minimum set of unique peptides distinguishing between GH families using sequence information from CAZyDB.....	95
4.3.2 Unique tryptic peptides selected for group of GHs in CAZyDB can separate groups of proteins based on their superkingdom origins	99
4.3.3 Narrowing down the number of unique peptides by constraining the proteome space	101
4.3.4 Taxonomic diversity of the 1401 high-quality (HQ) and medium-high-quality (MHQ) MAGs in the biogas microbiome	101
4.3.5 Distribution of GHs and other CAZymes in the proteomes of the 1401 MAGs.....	103
4.3.6 A unique set of peptides for GH families in in the biogas microbiome	107
4.3.7 Unique tryptic peptides selected for group of GHs in the gas microbiome can separate groups of proteins based on their taxonomical origins	110

4.3.8 Unique tryptic peptides selected for group of GHs in the gas microbiome can separate groups of proteins based on their enzymatic specificity	112
4.4 Conclusions.....	114
CHAPTER 5 An updated perspective on deconvoluting chimeric MS/MS spectra by liquid chromatography and precursor isolation and their subsequent assignment by a multi-peptide-spectrum match algorithm	116
5.1 Introduction	116
5.2 Experimental procedures.....	118
5.2.1 Standard samples for LC-MS/MS analyses.....	118
5.2.2 1D-LC-MS/MS runs of HeLa tryptic digest standards.....	118
5.2.3 2D-LC-MS/MS runs of HeLa tryptic digest standards.....	119
5.2.4 PSMs, peptide, and protein identifications by database search.....	120
5.2.5 Assessment of the performance of each 1D and 2D LC-MS/MS runs.....	120
5.3 Results & Discussions	121
5.3.1 Performance metrics across a range of nanoLC peptide separation techniques	121
5.3.2 Comparison of precursor isolation interferences across a range of nanoLC peptide separation techniques	122
5.3.3 Evaluation of the CharmERT mPSM search algorithm across different LC peptide separation techniques	124
5.3.4 Evaluation of spectra collected from LC peptide separation techniques with different isolation widths and searched with CharmERT.....	128
5.4 Conclusions.....	130
CHAPTER 6 LC-MS/MS based discovery and characterization of proteolytic cleavage products (PCPs) produced during the mutualistic symbiotic interactions between the perennial plant <i>Populus x canescens</i> and the ectomycorrhizal basidiomycete <i>Laccaria bicolor</i>	132
6.1 Introduction	132
6.2 Experimental procedures.....	134
6.2.1 <i>E. coli</i> background proteome sample preparation	134
6.2.2 Universal Proteomics Standard (UPS1) sample preparation	135
6.2.3 Heavy-labeled peptide standards.....	135
6.2.4 Preparation of master mixes for benchmarking the performance of a 10 kDa enrichment strategy of PCPs and their identification/quantification via <i>de novo</i> -assisted database search	135
6.2.5 1D-LC-MS/MS of master mixes samples	136
6.2.6 Preparation of <i>Populus X canescens</i> with or without <i>Laccaria bicolor</i> ectomycorrhizae samples.....	136
6.2.7 PCPs extraction from <i>Populus X canescens</i> and <i>Laccaria bicolor</i> samples.....	137

6.2.8 1D-LC-MS/MS of <i>Populus X canescens</i> with or without <i>Laccaria bicolor</i> samples ..	138
6.2.9 PSMs, peptide, and protein identifications by PEAKS.....	138
6.3 Results & Discussion	139
6.3.1 Benchmarking the identification and quantitation of PCPs.....	139
6.3.2 Identification metrics of PCPs from <i>Populus</i> and its ectomycorrhizal symbiont <i>L. bicolor</i>	144
6.3.3 <i>De novo</i> database-assisted searches enable the identification of non-functional and functional PCPs in the peptidome of different sample types	148
6.3.4 Comparative quantitative analysis of PCPs defines a set of significant analytes in ectomycorrhiza symbiosis between <i>Populus X canescens</i> and <i>L. bicolor</i>	150
6.3.4.1 <i>L. bicolor</i> endogenous peptides associated with symbiosis with <i>Populus</i>	150
6.3.4.2 <i>Populus X canescens</i> endogenous peptides associated with symbiosis with <i>L. bicolor</i>	153
6.3.4.5 Differentially abundant PCPs shared between <i>L. bicolor</i> proteins and <i>Populus X canescens</i> proteins	155
6.4 Conclusions.....	155
CHAPTER 7 Outlook, Challenges, and Conclusions.....	157
7.1 The potential of targeted proteomics to monitor environmentally relevant microorganisms in bioremediation.....	157
7.2 Measurement of groups of proteins by targeted proteomics to monitor metabolic processes in anaerobic bioreactors for biogas production	160
7.3 Getting more out of DDA-collected spectra in bottom-up proteomics experiments	162
7.4 Advantages of LC-MS/MS peptidomics to uncover bioactive peptides in different systems	165
7.5 Concluding perspective	168
List of references	170
VITA	200

LIST OF TABLES

Table 2.1 Performance metrics of the instruments used in this dissertation	45
Table 3.1 <i>Dhc</i> protein biomarkers used as initial targets in this study.....	74
Table 4.1 GH families selected for testing the <i>in-silico</i> development of a minimum list of unique peptides able to differentiate between groups of enzymes using sequence information from CAZyDB	96
Table 4.2 GH families selected from the biogas microbiome data for testing the <i>in-silico</i> development of a minimum list of unique peptides able to differentiate between groups of enzymes.....	108
Table 6.1 Heavy peptide standards (>97% purity) spiked into constant amounts of background <i>E. coli</i> proteome and UPS tryptic digest used to benchmark the <i>de novo</i> peptide identification and quantification pipeline	140

LIST OF FIGURES

Figure 1.1 Types of tandem mass spectrometry acquisition methods used in this dissertation.....	10
Figure 1.2 Chimeric spectra in LC-MS/MS bottom-up proteomics experiments	17
Figure 1.3 Biological origins of PCPs	20
Figure 2.1 Overview of a bottom-up proteomics experimental pipeline	26
Figure 2.2 1D-liquid chromatography setups.....	36
Figure 2.3 2D-liquid chromatography setup.....	39
Figure 2.4 Illustration of the electrospray ionization process	40
Figure 2.5 Mass resolution in mass spectrometry-based bottom-up proteomics studies impacts the correct identification of peptides	43
Figure 2.6 Schematic of the TSQ Quantum Ultra mass spectrometer used for targeted proteomics measurements via MRM-MS for the project presented in Chapter 3.....	47
Figure 2.7 Schematic of the Q Exactive Plus mass spectrometer used for global proteomics measurements via DDA in Chapters 3-5.....	49
Figure 2.8 Example of the most common fragment ions observed and used during the analysis of peptides/proteins by bottom-up proteomics.....	51
Figure 2.9 Peptide identification strategies used in this dissertation.....	52
Figure 3.1 Workflow for the selection of peptide signals in pure cultures of <i>Dhc</i> strains 195, FL2 and BAV1.....	75
Figure 3.2 Global proteomics metrics from the analyses of three pure cultures of <i>Dhc</i> ...	76
Figure 3.3 Number of protein groups (> 85% sequence identity) identified in each culture	77
Figure 3.4 LC-MRM-MS <i>Dhc</i> biomarker identification in a tryptic digest of the PCE-to-ethene dechlorinating consortium BDI	81
Figure 3.5 qPCR measurements of bacterial, <i>Dhc</i> , <i>Dehalobacter</i> and <i>Dehalogenimonas</i> 16S rRNA gene copy numbers.....	83
Figure 3.6 LC-MRM-MS <i>Dhc</i> biomarker identification in a tryptic digest of groundwater samples M17 and 33NA4.....	85
Figure 4.1 Example showing the process to select the minimum number of unique tryptic peptides and their associated number of protein seeds in different GH families	93
Figure 4.2 Minimum number of unique tryptic peptides and their associated number of protein seeds in different GH families.....	97

Figure 4.3 Examples of GH1 protein sequences analyzed with ProSite Scan	98
Figure 4.4 Peptide classification by taxonomy using sequence information from CAZyDB	100
Figure 4.5 Number of HQ & MHQ MAGs identified in the biogas microbiome project.....	103
Figure 4.6 Distribution of the sizes of the predicted proteomes from the HQ & MHQ quality MAGs from the biogas microbiome project.	104
Figure 4.7 CAZymes annotated in the proteomes of different phyla	105
Figure 4.8 Minimum number of unique tryptic peptides and their associated number of protein seeds in different GH families.....	109
Figure 4.9 Peptide classification by taxonomy using sequence information from the biogas microbiome MAGs	111
Figure 4.10 Functional classification of groups of proteins captured by numbers of unique peptides.....	113
Figure 5.1 Results from the LC-MS/MS analyses of 2µg of HeLa digest using different LC configurations	123
Figure 5.2 Results from the LC-MS/MS analyses of 2µg of HeLa digest using different LC configurations and MSMS spectra search algorithms.....	125
Figure 5.3 Search depth achieved by CharmerT in the spectral data collected through specialized LC modes of operation	127
Figure 5.4 Results from the spectral data collected with a range of isolation windows from the best LC configurations	129
Figure 6.1 Computational and experimental strategies followed for the identification and quantification of small endogenous polypeptides.	141
Figure 6.2 Performance of 10 kDa membrane filtering and PEAKS software for peptide identification.....	143
Figure 6.3 Proteomics metrics of the <i>Populus X canescens</i> and <i>Laccaria bicolor</i> PMI system	145
Figure 6.4 Length and basic residue distribution of all identified PCPs (unique & shared) across datasets	146
Figure 6.5 Counts of the top 10 most represented PTMs in PCPs identified in different sample types	147
Figure 6.6 Visual representations of detected PCPs (highlighted in grey and in red letters) produced as results of known and hypothesized protein maturation mechanisms in <i>Populus</i> and <i>L. bicolor</i> samples	149

Figure 6.7 <i>L. bicolor</i> proteins from significant upregulated peptides associated with ECM	151
Figure 6.8 <i>Populus X canescens</i> proteins from significant upregulated PCPs associated with ECM	154

CHAPTER 1

Introduction to mass spectrometry-based proteomics and overview of research topics covered in this dissertation

1.1 The advent of MS-based proteomics in the biological sciences.

1.1.1 Functional genomics and the wholistic view of biology.

The development of large-scale DNA sequencing technologies of the 90's provided an unprecedented molecular understanding of biological systems at an incredible pace.¹ In 1995 the complete cellular genome of the bacterium *Haemophilus influenzae* was obtained,² and in a span of 8 years, the first draft of the human genome was essentially finished.³ Even model organisms such as *Saccharomyces cerevisiae* turned out to be more complex than initially thought *i.e.*, from the 2000 genes previously characterized with traditional biochemical experiments to over 6000 found by genomics analyses.^{4, 5} By expanding the scope of biological investigations, gene sequencing quickly introduced an era of discovery-driven research, in which the components of the system under investigation are collected irrespective of any hypothesis of how they might work.^{5, 6}

As time kept passing by, however, the amount of information provided by genome sequencing with its fast discovering rates, surpassed the time it took to functionally characterize new genes, and databases quickly started to fill with sequences lacking functional annotations. Thus, cell and molecular biology research needed of new experimental approaches that could keep up with the enormous and rapid amounts of information provided by genomics, while at the same time providing of creative ways of how to handle, analyze, store and share biological data. The responses that came out of these efforts led to the generation of several new “omics” research fields, mainly transcriptomics and proteomics.⁵⁻⁷

Transcriptomics studies have been considered as a starting point to survey gene expression as the most direct product of a gene is an RNA transcript. The comprehensive interrogation of the identity and quantity for the complete set of transcripts in a cell directly reflects the activation of functional elements in the genome in a certain developmental state or physiological condition. Even when the genome is static in nature,⁸ the activity of genes is dynamic, and in every organism, subsets of genes get transcribed for subsequent functional roles.

Amongst the two most employed approaches available for transcriptomics profiling, are hybridization (e.g. microarray) and sequencing (e.g. RNA-seq), the former being high-throughput

and of generally low-cost making it widely popular in studies of various biological systems.^{9, 10} However, next-generation sequencing has made it feasible to achieve deeper sequencing of the transcriptome with high resolution and dynamic range, alongside independence from existing genome information, and clear determination of transcription boundaries.¹¹

Proteomics developments started in the late 20th century with the introduction of matrix-assisted laser desorption/ionizations (MALDI) and electrospray ionization (ESI) by Koichi Tanaka and John B. Fenn, respectively.^{12, 13} These two “soft” ionization techniques were the main drivers that allowed the use of mass spectrometry (MS) to analyze proteins and then peptides due to their ability of ionizing large and labile molecules. Importantly, the introduction of ESI enabled the interfacing of liquid chromatography (LC) to MS for protein and peptide characterization. The most notable feat was that by the group of Donald Hunt at the University of Virginia, in which using LC-MS, 19 major-histocompatibility-complex (MHC) class I-bound peptides were partially sequenced by the combined use of LC-MS,¹⁴ an achievement that over the years would lead to a decrease in the use of Edman degradations,¹⁵ which are based on the stepwise reaction of the N-terminal amino acids in a protein or peptide (up to 50 amino acids long) with phenylisothiocyanate and cyclic cleavage of the labeled amino acids without disrupting other peptide bonds.

With the ability to detect and sequence femtomolar amounts of peptides via LC-MS/MS, the challenge was now to assemble them back to their original proteins. This task was achieved by some research groups around the world^{16, 17} which introduced the use of MS/MS spectra-to-protein search algorithms, a bioinformatics success that until now remains at the core of bottom-up proteomics approaches. Continual improvements in protein extraction and enzymatic digestions, peptide separations via LC instrumentation and mass spectrometers providing higher scan speeds, mass resolution and accuracy; the field of proteomics has opened an era of comprehensive characterization for protein sequence identification and quantification in which each protein is not studied individually but under the context of interconnected systems. Successful applications of LC-MS for proteomics analysis have provided protein-level functional view for metabolic activities in various biological systems, from single bacteria isolates, higher order eukaryotes, to complex microbial communities.¹⁸⁻²⁰

1.1.2 Proteomics gives meaning to genomics and transcriptomics information.

Two polymeric strands of complementary deoxyribonucleotides form the DNA double helix. The genetic information carried by the linear sequence of deoxyribonucleotide subunits is further transferred to an intermediary RNA molecule that is then translated to a protein consisting

of a sequence of amino acids. This flow of genetic information from DNA to proteins (DNA→RNA→Protein) is the well-known central dogma of biology.^{5, 21} However, with the advent of OMICS technologies, we have now realized that there are exceptions to the dogma. More complex organisms are known to possess smaller numbers of genes, yet they express twice or thrice more proteins than less complex organisms. For example, in higher eukaryotes more than 50% of genes code for more than one protein, whereas in yeast and filamentous fungi, ~90% of genes encode one protein per gene.⁸

Even when the genome content in an organism does not change, its products do. The differential expression of genes is required to form distinct transcripts and proteins that control distinct functions. Thus, in comparison to the genome, the transcriptome and proteome are considered dynamic entities, because they can differ from cell to cell and even within the same cell type, at different stages of activity or development.^{22, 23} The dynamic nature of the transcriptome and proteome provides valuable information when studying functional relationships between genes.²⁴

Although mRNA quantification has been used as a proxy for measuring changes of protein abundance and has informed about many cellular responses, proteins provide a closer look to the functional activities of cells.^{25, 26} In contrast to mRNA, proteins are the effectors and regulators of essentially all cellular processes in an organism.²⁷ Across species, proteins constitute about 50% of the dry mass of a cell and reach a remarkable total concentration of 2–4 million proteins per cubic micrometer or 100–300 mg per ml.²⁶ The extensive proteome network of the cell adapts dynamically to genetic perturbations and thereby defines the functional state and determines the phenotypes of a cell.

Interestingly, with the advances in OMICS technologies, the general consensus nowadays in cell biology research is that mRNA levels cannot be consistently relied upon to predict protein abundance.⁴ Amongst the reasons are the complexity of post-transcriptional regulation processes (*i.e.*, riboswitches) as well as the multitude of post-translational modifications (PTM) in a protein such as phosphorylation, acetylation, and glycosylation that act as signals in information processing or as marks that mediate protein associations.^{4, 28, 29} Although one can attribute the discrepancy between mRNA and protein levels to measurement variability and noise in data, some studies have shown that this correlation reflects underlying biological mechanisms.^{30, 31}

Examples of the correlation between transcriptomics and proteomics data have been registered across various domains of life. For example, studies in different developmental stages of the human blood fluke *Schistosoma japonicum* concluded that around 40%–60% of the proteins detected in *S. japonicum* were consistent with its transcripts.³² In yeast, it has been found that for

some genes, while the mRNA levels were of the same value, protein levels varied by more than 20-fold.³⁰ Conversely, for some other genes, the transcript levels varied by 30-fold, while proteins level remained invariant. In higher eukaryotes, such in the livers of mice treated for different periods of time with three different peroxisome proliferative activated receptor agonist, the results shown that the differential expression of mRNAs explained at most 40% of the differential expression of proteins.³¹ Even in the small genome containing organohalide respiring bacteria *Dehalococcoides mccartyi*, poor correlation of mRNA- and protein abundance for four tested genes was explained by varying ribosome content and cell decay.³³

These examples highlight the importance of proteomics data and how its integration with transcriptomics can provide better or new insights into the regulation of cellular processes at a genome-wide scale. Whereas a strong correlation between both transcriptomic and proteomic data can serve as confirmation for an induced response to a treatment, the lack of strong correlation can also help detect experimental errors or hint to the possibility of a biological uncoupling between mRNA and protein expression.^{34, 35} Studying the levels of proteins in an organism offers insights to analyze translational status or post-transcriptional regulations and even by itself, it can also be applied to observe effects that may not be apparent by considering mRNA levels exclusively. MS-based proteomics can reveal the protein content of subcellular structure and organelles, provides an alternative to discern higher-order structural features of protein complexes, and contributes to functional evidence that is helpful in the characterization of metabolic pathways.²⁸

1.1.3 The practical implications of proteomics in biological research.

Before the advent of system-wide technologies, individual genes and their products, or small clusters of related components in specific biochemical pathways, were studied by techniques including, for example, site-directed mutagenesis, Western blotting, *in vitro* protein assays, and *in vivo* cell imaging.³⁶ The contributions of these now considered “classical” experimental techniques has been invaluable to assess the presence, sequence, function, and biochemical mode of action of one or a few specific proteins;⁴ however, the data they provide open a small window into the mechanisms governing the organization of the cell and highlight only a small part of a large interconnected network of functionally and physically interacting proteins.

Instead, technological and experimental advances in mass spectrometry-based proteomics, comprising mainly of comprehensive protein sequence databases (derived from

genomics and transcriptomics); sensitive and accurate MS instrumentation; an emerging collection of software that can match MS spectra to protein databases; and protein separation technologies like liquid-chromatography,^{4, 28} have allowed this technique to be now at the center of systems biology studies, in which, not only one or a few distinct proteins are analyzed each time, but the interplay of multiple ones and their roles as members of larger systems or networks are studied.³⁷

The impact that modern proteomics methods have had in our understanding of proteins under the systems biology context can be summarized into four principal applications that are 1) mining, 2) protein-expression profiling, 3) protein-network mapping, and 4) mapping of protein modifications.⁴

In mining, proteomics has been used to make an inventory of the content of all or all the possible proteins in a sample.^{37, 38} Although it can be considered a “brute-force” approach, the advantage of complete and high quality protein databases with computational algorithms able to match spectral data to peptides and proteins with adequate control of false positive hits, makes proteomics mining a feasible strategy to confirm what could only be inferred from gene-expression data or generate comprehensive proteome maps that can later be used as references for more in-depth investigations. Early examples of mining involve large-scale proteome maps of microorganisms such as yeast³⁹ or the bacterium *Deinococcus radiodurans* with products from more than 60% of the genes identified.⁴⁰

The most common and versatile use of large-scale MS-based proteomics has been to document the expression of proteins as a function of a cell or tissue state.^{4, 41} In protein-expression profiling studies relative or absolute quantitative values of the observed proteins at different states (*i.e.*, developmental, disease, physical stimulus, etc.), are measured to infer, for example, statistically significant changes of protein A to protein A in condition A to B, and attribute the change in expression to that particular condition(s). Expression profiling studies have found an appeal for biomedical applications as a means of detecting potential targets for drug therapy in disease, where normal and diseased cells or tissues are compared to determine which proteins are expressed differently in one state compared to the other.^{4, 24, 42-44}

Proteins rarely function alone, instead, they depend on the association of various components into macromolecular complexes. The interactions between proteins determine the functions of signal-transduction cascades and biosynthetic or degradation pathways.⁴ One of the most commonly employed experimental techniques coupled to MS detection is affinity-capture, where taking advantage of relative abundance levels of “prey” proteins and the endogenously expressed “bait” ones, complete interactome maps in cells can be drafted.^{8, 37, 45} For example, a

study with 11 human cell types, determined the relative changes of the protein components of 182 complexes under five different temporal states.⁴⁶ Other more specialized and elegant ways of studying complexes involve native mass spectrometry or MS based identification of crosslinked residues (XL-MS),^{37, 47} which has been used in combination with cryo-electron microscopy to identify substrate binding sites in molecular chaperones or to provide additional data to complement molecular dynamics simulations.⁴⁸⁻⁵⁰

Finally, the dynamic nature of the proteome in any organism is due in part to the variety of post-translational modifications (PTMs) that act to promote or repress the targeting, structure, function, and turnover of proteins.^{4, 41} Although modified proteins can be detected with antibodies, one question that is not always answered in such experiments is the specific location site of the particular modification.⁴ MS based proteomics is ideally suited to study PTMs because changes in the labeling of amino acids are reflected by shifts in mass that can be located with the resolution of a single amino acid through peptide-fragment ion spectra.³⁷ Regularly, mapping of protein modifications via mass spectrometry requires of specialized PTM enrichment strategies to overcome the expected natural abundance variance of modified peptides and/or proteins in a cells or tissue.^{51, 52} The most frequently studied types of PTMs via MS/MS are phosphorylation, ubiquitylation, the addition of ubiquitin-like proteins, glycosylation, methylation, acetylation and other types of acylation.³⁷ The study of PTMs by MS has been particularly gaining traction in the last years, as new roles for “well-studied” PTMs are being found in common organisms. For example, the lysine acetylation and succinylation in rice (*Oryza sativa*) has recently been implicated in seed development.⁵³

1.2 Overview of MS-based proteomics for biological research.

1.2.1 Mass spectrometry bottom-up proteomics as the powerhouse for large-scale/high-throughput analysis of proteomes.

Mass spectrometry (MS) based proteomics research is usually conducted in two ways: “bottom-up” and “top-down”. Top-down approaches involve the MS analysis of intact proteins⁵⁴ and their proteoforms, encompassing all forms of genetic variation, alternative splicing of RNA, and PTMs.⁵⁵ Although in principle top-down approaches seem like the most straightforward way for the detection and quantification of multiple proteins in a sample, analytical and technical challenges such as the difficulty of solubilizing and ionizing entire proteins; the separation of protein mixtures prior to MS analysis; the low sensitivity of measurements;⁵⁶ and the complexity

of matching intact and fragmented proteoform MS data to a protein sequence,⁵⁷ have limited their application to the broad study of the proteome.

Instead, bottom-up proteomics approaches, which are based on the mass spectrometric measurement of peptides generated by the enzymatic digestion of proteins,³⁷ are nowadays the standard for large-scale and high-throughput analyses of complex samples. Although in principle the digestion of proteins to peptides increases the complexity of the samples under study; the mass ranges and predictive tandem MS (MS^2 or MS/MS) fragmentation patterns of peptides, compared to those of large intact proteins,²⁸ makes them more compatible with MS instrumentation capabilities.

Until today, hundreds of bottom-up proteomics studies across all domains of life have been published with the overarching goal of having a better understanding of the biological processes occurring in cells and tissues.^{18, 58-60} Depending on the scope of the study, proteomics analyses include not only the identification of proteins, but also the nature and position of any posttranslational modifications (PTMs); the interpretation of dynamic quantitative changes between conditions; and/or the study of protein conformations or interactions within larger protein complexes or in the context of biological networks or pathways.

Generally, there are three approaches of how to achieve the aforementioned research goals with bottom-up proteomics and these are:^{28, 37} (1) Global (also known as shotgun or discovery) proteomics by means of data-dependent acquisition (DDA), aimed at achieving unbiased and complete coverage of the proteome; (2) targeted proteomics using multiple reaction monitoring (MRM) or parallel reaction monitoring (PRM), usually performed after discovery proteomics experiments and that is aimed at the reproducible, sensitive and high-throughput identification and quantification of a subset of known peptides of interest; and (3) multiplexed fragmentation of all peptides that elute from liquid chromatography by data-independent acquisition (DIA), which is nowadays seeing an increase of use in biological applications and is aimed at generating comprehensive peptide fragment-ion maps for a sample. In this dissertation, global proteomics by DDA and targeted proteomics using MRM were employed in the studies presented in Chapters 3-5.

1.2.2 Global proteomics by data-dependent acquisition-mass spectrometry (DDA-MS).

In global proteomics studies, peptide and protein identifications are maximized to give a comprehensive catalogue of analytes in unknown samples or explaining the biological

process(es) under study. A complete and reliable MS based proteome map is also a prerequisite for the development of targeted mass spectrometry techniques, as well as for DIA strategies.

After enzymatic digestion of proteins to peptides, the mass spectrometer is used to identify the sequence of all peptide ions (charged peptides) through the generation of a diagnostic MS/MS peptide fragment ion spectrum.²⁴ The identified peptide sequences must then be assigned to the proteins they originate from, through a nontrivial process called protein inference which is most commonly performed with bioinformatic software that match peptides to proteins in a database (*i.e.*, a FASTA file with protein sequences).⁶¹

Traditionally, global or discovery proteomics are based on a spectra collection strategy known as data dependent acquisition (DDA)²⁸ (**Figure 1.1.A**). According to this approach, mass spectra of all the peptide ion species that coelute at a specific time from chromatography (precursor-ion spectra) are recorded at the MS1 (or full scan) level.⁶² Depending on the intensities of the recorded precursor ions, the instrument then determines groups of precursors that are isolated and then fragmented to generate fragment ion-spectra or MS2 level scans.⁶³ Because of its dependency on the intensity of the peptide precursor ions detected at the MS1 level, global proteomics by DDA is subject to preferentially fragment high-abundant precursors while low abundant ones may not be sampled.⁶³ However, one way to increase the sensitivity of these analyses is by limiting the redundant fragmentation of abundant precursor, so that ions selected for fragmentation are temporarily excluded in the following scans over a determined time. This instrumental technique, known as dynamic exclusion,⁶⁴ improves the number of MS/MS spectra collected and, consequently, the number of identified peptides and proteome coverage.⁶⁵ However, even with dynamic exclusion, low intensity ions coeluting with high intensity peptides may still not be identified making it one of the main limitations of global proteomics by DDA.

1.2.3 Targeted proteomics by multiple reaction monitoring-mass spectrometry (MRM-MS).

Complementary to global proteomics approaches, in targeted proteomics experiments, the goal is not to maximize peptide and protein detection, but to monitor and quantify, with sensitivity and throughput, selected proteins of interest.⁶⁶ Due to its focused and sensitive nature, targeted proteomics is the gold-standard in biomarker development studies by proteomics and it has lately been regarded as a technique that may replace Western-Blot or antibody-based assays in the near future.⁶⁷

Using pre-existing information, such as that derived from global proteomics, characteristic peptides ions from proteins of interest are selectively isolated and then fragmented over their chromatographic elution time. Most of the targeted proteomics assays are performed in triple quadrupole mass spectrometers (QQQ) via a spectral data collection method known as multiple reaction monitoring-mass spectrometry (MRM-MS).⁶⁸ In MRM-MS, distinct pairs of precursor ion masses and cognate fragmentation masses, derived from existing datasets or large-scale spectral libraries, are used for sensitive and specific determination of these pairs or transitions (**Figure 1.1 B**).⁶⁹

Although QQQs have lower resolution and mass accuracy compared to the high-end mass spectrometers used in traditional global proteomics experiments; QQQ instrumentation offers the ability to detect peptides (and metabolites) that can vary in their abundance ranges by up to three to four orders of magnitude in linear response, as well as high speed of spectra acquisition which leads to good ion statistics and sensitivity.^{69, 70} Due to these characteristics, large number of samples may be interrogated for quantitative abundance values of a predetermined number of peptides without the need for time-consuming global proteomics experiments. Alongside the robust and economical triple quadrupole instruments, high-resolution instruments such as quadrupole orbitraps are used increasingly for targeted analysis, a variant known as parallel reaction monitoring because it utilizes the entire MS2 spectrum as oppose to certain fragment ions from a precursor.⁷¹

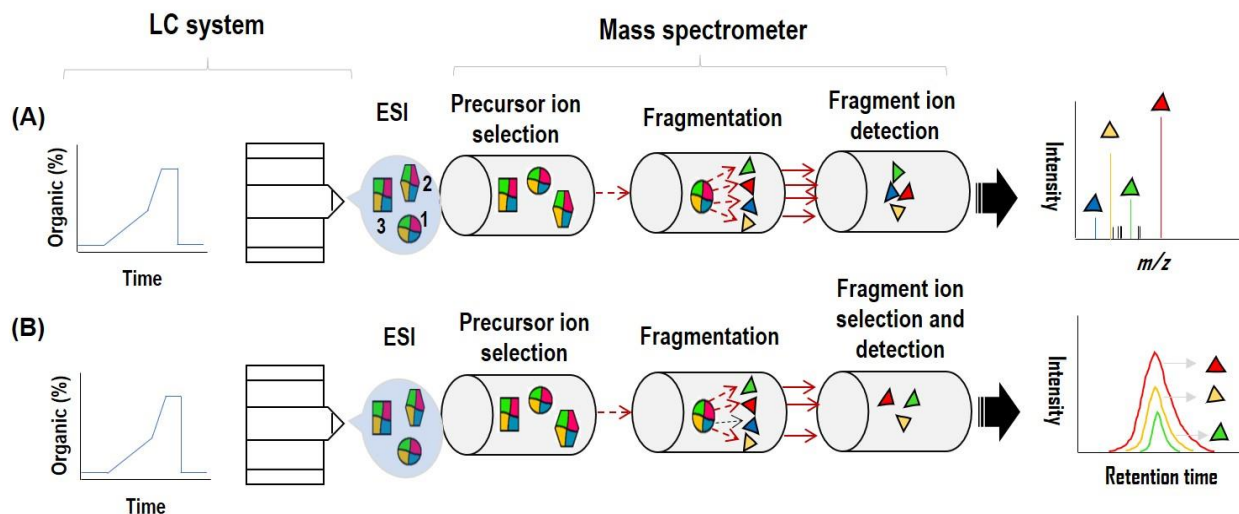


Figure 1.1 Types of tandem mass spectrometry acquisition methods used in this dissertation.

For clarity of presentation, three peptides are shown in each case (square, circle and hexagon shapes). **(A)** In data-dependent acquisition (DDA), peptides are selected for fragmentation based on intensity thresholds. Common setups include the analysis of the top 5, 10 or 15 most abundant peptides eluting at a given time. In this example, the peptide marked with “1”, as representation of abundance, is selected in the first mass analyzer of an instrument (MS1), fragmented in a collision cell, and the resulting fragments analyzed by a second analyzer (MS2). The process is then repeated for different precursors (*i.e.*, second and third place in abundance). MS/MS spectra recordings collected in DDA are better visualized using the mass-intensity dimension. **(B)** In multiple reaction monitoring (MRM) mode, the instrument cycles through a series of predefined sets of precursor-fragment ion pairs (also known as transitions). In this example, a precursor ion is selected in the first mass analyzer, fragmented in the collision cell, but instead of measuring every possible fragment ion, just three fragments are monitored. MRM signals are better visualized as a series or overlapping peaks from the measurement of each transition in the time-intensity dimension.

1.2.4 Relative label free and absolute quantification of proteins in bottom-up proteomics.

Quantification is at the heart of every bottom-up proteomics study today. However, peptide and protein quantitation by LC-MS/MS is not a trivial task. Although it would seem intuitively to quantify peptides in a spectrum based on their MS-derived spectral peak intensities, these do not often correlate with their relative abundance within a given analyzed sample.⁷² This scenario is even true for peptides originating from the same protein after enzymatic digestion and that in theory are present in equimolar amounts. Several factors impact the way how peptides ionize before entering the MS instrument and thus, their signal intensities. Amongst these factors are ion suppression events that can be a consequence of the coelution of components from the solvent and sample during chromatography;⁷³ or the varying physicochemical properties of each peptide that make them amenable to ionization (*i.e.*, hydrophobic peptides ionize more efficiently than hydrophilic peptides).⁷⁴

Therefore, to achieve an adequate level of quantitative accuracy, each individual peptide and/or protein are compared between different samples or experimental conditions.⁷⁵ Hence, several MS based strategies have been developed to achieve this form of relative quantitation. These strategies can be distinguished by the goal and scope of each experiment. Broadly, relative quantification strategies in proteomics are classified into two groups: Those using stable isotope labeling and those that are label-free (LFQ), the latter, being the focus of this dissertation.

As the name implies, relative LFQ techniques do not employ any stable isotopes to label different protein/peptide samples. In comparison to label-based approaches, LFQ approaches are easier to implement, cheaper, and offer the ability to quantify a larger number of proteins with a wider dynamic range, albeit, at the cost of sacrificing some level of quantitation accuracy.^{65, 76} The two most commonly used LFQ techniques are based on spectral-counting (also known as ion-counting) and peak-intensity principles.⁷⁷

In the spectral counting (SC) method, the number of MS/MS spectra collected for a given protein across multiple LC-MS/MS runs provides the values necessary for relative quantification.⁷⁸ This method is based on the empirical observation that the more abundant proteins are in a sample, they yield a larger number of proteolytic peptides which are more likely to be sampled and consequently, provide higher numbers of spectra.⁷⁸ Relative quantification of proteins via SC is then achieved by comparing the number of spectra between a set of runs. Although this approach seems intuitive and attractive in practical terms, SC is nowadays considered more as a “semi-quantitative”⁷⁹ approach because in reality, the spectrum count of each peptide in a run varies according to their chromatographic elution profile (*i.e.*, peak width and retention time);⁷⁵ in

addition, it has been reported that the method is not sensitive enough to detect subtle fold changes and that it tends to overestimate low-abundance peptides.^{80, 81}

In the case of relative quantification by peak intensity, peptide chromatographic peak areas or heights in different samples are obtained from extracted ion chromatograms (XICs) and these values are used for quantification. Initially, peaks need to be filtered from background noise and neighboring peaks and then isotopic patterns assigned by spectra deconvolution.⁷² Subsequently, retention times are adjusted to correctly match corresponding m/z peaks of different LC-MS/MS runs. Finally, the normalized peak area or height enables accurate matching and quantification values which can then be treated by statistical tests to measure significant abundance changes. XIC-based methods are consistently more sensitive and accurate than SC when using high-resolution mass spectrometers because these instruments can discriminate between coeluting peptides of similar mass while allowing the accurate mapping of XIC to a peptide.⁷⁵ XIC-based quantitation is also capable of accurately discerning fold changes as low as 1.1 compared to the SC limit of ~ 1.4 .^{82, 83}

Besides the relative quantification of proteins, sometimes it is necessary to quantify low-abundance proteins in a sample or determine absolute amounts of proteins in complex mixtures (*i.e.*, molar amounts). These types of quantitation are frequently applied in a more targeted context, rather than proteome wide and using specialized forms of MS, such as multiple reaction monitoring (MRM) via tandem MS.⁶⁸ Absolute quantification relies on the addition of known concentration of isotopically labeled peptide standards to a sample. These standards are identical to the endogenous peptides in the sample but are distinguished by a shift in mass as they have an incorporated stable isotope in their sequences (*i.e.*, ^{13}C , ^{15}N).⁸⁴ Stable isotope-labelled standards and endogenous peptides coelute during chromatography and absolute quantification is achieved by comparison of the peak area abundances of the internal standard peptide with the corresponding native counterpart.^{70, 84}

Absolute quantification via isotopically labeled standards, also known as AQUA,⁷⁰ expands the quantitative capabilities of MS proteomics beyond relative comparisons, to the calculation of molar amounts of specific proteins in complex mixtures allowing one to even determine the stoichiometries of individual members in protein complexes. AQUA peptides are the gold-standard for absolute quantification via bottom-up proteomics and have found a niche in research dealing with protein biomarker development, where accurate quantification is critical to establish limits of detection and quantitation.

1.3 Bottom-up proteomics in environmental microbiology applications.

1.3.1 The role of bottom-up proteomics in microbial environmental bioremediation.

Anthropogenic activities cause the release of as many as 100,000 compounds in the environment including petroleum derived and polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), pesticides, therapeutic drugs, transition metals, nitroaromatics, quinones and other electrophiles.⁸⁵ These molecules have hazardous health effects not only in humans but in the flora and fauna of ecosystems, and at the same time, can be effectors behind the disequilibrium of microbial diversity that is essential in the regulation of global nutrient cycles.⁸⁶

Although ways to combat the negative impact of contaminants in the environment involve the use of chemical cleanup methods, more recent technologies exploit the knowledge of the metabolic potential of living organisms in removing natural and xenobiotic pollutants. These bioremediation techniques have the broad goal of transforming or degrading contaminants into nonhazardous or less hazardous chemicals and act as an ecofriendly and cost-effective alternative to more traditional cleanup methods.^{87, 88}

Importantly, the metabolic activities of bacteria, as well as their dynamic behavior and ability to survive under extreme conditions, makes them the most important group of organisms that are investigated for their potential in bioremediation.⁸⁸ In fact, a large number of bacteria have been reported for the degradation of different industrial wastes such as dyes, hydrocarbons, chlorinated aromatics, pesticides and heavy metals.

The use of gene-centric approaches to interrogate bacterial communities in the environment has led to the discovery of unique bacteria with bioremediation potential that were not accessible by traditional microbiology techniques, and in certain cases it has facilitated their subsequent isolation and culture-dependent characterization; untargeted-nucleic acid measurements of bacterial communities can be benefited from global proteomics profiling,⁸⁹ because the catalysts responsible for contaminant transformation are proteins expressed by the bacteria. Furthermore, when trying to determine functionality in environmental bacterial samples, mRNA and protein levels are generally not strongly correlated;²⁹ and this case is especially true for bacteria living in perturbed systems.⁹⁰

Thus, the mere study of individual genes and their regulation is not enough to fully understand microbial adaptation strategies of bacteria living in polluted environments and post-genomic/transcriptomic analyses involving proteomics, are needed to investigate the physiology

of complex microbial consortia at a molecular level. The functional insights provided by global proteomics studies are useful to fill in the gap between the genetic potential and functional diversities in microbial communities. In terms of bioremediation, proteome analysis can be used to confirm the occurrence of specific metabolic activities of candidate bacteria with detoxification abilities, tolerance to extreme environments and respiratory versatilities. In addition, the detection of new catalytic enzymes in novel metabolic pathways, or of known proteins with marked changes in their relative abundances in the proteomes of microbes in response to contaminants, can assist in finding protein biomarkers or bioindicators that have the potential to monitor the dynamics and sustainability of environmental quality.⁸⁷

1.3.2 From the clinic to the river: Environmental monitoring of protein biomarkers with targeted mass spectrometry.

Over the last decade, the emerging field of environmental microbial proteomics has contributed to the determination of molecular mechanisms of microorganisms living in various natural and polluted soil types, aquatic bodies and other environmental matrices, such as organic wastes and mine spoils. The global profiling of proteins in microbial isolates or mixtures in the environment has revealed taxa contributing to the expression of proteins and functional pathways between enriched and control ecosystems.⁹¹ Comprehensive proteomics approaches have provided the identification of key proteins involved in molecular responses related to xenobiotic toxicity, homeostasis or contaminant degradation.^{87, 91-93} For example, proteomic analysis of toluene-grown *Pseudomonas putida* DOT-TIE revealed the participation of four stress proteins such as GroES, cold shock protein CspA2, translational elongation factor EF-Tu and a xenobiotic reductase (XenA) in the tolerance of the bacteria to the solvent.⁹⁴ Another case includes the citrate synthase (gltA) enzyme in *Geobacter* which abundance correlates with the activity of this bacteria during *in situ* bioremediation of soluble uranium U(VI) to insoluble U(IV).⁹⁵

Such information is not only useful to understand the catalytic roles of microbes, but is also helpful in determining sets of biomarker proteins which could be monitored in environmental samples to provide evidence of the exposure or effects of one or more chemical pollutants, or to determine the presence and contaminant degrading activity of an organism of interest.⁹⁶ Indeed, the identification of key proteins and catabolic enzymes from bacteria in the environment has been envisioned as a promising way, other to gene-based approaches, to show that specific bioprocesses are occurring in a sample and/or that a particular set of conditions that favor the

degradation of a contaminant or class of contaminants is present.⁹⁷ However, challenges in the detection of specific enzymes in complex environmental samples exist.

One of the biggest challenges is the detection of key proteins that can be of very low abundance in an environmental matrix compared to others. Global bottom-up proteomics techniques are generally biased towards finding peptides of the most abundant proteins in a sample. Thus, due to the wide dynamic range of proteins that can be present at any time in an environmental sample, biomarker proteins can be missed in global or discovery-based proteomics analyses, especially if the conditions for their expression are not optimal. Besides, other interfering substances that can be present in environmental matrices can compromise the quality of protein identification by mass spectrometry (*i.e.*, presence of humic acids⁹⁸). Ideally too, absolute quantitative information from the measurement of these proteins could be more valuable to environmental diagnoses comparing levels of biomarkers within a sample at different conditions or derived from different sampling sites than the relative abundance and largely descriptive information provided by global proteomics approaches.⁹²

For these purposes, targeted proteomics approaches via multiple-reaction monitoring or parallel-reaction monitoring (explained in section 1.2.3) provide technical advantages to the environmental monitoring of proteins of interest compared to global qualitative approaches. Not only are the latter techniques more sensitive,⁶⁸ thus increasing the chances of circumventing the issues of low protein abundance and recoveries that are characteristic from environmental samples; but in addition, they allow to test multiple samples in a fraction of the time that it takes to run and analyze samples by global proteomics which could be useful when monitoring multiple samples for environmental hazard assessment or bioremediation effectiveness. Targeted proteomics also provides an interesting advantage to other protein-based identification methods like western-blot in which multibiomarker tests targeting up to 100 proteins could be developed to monitor a wide range of biological responses without the need of developing specific antibodies for every target. Absolute quantitative information from proteins in the attomole-nanomole range can also be obtained from targeted proteomics techniques with the use of heavy-labeled peptide standards.⁷⁰

Although most of targeted proteomics research has been developed and conducted with microbial isolates^{99, 100} or clinical samples (*i.e.*, plasma),^{101, 102} the selectivity, dynamic range and reproducibility of targeted proteomics assays make them ideal for quantification of specific peptides and proteins in complex environmental samples.¹⁰³ Chapter 3 provides a glimpse of the combined uses of global proteomics to selected specific peptides from reductive dehalogenases expressed by different strains of the anaerobic bacteria *Dehalococcoides mccartyi* (*Dhc*) during

dechlorination of polycyclic aromatic hydrocarbons,^{104, 105} and for the first time, their identification in contaminated groundwater via targeted proteomics which is the first step towards the development of an absolute quantitative assay to measure these proteins in sites undergoing active *Dhc*-mediated bioremediation.

1.4 Isolation interference in bottom-up proteomics experiments.

1.4.1 The occurrence of chimeric spectra in DDA bottom-up proteomics experiments.

In discovery-based bottom-up proteomics experiments where MS spectra data is collected via data-dependent acquisition (DDA), peptides derived from the enzymatic digestion of proteins in a sample are separated with liquid chromatography (LC), ionized via electrospray and then detected by MS scans which report the mass-to-charge (m/z) values of the ionized molecules. The peptide precursor ions (or peptides with a charge state) that enter the mass spectrometer are further isolated based on their MS^1 scan intensities and then fragmented to generate MS^2 or MS/MS fragmentation spectra. The combination of both MS^1 and MS^2 information is then used to identify peptides and proteins in sequence databases.

Thus, although it can be said that the most decisive parameter that determines which peptides are observed in a DDA bottom-up proteomics experiment is their natural abundance in a sample, this is not always the case. Since thousands of peptides are separated with LC before MS detection, it is expected that more than one peptide ion could enter the mass spectrometer at any given time. The coelution of peptides in LC introduces an element of randomness in each run, as the appearance of a precursor ion spectrum at a given chromatographic retention time will never be completely identical between runs (including biological and technical replicates).

Besides chromatographic conditions (discussed in more detail in Chapter 2), another DDA setting that impacts the collection of MS/MS spectra is the precursor isolation width. The isolation width is defined as the m/z range that the MS detector uses to isolate the precursor ions. In simple terms, this works as a window that sits on top of the most abundant precursor monoisotopic peak and which results on its isolation along with some of its isotope peaks, but potentially of other coeluting peptides with a similar m/z as well. Setting the isolation width too narrow will result in loss of sensitivity, and in extreme cases cut out a portion of the ion packets, thus resulting in inaccurate measurements. In contrast, a wide isolation window may result in co-isolation and co-fragmentation of even more neighboring peptides resulting in unidentifiable or low scoring spectra.

In a typical DDA setup, isolation widths may range from 1 to 3 m/z , and it has been suggested that these values give the best tradeoff between sensitivity and selectivity. However, despite advances in LC-MS/MS instrumentation, and due to the complexity of the samples analyzed in bottom-up proteomics, the co-elution, co-isolation and co-fragmentation of two or more peptide ions with similar m/z at any given retention time still occurs. The MS/MS spectra that results from these cases are known as chimeric or mixture spectra (**Figure 1.2**). The more conservative estimates of chimeric spectra in proteomic DDA-based research are calculated as over 11.2% of all MS² spectra acquired,¹⁰⁶ but others have calculated this to be as high as 50% or more.¹⁰⁷⁻¹⁰⁹

Chimeric spectra in proteomics studies has negative effects that compromise peptide and protein identification as well as relative quantification. First, the identification of peptides based on their MS/MS spectra becomes more difficult because, for example, many database search engines use the number of unassigned fragment ion peaks (which increases for chimeric spectra) for scoring. Insufficient spectral quality due to low signal intensity and the presence of unexpected modifications also exacerbates this issue. In addition, chimeric spectra reduces the accuracy of quantitative isobaric tagging-based quantification methods such as isobaric Tags for Relative and Absolute Quantifications (iTRAQ) or Tandem Mass Tag (TMT) strategies in which the contribution of reporter ion intensities from co-fragmented peptides, causes the under-estimation of protein/peptide abundance differences (a phenomenon termed as “ratio compression”).¹¹⁰⁻¹¹²

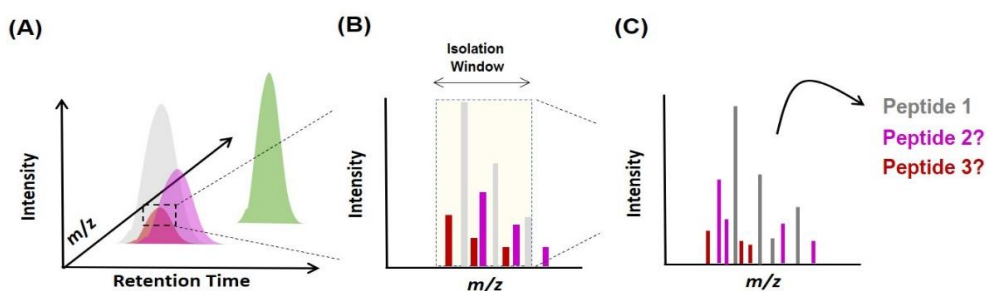


Figure 1.2 Chimeric spectra in LC-MS/MS bottom-up proteomics experiments.(A) A three-dimensional map of peptide precursor ions with signals in gray, pink, red, and green. The grey, red and pink precursor coelute at a given retention time and have close m/z values. (B) MS1 scans of the three coeluting precursors. By DDA, the grey precursor and its isotopes would be targeted for fragmentation, however, due to the presence of the other interfering precursor, more than one precursor ion and their isotope peaks within the isolation window are sent to fragment. (C) The MS/MS spectrum contains fragment ions from the three precursor ions. Fragment ions from the grey precursor have higher chances of being assigned to a peptide sequence compared to the fragments from the other contaminating precursors.

1.4.2 Exploiting chimeric spectra in database-based peptide searches to increase proteome depth.

The complexity of the peptide mixtures and the limitations of the current MS technology make chimeric spectra unavoidable in bottom-up proteomics experiments.¹⁰⁷ However, various approaches have been proposed to overcome their occurrence. These approaches can be categorized into experimental and computational. On the experimental side, specialized methods of acquiring data such as triggering MS/MS at or near chromatographic elution peak to minimize co-selection and fragmentation, or reduce the size of isolation windows, have been tested, however, at the expense of sensitivity. Another alternative is to improve the chromatographic resolution of peptides eluting into a mass spectrometer to minimize peptide coelution, such as using orthogonal chromatographic separations or separations based on ultra-high performance liquid chromatography (see section 2.3), with the former having a penalty of increased measured times and the latter of affording costly equipment.

Computationally, although chimeric spectra in bottom-up proteomics is known to reduce peptide identification rates of algorithms used in database searches, they also present the opportunity to identify more than one peptide from a MS/MS spectrum. In addition, software that can deconvolute chimeric spectra is more attractive in terms of sensitivity, cost and analysis time. The new approach used by these algorithms of finding “more than one peptide per MS/MS spectrum” challenges the “one peptide per MS/MS spectrum” dogma of conventional database-driven algorithms and they require further adaptation of computational workflows, with the biggest challenge being the scoring and validation of chimeric spectra.¹⁰⁸ The confident score of the most abundant peptide in a spectrum is not easily comparable to the score of a second coeluting peptide, for example. Even in regular database searches, some spectra may receive higher scores than others because they have more peaks or because their precursor mass results in more peptide candidates from database for consideration.^{113, 114} Therefore, it has been suggested that scoring functions that account for spectrum or peptide-specific effects can make the score more comparable and thus help assess the confidence of identifications across different spectra.¹⁰⁸

In theory, methods that consider more than one peptide per MS/MS spectrum can potentially double the sensitivity in recovering peptide information from chimeric MS/MS spectra. Although this goal has not been achieved, improvements in scoring functions as well as better management of computational resources, have enabled the development of several software options for chimeric spectral deconvolution that are able to identify more peptides than their single search counterparts. Early software like ProbiDTree, for example was able to increase peptide

identification rates between 20-30% compared to the popular search engine SEQUEST.¹¹⁵ Andromeda, another widespread used search engine in bottom-up proteomics, was able to increase peptide identifications by ~ 10% when enabling the identification of the second least abundant precursor ions in chimeric spectra to a dataset published originally without it.¹¹⁶ The DeMix algorithm demonstrated peptide identification efficiencies that were 1.5 times higher than that of Andromeda and 2.5 times higher than Mascot, albeit at reduced computational speeds.¹¹⁷ More recently, the search engine Amanda, which is optimized for high resolution and high mass accuracy data, has been modified to deconvolute chimeric spectra via an optimized algorithm that considers retention time for scoring and validation of spectra, showing impressive gains of up to 63% more unique peptides compared to a conventional search strategy.¹¹⁸

With all the available tools, what is then limiting their widespread use by the bottom-up proteomics community? A limitation is that they typically replace the search engine in the data analysis, potentially disrupting pipelines already established and relied upon in laboratories; while another one is to methodologically demonstrate the identification gains that these tools provide in a wide array of LC-MS/MS workflows, such as those considering less advanced HPLC peptide separations to more sophisticated ones, or in other words if it is worth exploring the idea of using a multiple-peptide matches-per-spectrum (mPSM) database search strategy. In this dissertation, the research presented in Chapter 4 tested the capabilities of the most recent iteration of the search engine Amanda across several LC setups and other factors influencing the acquisition of chimeric spectra in the analysis of HeLa cell tryptic digests and suggest that this can be a strategy to increase throughput and depth in bottom up experiments.

1.5 LC-MS/MS based peptidomics.

1.5.1 Small and mighty: Proteolytic cleavage products (PCPs) across different domains of life.

Amongst the group of low molecular weight organic analytes (defined arbitrarily by a MW cutoff <18 kDa) mediating different biological processes in organisms are peptides or proteolytic cleavage products (PCPs) of less than 100 amino acids in length. These molecules are the products of either native protein turn-over or protein maturation events (**Figure 1.3**). PCPs in the latter group can be further classified as functional, bioactive peptides or non-functional remnants of the protein maturation process.

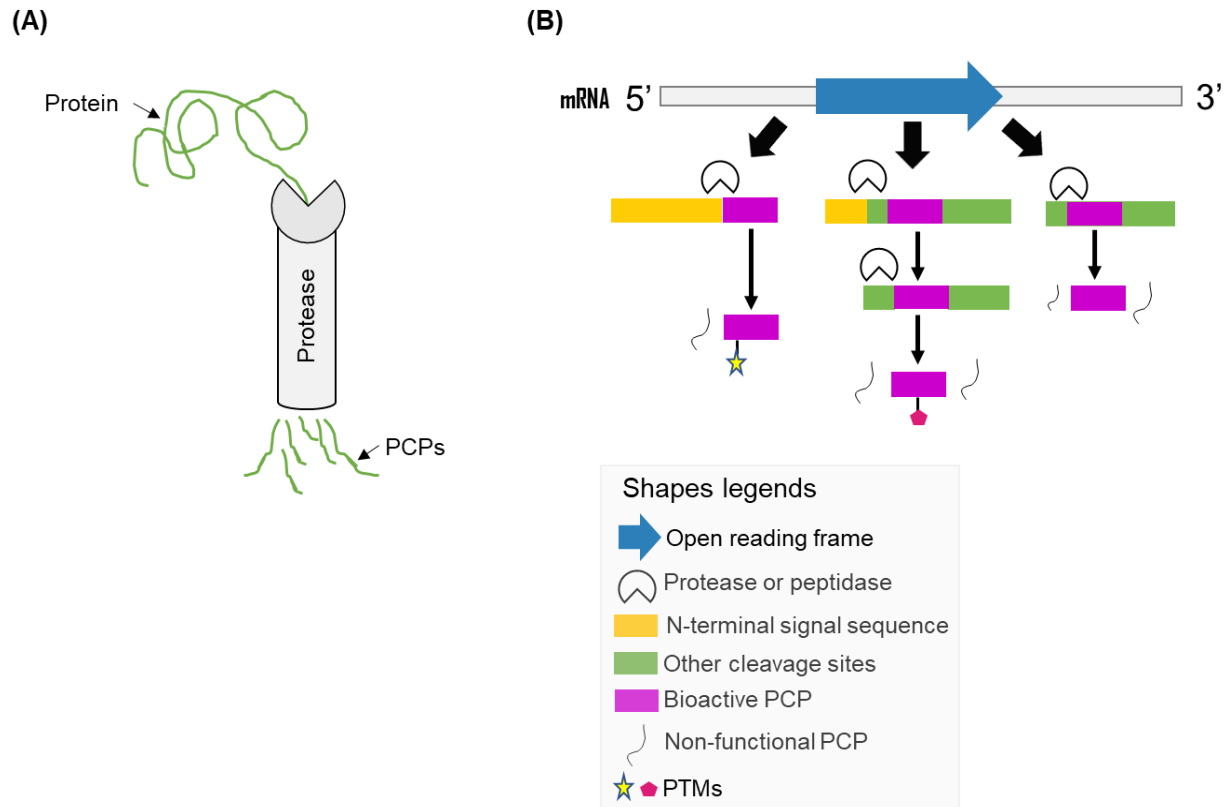


Figure 1.3 Biological origins of PCPs. (A) PCPs can be the products of protein degradation/turnover mechanisms. (B) Bioactive and non-functional PCPs can be produced from the maturation of a protein precursor, for example, a precursor with a N-terminal signal sequence; a precursor protein with NSS and other domains, or from a protein precursor without a NSS. The mechanisms of how these PCPs are produced may or may not be understood. Bioactive PCPs in different organisms have structural characteristics such as post-translational modifications (PTMs).

Particularly, bioactive PCPs in prokaryotes and eukaryotes have been linked to cell development and division,¹¹⁹⁻¹²¹ cell-to-cell communication¹²² and environmental response.¹²³ Intriguingly, bioactive PCPs have also become the targets of bioanalyses for biomedical and agricultural applications.¹²⁴⁻¹²⁷

Protein turnover is an indispensable process for cell maintenance and regulation that is coordinated by the action of proteases and peptidases. For example, the mammalian proteasome cleaves ~70% of intracellular proteins and is recognized as a dynamic complex that modulates cellular function in health and disease.¹²⁸ In the same manner, for plants, different proteases and peptidases participate in events like protein remobilization on leaf senescence,^{129, 130} the breakdown of storage proteins during seed germination,^{131, 132} and the recycling of misfolded and damaged proteins.¹³³ Thus, PCPs derived from protein turnover events contribute to the total pool of native peptides in a cell. Although the products of protein turnover can act as sources of nutrient recycling in cells (*i.e.*, in plants) many of their functions remain to be characterized.

Bioactive PCPs instead are the products of protein maturation and their functions have been studied with more detail. These types of PCPs can be synthesized by the removal of a N-terminal signal sequence (NSS) in a nonfunctional precursor or, alternatively, a precursor can be enzymatically modified in other regions to yield a mature form.^{134, 135} Most bioactive PCPs in cells are secreted to the extracellular space where they have regulatory functions mostly through the interactions with specific receptors. Notable examples of bioactive PCPs from protein precursors include plant antimicrobial peptides (AMPs)^{133, 135} and neuropeptides in mammals which signal between neurons and influence a wide range of physiological processes.^{134, 136} Bioactive PCPs differ vastly in amino acid composition, length, 3D structure, and function.

Discovery and characterization of PCPs has benefited from the application of MS techniques. These techniques are part of a specific subdiscipline of proteomics termed peptidomics. Peptidomics is defined as the comprehensive characterization of the total peptide content in a biological system using MS.^{134, 137} Typically, peptidomics studies are focused on bioactive molecules but products from the transient degradation of proteins are also captured on them. Although both peptidomics and proteomics are related, there are fundamental differences between these fields. For example, experiments in the former field use enrichment strategies to optimize the identification of as many endogenous peptides as possible, and the absence of an enzymatic digestion step is also a common characteristic of them that helps to capture native proteolytic mechanisms. Although it may seem that peptidomics experiments are easier to conduct than proteomics ones there are several factors that make peptidomics a bit more challenging than proteomics, these are explained in the next section.

1.5.2 Challenges in the identification of PCPs by LC-MS/MS: Natural protein dynamic ranges, databases and PTMs.

Several optimization strategies to improve PCPs detection by MS-based global peptidomics have been deployed and each one covering different biological and methodological challenges that are consequences of their low molecular weights (4–10 kDa), low abundances and transient expressions, sequence diversity, as well as complications in data analysis.

Regarding sample preparation, the separation of PCPs both from small molecules (*i.e.*, metabolites, salts, detergents) as well as from larger proteins needs to be addressed before analysis by MS. For example, it is well known that in blood serum and plants tissues, where the dynamic range of protein concentrations varies by ~10 orders of magnitude,^{138, 139} the natural abundances of proteins like albumin and RuBisCO, respectively, can hinder the detection of low-abundance proteins and peptides.^{140, 141} Thus, experimental strategies that enrich for PCPs in different biological matrices have been successfully applied. Methods include 2D separations with gels,¹⁴² differential solubilization,¹³⁹ organic solvent precipitation^{138, 143} and the use of molecular weight cut-off filter membranes.^{144, 145} The parallel combination of several separation schemes has also been reported.⁹

Endogenous peptides range considerably in size and properties, which makes them technically challenging to work in mass spectrometry. Because of their size, PCPs may not contain a cleavage site for protease digestion, or they can yield short fragments which in consequence reduces the fidelity of identifications mapped to protein databases. Therefore, peptidomics experiments usually omit steps of enzymatic digestion which is also helpful in capturing native proteolytic mechanisms. Native peptides can also yield charged molecules with +5 and +10 charge states, which complicates the deconvolution of data by MS software.¹⁴⁶

Another challenge in peptidomics research is how to map the identified peptides to protein precursors. In proteomics there is no need to identify every possible peptide from a protein to positively identify it, and proteins with single identifications are usually discarded; but in peptidomics, it is possible to have identifications based on single peptides.¹⁴⁷ If doing regular database searches, that is, mapping of collected spectra to peptides and proteins in a database, peptidomic analyses need to allow for a range of cleavage sites and post-translational modifications (PTMs).¹⁴⁸ For example, several bioactive peptide families from plants contain PTMs on their Pro and Tyr residues,¹³⁵ while neuropeptides in mammals may present a range of modifications in Ser, Thr, Tyr or Asp residues.¹³⁴ Without the constraints of sites for trypsin or another protease in the data analysis, database searches can become a real challenge, and the

false positive rate of peptidomics identifications compared to proteomics has been estimated that it could be 100 times higher.¹⁴⁷

To alleviate these issues, researchers have employed samples known to be enriched in endogenous peptides and reduced databases with known families of bioactive peptides as well as their potential variants. However, this sort of “targeted” approach hinders the identification of novel or less characterized peptide species. Besides the use of specific protein databases for MS-based proteomics identification, technical improvements in *de novo* peptide sequencing algorithms, which derive peptide sequence information directly from the mass spectral peaks and not from protein databases, has now become an attractive alternative to broadly profile the presence of endogenous PCPs. This resource has also been coupled to database searches in a hybrid approach called *de novo*-assisted peptide database search.¹⁴⁹ Chapter 5 describes in detail LC-MS/MS peptidomics workflow that was applied for the comprehensive identification of PCPs in the symbiotic interactions between a species of the perennial plant *Populus* and the ectomycorrhizal basidiomycete *Laccaria bicolor*. Such workflow involved some of the experimental techniques and the *de novo*-assisted peptide database search strategy mentioned here.

1.6 Dissertation summary.

Peptide identification is at the core of bottom-up proteomics workflows. Although the identification of peptides is more amenable to the technological capabilities of liquid chromatography and mass spectrometry when compared to intact proteins, the complexity of the samples analyzed in bottom-up studies make it unavoidable for some peptide identification information to be lost. In this dissertation, different MS based proteomics methodologies and technological advances were explored to detect and rescue peptides of interest in complex biological matrices, akin to the analogy of finding peptide mass “needles” in a haystack of protein diversity.

Chapter 3 shows the first published demonstration of the feasibility of deploying a targeted bottom-up proteomics approach in environmental samples for the identification of specific peptides that can serve as surrogates to confirm the presence or absence of bacterial proteins with bioremediation potential. Chapter 4 evaluates the *in-silico* selection of minimum lists of unique peptides found amongst groups of functionally related glycoside-hydrolase proteins, and which could be used in a targeted proteomics experiment to determine the presence of these groups of proteins in anaerobic bioreactors, instead of individual proteins that are part of a group.

Chapter 5 evaluates the performance of a new software package to improve proteome depth and coverage by rescuing peptide identifications buried under chimeric spectra and suggests a new experimental strategy to employ in bottom-up global proteomics experiments. In Chapter 6 some of the recent experimental and bioinformatics approaches to detect and quantify endogenous produced peptides in organisms by MS/MS were evaluated. Specifically, the diversity of peptides expressed in a symbiotic plant-fungal root system was characterized and based on relative abundance values, analytes with potential roles in the establishment of the association between organisms were identified. Finally, Chapter 7 presents a perspective of the newest technological advances in the field of bottom-up MS proteomics and describes how these may be applicable to the studies presented here and the benefits they would provide.

Herein, broad explanations of the experimental methods including protein extraction and digestion protocols; LC-MS/MS instrumentation for global and targeted proteomics, as well as computational algorithms used to develop Chapters 3-5 are presented. Detailed Materials & Methods sections can be found in each chapter.

CHAPTER 2

Experimental methodologies, mass spectrometry instrumentation and bioinformatic approaches used for proteomics analyses in this dissertation

2.1 Overview of a bottom-up proteomics workflow.

In general, every bottom-up experiment consists of five major stages which are adapted each time according to the sample under analysis and these are (1) protein extraction, (2) protein digestion, (3) peptide analysis by nano-liquid chromatography (LC) and MS/MS detection, (4) database searches, and (5) downstream processing of data (**Figure 2.1**).^{24, 36, 37, 41}

Sample preparation for bottom-up proteomics experiments can be basically summarized into protein extraction and digestion. Cells are lysed using a combination of detergents like SDS, sonication or bead beating. After a round of centrifugation, proteins are recovered and then submitted to denaturation with a chaotrope, disulfide bridge blocking and protease digestion.³⁶ In general, trypsin digestion is preferred due to its ability of generating peptides that are on average ~ 14 amino acids in length and ~1.8 kDa in molecular weight and which are amenable to the subsequent electrospray ionization process.^{150, 151} In addition, after extracting proteins from cells or tissues, some protocols employ SDS-polyacrylamide gel electrophoresis (SDS-PAGE) fractionation to analyze fractions of interest.⁸ Gel bands containing the proteins of interest are excised and then digested *in situ* using trypsin or any other protease, and then extracted.

Before peptide mixtures are introduced into the mass spectrometer, they are separated by one or more steps of high-pressure liquid chromatography in very fine capillaries which are connected in line with the MS instrument. This step is crucial as it improves the detection capabilities of the instrument by de-complexing the samples under analysis. As a hypothetical example, it has been described that any typical human cell may be able to yield ~ 6 million tryptic peptides.⁴ In addition, the small internal diameter (~3-5 μ M) of the LC nanocolumns ensure that the best sensitivity and efficiency of separation are achieved.^{36, 152} As peptide mixtures become more complex, orthogonal chromatographic separations that exploit different physicochemical characteristics of peptides such as hydrophobicity (*i.e.*, reverse phase C18) and charge (strong cation-exchange, SCX) can improve the resolution of the separations. Multiple solvent gradient strategies are used to elute the peptides from the chromatographic columns into the MS instrument, but the most common ones employ chromatographic gradients that change the percentage of organic solvent (*i.e.*, acetonitrile) from low to high.

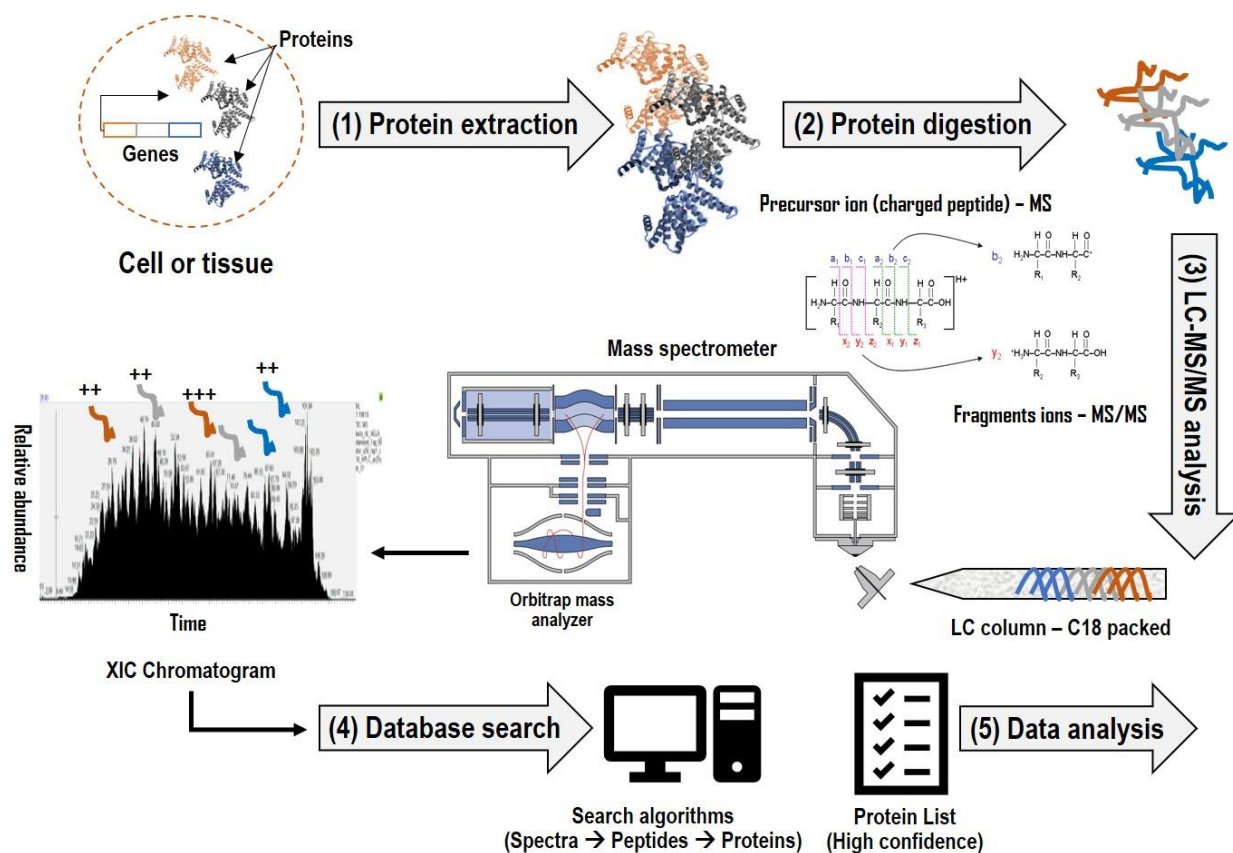


Figure 2.1 Overview of a bottom-up proteomics experimental pipeline. A typical bottom up MS-based proteomics experiment consists of five stages. First, proteins are extracted from the biological system under investigation (*i.e.*, cell lysate or tissues). Proteins are then enzymatically digested to peptides, as these are more amenable to the operational capabilities of a mass spectrometer. Peptide mixtures are then separated by one or more steps of liquid chromatography and eluted into an electrospray ion source where they are nebulized in small, highly charged droplets. Peptides enter the mass spectrometer as protonated molecules where they are analyzed using different modes of spectra data acquisition. The MS and MS/MS spectra are stored and matched against protein sequence databases using computational search algorithms. Finally, post-processing of the data with adequate statistical testing highlights groups of proteins relevant to the biological problem under investigation.

During peptide elution, an electric potential is placed on the liquid flowing from the LC column causing the solution to spray.³⁶ The spray is a fine mist of droplets that encompass the peptide ions as well as components from the LC mobile phase.⁴ These droplets are then desolvated by either applying heat or a curtain of nitrogen gas from which peptide ions are liberated. Peptide ions pass from the electrospray source into the mass analyzer, where the majority of the remaining solvent from the droplets is pumped away by the vacuum system.⁴ Different theoretical models dealing with the formation of droplets and the liberation of ions have been proposed (explained in more detail in section 2.4), but both agree that the efficiency of how each molecule ionizes, depends on their physicochemical properties.³⁶

Inside the mass spectrometer, mass spectra of the peptides eluting at a certain time point are taken, which in literature are commonly referred as MS1 spectra or “normal mass spectra”. Through a series of real-time software decisions, MS1 peptide precursor ions are prioritized for fragmentation and a series of “tandem MS”, MS/MS-, MS2- or fragmentation spectra are captured for each precursor ion.^{24, 37} As mentioned in Chapter 1, the most common spectral acquisition mode run in bottom-up proteomics experiments is by data-dependent acquisition (DDA) which is a process that consists of the isolation of a given peptide precursor ion, fragmentation with energetic collision with gas, and recording of the MS/MS spectrum.²⁴ The MS and MS/MS spectra are typically acquired for about one second each and stored for matching against protein sequence databases via software algorithms.²⁴

The databases for protein identification can be derived from public repositories of protein sequences like NCBI or Uniprot or be sample specific and generated from genomic or transcriptomic sequencing information. Proteins in the database are digested *in-silico* to generate predicted peptides which are subsequently fragmented *in-silico* to obtain predicted fragmentation patterns. The experimental precursor mass is used to search for predicted peptides, and the experimentally generated MS/MS spectra can be matched to predicted fragmentation patterns to determine peptide sequences. Once peptide sequences are determined, they are traced back to the source proteins. Finally, post-processing of the data employing different statistical analyses, ensures that groups of observed proteins are relevant for the biological phenomenon under investigation.

Here in, background information pertinent to the workflow of the different bottom-up mass spectrometry-based proteomics techniques employed in the development of Chapters 3-5, including sample preparation, and mass spectrometry measurement and data analysis, is presented while detailed protocols are outlined in the “Experimental procedures” sections of each chapter.

2.2 Protein extraction and digestion.

2.2.1 Being mindful about the sample: Considerations before starting a bottom-up proteomics experiment.

Modern bottom-up mass spectrometry techniques have been deployed to analyze a wide range of samples with environmental and clinical importance. The current technology used for bottom-up proteomics studies is nowadays sufficiently advanced to provide the in-depth characterization of very distinct biological matrices,³⁷ including, for example, acid mine drainage biofilms¹⁵³ and cerebrospinal fluid.¹⁵⁴ However, an initial understanding of their complexities helps informing the steps to take in an experimental workflow. Thus, although every bottom-up proteomics pipeline consists of the four steps described in section 2.1, each one has to be adapted according to the samples under analysis and the overall goal of the study.⁴¹

Prokaryotic and eukaryotic cells are very different in size and complexity. The genome sizes of bacteria, plants and humans, or even between different species of bacteria and plants, may translate into very different numbers of proteins that can be expressed at any given time or condition in a cell. When proteins are digested, and depending on which protease is used, the number of peptides analyzed by LC-MS/MS can be in the hundreds of thousands or even more. For example, a theoretical calculation in a human cell puts the number of peptides that could be produced with a trypsin mediated digestion near the 6 million.⁴ Depending on how complex the peptidome background is, the type of LC separation to employ is critical in order to minimize coelution of peptides and ion suppression events.

The types of organic structures that are part of cells also determine the procedures to conduct for protein recovery. Due to the presence of multiple layers of cellulose, plant cells are particularly strong and difficult to disrupt; while bacterial cells are easier to break,¹⁵⁵ but within them, Gram-positive bacteria cells are more troublesome to work with due to the presence of a thick layer of peptidoglycan and lipoteichoic acid surrounding the membrane.¹⁵⁶

The natural abundances of proteins in a given sample impacts the ability to detect targets of interest by bottom-up proteomics. Proteins can be expressed at levels as low as 10^1 - 10^2 proteins per cell, although some are expressed at much higher levels as 10^4 - 10^6 proteins per cell.⁴ Depending on the goal of the study, one may be interested in the identification of less abundant proteins. For instance, bottom-up proteomics pipelines that study proteins in plasma and try to determine biomarkers of clinical interest, are always confronted with the challenge that the dynamic range of protein abundances is 10 orders of magnitude in concentration from albumin to

the least abundant proteins.¹⁵⁷ In these cases, protein depletion or enrichment strategies can be employed to increase the chances of identifying targets of interest.¹⁵⁸

Recovery of biomass from bacterial cell isolates in suspension to bacterial cells from environmental samples is different. Low complexity samples like microbial pure cultures, defined enrichments, or symbioses (*i.e.*, co-cultures) usually comprise one or few species that are present in high cell numbers, whereas environmental samples such as those derived from soils, sediments or aquatic sources may comprise different species and be characterized by low microbial biomass levels.⁷² For example, in marine water columns, cell numbers range from 5×10^4 to 5×10^5 cells per mL.¹⁵⁹ In order to obtain sufficient cell material, water volumes from 1L to 20L need to be processed by differential filtration.⁷² Aside from their characteristic low protein abundance, environmental samples can contain perturbing substances that impact the measurement of peptides by MS. In soils, the most representative case is the presence of humic acids which are known to cause unwanted modifications in proteins,¹⁶⁰ suppress the electrospray ionization of analytes,¹⁶¹ and make loading samples to LC columns harder due to flow obstructions.¹⁶²

These initial sample considerations are examples that can dictate the success of bottom-up proteomics experiments.

2.2.2 Cell lysis and protein extraction via the SDS-TCA and CME methods.

The goal of cell lysis and protein extraction protocols is to recover as much protein as possible while minimizing contamination by other biomolecules that are present in the sample (*i.e.*, lipids, cellulose, nucleic acids, etc.).⁴ Depending on the type of samples, some protocols may need to homogenize tissues first, employing treatments with beads, sonication and/or liquid nitrogen.¹⁵⁵ After tissue disruption, cells are lysed with physical and/or chemical-based methods, but most commonly with a combination of both so as to physically break apart cell walls and membranes and then allowing the chemical agent to solubilize proteins.^{155, 163} In addition, some protocols may employ protease inhibitors like phenylmethylsulfonyl fluoride (PMSF),⁴ or ethylenediaminetetraacetic acid (EDTA),¹⁶⁴ to increase the yield of protein extraction, as the rupture of cellular structures provokes the liberation of hydrolytic enzymes that degrade proteins slowly.¹⁵⁵

Briefly, physical methods of cells lysis include ultra-sonication, bead osmotic shock, freeze/thawing treatments and mechanical blending,¹⁶⁵ while chemical lysis methods commonly employ detergents, which aid in destabilizing the structure of cell membranes, breaking lipid-

protein interactions and solubilizing proteins.¹⁵⁵ The choice of detergent is a critical decision step in bottom-up proteomics sample preparation protocols as some of them are known to interfere with enzymatic digestions, and most of them interfere with LC-MS/MS, sometimes damaging mass spectrometers and clogging LC columns.¹⁶⁶ Thus, factors such as the critical micelle concentration (CMC), or the ability of detergents to form micelles that can affect their efficacy and posterior removal from a sample, are important considerations when preparing a lysis buffer.¹⁶³

It is worth mentioning that lysis, extraction, and protein denaturation can all happen in the same step with certain experimental protocols. In this dissertation, the projects developed and presented in Chapters 3-5 all used a combination of either sonication or bead beating of cells in lysis buffer containing between 2-4% SDS. SDS has reported CMC values of 6-8 mM and its micelles have an approximate molecular weight (MW) of 17.8 kDa,^{163, 167} which means that it can be removed easier from a solution by buffer exchange or dialysis when compared to other detergents like Triton X-100 or Tween 80.¹⁶³

Once the cells are lysed, they need to be separated from any other biomaterials and detergents. Since the presence of lipids and genomic DNA is known to cause MS signal suppression and chromatographic interference,¹⁶³ their removal is important. This can be achieved by precipitating proteins out of the cell lysate solution using trichloroacetic acid (TCA)¹⁶⁸⁻¹⁷⁰ and rounds of centrifugation, which is one of the most popular procedures to eliminate nucleic acids, polysaccharides and salts from the cell lysate solution.¹⁵⁵ In general, getting rid of other biomolecules helps in acquiring “cleaner” spectra from which to obtain protein identification data. TCA precipitation can then be followed by detergent removal using cold acetone (0 to -20 °C).¹⁶³ Acetone is effective in removing SDS from proteins precipitated by TCA and is added to sample mixtures to a composition of 80%.¹⁷¹ This method was employed for samples in Chapter 3.

Another form of protein precipitation and clean up technique is methanol-chloroform precipitation (CME).^{163, 172} This technique is particularly efficient in recovering membrane proteins.^{163, 173} In the CME methodology, methanol, chloroform and water in volume ratios of 4:1:3 are added to cell lysate solutions followed by three additional volumes of methanol. Briefly, proteins are first separated from other unwanted materials on the water-organic interface and are then pelleted by “washing” the sample with methanol. The efficiency of this procedure has been described to approach 100% for a variety of protein concentrations and detergent solutions.¹⁶³ CME was the main protein recovery strategy employed for samples prepared in Chapter 5 of this dissertation.

At the end of protein precipitation methods, it is common to have a protein pellet that is often dried to some extent. This situation, added to the high degree of protein heterogeneity, and

the stability of protein-protein interactions can complicate their downstream processing by any enzymatic digestion method.¹⁵⁵ Hence, protein pellets need to be solubilized back in an adequate buffer containing several additives that prevent protein aggregation.

2.2.3 Protein solubilization with urea or SDC buffers and DTT and IAA.

Strategies for fully resolubilizing protein pellets involve the use of buffers with additives that help in denaturing proteins and maintaining them in solution, as well as other physical treatments like vigorous vortexing^{174, 175}, ultra-sonication^{174, 176} or shaking. The three main classes of additives added to re-solubilization buffers are: (1) chaotropes, which are used to prevent protein aggregation; (2) detergents, that increase the solubility of several proteins; and (3) reducing agents, that “break” disulphide bonds in proteins.^{4, 155, 177}

Chaotropes disrupt hydrogen bonds and inter- and intra- molecular hydrophobic interactions in proteins by altering the solution ionic strength and pH. Thus, these agents destabilize or denature the native structures of macromolecular protein assemblies.^{4, 155, 177} In particular, urea has been widely used in proteomics experiments at concentrations between 8-9 M to help maintain proteins in solution.¹⁵⁵ One consideration when using urea as the main chaotrope, however, is that heating must be avoided in order to prevent its hydrolysis and other non-desirable side reactions;¹⁵⁵ as such, physical treatments like ultra-sonication and incubations in thermomixers are not recommended to be used in concert. In this dissertation, urea was primarily used for protein denaturation in the development of Chapter 3.

As an alternative to chaotropes, detergents can be used typically in concentration ranges that vary between 1-4%.¹⁵⁵ Detergents are useful in preventing hydrophobic interactions, but careful selection of which of them to use and how compatible are they with remaining experimental steps are important considerations. In particular, sodium deoxycholate (SDC), which is a soft ionic detergent,¹⁶³ has been employed in recent bottom-up proteomics approaches as a substitute of SDS, due to observed reductions in technical variation during sample preparation,¹⁷⁸ its ability to increase the solubility of hydrophobic proteins, and compatibility with downstream trypsin digestion.¹⁷⁹ SDC is also easily removed after peptide digestion by crashing it out of solution with the addition of formic acid and then adding a water-immiscible organic solvent such as ethyl acetate,¹⁸⁰ into which the detergent is transferred while digested peptides can be recovered in the aqueous phase. This strategy was employed for the experiments in Chapter 5.

To prevent protein renaturation before enzymatic digestion, a reduction step of disulfide bridges in proteins is conducted with agents like dithiothreitol (DTT) or dithioerythritol (DTE).^{4, 177}

This step is often followed with the addition of reagents such as iodoacetamide (IAA), that are used to alkylate free sulfhydryl groups in proteins,¹⁵⁵ in order to protect them from re-oxidation events. Through the entirety of this dissertation, the combined actions of DTT and IAA were used in the steps of protein solubilization. Importantly, one must also be aware that some of the agents mentioned here (*i.e.*, IAA¹⁸¹) may produce unwanted effects in the sample if left for extended periods of time or used under the wrong temperatures; but it is through experimentation as well as tricks learned in the art of sample preparation that they are used in different degrees and combinations.

2.2.4 Protein digestion in MS bottom-up proteomics with trypsin.

If there is one protein to which proteomic scientists own their career to is trypsin. Trypsin is considered as the gold standard of protein digestions in bottom-up proteomics protocols due to its robustness, specificity and broad availability at a reasonable cost.^{4, 28, 150, 177, 182} Specifically for LC-MS/MS applications, companies sell trypsin as a highly purified enzyme that is extracted from either porcine or bovine pancreas. Trypsin is a serine protease that cleaves exclusively at the carboxyl side of arginine (R) or lysine (K) residues in a protein, unless either of these are followed by a proline residue. Compared to other proteases that act only on one amino acid, the dual specificity of trypsin means that it acts more frequently in proteins in theory producing more tryptic peptides for analysis. For example, a 50 kDa protein will yield about 30 tryptic peptides.⁴

Because of the spacing of R or K amino acids in a protein, which is approximately one residue per every 10–12 amino acids,¹⁵⁵ trypsin generates peptides that are of suitable lengths (~10-15 amino acids)¹⁸³ and MWs (~ 5-30 kDa)^{150, 155} to chromatographic separation and that fall within the ideal mass range of mass spectrometers.¹⁸² In addition, the presence of C-terminal lysine or arginine residues is beneficial, as they efficiently protonate under acid conditions and this in turn helps peptides to ionize and to produce rich fragmentation spectra that enhances protein identification.^{28, 155, 182} In fact, a protein can be confidently identified with the information produced by only one tryptic peptide.⁴ Other advantages of trypsin for bottom up proteomics work is that the enzyme displays good activity both in solution and in in-gel digestion protocols.⁴

As with any other reagent used in bottom-up pipelines, some considerations are important when using trypsin for protein digestion. In this sense, digestion conditions like time, temperature, pH and enzyme-to-protein ratios can influence the specificity and efficiency of trypsin.¹⁵¹ It has been recommended that in order to maximize protein identification in global proteomics studies, a digestion of pH 7, temperature of 37 °C and enzyme-to-protein ratio of 1:50 can maximize

protein identification.¹⁸² The latter parameter varies in published articles where the values can range from 1:100 to as high as 1:2.5,¹⁸³⁻¹⁸⁸ with targeted proteomics studies using lower enzyme-to-protein ratios to maximize peptide yield. Regarding time, this factor also changes from study to study, but it is common practice to incubate protein samples with trypsin for 1-3 hours and then conduct a second round of overnight digestion.¹⁵⁵

Importantly, although it has many advantages for bottom-up proteomics, trypsin is not free of flaws. For some applications, it has been reported that the short length of tryptic peptides (≤ 6 residues) are not optimal for MS fragmentation and can restrict proteome coverage.¹⁸⁹ Other studies have reported that whenever there are cases of adjacent K or R residues in a protein, trypsin will prefer to act on those that are located on its C-terminus side, and will fail to produce peptides from its N-terminus.¹⁸³ PTMs can also inhibit trypsin cleavage.^{190, 191}

In this dissertation, trypsin digestion was used to produce peptides for subsequent measurement through LC-MS/MS although, due to the different applications presented in Chapter 3-5, with digestion parameters adapted each time.

2.2.5 Peptide clean up before loading into a LC column.

Peptide mixtures should ideally be loaded into a LC column without any compounds that may compromise their MS/MS detection. Excess salts or trace amounts of detergents are known to interfere with peptide ionization and add chemical noise in the mass spectra.¹⁵⁸ Additionally, detergents can clog LC columns, and salts can accumulate in the ion transfer capillary of the mass spectrometers.¹⁹² For these reasons, peptide sample cleanup procedures are important to perform before LC-MS/MS.¹⁹³ These protocols can be either on-line using trap columns or off-line using several commercially available products.¹⁵⁸

On-line cleanup/desalting occurs as the sample moves through a pre-column that is connected to the analytical LC column where peptide separation takes place; however, the initial flow-through going through the pre-column is diverted to waste. After peptides have been washed extensively, they can now be eluted into the analytical column. For off-line cleanup/desalting, devices that employ reversed-phase stationary resin (*i.e.*, C₁₈) packed in pipet tips, spin columns, or syringes are used. These products have varying binding capacities ranging from 1 to 25 μg protein.¹⁹² For example, StageTips, are ordinary pipette tips with inserted small disks made of beads with reverse phase surfaces embedded in a Teflon mesh.¹⁹⁴ The fixed nature of the beads allows flexible combination of disks with different surfaces (*i.e.*, cation-exchange or anion-

exchange) to obtain multi-functional tips.¹⁹⁴ C18 stage tips were used in Chapter 3, while C₁₈ spin columns were employed for peptide cleanups in Chapters 4 and 5.

In addition to peptide clean-up, all the peptide mixtures obtained at the end of every experimental protocol in Chapters 3-5 were concentrated using a SpeedVac and then resuspended in a solvent mixture whose organic content and pH matched the starting chromatographic conditions in each case.

2.3 Peptide separation by nanoLC.

2.3.1 Basics of peptide chromatographic separation by nanoLC.

The number of peptides that can be produced from a proteolytic digestion protocol presents an analytical challenge to any modern mass spectrometer. For instance, even in a common bacterial system, theoretical calculations have predicted that the number of peptides produced could be higher than 62,500 and these could have abundance differences higher than 10⁵.¹⁹⁵ Thus, ways to de-complex, or simplify a peptide mixture, before it is measured with a mass spectrometer are necessary. This goal has been achieved with nano liquid chromatography (nanoLC), which is a fully automated form of liquid chromatography, that allows to separate peptides with high speed, resolution and sensitivity.¹⁹⁶

Similar to the more traditional high-performance liquid chromatography (HPLC) at the microscale level, in nanoLC, chromatographic columns with small internal diameters $\leq 100\ \mu\text{M}$,^{36, 197, 198} and packed with a stationary phase, are used to inject peptide mixtures. Peptides bind to the stationary phase and are then separated by flowing a mobile phase at nL min⁻¹ flow rates that changes its composition over time either in variable steps or continuously.^{36, 197} The low volumes and flow rates of nanoLC are beneficial to the electrospray ionization (ESI) technique, which acts like a concentration-dependent detector in mass spectrometers, and help improve the detection sensitivity of the measurements.^{36, 196}

By exploiting the physicochemical properties of the naturally occurring L- α -amino acids found in proteins,¹⁹⁶ optimal stationary phases as well as mobile phase compositions may be selected to achieve a successful chromatographic separation of peptides in bottom-up proteomics experiments. For example, the number and distribution of charged groups on the side chains of amino acids influences the polarizability and ionization status of a peptide, as well as their hydrophobicity.¹⁹⁶ Thus, chromatography modes that make use of peptide hydrophobicity, like in reverse phase (RP) chromatography, or charge, like in ion exchange (IEX) chromatography; can

be used independently or in combination to deconvolute samples reaching a mass spectrometer. Other separation modes in bottom-up proteomics research are hydrophilic interaction (HILIC) chromatography; and affinity chromatography (including immobilized metal ion-affinity and biospecific chromatography).^{8, 72, 198}

In this dissertation work, RP and strong cation exchange (SCX) separations were used for the diverse proteome measurements presented through Chapters 3-5. Both types of separation are described in more detail in the following subsections.

2.3.2 HPLC vs. UHPLC

With the goal of improving analyses times, while at the same time maintaining acceptable chromatographic resolution in a complex peptide mixture, improvements to conventional nanoHPLC strategies have been tested in bottom-up proteomics experiments. One of those developments is ultra-high-pressure liquid chromatography (UHPLC). Compared to conventional nanoHPLC which uses pumps and plumbing that can support flow backpressures up to 450 bar, nanoUHPLC uses instrumentation that can tolerate up to 1400 bar^{197, 199} allowing the use of smaller inner diameter columns packed with stationary phases made of particles sizes with diameter sizes < 2 μM ,²⁰⁰ instead of stationary phases consisting of particles having diameter sizes of 5 μM that are employed in nanoHPLC.

The decrease in stationary phase particle sizes and internal diameter of analytical columns and plumbing has brought many benefits to peptide detection with tandem mass spectrometry. One of those is greater column efficiency, measured by the number of theoretical plates, which can be described as a number of imaginary layers across the length of column in which an analyte is in equilibrium with the stationary and mobile phases. The plate model theory states that the higher the number of theoretical plates in a column at a given length, the better the quality of the column and the narrower the peaks.¹⁹⁶ Columns packed with stationary phases < 2 μM and used in combination with UHPLC pumps, have many more theoretical plates than columns used in HPLC separations (**Figures 2.2 A-B**). The former translates into narrower chromatographic peak widths at baseline that are on average 2-4 seconds compared to 10 seconds in HPLC,²⁰¹ which in turns improves the resolution of peptide separation, with some studies reporting improvements by a factor of 3 when conducting UHPLC versus HPLC under similar analysis times.^{202, 203}

Better chromatographic resolution also means that regions in a chromatogram that look crowded in HPLC can be better differentiated in UHPLC and this improves peptide identification by computational algorithms. Furthermore, the narrower peak elution profiles also improves

column capacities, or the number of peaks that can be accommodated in certain column segments, which consequently brings a reduction in the number of analytes that coelute at a given time and hence, less chance of chimeric spectra occurrence²⁰⁴ which is an ever present problem in bottom-up experiments and is described in more detail in section 1.4.

The sharpening of chromatographic peaks in UHPLC also boosts the sensitivity of the analysis.²⁰⁴ Peptides that were undetectable in HPLC due to broader chromatographic elution profiles, can now present higher chromatographic peak areas in UHPLC that allow for their sequencing via MS/MS. Increases in sensitivity when compared to traditional HPLC have been reported. Other advantages of UHPLC compared to HPLC runs are shorter analysis times²⁰⁵ and the use of small sample volumes ($\geq 0.5 \mu\text{L}$).²⁰⁶ Interestingly, the introduction of UHPLC to bottom-up proteomics analyses have enabled “single-shot” proteomics studies, in which complex proteomes are analyzed in-depth with the aid of long 1D chromatographic columns (25-50 cm) and LC gradients (300-500 mins).^{152, 200, 204, 207}

However, the main disadvantage of UHPLC is that the equipment to run these types of experiments is more expensive than that used by HPLC and its correct implementation requires of delicate handling and expertise. Thus, alternatives to increase throughput and depth in bottom-up analyses using regular HPLC have been tested. In fact, one of those ways was explored in more detailed in the research presented in Chapter 4.

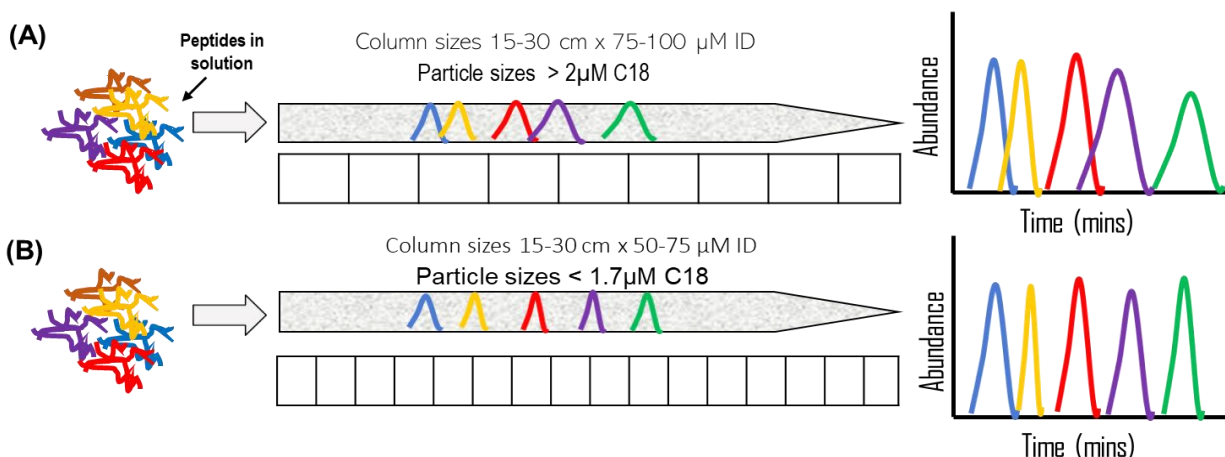


Figure 2.2 1D-liquid chromatography setups. The square grids below each column represents the number of theoretical plates. **(A)** One-dimensional (1D) separation with a C18 packed column using particle sizes $> 2 \mu\text{M}$ employed with HPLC pumps. **(B)** 1D separation employing C18 particle sizes $< 1.7 \mu\text{M}$ and typically employed with UHPLC pumps due to the high backpressure caused by the decrease in particle sizes.

2.3.3 Reverse phase liquid chromatography (RP-LC).

Reverse phase liquid chromatography (RP-LC) exploits peptide hydrophobicity to separate peptides in a column.^{4, 8} In this type of chromatography, the stationary phase consists of a nonpolar phase, which in bottom-up proteomics is usually made of C18 chains that are bound to base silica material.⁷² Peptides are first loaded to the analytical column in an acidic mobile phase containing a low percentage of organic solvent. After peptides are retained, elution happens by gradually increasing the content of organic solvent in the mobile phase over time. The retained peptides elute in order of increasing hydrophobic interaction with the stationary phase; with peptides having lower hydrophobicity eluting first.⁴

In case of samples that are suspected to have low amounts of peptides, trap columns (~ 10–20 mm length) with larger inner diameters (usually 300 μm) can be used for enrichment;⁷² alternatively, they can be used to perform desalting washes of peptides retained in the stationary phase. These trap columns can be operated either off-line or on-line¹⁹⁶ with the main C18 separation column. In the off-line mode, peptide mixtures can be loaded to the trap column and subsequently desalted using a pressure cell to then be interfaced with the separation column in the LC instrument. Instead, in the online mode, the LC instrument is configured to do automatic valve switching which allows the selection of mobile phase pathways for first loading the peptides into the trap column (without loading them into the separation column), then do desalting (without sending the flow through to the separation column) and finally conduct the analytical separation.

Although RP-LC separation is reproducible, selective, and effective for the separation of peptides in bottom-up proteomics,¹⁹⁸ it may reach its limits with samples of high complexity. After enzymatic digestion, the thousands or hundreds of thousands of peptides in a sample may exceed the peak capacity of RP-LC columns. The co-elution of too many peptides as well as their large ion abundance differences may also cause undesirable ion suppression events during electrospray ionization. Hence, separations that can provide increased peak capacities and resolving power can improve peptide detection, thereby increasing proteome coverage and dynamic range.²⁰⁸

2.3.4 Two-dimensional liquid chromatography (2D-LC).

Before the adoption of UHPLC systems for reverse phase-based separations, multidimensional separation techniques were introduced as an alternative to increase the separation power provided by RP-HPLC. Multidimensional LC couples two (or more) different consecutive LC chromatographic separation methods (or dimensions, hence the name

“multidimensional”) to improve peptide separation.²⁰⁹ To increase the resolving power of any multidimensional LC system, the chromatographic phases need to be “orthogonal” to each other or, in other words, exploit different physical properties of peptides such as charge, hydrophobicity or molecular weight²¹⁰ while at the same time being technologically feasible to implement.

For bottom-up proteomics in particular, two-dimensional strategies that combine the use of SCX, as first dimension, and RP, as second dimension, have been widely employed and adapted over the years.²⁰⁸ Especially with the demonstration conducted by the Yates group in the seminal paper introducing the applicability of the multidimensional protein identification technology (MudPIT), as an alternative to “gel-free” separation method, the strategy of combining SCX with RP became widely adopted. This application demonstrated that it was capable of identifying ~25% of the complete yeast proteome within a single experiment.³⁹

In a 2D-LC system including SCX as first dimension and C18 as second dimension (**Figure 2.3**), peptides are first loaded into a SCX packed column in acidic mobile phase (pH < 4.0). This causes amino acid residues in peptides to be protonated and able to interact with the negative charged sulfonyl groups of the SCX resin.⁷² Each peptide adsorbs to the SCX resin with affinities that are proportional to the overall number of positive charges on them.⁴ Once peptides are loaded, elution happens gradually or stepwise with mobile phase “pulses” or “steps” at increasing salt concentrations.^{4, 72} Each salt pulse releases different groups of peptides that are captured on a RP column which is positioned downstream of the SCX column. In the RP column, peptides are separated based on hydrophobicity in similar fashion to a regular one-dimensional RP-LC strategy.

The fractionation based on electrostatic interaction is highly orthogonal to RP.⁷² Similarly as using a trap column for 1D-RP-LC, chromatographic separations in a 2D strategy can be executed off-line, for example, by manual collection of peptide fractions from SCX; or on-line, involving a valve switching mechanism with an adequate setup so as not damage any component of the system with the different mobile phases used.²¹¹ The performance of a 2D separation was compared to a RP-UHPLC strategy in Chapter 4.

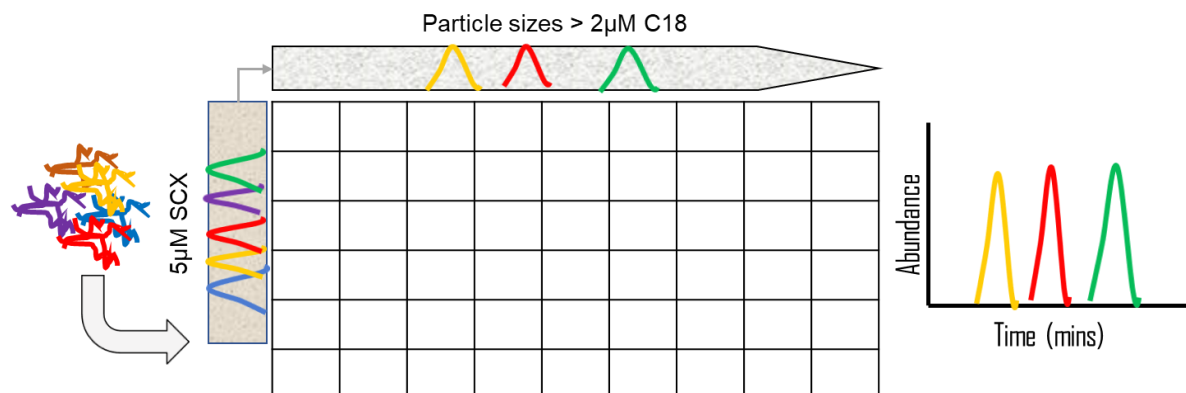


Figure 2.3 2D-liquid chromatography setup. A 2D separation strategy exploiting the orthogonality of a strong cation exchange (SCX) packed column interfaced with a C18 packed column is shown. The square grids below each column represents the number of theoretical plates.

2.4 Fundamentals of electrospray ionization (ESI).

ESI is the process by which gas-phase peptide ions are produced in the source of a mass spectrometer to enable subsequent mass analysis.⁴ ESI is classified as a soft-ionization technique, which means that hardly any internal energy is transferred to the ions thus minimizing in-source fragmentation significantly.⁶² In this technique, a protonated or deprotonated molecule, represented by the symbols $[M+H]^+$ or $[M-H]^-$, can be generated, respectively. In particular, for the analysis of peptides in bottom-up LC-M/MS, $[M+H]^+$ peptide ions are produced and measured as their positive charge is favorable for fragmentation via MS/MS and several of their LC chromatographic characteristics are improved at acidic pH.⁴ The presence of multiple proton-accepting sites in peptides (*i.e.*, amine groups), and in particular of those derived by tryptic digestion (bearing lysine or arginine residues), causes them to exist as a mixture of singly, doubly or triply charged ions in solution.⁴

The ESI process involves the placement of an electric potential of 2-3 kV between the liquid flowing from the LC column through a fused silica column or needle, and the counter-electrode in the mass spectrometer (**Figure 2.4**).²² In the positive ion mode and for ionic analytes, such as positively charged peptides, the electric field, when turned on, goes inside the solvent carrying the molecules. This causes the polarization of the charged molecules in the liquid. In the positive mode, an enrichment of positive ions occurs near the surface of the solvent meniscus leaving the fused silica or needle, while negative ions, if present, try to get away from it.²¹² These polarizing forces cause the formation of the “Taylor cone”²¹³, in which the distortion of the meniscus creates a cone-shaped spray that points towards the counter electrode of the mass

spectrometer. The increase of surface due to the cone formation is resisted by the surface tension of the liquid.²¹²

If the applied field is sufficiently high the spray tip becomes unstable and a fine jet emerges from it. The surface of the jet is charged by an excess of positive ions. The repulsion between the charges on the jet causes the jet to break up into small charged droplets.²¹² The size of the droplets formed from the cone jet is dependent on its diameter and studies have proposed and observed that the majority of the droplets produced have approximately the same size.^{214, 215} The droplets are positively charged due to an excess of positive electrolyte ions at the surface of the cone and the cone jet. As charged droplets get expelled and travel towards the mass spectrometer, the solvent that they carry gets evaporated, causing their shrinkage and an increase in the electric field the electric field normal to their surface.

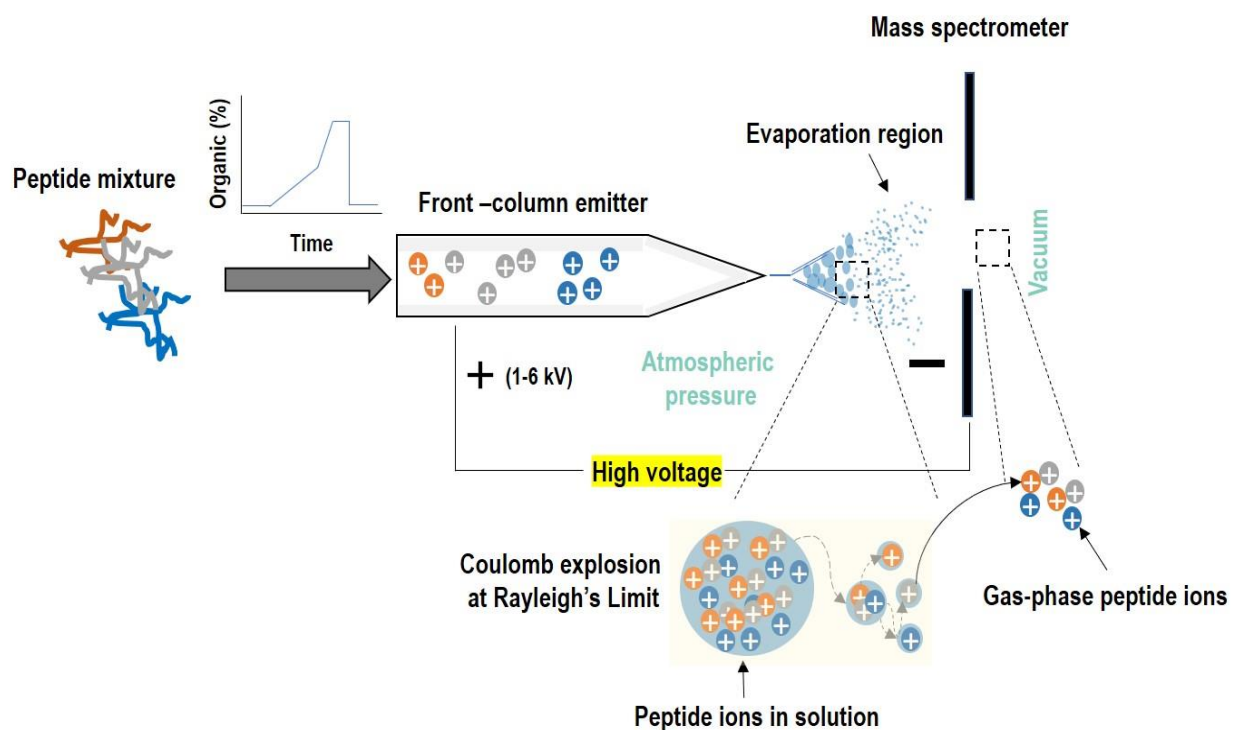


Figure 2.4 Illustration of the electrospray ionization process. Protonated peptide mixtures are separated by liquid chromatography and then sprayed from the front-column emitter towards the mass spectrometer. Initially, large solvent droplets containing hundreds, or thousands of peptide ions are sprayed. As these droplets move towards the heated capillary in the spectrometer, solvent evaporates from the surface of the droplets, and ions are forced together until charge repulsion causes peptide ions to leave the droplet. The high voltage potential placed between the emitter and the mass spectrometer pulls the ions into the high vacuum region of the latter

As the droplets get smaller, the increasing repulsion between the charges at their surfaces increases and it overcomes the cohesive force of their surface tension. This phenomenon leads to the fission of the droplets that releases a series of small and charged secondary or progeny droplets. This condition of instability, is also called Coulomb fission or “explosion”, and is explained by the Rayleigh equation²¹⁶:

$$Q_{Ry} = 8\pi(\varepsilon_0\gamma R^3)^{1/2}$$

Where:

Q_{Ry} = Charge of the droplet

γ = surface tension of the solvent

R = radius of the droplet

ε_0 = electrical permittivity

The shrinkage of the droplets at constant charge fission at or near the Rayleigh limit, and the release of a jet of small, monodisperse charged progeny droplets has been confirmed by a number of experiments.²¹² Further evaporation of the solvent leads to a higher occurrence of droplet fissions. Ultimately, very small charged droplets enter the mass spectrometer and lead to the formation of gas-phase ions by processes which are mainly explained by two theoretical models.

The first model is the charge-residue model of Dole.²¹⁷ In this model, the sequence of droplet disintegration proceeds until the microdroplets contain only one preformed analyte ion per droplet. By evaporation of the solvent, the preformed analyte ion is released to the gas phase. The second model is the ion-evaporation model proposed by Iribarne and Thomson,²¹⁸ in which gas-phase ions are released from highly charged microdroplets, because the local field strength is high enough for preformed ions to be emitted into the gas phase.⁶² Both of these models are to some extent complementary and there is no consensus of which one is actually more important in the actual ion production of a particular analyte.

Given the importance of droplet evaporation during ESI, the generation of smaller droplets is more favorable in term of sensitivity and the ability to preserve noncovalent molecular associates. This can be achieved by using nanoLC, where the analyte is sprayed, for example, from silica capillary with tip diameters of 1–5 μm rather than from the 100 to 150- μm tips that are used in conventional ESI. This mode of operation is called nanoESI and usually allows the use of gentler operating conditions such as temperature, gas flows, and needle voltages.²¹⁹ Compared

to conventional ESI, decreasing the flow rate to nL/min rates, leads to increased ionization efficiency by 50~60%, greater sensitivity and better tolerance to the salts and other impurities. nanoESI has become the most used ionization technique in peptide and protein analysis and as such all the MS instruments used for the experiments in Chapters 3-5 used a nanoESI source.

2.5 Fundamentals of mass spectrometric detection.

2.5.1 Figures of merit in mass spectrometric detection.

There are several types of mass spectrometers available for the detection and measurement of peptides in bottom-up proteomics; however, the type of results obtained are heavily influenced by the capabilities of each instrument as well as the complexity of the sample under study. Hence, before introducing the most common types of spectrometers used in bottom-up proteomics work, and describing in more detail the specific ones employed in this dissertation, it is important to first define the analytical figures of merit that are used to evaluate the performance of each type of instrument. These figures of merit are: mass resolution, mass accuracy, and the acquisition speed, or the time needed to acquire one data point in a chromatogram.⁶²

The resolution, also known as the “resolving power” of an instrument, is defined as the ability to distinguish two ions with similar mass-to-charge (m/z).⁷² More specifically, and according to the International Union of Pure and Applied Chemistry (IUPAC) definition, mass resolution is a dimensionless unit that is the product of $m/\Delta m$, where m is the mass of the ion and Δm is either the peak width at full-width-half-maximum (FWHM) between two equal-intensity peaks, or at a 10% valley definition between two equal-height peaks.^{8, 220, 221} Typically, mass spectrometers (or more rigorously mass analyzers) can provide unit-mass resolution or high-mass resolution.⁶² Instruments that provide high-mass resolution are able to differentiate the isotope clusters of a compound which enables the detection of the “monoisotopic ion”, or the ion that is composed of only one isotope of each element (*i.e.*, ^1H , ^{12}C , ^{14}N and ^{16}O).⁷² An example of how the resolving power of a mass analyzer impacts peptide identification in bottom-up proteomics is shown in **Figure 2.5**.

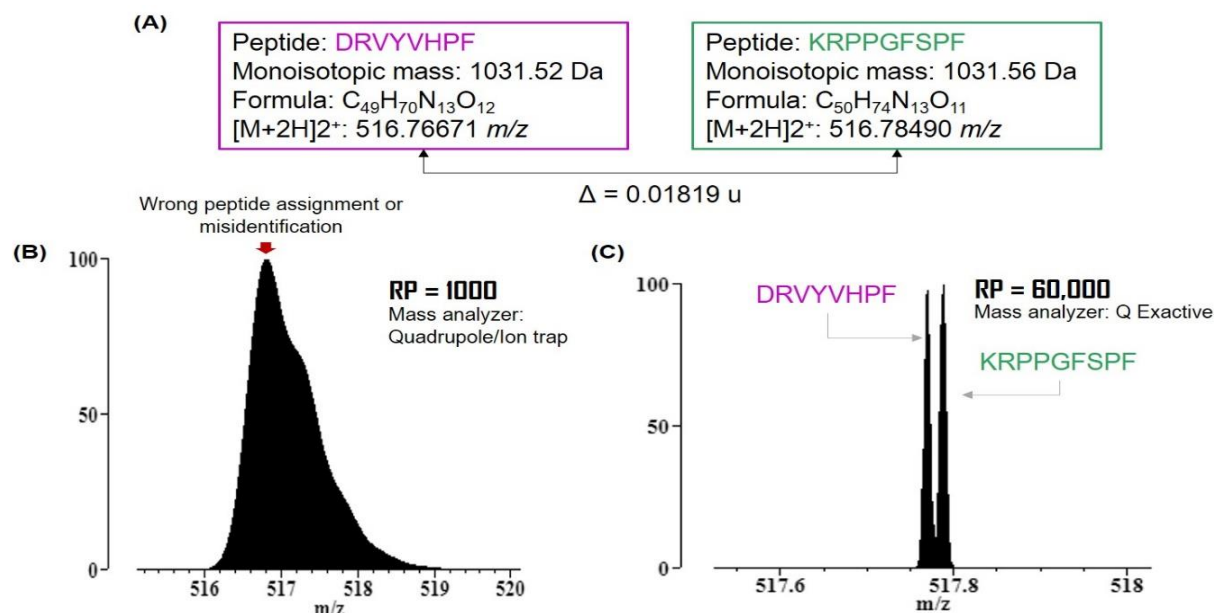


Figure 2.5 Mass resolution in mass spectrometry-based bottom-up proteomics studies impacts the correct identification of peptides. (A) Two peptides with similar monoisotopic mass generate doubly charged ions with close m/z values. (B) Simulation of isotope profiles of both peptides in (A) using a resolving power (RP) of 1000 which is typical of standalone quadrupoles or ion trap mass analyzers. (C) Simulation of isotope profiles of both peptides in (A) using a RP of 60,000 which is available in Thermo Scientific Q Exactive hybrid mass spectrometer. The MS-Isotope program of the Prospector suite of proteomics tools was used to generate the isotope profiles.

The mass accuracy of an instrument determines how close the experimental mass values for ions (accurate mass) in a sample are to their theoretical ones (exact mass).⁴ Low-resolution instruments generally allow the m/z determination for single-charge ions with an accuracy of ± 0.1 units (in nominal mass).⁶² While instruments that provide high-mass resolution, as mentioned before, enable the determination of a monoisotopic mass with accuracies that are better than ± 0.01 Da,⁷² referring to the masses of the most abundant natural isotopes of the elements present in the ion or molecule. In practice, to achieve mass accuracies that are between the operational parameters of each instrument, these need to be m/z calibrated as needed with special solutions recommended by the manufacturers. For high-resolution instruments, the accuracy is expressed either as an absolute mass error or as a relative error (in ppm), and this is calculated from:⁶²

$$\frac{(\text{accurate mass} - \text{exact mass})}{(\text{exact mass})} \times 10^6$$

Mass accuracy is an important parameter for the identification of biomolecules in mass spectrometry, because the experimental mass values of ions are often compared to exact mass values calculated by software such as those used in peptide/protein database searches in bottom-up proteomics.⁴

Scan speed or the acquisition rate of an instrument,²⁰⁰ refers to the number of spectra that is acquired per unit time in a particular m/z range. Fast scan speeds are required for rapidly changing events such as to monitor peptides coming out from chromatography.²²¹ For example, in global proteomics studies using data-dependent acquisition, mass spectra are continuously acquired between a low m/z and a high m/z range within a preset period of time (mostly ≤ 1 s).⁶²

Sensitivity in mass spectrometry measurements has two widely accepted definitions: (1) It can refer to the change in signal per unit change in concentration of analyte;²²² or (2) it can be described as the smallest amount of an analyte that can be detected at a certain defined confidence level or from background noise.⁴ Considering this latter definition, the limit of detection (LOD) is determined from the analyte signal-to-noise ratio (S/N) and is the lowest concentration of a molecule where its signal can be distinguished from system noise.²²³ Instead, the limit of quantification (LOQ) is defined as the lowest amount of an analyte in a sample that can be quantitatively determined with precision and accuracy.²²² Both LOD and LOQ values are usually plotted in calibration curves which depict a range of concentration-response values that can be measured with reliability and reproducibility.

2.5.2 Overview of tandem mass analyzers and detectors used in bottom-up proteomics applications.

Mass spectrometers measure the mass-to-charge ratio (m/z) of gas phase ions,²³ or in proteomics, of peptide ions. The three main components that make up a mass spectrometer are the ion source, which produces the ions from the sample (*i.e.*, the nanoESI source); the mass analyzer, which resolves ions based on their m/z ratio; and finally, the detector that as the name implies, records the number of ions that are resolved by the mass analyzer.^{4, 23, 72} All the data or spectra are plotted as a function of ion abundance versus m/z which can then be retrieved for manual or computational-assisted interpretation. From these m/z values the mass of the analytes can later be determined.

Mass analyzers are central to MS technology. At the most fundamental level, analyzers use some type of static or dynamic electric and magnetic fields to selectively separate ions. For bottom-up proteomics applications, and importantly, workflows that use nanoESI sources, the

most common types of analyzers employed are quadrupoles, ion traps (quadrupole ion trap, linear traps, Orbitraps) and time-of-flight (TOF) analyzers.⁴ Although most of these analyzers may be used as stand-alone systems, most often they are combined into hybrid systems that allow the collection of tandem MS data from which peptide sequences can be deduced.⁶² All of them vary in their physical principles and analytical performance (Table X), however, they all are used in bottom-up proteomics to perform the measurement of peptide ions through tandem MS/MS.

The high-speed data acquisition in mass spectrometers needs to be backed by fast electronics, including analog-to-digital converters (ADCs).²²⁴ Electron multipliers are the most widely used types of detectors used in combination with quadrupole and ion-trap instruments.⁶² Detection of ions in multipliers is based on the repeated emissions of secondary electrons, resulting from the repeated collisions of energetic particles at a suitable surface. With TOF instruments, microchannel plate (MCP) detectors, which are arrays of miniature electron multipliers oriented parallel to one another, are applied, as they are more suitable for ion detection when the ion beam is more spread in space.²²⁵ In orbitrap MS systems, the signal of all ions with different m/z values is based on the detection of high-frequency image currents.²²⁶

The projects presented in Chapters 3-5 of this dissertation employed a triple quadrupole and a hybrid Orbitrap analyzers to conduct targeted and global proteomics analyses of the samples under study, respectively. Their operational principles are described in more detail in the following sections and their performance metrics summarized in **Table 2.1**.

Table 2.1 Performance metrics of the instruments used in this dissertation

Instrument	Mass resolution	Mass accuracy†	Mass range	Sensitivity	Data acquisition speed
TSQ Quantum Ultra Discovery Max	Up to 7500 (FWHM) at m/z 508 of polytyrosine	300- 500 ppm	10-1500 Da	Picomoles-Femtomole	Up to 5000 Da/sec
Q Exactive Plus	Up to 140,000-280,000 at m/z 200	1 – 3 ppm	50-6000 Da	Attomole-Femtomole	Up to 12 Hz resolution setting of 17,500 at m/z 200

† Mass accuracy range calculated for molecules at m/z 300-500.

2.5.2.1 Targeted proteomics via multiple reaction monitoring MS on a triple quadrupole (MRM-MS).

The triple quadrupole is a hybrid MS system that is well suited for targeted proteomics studies that collect spectral data via MRM-MS. To understand its basic mechanism of operation, it is important to first describe a quadrupole and how it works. A quadrupole is a mass analyzer that is made of four circular-shaped rods that are arranged in parallel.^{4, 62} By applying direct current (DC) and radiofrequency (RF) voltages to the rods, a dynamic electromagnetic field is produced which results in a complex oscillating movement of the ions from beginning to end of the analyzer.^{4, 72} Depending on the DC/RF combinations applied to the rods, ions with different m/z values can be transmitted to the detector one after another. The transmission of distinct ions by continuously changing the voltages can be mathematically modeled with the Mathieu differential equations.²²⁷

Quadrupoles are low resolution mass analyzers that provide mass accuracies <0.3 Da and scan speeds as high as 10 000 u/sec in a mass range of 10–2000 Da.^{62, 72} When three quadrupoles are combined in series, they form a triple quadrupole instrument (also known as “triple quad” or “QqQ”). A QqQ is composed of two quadrupoles, “Q1” and “Q3”, that are controlled by DC and RF potentials and act as mass filters that transmit ions of given m/z ratios; both of them are separated by a quadrupole “q2”, which is operated in the RF-only mode.⁴ In q2, fragmentation of ions is achieved by collision-induced dissociation (CID), in which neutral gas atoms, for example argon, are used to impart internal energy to the bonds of the ions under analysis, causing them to break.³⁷ These fragment ions can then be separated according to their m/z by Q3 and then passed to the ion detection system.

The research project presented in Chapter 3, used a TSQ Quantum Ultra instrument manufactured by Thermo Fisher Scientific (**Figure 2.6**). The TSQ mass spectrometer consists of an ion source interface, ion optics, mass analyzer section (quadrupoles), and an ion detection system. The instrument can be interfaced with several ionization sources, but for the experiments in Chapter 3 a nano-ESI source was employed. The ion source interface consists of elements that help in ion desolvation and maintain vacuum inside the system. The ion optics focus the ions produced in the electrospray source and transmit them to the mass analyzer, these are divided in Q00 ion optics, which are located closer to the ion source interface, and Q0 ion optics. The mass analyzer region consists of quadrupoles Q1, Q2, Q3 and three lens sets with diverse functions including, for example, minimizing collision gas entering to Q1 (lens L22 and L21) and Q3 (lens L32 and L33) from Q2; retain collision gas in Q2 (lens L23 and L31); shield from the different voltages applied to the quadrupoles and focus the ion beam. According to the manufacturer, the

design of the Q2 quadrupole, which are quadrupole rods bent through a 90-degree arc, is helpful in preventing the transmission of unwanted neutral species to the detector, thus lowering the noise level in the data. Finally, the TSQ instrument is equipped with a high-sensitivity detection system, located at the rear of the vacuum manifold behind the mass analyzer which has been designed to increase signal and decrease noise levels.

The TSQ quantum can be operated to collect MS1 signals from all ions coming the ionization source or in the “full-scan” mode, but alternatively, and more frequently, it is used to selectively induce the fragmentation and detection of ions coming to the ion source. This type of spectral data acquisition is known as multiple reaction monitoring-mass spectrometry (MRM-MS) and is the basis of targeted proteomics measurements. On MRM-MS, the voltage settings applied to Q1 only allow peptide ions with specific m/z to pass through it. Ions then enter q2 where they collide with the neutral gas to produce fragment ions and neutral fragments. The fragment ions are then analyzed based on their m/z by Q3, which scans repeatedly over a mass range. Although they lack mass accuracy and resolution, MRM-MS measurements are more reproducible and sensitive than DDA methods used for global proteomics applications. Another advantage is the ability to determine absolute abundance values from the peptides and proteins under study using,²²⁸ for example, the AQUA strategy described in section 1.2.2. In theory, up to 1000 proteins can be targeted in a single LC-MRM-MS run after careful selection of ion transitions.²²⁸ Due to their consistency and quantitative accuracy, targeted proteomics measurements provide an alternative to test biological hypotheses generated from global proteomics approaches.²²⁹

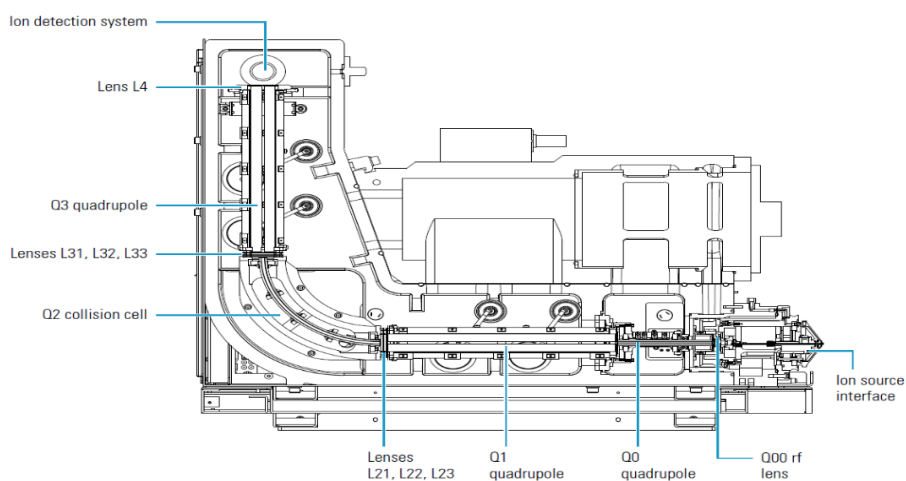


Figure 2.6 Schematic of the TSQ Quantum Ultra mass spectrometer used for targeted proteomics measurements via MRM-MS for the project presented in Chapter 3. Image taken from the hardware manual provided by Thermo Scientific.

2.5.2.2 Global proteomics through spectral data-dependent acquisition on a Q Exactive hybrid quadrupole-Orbitrap.

The Orbitrap mass analyzer was developed 20 years ago²³⁰ and it is considered a technical improvement to linear ion traps. Different than quadrupoles, where ions are measured and fragmented “on the fly”,⁴ in ion traps, ions are collected and stored over time in order to perform MS/MS analyses. The Orbitrap is an electrostatic trap that consists of two outer electrodes and a central electrode to which a constant electrostatic potential is applied.⁷² Ions are attracted to the inner electrode where they orbit around it while at the same time maintaining an axial oscillation with a frequency that is characteristic of their m/z values.²³ Ions of different velocities spread into rings with different rotational frequencies, but same axial frequency.⁷² The image currents that are generated from these movements are captured by two outer electrodes and Fourier transformed into the time domain to produce mass spectra.²³

Orbitraps provide high mass resolutions of up to 240,000 that contribute to measure analytes with mass accuracies < 3 ppm of up 4000 m/z .⁷² The operational characteristics of Orbitraps make them ideal instrument for experiments that try to identify all possible proteins in a complex sample and those global proteomics strategies using spectral data-dependent acquisition (DDA). Although stand-alone orbitrap systems have been produced, Orbitraps are combined in hybrid systems to perform tandem MS analyses of peptides.⁶² In fact, the first commercial introduction of the Orbitrap analyzer was as a hybrid instrument combining a low resolution linear-ion trap with the high-resolution Orbitrap (LTQ-Orbitrap).²³¹

Nowadays, however, of the many possible Orbitrap configurations, hybrid quadrupole–orbitrap analyzers are widely used in bottom-up global proteomics research.³⁷ These type of instruments were first introduced by Thermo Scientific with its Q Exactive line of instruments. In particular, the projects presented through Chapters 3-5 used a Q Exactive Plus spectrometer to acquire global proteomics data via DDA. This instrument consists of six main components, including the ion source, an injection flatapole that pre-filter ions according to their m/z ratios, a quadrupole mass filter for precursor ion selection, curved linear trap (C-trap), a high energy collision cell (HCD cell), and Orbitrap analyzer (**Figure 2.7**).

Briefly, on DDA acquisition, ions coming from the electrospray source are transmitted into the quadrupole via the flatapole that acts like an ion focus device and as an ion-prefiltering device. The proprietary technology by Thermo is called the “Advanced axial beam guide” and uses an axial bent flatapole that reduces noise by preventing neutrals and high-velocity clusters from entering the quadrupole. The segmented quadrupole, known as the “HyperQuad”, transmit and filter ions according to their m/z ratios but with improved transmission efficiencies to that of a

regular quadrupole. Upon finding the presence of the most intense or abundant analytes due to the nature of the DDA mode, the HyperQuad isolates sends them into the “C-trap”, which is a “C”-shaped RF-based quadrupole where the ions are cooled and aligned using nitrogen gas.^{72, 232} The ions are then injected into the Orbitrap analyzer to acquire MS1 spectra. Besides, to acquire MS/MS data, the unit-mass resolved peptide precursor ions selected by the quadrupole are also passed through the “HCD collision cell”,⁶² which consists of higher-energy RF-only collision octapoles where the precursor ions are fragmented. The resulting product ion packets are then injected tangentially⁷² into the Orbitrap mass analyzer for detection. This process keeps repeating until the most abundant precursor ions eluting at a given time are fragmented and mass analyzed.

The Q Exactive Plus offers mass resolutions up to 140,000 at m/z 200, with mass accuracies < 3 ppm in a range between 50-6000 m/z . Furthermore, because high-collision energies generate more fragment ions that can be measured with high mass resolution and accuracies in the Orbitrap, the identification of peptides and proteins in a complex mixture by database search algorithms is improved.

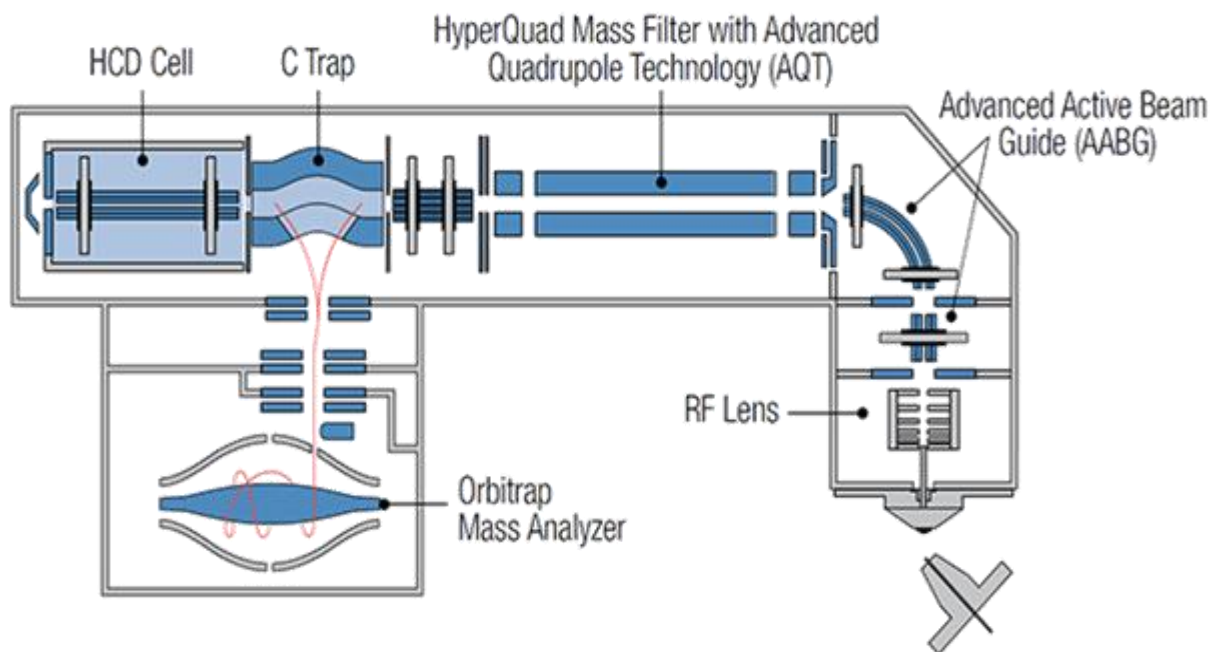


Figure 2.7 Schematic of the Q Exactive Plus mass spectrometer used for global proteomics measurements via DDA in Chapters 3-5. Image taken from the hardware manual provided by Thermo Scientific.

2.6 Fundamentals of tandem MS (MS/MS).

Tandem mass spectrometry, commonly abbreviated as MS/MS or MS²,^{8, 38} involves the mass analyses of precursor ions as well as of their fragment or product ions in regular mass analyzers like ion traps, or in hybrid instruments like a Q Exactive. The process is also described as the acquisition of MS1 spectra from precursor ions of interest and of their fragment ions or MS2 spectra. MS/MS improves the specificity of peptide detection in bottom-up proteomics experiments and enhances the signal-to-noise ratio of the collected spectra. When MS/MS spectra is acquired in mass analyzers that allow the collection of ions over time and their fragmentation like ion traps, the process is referred as “tandem-in-time”; whereas in instruments like QqQ, in which ions are not accumulated over time and fragmentation occurs in a different mass analyzer, the process is referred as “tandem-in-space”. There are multiple methods to induce the fragmentation of peptide precursor ions in a mass spectrometer but the most common ones used for structural elucidation of peptides obtained from protease digests are: (1) collisional induced dissociation (CID),⁶² in older instruments like ion traps or TOF analyzers, and (2) higher energy collisional dissociation (HCD) in more newer instruments like the hybrid Q Exactive Plus.

In CID, precursor peptide ions are allowed to collide with inert gas atoms (*i.e.*, Helium, nitrogen or argon) in an ultrafast collision event ($\sim 10^{-15}$ sec).^{62, 233} Some of the kinetic energy is converted into internal vibrational energy, which ultimately (and preferentially) breaks the amide bonds in peptides. This type of fragmentation produces series of *b* and *y* fragment ions on the structures of peptides that can be useful to determine their amino acid sequences (**Figure 2.8**).^{38, 62, 234} In HCD fragmentations, helium is preferentially used as the collision gas and the collisions impart several thousands of electron-volts of energy instead of the <100 eV in CID.⁶² Generally speaking, more informative and complex spectra is obtained from HCD, because of a wider range of potential fragmentation pathways.⁶²

Other types of fragmentation that are employed in bottom-up proteomics but are used for specialized applications such as to determine the location of PTMs in peptide sequences or proteins are electron-capture dissociation (ECD) and electron-transfer dissociation (ETD). Both of these techniques are based on the interactions of multiple-charge peptides ions with low energy electrons (~ 1 eV) to produce the cleavage of the N-C α bond in peptides.²³⁵⁻²³⁷ In ECD, the low energy low-energy free electrons are generally generated from an electron-emitting surface, while in ETD, electrons are transferred to protonated peptides from radical anions of compounds such as fluoranthene.^{62, 72} ETD and ECD produce *c* and *z* fragment ion structures from precursor peptides while keeping PTMs intact.⁶²

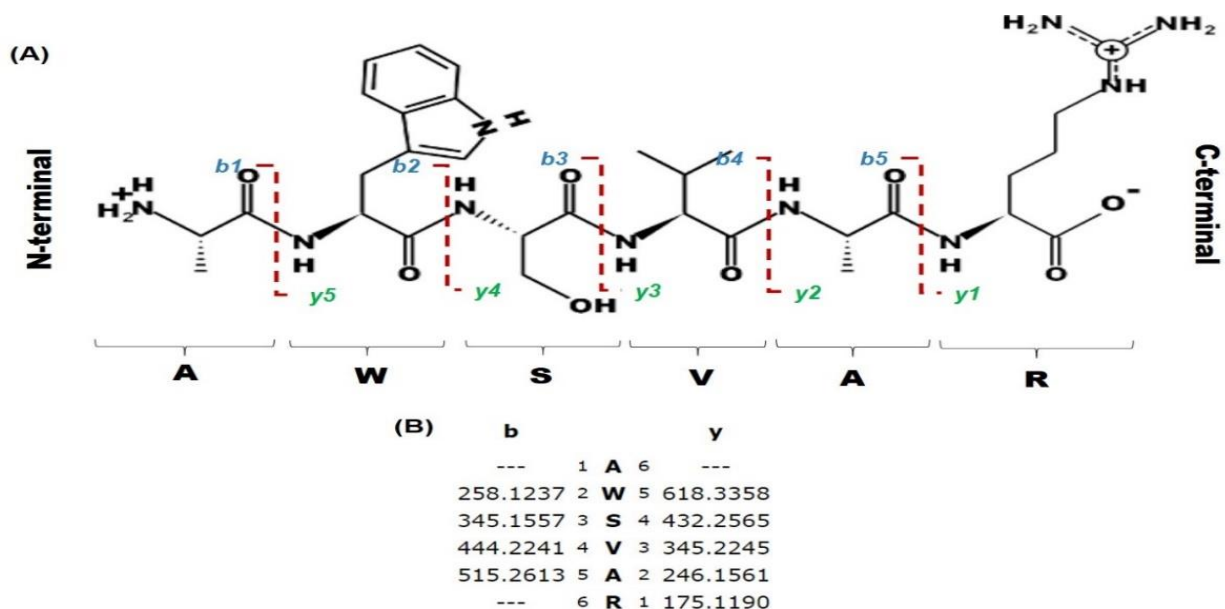


Figure 2.8 Example of the most common fragment ions observed and used during the analysis of peptides/proteins by bottom-up proteomics. (A) The backbone of the bovine serum albumin tryptic peptide “AWSVAR” shows the corresponding C-terminal y-ions and N-terminal b-ions that can be generated during fragmentation using collision induced dissociation (CID) or high-energy collision induced dissociation (HCD). (B) Shows the sequence and monoisotopic mass of the different fragment ions illustrated in (A), for example, the y4 fragment ion is made of amino acids “S-V-A-R” and has a mass of 432.256 Da.

2.7 Downstream processing of MS/MS data in global and targeted proteomics.

2.7.1 Introduction to database search driven algorithms in bottom-up global proteomics research.

In this dissertation, two approaches to analyze global proteomics data from bottom-up proteomics pipelines were used: Database search strategies conducted through the application of different software for the projects described in Chapters 3 and 4; and *de novo*-assisted database searches through the PEAKS software, used in Chapter 5 (**Figure 2.9**). Descriptions of each strategy and software package employed are presented in the following sections.

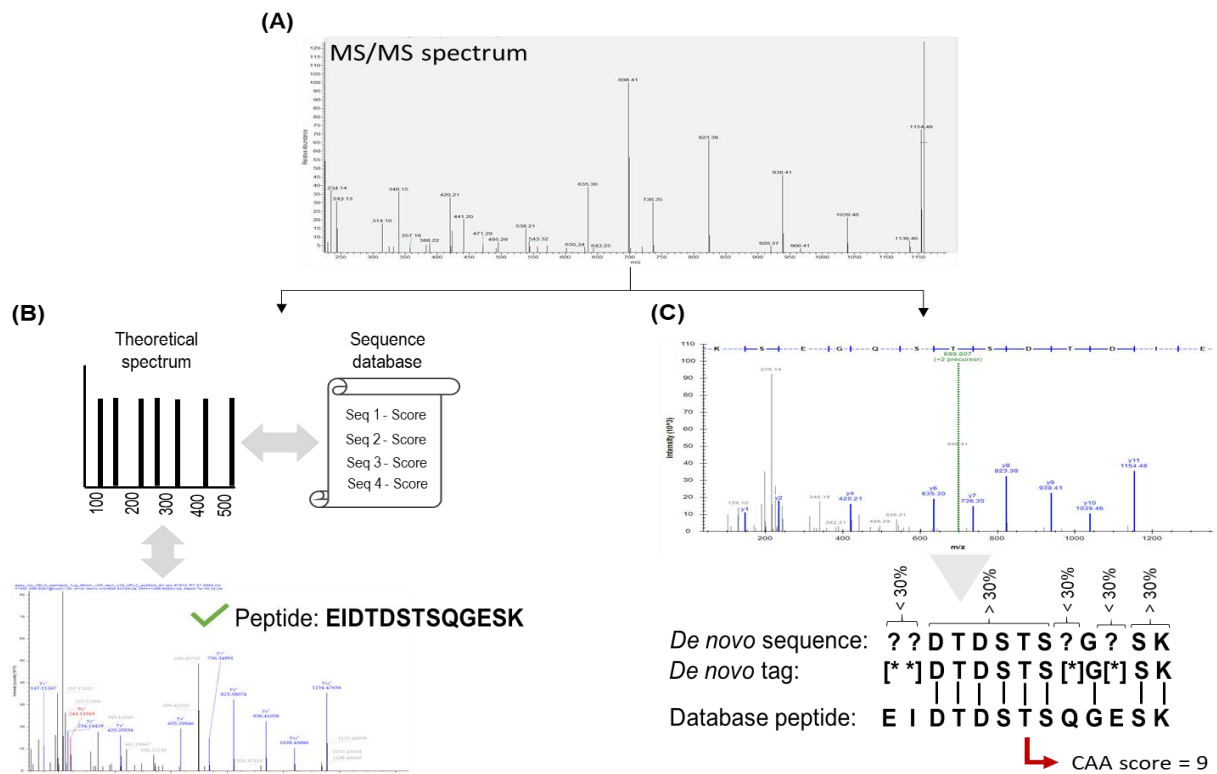


Figure 2.9 Peptide identification strategies used in this dissertation. (A) An acquired MS/MS peptide spectrum in a bottom-up proteomics run. (B) Peptide identification performed by database searching where experimental MS/MS spectra is correlated with theoretical spectra predicted for each peptide contained in a protein sequences database. This strategy was employed in Chapters 3 and 4. (C) The *de novo*-assisted database searching strategy of the software PEAKS was used in Chapter 5. Through a sophisticated algorithm, PEAKS computes a *de novo* sequence with local confidence score on each amino acid, represented here by the percentages shown above each residue. Setting up a confidence threshold of < 30%, for example, low confidence amino acids are substituted by their mass values, represented here with *. The sequence tag is then mapped to a protein database. The CAA score reported by the software is the number of common amino acids in the alignment.

2.7.1.1 Peptide identification by database searching.

The manual interpretation of peptide tandem MS spectra to derive peptide sequences is an error-prone and time-consuming process that is exacerbated with the hundreds of thousands of spectra that can be collected in global proteomics analyses of cell cultures or communities of microorganisms. These reasons have enabled the development of computational algorithms or programs that are used to directly correlate peptide ion spectral MS1 and MS/MS (MS²) data, as well as their respective intensities, to candidate peptide sequences in a database without having to interpret each MS/MS spectrum manually.

Today, more than a dozen database search algorithms, each providing technical advantages such as faster computational times, improved confidence scores, and better accessibility to users without strong computational backgrounds are available for large-scale protein identification. In addition, the database search approach fits well with the emerging database resources brought about by genome sequencing which each day keeps providing of more accurate and complete protein databases.⁴

The first algorithm or engine to identify proteins by matching MS/MS data to database sequences was SEQUEST, introduced by the Yates group in 1995.¹⁶ SEQUEST core functionalities have been adapted and improved in many other programs developed over the years,^{116, 238-242} including the ones used in this dissertation. Thus, a brief description of these core functionalities is warranted to have a better understanding of the new capabilities provided by other software.

In brief, SEQUEST (as well as other database search algorithms) take three main inputs from the user: a peak list containing the spectra to search, a sequence database (*i.e.*, a FASTA file), and the search settings to match the search to the experimental setup (*i.e.*, protease used and type of instrument employed).²⁴³ The software then performs the following steps: *in-silico* digestion of the protein database, spectrum preprocessing, matching, preliminary scoring and cross-correlation analysis.

Before the search starts, the protein database is digested *in-silico* to produce a candidate peptide list.⁴ The user has to specify digestion parameters such as maximum amino acid length and the protease that was used in the experiment, for example. Once the search starts, each experimental MS/MS spectrum in the form of instrument proprietary files is “cleaned” by keeping the top 200 most abundant *m/z* values. This process ensures that noisy spectra are removed from further analysis thus improving the performance and accuracy of the subsequent steps. Next, the algorithm uses peptide precursor *m/z* values and their MS/MS data, detected in the experiment

and registered in the RAW file, to select candidate peptides in the database. Importantly, this process is constrained to a user defined mass tolerance which impacts the number of peptide candidates that the algorithm considers for matching. *In-silico* MS/MS spectra are then generated from each of the candidate peptide sequences in the database and compared to the experimental MS/MS spectra. A preliminary score to determine how well a theoretical versus experimental spectrum match one another is computed taking into consideration the (1) the summed intensity of matched ions, (2) the continuity of each fragment ion series and (3) the percentage of ions found versus those that were expected. This preliminary score acts as a filter to select the top five hundred peptide candidate sequences to compute a cross correlation score (Xcorr) of them, which represents an average of differences between the experimental to the *in-silico* generated *m/z* values. In addition, SEQUEST also uses another score, named DeltCn, that captures the scoring difference between the lowest ranked peptide scores and the XCorr value of the best match which informs the user how well the software was able to separate the top peptide-spectrum match (PSM) compared to the second best PSM.

Around the time that SEQUEST was released there was no control of the quality of the matches that were assigned by the software. Although the algorithm was scoring the best peptide sequence match to a given experimental spectrum, it could not efficiently distinguish matches of poor quality.^{4, 244} Thus, the user had to go back to the raw spectral data and use a combination of knowledge and visual inspection to decide which matches to accept and which ones to reject. This type of approach, however, made comparisons between datasets subjective and once again time-consuming. Therefore, more practical ways to classify correct and incorrect PSMs assignments were developed.

Today, proteomic studies use statistical approaches to provide a global assessment of confidences and estimate a false discovery proportion (FDP) of false positive hits in a dataset, or those PSMs accepted by the algorithm, but that are not real hits. This FDP is more commonly known in proteomics research as the false discovery rate (FDR), although this has been regarded as an oversimplification of the term.²⁴⁵ FDRs can reflect the overall credibility of the identified PSMs in a study.²⁴⁶ In bottom up proteomics, the most commonly used and accepted method for computing the FDR is the target-decoy strategy (TD).²⁴⁷ Simply, the strategy requires that all experimental MS/MS spectra are searched against the reference protein database (target) appended with a reversed, shuffled, or randomized version of itself (decoy).²⁴⁸

The underlying assumption in TD is that the number of false PSMs in a decoy search will follow the same search score distribution as sequences matches from the target database.²⁴⁸ Therefore, an entire data set can be filtered at various search score cut-offs, and a corresponding

FDR can be computed as $\left(2 * Nd / Nt\right) * 100$, where Nd is the number of decoy peptide matches and Nt is the total number of matches.²⁴⁹ An acceptable FDR should be conservative such that it ensures that the number of false discoveries selected does not exceed the number of estimated and accepted ones²⁴⁵ with these values ranging in literature from 1% to 5%. The FDR calculation can be performed by the same PSM software or by taking the results to standalone programs like Percolator, which is a popular semi-supervised machine learning algorithm that dynamically learns to separate target from decoy peptide-spectrum matches.²⁵⁰ Once an acceptable FDR is achieved and the PSMs have been statistically validated, protein identifications can then be inferred, and if desired, quantified, from the detected peptides.

With each passing day, new search engines are being developed and improvements in computational speed and lower FDRs, especially in PTM analyses, are being implemented or have been implemented in old ones.²⁴⁴ For example, one of the newest iterations of SEQUEST, SEQUEST-HT, takes advantage of multi-core CPUs to improve its informatic speed as well as of a new algorithm to optimize peptide sequence database management, in addition, the program is able to handle fragmentation data from ETD or HCD approaches²⁵¹ bringing its capabilities up to date.

2.7.1.2 Protein inference.

After a list of confident PSMs has been obtained, the next step is to assemble the identified peptides back into proteins and provide statistical confidence levels for each of them to arrive to a meaningful biological explanation. However, this task is not trivial. Because the connectivity between peptides and proteins is lost in bottom-up proteomics pipelines, the assembly of peptides back into proteins is challenging.

The diversity of proteins sizes in any given organism, may led to a very small number of confident peptide identifications available for small proteins. In addition, and because of their different physicochemical properties, even peptides originating from the same protein are not equally likely to be identified in proteomics experiment. Although with proper statistical control at the PSM level the identification of a single peptide can be enough to provide evidence of the presence of a certain protein in a sample, this is not enough to discriminate between two proteins sharing extensive homology which is a common situation in the proteomes of higher eukaryotes organisms like plants and humans.

The presence of shared or degenerate peptides²⁵² in a proteome is particularly troublesome and has not found common ground in the proteomics community. To date, groups

around the world use a diverse set of approaches to tackle the problem of protein inference and these can be grouped in:^{61, 252} (1) Rule based strategies, such as methods that disregard shared peptides from the dataset and only consider confident peptides that are unique, or those mapped to only one protein;²⁵³ (2) combinatorial optimization algorithms, or methods that rely on constrained optimization formulations of the protein inference problem resulting, for example, in the minimal protein lists that cover some or all confidently identified peptides; and (3) probabilistic inference algorithms that formulate the problem probabilistically and assign identification probabilities for each protein in a database.

Each of these methods have their advantages and disadvantages. For example, those that only consider unique peptide identifications may see a significant loss of protein information in datasets from eukaryotic organisms which can lead to underestimations of protein content due to the presence of shared peptides originating from whole-genome duplications, protein families or alternative splicing variants. Because of this, several suggested guidelines on how to deal with the protein inference problem have been published^{254, 255} and by far the most commonly suggested strategy is the minimum set cover formulation (MSC).

In the MSC, a minimal list of proteins based on the principle of parsimony coupled with an Occam's razor constraint, is used to explain all the identified peptides in a dataset.⁶¹ This approach classifies proteins by their level of ambiguity:⁶¹ proteins that consist of only distinct peptides are classified as distinct proteins; proteins are classified as differentiable when they contain at least one peptide that is unique to that protein, as well as one or more peptides that map elsewhere in the proteome; indistinguishable proteins consist of only shared peptides and in these scenarios it is recommended to collapse all of these proteins into a single entry in the protein summary report as there is often no basis to eliminate any of them.

Like FDR software, multiple protein inference software options are available. In this dissertation, MSC approaches were used in all the programs used to derive proteins. Minimal lists have several advantages that allows a consistent calculation of the number of proteins identified in the experiment and simplifies the interpretation to only those proteins that are conclusively determined to be present in the sample.

2.7.2 Overview of database search engines used in this dissertation.

2.7.2.1 Myrimatch-IDPicker.

A computational pipeline consisting of the open-source software Myrimatch and IDPicker, were used to obtain global proteomics information from the experiments presented in Chapter 3. Peptide sequences were derived from the database search algorithm MyriMatch; while proteins were assembled with IDPicker, an algorithm which assembles a minimum protein list from peptide identifications filtered to a specified FDR.

Myrimatch²³⁸ was developed based on the foundations of SEQUEST but improved the matching criteria of experimental tandem spectra to theoretical spectra to provide more confident results. Its first step is a tunable preprocessing step that ranks fragment ions by their intensity and retains, by default, only 98% of ions in each scan. All the retained peaks are divided into three classes based on their intensities (*i.e.*, high, medium, and low). As such, the highest intensity class is expected to have the least number of peaks, while the lowest intensity class will have many more peaks.

The second step consists on the *in-silico* generation of theoretical MS/MS spectra from all possible peptides in a protein database. To achieve the latter, MyriMatch employs basic fragmentation rules but also a novel system for modeling fragments by considering ion charge state differences. The third step is the comparison of acquired experimental spectra to the predicted one. In this process, the measured m/z value are first matched to the predicted m/z values based on the type of instrument used and by coupling the intensity class information. For each experimental spectrum, the software examines each m/z location and computes two probabilistic scores. The “MVH score” is based on a multivariate hypergeometric distribution of hits, while the “mzFidelity score” is based on a distribution of fragment mass errors. The algorithm then uses the MVH scores to rank the candidate peptide sequences and, when needed, it uses the mzFidelity score as a tiebreaker. Finally, a cross-correlation (Xcorr) value is computed to independently validate the best PSM ranked by the MVH score.

The list of best scoring PSMs from Myrimatch is then imported to IDPicker²⁵⁶ to assemble a list of proteins. IDPicker contains a module for calculating confidently identified peptides using an FDR-based approach. The peptides that are accepted based on an FDR threshold are used to compute a minimal list of proteins.²⁵² The algorithm starts by collapsing a peptide-protein bipartite graph such that all peptides and proteins connected to the same proteins and peptides are grouping into nodes containing multiple peptides or proteins. It then finds a set of disconnected

subgraphs within a bipartite graph using a depth-first search. Finally, it performs an MSC optimization in each of those subgraphs.

2.7.2.2 CharmeRT.

Chapter 4 evaluated the performance of the computational framework CharmeRT which consists of the peptide identification algorithm MS Amanda²⁵⁷, developed specifically for high accuracy/high-resolution mass spectrometry data, and Elutator, which is a mPSM validator that improves upon the principles of Percolator. Both tools are included in Proteome Discoverer, a commercial software licensed by Thermo Fisher Scientific that comprises a selection of mass spectrometry-related tools.

MS Amanda comprises three major steps: spectrum preprocessing, matching of theoretical-to-experimental spectra, and probabilistic scoring. In the preprocessing step, precursor ion peaks are removed from the list of experimental spectra, and MS/MS spectra are cleaned by only selecting between 1 to 10 of the most intense peaks in segments of 100 Da. Afterwards, *in-silico* generated fragment ions of each candidate peptide precursor (targets and decoys) from the protein database are matched to the best experimental precursors and their fragments within specified MS1 and MS2 mass tolerances. The software then proceeds to probabilistic scoring of PSMs, which according to the authors is based on a binomial distribution function that incorporates peak intensities to determine favorable outcomes (successes) and possible outcomes (sample space) in a specific manner. Importantly, the version of MS Amanda used in the CharmeRT framework includes a “second search” approach that assigns more than one peptide precursor ion to experimental spectra and in this way, deconvolutes chimeric spectra.¹¹⁸

The validation of the multiple PSMs (mPSMs) that can be identified by MS Amanda is done through Elutator. This tool utilizes additional information not considered in Percolator, such as peptide elution retention time predictions, as well as recalibrated masses for precursors and fragment ions to improve the validation of mPSMs.¹¹⁸ According to the original paper, the combined use of Elutator with MS Amanda increased the numbers of identified PSMs by 17-51% for the first search and between 106-149% for PSMs derived from second searches in a HeLa proteome dataset compared to a version of CharmeRT in which no retention time prediction was used in Elutator. This means that the impact that auxiliary information used in Elutator has on the identification of mPSM benefits spectra of lower quality or that derived from chimeric spectra.

2.7.2.3 *De novo*-assisted peptide database searches with PEAKS.

Alternative to the popular database searching algorithms to identify peptides and proteins in a bottom-up global proteomics experiment, the deduction of amino acid sequences from MS/MS spectra can also be conducted via *de novo* algorithms. *De novo* peptide sequencing is independent of protein databases and is analogous as to search for a peptide in a space containing all possible peptides.²⁴⁹ Historically, *de novo* algorithms have been deemed to give less accurate results than their database counterparts, as well as to be less efficient due to the non-restricted search space that comes with a cost in computational time and resources. However, with the introduction of high-resolution MS instruments, the performance of *de novo* sequencing software has been significantly improved.

One of those software packages is PEAKS,¹⁴⁹ a commercial software released in 2002 by Bioinformatic Solutions and that has become an industrial standard for automated *de novo* sequencing.²⁴⁹ PEAKS has the capability of incorporating *de novo* sequencing results into database searches.¹⁴⁹ According to its developers, the combination of *de novo* sequencing to improve the creation of a preliminary list of protein or peptide candidates with an improved scoring function results in significantly improved sensitivity and accuracy. This mode of operation, more commonly known in the field as *de novo*-assisted database search, was employed in Chapter 5 of this dissertation to identify small endogenous proteins (≤ 100 amino acids) and/or their peptide products in a plant-microbial interface system. Although more details are described in Chapter 5, it is important to state right now that these biomolecules were not submitted to any *in vitro* enzymatic digestion treatment before (*i.e.*, with trypsin). Avoiding this step means that their identification by MS/MS captures important cellular processes such as endogenous posttranslational processing.

In particular, combining the predictability of *de novo* searches to identify small endogenous proteins or peptides in any biological system is advantageous due to the following reasons: (1) There is no need to generate an *in-silico* peptide library via regular database search that can cover potential or expected enzymatic cleavages occurring naturally in a system. Database search algorithms are particularly not efficient in the computational resources they use when conducting an “open-search” to generate multiple versions of proteins (*i.e.*, digested with trypsin, chymotrypsin, LysC, etc); and (2), since small protein products may not be well represented in a database, having information that is not biased towards the use of one may help identify novel proteins for analysis.

In total, PEAKS DB performs six algorithm steps to identify peptides from tandem MS spectra and this has been summarized with ease and described in more detail in the original publication to which the reader is referred to.¹⁴⁹

2.7.3 Targeted proteomics data analyses with Skyline.

In Chapter 3, a targeted proteomics methodology was developed and applied to groundwater samples to investigate the presence of bacterial biomarkers of interest. Targeted proteomics has gained significant attention in proteomics research due to the specificity, sensitivity and robustness that it brings to the identification and, most importantly, absolute quantification of predetermined peptides acting as surrogates for specific proteins (targets) that are present within a large dynamic concentration range in complex matrices.^{103, 258} Thus, even if bottom-up global and targeted proteomics experiments measure peptides in a sample, the type of data acquired from each experiment looks quite different and needs to be processed and interpreted by computational pipelines adapted to their particular requirements.

Technological advances in mass spectrometry instrumentation and the promise that global proteomics had for systems biology investigations, led to the early emergence of software packages that were specifically intended to deal with the complexity of the spectral data acquired through DDA approaches. Thus, when targeted proteomics investigations started to have an active role in research, software tools to analyze the spectral data collected from these types of experiments lagged behind. Although early efforts tried to breach the gap in bioinformatic developments, the tools that came out of it were limited in scope and the field was mainly dominated by proprietary software specific to a particular instrument vendor.²⁵⁹ Hence, to tackle these issues and provide users with an open-source alternative that at the same time could facilitate sharing of data files and methods between multiple instruments, the MacCoss Laboratory at the University of Washington started the development of the Skyline project.

Skyline²⁵⁹ is a windows client application for building and analyzing selected reaction monitoring (SRM)/multiple reaction monitoring (MRM), parallel reaction monitoring (PRM – targeted MS/MS and DIA/SWATH), and targeted DDA data. Over the last ten years, the project commanded by Brendan McLean, its main developer, has become a standard in commercial and academic targeted proteomics research around the world. The appealing of the software is its focus on end user design which allows to create complete targeted proteomics pipelines all the way from method development (reviewed in Section 1.2.1.2) through data analysis, with ease.²⁶⁰ Due to its capabilities and broad adoption by the community, Skyline was used in

Chapter 3 of this dissertation to test MRM-MS methods and analyze the results obtained from them.

Without going into many details in MRM-MS method building, Skyline works like a text editor document in which a user can input protein sequences, digest them *in-silico* with a protease and generate lists of transitions, consisting of peptide ion precursors and fragment ions, that can be imported to an instrument (for this dissertation, a triple quadrupole) for acquisition via MRM-MS. Skyline also provides additional support for more advanced techniques like “scheduling” which enables the measurements of tens to hundreds of individual peptides by allowing only a subset of the targeted peptides to be measured in a given time window.²²⁸ Options to optimize collision energies for each measured peptide, so as to get adequate fragment ion intensities, are also available.

Once an MRM-MS run or sample campaign is over, Skyline can take as input multiple raw spectral files in either instrument-vendor specific or generic formats and process them sequentially or in parallel to identify the targeted peptides and calculate their peak areas. Peak detection and integration in Skyline is composed of six algorithm steps that are:²⁶⁰ (1) Chromatographic extraction, (2) Resampling, (3) Peak detection, (4) Peak grouping, (5) Peptide identification and, (6) Peak area calculation. These steps, specifically for processing of MRM-MS data, are summarized below.

First, Skyline extracts the necessary information like m/z , retention times and intensities from the measured chromatograms in each MRM-MS raw file. The success of this step depends on the adequate initial setup of the instance of Skyline that will be used to import the results, such as retention time window width or the mass accuracy of the instrument that was employed to take the measurement. Then, Skyline conducts resampling, in which it calculates an interval in a peak that captures as much MS2 information as possible. This is done through linear interpolation and normalizes the irregular collection of MS2 scans that are common in all tandem mass spectrometry data acquisition types.²⁶⁰ The resampled data are then searched for areas that represent peaks and this is performed by the Chromatogram Retention time Alignment and Warping for Differential Analysis of Data (CRAWDAD) Peaks algorithm,²⁶¹ which finds a set of local maxima, local minima and inflection points forming a viable peak.

The next step is to create peak groups for each targeted peptide and/or its modified versions (*i.e.*, heavy labeled standards). Due to the nature of the MRM data acquisition type, specific pairs of peptide precursor-fragment ions or transitions are measured over a certain period of time (*i.e.*, when it elutes from the LC), so the extracted ion chromatogram of a peptide resembles a collection of peaks that coelute at a specific time. During peak grouping, Skyline

considers similarity with apex retention time (RT), start RT, and end RT that are drawn from the local maxima and inflection points from the resampling step.²⁶⁰ After groups of peaks have been defined, the program now performs “peak picking” which is the evaluation of multiple peak groups that can represent the peptide of interest. During this step, the top 10 peak groups are assessed for seven different features (*i.e.*, log intensity, coelution count, shape score, etc) that are weighted with particular coefficients and summed to give a final score to the peak group. Finally, Skyline calculates peak areas, or areas under the curve (AUCs) for the selected peak groups and reports them in ion counts units. In this process, the software integrates the area within the peak boundaries and subtracts the background areas from it. Total area values sum the AUC values of coeluting transition chromatograms. In addition, and particularly important for absolute quantification, the software can calculate the amount of endogenous peptide present in a sample by single point calibrations to a heavy standard or by a calibration curve.

Once raw data are automatically processed, Skyline creates visual displays of the data that allow researchers to quickly inspect its quality and identify issues with it. Although automated peak detection and boundary settings are generally reliable in Skyline, and with each software update there have been refinements to the peak picking process or bug fixes, manual curation of the data ensures reliable identification and quantification of each targeted peptide. For example, the software can show detailed chromatograms for each peptide with peak boundaries and indicators for retention time and dot product correlation scores of MRM spectral data of each picked peak compared to DDA spectral data captured in the same or other instrument. Retention times and peak areas can be plotted and grouped by proteins or by replicate groups allowing to check the reproducibility of the measurements. The mass error of each selected peak, which is calculated in Skyline as a weighted mean of the mass error in all the integrated points across the annotated chromatogram,²⁶⁰ can be inspected to detect interference at the transition level. Certainly, the manual curation process of targeted proteomics data in Skyline can be a bit time consuming, however, the software makes it more amenable to the end user. Examples of this process will be shown in Chapter 3.

2.8 Post-processing of bottom-up proteomics data.

The amount of information obtained from bottom-up proteomics experiments, especially in global proteomics studies, can be overwhelming. The dissection or interpretation of the matrices or lists of peptides and proteins obtained in proteomics experiments is a multifaceted process that integrates different aspects of bioinformatics and statistics.²⁶² This is particularly true

for those studies that seek to answer questions of primary scientific interest through the interrogation of protein abundance values that can be considered as a true reflection of the state between samples. In global studies, normalization and missing value imputation are two steps that occur before applying any statistical inference test to a dataset.²⁶³

Normalization is the process of removing systemic bias in LC-MS/MS data that are introduced during the various steps of sample processing and data generation.²⁶⁴ For example, differences in the amounts of sample loaded for analysis, ionization efficiencies in complex samples, degradation of packing material in LC columns, and drifts in the performance of mass spectrometers over time,^{263, 265} all have the potential to influence the measured peptide and protein expression levels leading to erroneous conclusions from a quantitative perspective. Data normalization reduces the effect of these sources of bias. Normalization strategies have been reported to be dataset dependent,²⁶⁶ and each address systematic bias differently. For example, global adjustment techniques try to center intensity values of peptides around the global mean, median or some other fixed value in a dataset and they can correct for differences in the amount of sample loaded for analysis, but it cannot capture more complex biases.²⁶³ Instead, scatterplot smoothing techniques such as lowess regression²⁶⁷ are able to capture non-linear intensity-dependent biases and are therefore more flexible than global adjustments.²⁶³

After normalization, another key challenge in quantitative proteomics studies is to address missing peptide abundance values. These cases are more prevalent in global proteomics studies than in targeted ones, where as much as 50% of peptide values can be missing in datasets.^{268,}
²⁶⁹ There are multiple reasons behind peptide missing values but, broadly, they can be caused by biological, analytical or unknown effects. In fact, the underlying mechanisms of these types of missing values have been categorized by statisticians as independent of the value itself or missing at random (MAR) or dependent on the data, also known as not missing at random (NMAR) values,²⁶⁹ and they are not particular to proteomics but can occur in other types of techniques yielding matrices of experimental data such as in microarray-based gene expression.

Missing information in global proteomics datasets can have negative effects in the interpretation of meaningful biological data as it prevents the complete and accurate extraction of quantitative protein values and functional information.²⁶⁹ Although one can ignore missing values from the data, this option would dramatically reduce the size and completeness of it and some important biological information may escape analysis. Fortunately, however, as with the case of transcriptomics and microarray data analysis, numerous methods of data imputation for proteomics have been reported to tackle the issue of missing peptide values and have been

suggested to be contextualized according to each dataset, proportion of missing values and their nature.²⁶⁸

In the present work, the random tail imputation (RTI) approach²⁶⁹ was employed to treat missing values in global proteomics data. RTI performs well in the presence of left-censored missing values,²⁶⁸ which for MS based proteomics analysis are frequent, as peptides whose abundance are close enough to the limit of detection of the instrument have a higher rate of presenting missing values.²⁶⁸ The assumption is that the proteomics dataset can be modeled by an empirical abundance distribution and that the majority of the missing data are left-censored and can be drawn from the tail of the distribution.^{263, 269} For the global proteomics data presented in Chapters 3-5, distributions with mean and standard deviations simulating low abundance values below the noise level were computed first, and then missing values were imputed with random numbers generated from this distribution. The values imputed in such way vary thus diminishing the effects of variance underestimation.

After data has been normalized and imputed, the next step is to find protein abundance changes that can provide meaningful information to the goal of the study. For this purpose, a significance test can be carried out. Statistical tests are based on the idea that false discoveries are extremely frequent, unlike true discoveries, which are scarce. Thus, if enough data is available, a “null hypothesis”, or a standard based on false discoveries in the data, can be established and based on this null hypothesis and its distribution, the similarity between a putative discovery and the null distribution can be quantified by a test statistic. Concretely, for any given quantified protein, the null hypothesis is that of nondifferential abundance or at last that the difference in abundance can be explained by random fluctuation. Thus, by rejecting a null hypothesis and accepting the alternative hypothesis or that a protein exhibits differential expression not explained by random fluctuations, a list of significant proteins can be identified. Accordingly, the similarity between the abundance changes of a protein compared to a null distribution of abundances can be measured using any statistic of the Student family. Because the Student's statistics are difficult to interpret, they are replaced by p-values which for proteomics, values below 0.05 or 0.01 are values used to reject the null hypothesis. Depending on the comparisons, other statistical tests like ANOVA can be implemented, such as with the data presented in Chapter 5.

Finally, there are many ways to visualize the curated results from a bottom-up proteomics experiment. Qualitative data such as the number of peptides and proteins identified within replicates and conditions can be shown in bar charts,²⁷⁰ and common and/or unique sets of proteins identified in one or up to four conditions²⁷¹ can be clearly illustrated with Venn diagrams.

Quantitative proteomics data is commonly displayed through volcano plots in which the fold change ($\log_2$ transformed) is plotted versus the p-value ($-\log_{10}$ transformed).²⁷² This kind of plot highlights proteins with high fold changes and low p-values and is well suited for illustrating changes in large datasets. Individual expression levels of hundreds or even thousands of proteins across multiple conditions can be visualized by color intensity in heatmaps. Unsupervised hierarchical clustering of the data in heatmaps enables the visualization of groups of proteins which expression can be up- or downregulated in certain conditions. PCA plots are also ways to investigate underlying differences between replicates and conditions in quantitative proteomics results.²⁷³ Multiple software packages are available to researchers such as Perseus²⁶² or InfernoRDN²⁷⁴ to visualize proteomics results.

CHAPTER 3

Evaluation of the application feasibility of a LC-MRM-MS targeted proteomics approach for the detection of *Dehalococcoides mccartyi* protein biomarkers in chloroethene contaminated groundwater

Text and figures were adapted from the following published journal article:

Villalobos Solis, M.I., Abraham, P.E., Chourey, K., Swift, C.M., Löffler, F.E. and Hettich, R.L. (2019) Targeted detection of *Dehalococcoides mccartyi* microbial protein biomarkers as indicators of reductive dechlorination activity in contaminated groundwater. *Scientific Reports*, **9**.

Authors contributions: Manuel Ivan Villalobos Solis (M.I.V.S.) planned the study alongside Robert L. Hettich (R.L.H.) and Frank E. Löffler (F.E.L.). M.I.V.S. and Karuna Chourey (K.C.) collected biomass from groundwater samples. M.I.V.S. and K.C. prepared samples for global and targeted proteomics. M.I.V.S. developed the MRM assay, performed proteomics data analyses and wrote the main manuscript. Paul E. Abraham provided critical reviews and proof-read the manuscript. All authors reviewed and approved the final version of the manuscript.

Notes: Supplemental figures and tables mentioned in text are available at <https://doi.org/10.1038/s41598-019-46901-6> and their numbering reflects that of the online material.

3.1 Introduction.

Dehalococcoides mccartyi (*Dhc*) bacteria are key players in bioremediation strategies for groundwater aquifers contaminated with the industrial solvents tetrachloroethene (PCE) and trichloroethene (TCE), which are common groundwater pollutants classified as toxic and carcinogenic to humans. Specialized *Dhc* bacteria grow under anoxic conditions by deriving energy from the reductive dechlorination of PCE, TCE, and the degradation products *cis*-1,2-dichloroethene (*cis*-DCE) and vinyl chloride (VC) to ultimately yield environmentally benign ethene.²⁷⁵⁻²⁷⁹ The ability of some *Dhc* strains to detoxify chlorinated ethenes makes them also functionally unique compared to other bacterial groups such as *Geobacter*, *Dehalobacter*, *Desulfitobacterium*, *Sulfurospirillum*, which are not able to reduce PCE beyond *cis*-DCE.²⁷⁶

Various *Dhc* strains have been maintained in axenic cultures or in consortia supplied with a chlorinated ethene as electron acceptor, and several reductive dehalogenase (RDase) genes and their products have been identified as biomarkers of dechlorination activity.^{33, 278, 280} Quantitative polymerase chain reaction (qPCR) measurements of the *Dhc* 16S rRNA gene and/or RDase genes in contaminated groundwater enabled comparative studies of the distribution and abundance of *Dhc* strains and RDase genes in response to bioremediation treatment (*i.e.*, bioaugmentation and/or biostimulation).²⁸¹⁻²⁸³ However, as with other existing nucleic acid-based

measurement approaches, the challenge of qPCR measurements of 16S rRNA gene and/or specific RDase genes is their inability to reveal the actual metabolic activity. Several studies have demonstrated that the sole presence of RDase genes and also their transcripts does not necessarily correlate with dechlorination activity.²⁸⁴⁻²⁸⁷

Consequently, proteomics approaches to measure the expression levels of proteins involved in the reductive dechlorination processes of *Dhc* has been gaining traction in recent years. One of these approaches is targeted proteomics via liquid chromatography-multiple reaction monitoring-mass spectrometry (LC-MRM-MS), which enables the absolute quantification of proteins of interest by measuring proteotypic peptides derived from their enzymatic digestion. Although the application of LC-MRM-MS to pure and mixed cultures of *Dhc* has been demonstrated, its utility to groundwater collected from sites impacted with chlorinated ethenes remains to be proven.^{33, 103, 288} Herein, we aimed to test the feasibility of developing and implementing a targeted proteomics approach via LC-MRM-MS for the detection of *Dhc* proteins to inform about *Dhc* reductive dechlorination activity in groundwater samples from sites impacted with chlorinated ethenes.

To effectively track the presence of the targeted *Dhc* biomarker proteins, we first selected candidate proteotypic peptide sequences observed in high-mass-accuracy/high-resolution global proteomics datasets of actively dechlorinating pure cultures of *Dhc* strains 195, FL2, and BAV1. After signal evaluation of the selected peptides by LC-MRM-MS on a triple quadrupole mass spectrometer, the most robust and reproducible transitions (pairs of peptide precursor and fragment ions m/z values) were used to detect the targeted *Dhc* proteins in groundwater collected from six geographically distinct locations. Peptide identifications in groundwater samples were supported by comparing their fragmentation profiles to those obtained from pure cultures or, in several cases, by comparing fragmentation profiles and retention times to spiked-in, unlabeled, synthetic peptide standards. Furthermore, 16S rRNA gene qPCR and global proteomics measurements performed for each groundwater sample allowed a comparative assessment with the LC-MRM-MS data.

3.2 Experimental procedures.

3.2.1 *Dehalococcoides mccartyi* (Dhc) cultures and growth conditions.

Biological duplicates of actively dechlorinating axenic cultures of *Dhc* strains 195 and FL2, known to express the RDase TceA (TCE→VC and ethene), as well as strain BAV1, which

expresses the BvcA RDase (DCEs→VC→ethene), were prepared and used to monitor the abundances of the targeted *Dhc* proteins in both global and targeted proteomics measurements. For the purpose of method development, the targeted proteomics approach also used a culture of the PCE-to-ethene dechlorinating Bio-Dechlor INOCULUM (BDI) consortium known to contain several *Dhc* strains and a PCE-to-*cis*-DCE-dechlorinating *Geobacter lovleyi* strain,²⁷⁷ amended with PCE as electron acceptor. Cultures were grown in completely synthetic, defined mineral salts medium as previously described.²⁸⁹ Approximately 100 mL of culture ($\sim 1 \times 10^{10}$ cells) were passed through Sterivex 0.22 μm filter units (EMD Millipore Corporation, Billerica, MA, USA) to collect the biomass. Filters were stored at -80°C prior to protein extraction and digestion. *Dhc* cell numbers were calculated by qPCR measurements of 16S rRNA genes as previously described.²⁸⁰ The known *Dhc* genomes contain a single copy of the 16S rRNA gene and RDase genes, and the gene copies were measured with qPCR equal the *Dhc* cell numbers.²⁹⁰

3.2.2 Biomass recovery from groundwater samples.

Biomass was obtained from volumes of groundwater samples collected from sites impacted with chlorinated ethenes. Sample 33NA-4 (360 mL) was collected from a contaminated site in Brazil. Samples M17 (745 mL), M18 (1,350 mL), 97 (1,000 mL), 116 (962 mL), and 129 (964 mL) were collected from contaminated sites in the United States. The entire volumes of groundwater were filtered through Sterivex 0.22 μm pore-size filter units to concentrate the biomass and then stored at -80°C prior to protein extraction and digestion.

3.2.3 Sample preparation for global and targeted proteomics analyses.

Filtered cells from axenic cultures of *Dhc* strains 195, FL2, BAV1 ($n=2$ biological replicates), the BDI Consortium, as well as the M17, M18, 97, 116 and 129 groundwater samples ($n=1$) were processed by adding 2 mL of SDS lysis buffer (4% SDS in 100 mM Tris-HCl, pH 8.0) to the Sterivex cartridges followed by incubation in a water bath at 97°C for 15 minutes and incubation at room temperature for 1 hour. The SDS lysis buffer was recovered and the filters rinsed once more with fresh lysis buffer. As previously described, proteins were extracted from cell lysates by trichloroacetic acid (TCA) precipitation and proteolytically digested with trypsin following denaturation and disulfide bonds being reduced and blocked.²⁹¹

Frozen filter membranes with biomass from the 33NA4 groundwater sample ($n=1$) were removed from the cartridges and cut into ~ 1 cm pieces using a sterilized razor blade and then

suspended in 5 mL of SDS lysis buffer (5% SDS in 50 mM Tris-HCl, pH 8.5; 0.15 M NaCl, 0.1 mM EDTA; 1mM MgCl₂; 50 mM DTT). Cells were heat-lysed as described earlier²⁹² and the supernatant containing the whole cell lysate transferred to new tubes. Proteins were then precipitated by TCA. Lysate mixes were centrifuged at 21000 g x 20 min to obtain a protein pellet which was washed with chilled acetone, air dried and solubilized in 6M guanidine buffer²⁹³. Following protein solubilization, proteolysis was initiated using trypsin. All peptide solutions were desalted on 200µL C₁₈ stage tips (Thermo Scientific, Waltham, MA) and stored at -80°C prior to global proteomics analysis. For targeted proteomics runs, volumes of processed samples were loaded directly onto capillary back columns and desalted off-line.

3.2.4 Global proteomics of axenic cultures and groundwater samples.

Global proteomics measurements of the axenic cultures of *Dhc* strains 195, FL2 and BAV1 ($n=2$ biological replicates), the BDI consortium ($n=3$ technical replicates), as well as groundwater samples ($n=3$ technical replicates) were obtained with an Orbitrap Q Exactive Plus mass spectrometer (Thermo Fisher Scientific, San Jose, CA) equipped with a nano-electrospray (ESI) source and interfaced with a Proxeon EASY-nLC™ 1200 system. Proteolytic peptide aliquots from pure cultures (1 µg), consortium BDI (2 µg), and groundwater samples (2 µg) were suspended in solvent A (0.1% formic acid, 2% acetonitrile) and injected onto a 75 µm inner diameter microcapillary column packed with 35 cm of Kinetex C₁₈ resin (1.7 µm, 100Å, Phenomenex, Torrance, CA). Peptides were separated using a 90 minutes gradient at a flow rate of 250 nL/min from 2 to 30% solvent B (0.1% formic acid, 80% acetonitrile), followed by an increase to 40% solvent B within 10 minutes and a 10-minute equilibration with 98% solvent A. Specific details of the MS/MS data acquisition parameters have been reported previously.²⁹⁴

Tandem MS spectra from pure cultures of *Dhc* strains 195, FL2, BAV1, and the BDI consortium culture were searched against individual or concatenated databases of *Dhc* strains downloaded from UniProt (for strains 195, GT, VS, CBDB1, BAV1 02/2017). The IGS Annotation Engine was used for structural and functional annotation of the *Dhc* strain FL2 protein sequences (<http://ae.igs.umaryland.edu/cgi/index.cgi>, Reference: PMID:21677861) and the web-based tool Manatee was used to view and download protein annotations (<http://manatee.sourceforge.net/>). The tryptic digest of the BDI consortium was searched with a database assembled from the proteomes of the six strains of *Dhc* and *Dehalobacter restrictus* DSM 9455 (**Supplementary Table S5**). Spectral data collected from groundwater samples were searched against a database

encompassing the proteomes of bacterial isolates known to coexist with *Dhc* or known to inhabit aquifer and sediments (**Supplementary Table S7**).

In addition to common contaminant proteins, the reversed protein sequences were appended and used as decoys to discern the false-discovery rate (FDR) at the spectral level. For standard database searching, the tandem fragmentation spectra (MS/MS) were searched with Myrimatch v2.2 algorithm⁴ set to parameters described before.⁵ Resulting peptide spectrum matches were then imported, filtered and organized into proteins with IDPicker v.3.16 software. To achieve a final peptide-level confidence > 99% (or false discovery rate FDR < 1%), proteins were identified with at least two distinct peptides sequences and a minimum spectra of 2 per protein.

Protein intensity values from each global proteomics dataset were calculated by summing together the MS1-level intensities of peptide precursors that were derived from IDPicker using IDPQuantify.²⁹⁵ Extracted ion chromatograms (XICs) were identified using ± 30 s lower and upper retention time tolerance and ± 10 ppm lower and upper chromatogram tolerance. Protein abundance values were normalized by dividing the protein intensity values by their length (*i.e.*, number of amino acids), performing a log2 transformation, and mean central tendency adjusted with the software platform Inferno RDN (<https://omics.pnl.gov/software/infernordn>).

Using the Perseus software,²⁶² we removed proteins in pure cultures of *Dhc* strains 195, FL2 and BAV1 that were stochastically sampled by requiring quantified proteins to be observed in both biological replicates per strain. For the BDI consortium and groundwater sample sets ($n=2$ and 3 technical replicates, respectively) proteins observed in at least one run were considered for comparison to targeted results as their sporadic identification by global proteomics may have been due to their low biological abundances and thus we could have a probability of observing them employing LC-MRM-MS. Missing values were then imputed with random numbers from a simulated Gaussian distribution of low abundant proteins (down-shift value of 2.5 and width of 0.3). All proteins identified by LC-MS/MS were clustered at > 85% amino acid sequence identity with the UClust algorithm of the analysis tool USearch v10.0.²⁹⁶ Venn diagrams were generated with the web application jvenn (<http://jvenn.toulouse.inra.fr/app/index.html>).

3.2.5 RDase phylogenetic tree construction.

To provide insight into the diversity of the RDases sequences present in the proteomes of *Dhc* strains 195, FL2, and BAV1, their phylogenetic relationships were evaluated with the software MEGA 7.²⁹⁷ A total of 52 RDases and two outgroup RDase sequences from *Desulfitobacterium*

hafniense strain Y51 (Q8L172) and *Dehalobacter restrictus* DSM 9455 (AHF10441) were aligned with the MUSCLE algorithm.²⁹⁸ All columns in the alignment of the protein sequences containing gaps and missing data were eliminated, leaving a total of 56 amino acid positions in the final dataset. A phylogenetic tree using the Maximum Likelihood algorithm based on the JTT matrix-based model was then constructed.^{297, 299} Initial trees for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distance estimates using a JTT model, and then selecting the topology with superior log likelihood value. Estimation of the relative confidence scores in phylogenetic groups were determined by using 1000 bootstrap replications of the data set.³⁰⁰ The tree was rooted with the outgroup RDase sequences from *Desulfitobacterium hafniense* strain Y51 and *Dehalobacter restrictus* DSM 9455.

3.2.6 LC-MRM-MS method development.

Initial lists of peptides (7-18 amino acids, without Methionine residues) and their transitions (+2 charged precursors, singly charged y_3 to terminal y -fragment series) from the targeted proteins identified in data-dependent global proteomics analyses of axenic cultures of strains 195, FL2 and BAV1, were evaluated by analyzing 500 ng, 2 μ g and 8 μ g of total tryptic digests via LC-MRM-MS.

For each measurement, peptides were loaded onto capillary back-columns (150 μ m x 120 mm) packed with ~ 50 mm Kinetex 5 μ m C18 resin and chromatographically separated on in-house pulled nanospray emitters (100 μ m x 170 mm) packed with ~ 160 mm of Kinetex 5 μ m C18 resin. Chromatographic separation consisted of a linear gradient of solvent B (70% acetonitrile, 0.1% formic acid) at 300 nL/min from 2 to 60% within 90 minutes. After each sample run, wash/re-equilibration runs were queued. The TSQ instrument was operated with a dwell time of 20 ms, scan width set at 0.002 m/z, and Q1/Q3 at 0.70 full width at half maximum (FWHM). Spray voltage and capillary temperature settings in the ion source were set at 1.75 kV and 270°C. Collision energies for each peptide were calculated using the default linear equation specific to a Thermo Scientific TSQ Ultra instrument provided in the Skyline environment. Raw LC-MRM-MS spectral data collected were imported into the software package Skyline v3.7 (<http://skyline.maccosslab.org>)^{260, 301} and the signals were manually analyzed to determine the quality of the peptide signals in an extracted ion chromatogram (XIC).

In addition, peptide sequence specificities were also assessed *in-silico* with the Tryptic Peptide Analysis tool of Unipept 3.2 and Protein BLAST searches. Briefly, The Peptidome Clustering tool of the web application Unipept 3.2³⁰² was used to compare the percentages of

pairwise similarity between the *in-silico* generated peptidomes of six *Dhc* strain proteomes databases (*i.e.*, *Dhc* strains 195, FL2, VS, GT, BAV1, and CBDB1) against the peptidomes of representative bacterial isolates that have been obtained from groundwater, aquifer, sediment, or soil (see SI for additional information). Peptidome similarity percentages were calculated based on the minimum similarity method and then clustered by the UPGMA algorithm. To assess if other protein records stored either at UniProt or NCBI could produce the selected peptides before monitoring in groundwater, *in-silico* specificities were evaluated using the Tryptic Peptide Analysis tool of UniPept 3.2 (equating isoleucine and leucine residues) and Protein BLAST searches against non-redundant protein sequences (replacing the N-terminus of each peptide with either K or R residues, respectively). Peptides were deemed as *Dhc*-specific if they were not found in the proteins of any other bacterial species by means of both *in-silico* searches.

3.2.7 Analyses of the commercial bioaugmentation culture BDI and groundwater samples by LC-MRM-MS.

Peptide and transitions signals selected before were monitored in technical triplicate runs of consortium BDI and groundwater samples using the same LC-MRM-MS setup as for pure cultures. Amounts of tryptic digests analyzed were 4 µg of BDI sample; 10 µg of M17, M18, 97 and 129 groundwater samples; and 20 µg of groundwater samples 33NA4 and 116. To validate peptide identifications in complex samples, we required the following criteria: (A) Co-elution of all selected transitions per peptide; (B) average dot-product (dotp) correlation scores > 0.80 for transition intensity ratios between the signals detected in groundwater to those observed in the respective pure culture; and (C) peptide signal reproducibility in all technical runs. (D) For a subset of the target proteins (n=6), we required strong agreement between the average transition ratios (dotp > 0.80) and retention times (≤ 3 mins differences) of the endogenous peptides with spiked-in synthetic unlabeled peptide standards.

For the six target proteins, a collective set of 11 synthetic unlabeled peptide standards were purchased as purified lyophilized solids (>95%, Thermo Scientific, Waltham, MA) and reconstituted to standard solutions in solvent A (peptides marked with Δ in **Supplementary Table S3**). A total of 5 pmol of each peptide standard were spiked to 4 µg of BDI sample; 10 µg of M17, M18, 97 and 129 groundwater samples; and 20 µg of groundwater samples 33NA4 and 116. For additional validation, high-mass-accuracy / high-resolution global proteomics data filtered at a peptide false discovery rate (FDR) level < 1% were used to verify the presence of the target peptides and proteins in the groundwater samples.

3.3 Results & Discussions.

3.3.1 Global proteomics exploration of axenic *Dhc* cultures enables selection of candidate peptides for targeted proteomics.

The global proteomics datasets (**Supplementary Table S1**) collected from measurements of pure cultures of *Dhc* strains 195, FL2, and BAV1 provided a set of candidate peptide sequences from housekeeping and reductive dechlorination biomarker proteins (**Table 3.1**) that were the starting point for the targeted method development (**Figure 3.1**).^{33, 103, 288}

Table 3.1 *Dhc* protein biomarkers used as initial targets in this study.

Targeted biomarker [Designation]	Biomarker description	Strain 195 ^a	Strain FL2 ^b	Strain BAV1 ^a
60 kDa chaperonin [GroEL]	Housekeeping protein. Informs presence of <i>Dhc</i> .	Q3Z6L3	demc_1274	ABQ17815
Formate dehydrogenase, alpha subunit [FdhA] ^c	General marker of active dechlorination processes.	Q3ZA14	demc_808	ABQ16756
Trichloroethene reductive dehalogenase [TceA]	Process specific marker of active dechlorination (TCE→VC)	Q3ZAB8	demc_738	✕ ^d
Vinyl chloride reductive dehalogenase [BvcA]	Process specific marker of active dechlorination (DCEs, VC → Ethene)	✕	✕	ABQ17429
Elongation factor Tu [EF-TU]	Housekeeping protein. General activity/presence of <i>Dhc</i> .	Q3Z7S9	demc_108	ABQ17463
Ribosomal protein L7/L12 [rpL7/L12]	Housekeeping protein. General activity/presence of <i>Dhc</i> .	Q3Z7T6	demc_114	ABQ17470
BNR/Asp-box repeat domain protein [S-layer]	Structural protein. Presence of <i>Dhc</i> .	Q3Z6N3	demc_1296	ABQ17793

^aProtein database from *Dhc* strains 195 and BAV1 were downloaded from Uniprot (IDs. UP000008289 and UP000002607, respectively).^b The IGS Annotation Engine was used for structural and functional annotation of the *Dhc* strain FL2 sequences (<http://ae.igs.umaryland.edu/cgi/index.cgi>, Reference: PMID:21677861) and the web-based tool Manatee was used to view and download protein annotations (<http://manatee.sourceforge.net/>).^c *Dhc* bacteria are unable to grow using formate. Cells extracts lack any formate dehydrogenase ability. Recent work has assigned the FDH protein and its subunits an electron transfer role to the RDases during reductive dechlorination reactions. See Kublik *et al.*, (2016) for more details.^d ✕ - protein is not present in the respective proteome.

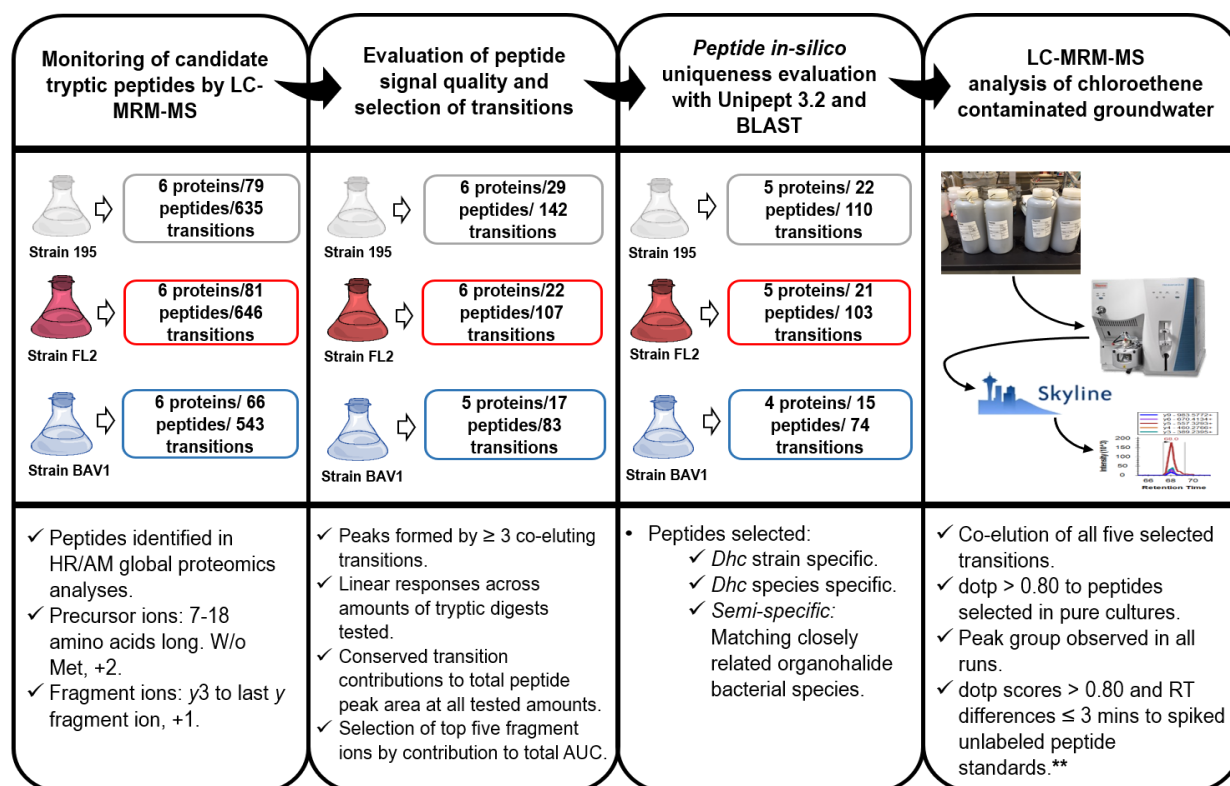


Figure 3.1 Workflow for the selection of peptide signals in pure cultures of *Dhc* strains 195, FL2 and BAV1. Each peptide peak group was submitted to a series of validation and refinement steps to identify peptide candidates having consistent fragmentation patterns, linearity in AUCs and ability to be generated upon tryptic digestion in groundwater monitoring. ** Only available for 11 targeted peptides.

Global proteomics analyses resulted in proteome coverages of 59%, 57%, and 60% for *Dhc* strains 195, FL2, and BAV1, respectively. These percentages are close to the ~60% that has been obtained before in shotgun proteomics studies of *Dhc* strains 195, CBDB1, and DCMB4.^{103, 303-305} Overall, the analytical dynamic range of the proteome measurements spanned ~5 orders of magnitude. All the targeted housekeeping and structural proteins (*i.e.*, chaperonin GroEL, S-layer associated protein) and those indicative of active dechlorination (*i.e.*, FdhA, TceA, and BvcA) ranked amongst the top 50% most abundant proteins (**Figure 3.2A**). Each biomarker was also found with similar normalized intensities within each strain dataset (**Supplementary Figure S1**). The resulting percentages of sequence coverages and the number of peptide spectrum matches per target protein demonstrated comparable efficiencies of tryptic digestion achieved between the *Dhc* strains included in the analysis (**Figures 3.2B & 3.2C**).

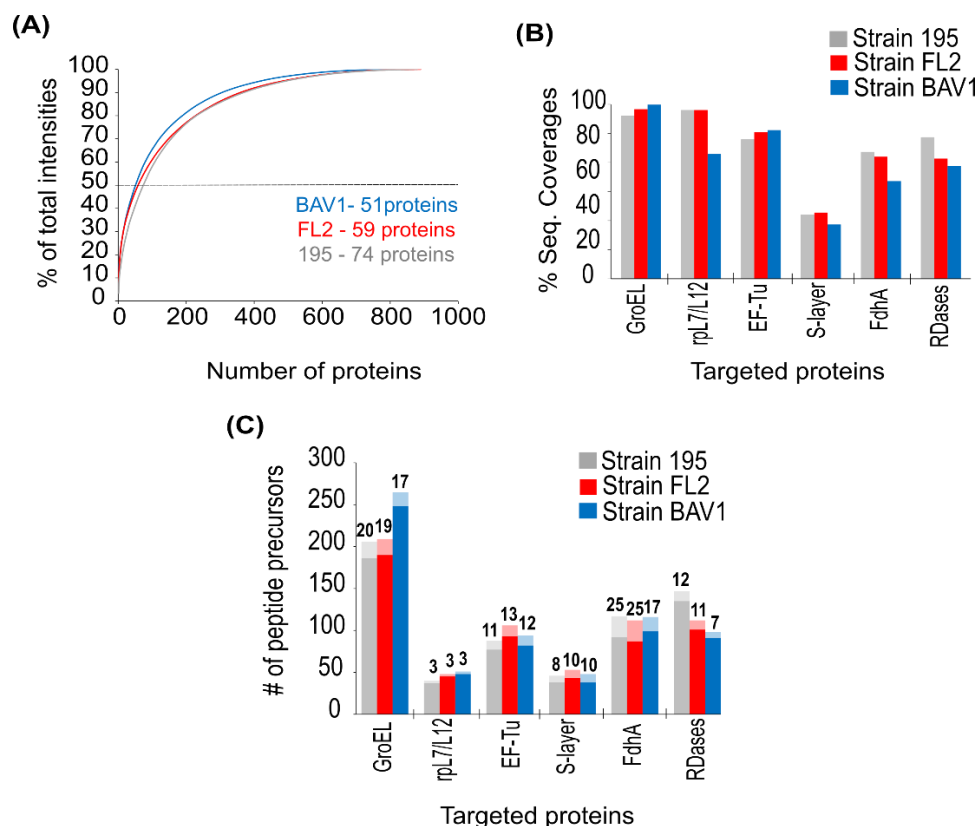


Figure 3.2 Global proteomics metrics from the analyses of three pure cultures of *Dhc*. Relative percentage of contribution to the total intensity of the proteins identified with a peptide-level confidence > 99% in the global proteomic analysis of axenic cultures of *Dehalococcoides mccartyi* strains 195, FL2 and BAV1. Targeted proteins ranked amongst the numbers of proteins contributing to half of the total measured intensities (below the dashed line, numbers next to strain names). (B) Targeted protein sequence coverages obtained from the cultures processed in this study. (C) Total number of peptide spectrum matches (PSMs) for each targeted protein. RDases per strain are homologues of TceA in strain 195 and FL2, and BvcA in strain BAV1. The numbers in the bar chart (represented by light colors) are the fraction of +2 peptide charged precursors meeting the selection criteria for LC-MRM-MS analysis described in Materials & Methods.

The expression of FdhA proteins observed in cultures of strains 195, FL2, and BAV1 is in agreement with studies that have reported high abundance values of these proteins compared to RDases and hydrogenases in *Dhc* pure and *Dhc*-containing mixed cultures that are actively dechlorinating.^{104, 105, 306-308} In addition, measured mRNA levels of the Fdh subunits have been reported to be dependent on the presence of a chlorinated electron acceptor but not on the presence of hydrogen.³⁰⁹ Recent studies of the Fdh complex of *Dhc* strain CBDB1 (iron-sulfur molybdoenzyme complex I (CISM)) revealed a tight spatial association between FdhA and the RDase CbrA (ID. CbdbA194).³¹⁰ Supported by *in-vitro* dehalogenation activity assays, these observations suggest that FdhA serves an integral role in the respiratory chain of *Dhc* (*i.e.*, FdhA

serves as an electron-channeling module between the Hup hydrogenase and the RDase),³¹¹ and as such, can serve as a general biomarker of *Dhc* dechlorination activity.

RDases enzymes are biomarkers of active dechlorination processes and can provide additional information regarding specific chlorinated compounds that undergo reductive dechlorination. The types of chlorinated compounds used by RDases makes the *Dhc* group functionally diverse.^{306, 312-314} The TceA RDase in the proteomes of strains 195 and FL2 and the BvcA RDase in the proteome of strain BAV1 were observed amongst the top five most abundant proteins in each dataset, respectively.

Sequence identities of the protein biomarkers selected in each strain were also evaluated. In total, 617 protein groups (> 85% amino acid sequence identity) were common between the three *Dhc* strains analyzed (**Figure 3.3**). These protein groups encompass homologues of the targeted GroEL, EF-TU, rpl7/L12, and FdhA proteins. Interestingly, the putative S-layer sequence of strain BAV1 and the annotated S-layer proteins of strains 195 and FL2 did not group together. The TceA homologues of strains 195 and FL2 clustered at 99% identity, while RDase BvcA was found amongst the 61 unique protein groups of strain BAV1. These observations highlight that candidate peptide sequences from protein biomarkers can target multiple *Dhc* strains or can be potentially used as strain-specific targets when monitoring mixed cultures or environmental samples.

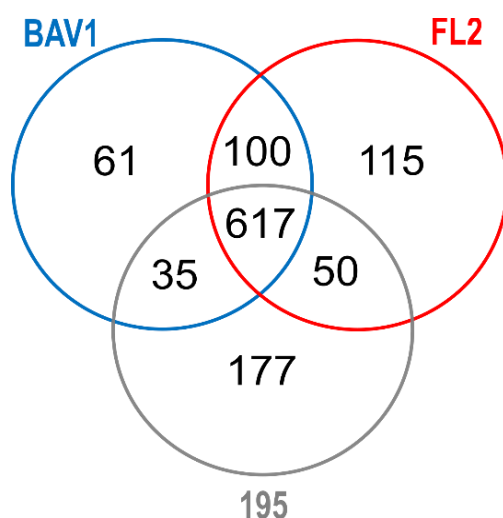


Figure 3.3 Number of protein groups (> 85% sequence identity) identified in each culture.

In addition to the TceA homologues and BvcA, 18 other RDases were identified in these shotgun proteomics measurements, albeit, at lower abundances (**Supplementary Figures S2A-C**). The protein sequence coverages of the other identified RDases were on average below 60%, except for two other RDases (demc_816 in strain FL2 and Q3Z6A6 in strain 195). The identification of multiple RDases in actively dechlorinating *Dhc* cultures is related to the various sets of RDase genes present in single *Dhc* genomes (*i.e.*, 17 *rdhA* genes in strain 195, 24 *rdhA* genes in strain FL2, and 11 *rdhA* genes in strain BAV1³⁰⁸). The co-expression of RDases by single *Dhc* strains has been reported before and has been hypothesized as a mechanism of adaptation to use naturally occurring and anthropogenic organohalogenes.^{290, 315}

To provide insight into the diversity of the expressed RDases and better validate the biomarker selection, we evaluated the phylogenetic relationships between the RDases present in the proteomes of *Dhc* strains 195, FL2, and BAV1. The selected TceA homologues from strains 195 and FL2 formed a subcluster, while the targeted BvcA did not group with any of the remaining RDases, nor did any of the second most abundant RDases in each dataset cluster with any of the targeted enzymes (Supplementary Figure S2). These results demonstrate the sequence conservation of TceA and BvcA as compared to other RDases expressed by other or the same *Dhc* strains. Moreover, the substrate ranges of TceA homologues and BvcA are known, while the participation of other RDases in reductive dechlorination reactions remains to be proven experimentally.³⁰⁸ The higher expression and sequence coverages obtained for the TceA and BvcA RDases (**Supplementary Figures S2A-S2C**) resulted on average in four times higher numbers of tryptic peptides than those obtained for other expressed RDases, which was helpful for the development of the targeted assay.

3.3.2 Selection of *Dhc* MRM-MS observable peptides and *in-silico* evaluation of their biological specificities.

Evaluation by LC-MRM-MS was conducted on 79, 81, and 66 peptides from the targeted proteins previously identified in the global proteomics datasets of cultures of *Dhc* strains 195, FL2, and BAV1, respectively. By examining three different loading amounts of total digested protein (*i.e.*, 500 ng, 2 µg, and 8 µg) and manually analyzing the data to determine the quality of the resulting peptide signals, 29 peptides and 142 transitions from the digest of strain 195, 22 peptides and 107 transitions from strain FL2, and 17 peptides and 83 transitions from strain BAV1, were selected. Examples of the type of signals chosen and discarded from these steps are shown in **Supplementary Figures S3 and S4**. From this initial selection, the top five transitions ranked by

contribution to total area under the curve (AUC) per peptide were preserved, resulting in a total of 55 peptides (unique and shared between strains) equivalent to 270 transitions.

Essential for targeted proteomics experiments is to ensure that the selected peptides uniquely identify the proteins of interest.²²⁸ The coexistence of multiple *Dhc* strains with other bacterial species in groundwater^{276, 289, 290, 316} creates a challenge for the selection of unique peptides. A preliminary *in-silico* comparison between several *Dhc* proteomes, as well as the proteomes of other organohalide-respiring bacteria commonly found in groundwater aquifers or sediments, demonstrated that *Dhc* strains shared greater similarities ($\geq 47\%$) amongst their peptidomes compared with those of other bacterial species ($\leq 4\%$ similarity between the proteomes of *Dhc* strains 195 and VS with *Dehalogenimonas lykantroporepellens* strain BL-DC-9) (**Supplementary Figure S5**). Although this analysis supported the development of a *Dhc* species-level targeted proteomic assay and its application to contaminated groundwater, the sequence specificities of each remaining peptide candidate selected were further assessed individually with the Tryptic Peptide Analysis tool of Unipept 3.2 and Protein BLAST searches. Peptides were deemed as *Dhc* specific when they were not found in any other bacterial protein sequence available in UniProt and NCBI *nr* databases, and when several strains shared the candidate peptide sequence by means of both-*in-silico* searches; as semi-specific, when they were found in proteins derived from related organohalide-respiring bacteria; and as non-specific, when they were found in proteins of non-organohalide respiring bacteria. Compiled results from these *in-silico* searches are presented in **Supplementary Table S2**.

Out of the seven peptides selected for monitoring the presence of the housekeeping chaperonin GroEL, peptide DGVITIEESR was the only one non-specific to *Dhc*, which was surprising considering that homologues of this protein are found in diverse bacterial classes.³¹⁷ From the targeted EF-TU proteins, peptide TTLTAAITR was found in more than 100 UniProt protein entries, and similar observations were made for peptide ELTSLGLK from the ribosomal protein L7/L12. The presence of peptides DGVITIEESR, TTLTAAITR, and ELTSLGLK in the proteomes of non-organohalide respiring bacteria prompted us to remove them from the list of selected peptides, which resulted in the loss of the rpL7/L12 marker protein from the assay.

Candidate peptides from the annotated FdhA (general biomarker of *Dhc* activity) and S-layer (structural housekeeping) proteins were specific to *Dhc* and in certain cases provided strain level resolution. For example, the *in-silico* analysis demonstrated that the FdhA peptides GTELISVDCR and SELEVISSLFSR were specific to *Dhc* strain 195, while peptide TDNNTNYSYINAIK was specific to the FdhA in *Dhc* strain BAV1. All peptides of the S-layer

protein were specific to five or less *Dhc* strain proteomes stored in UniProt, a useful characteristic for environmental monitoring of certain *Dhc* strains.

The expression of RDases from bacteria other than *Dhc* can complicate their *exclusive* use as specific biomarkers of *Dhc*-mediated reductive dechlorination in groundwater.²⁸⁷ For example, the *in-silico* searches of the six peptides selected in total for TceA and BvcA RDases revealed that these are also found in RDase sequences of other organohalide-respiring bacteria like *Dehalogenimonas*. Thus, the information that the shared RDase peptides selected here could provide in contaminated groundwater needs to be interpreted in concert with information from other biomarker proteins, such as FdhA, to have a more direct line of evidence that *Dhc*-specific biologically driven dechlorination is happening at a site.

Additionally, the identification of shared RDase peptides in an MRM assay, combined with other experimental measurements like *Dhc* 16S rRNA gene-targeted qPCR, can provide insights into the identity of the bacterial species carrying out dechlorination processes. Altogether, these observations suggested that a panel of protein biomarkers should be utilized for the most detailed characterization of *Dhc* mediated dechlorination processes in groundwater. **Supplementary Table S3** shows the complete list of peptides and their transition *m/z* values per protein used for LC-MRM-MS analysis of groundwater.

3.3.3 Application of the selected biomarkers for targeted proteomics analyses in the PCE-to-ethene dechlorinating consortium BDI.

Peptides and transitions selected in axenic *Dhc* cultures were initially tested in a tryptic digest of the nonmethanogenic PCE-to-ethene dechlorinating BDI consortium. BDI harbors multiple *Dhc* microorganisms, including strains BAV1, FL2, and GT.^{277, 318} The known microbial diversity of BDI allowed an easier validation of peptide identification with criteria that included, amongst others, the comparison of dot-product (dotp) correlation scores for transition intensity ratios between the signals detected in samples to those observed in pure cultures or to samples spiked with 5 pmol of internal standards (**Supplementary Table S4**).

Through LC-MRM-MS analyses, 13 peptides were identified out of the 37 that were targeted. Among these, GroEL peptides with high representation in the proteomes of multiple strains of *Dhc* were observed (**Figure 3.4**). However, we also detected the presence of the more conserved GroEL peptide LEGDEATGVSI_{VR}, which according to the UniPept searches, is only present in the proteomes of *Dhc* strains 195, KBTCE2, CG4 and KBTCE3 (**Supplementary Table S2**). Interestingly, in relation to the identification of peptide LEGDEATGVSI_{VR}, we also detected the EF-TU peptide NSFPGDEIPI_{VR}, which is specific to the proteomes of the same *Dhc* strains

as peptide LEGDEATGVSIVR, thus suggesting that these microorganisms are part of the *Dhc* population in BDI.

Active dechlorination activity was inferred through the presence of four FdhA peptides and one of the targeted TceA peptides. As with the previous cases of the GroEL peptide LEGDEATGVSIVR and the EF-TU peptide NSFPGEIPIVR, the FdhA peptide GSAGEYPVICTTVR found in the proteomes of strains 195, KBTCE2, CG4 and KBTCE3 also suggested the involvement of these strains in dechlorination processes. In fact, the additional detection of the TceA peptide YFGASSVGAIK, which acts as a marker of specific functional information (*i.e.*, TCE→VC, ethene transformations), and is also found in the proteome of *Dhc* strain 195, narrowed down the possibility that this strain was present in culture BDI.

Hence, to inquire in the evidence provided by targeted proteomics about the presence of additional but not yet recognized *Dhc* strains in consortium BDI, we decided to explore the microbial diversity of this culture by means of high-mass-accuracy /high- mass-resolution global proteomics analyses. By assembling a proteome database of other known strains of *Dhc*, the BDI spectral data indeed shown that organisms representing *Dhc* strain 195 were present this culture, as we were able to detect unique peptides matching proteins specific to certain strains (*i.e.*, to the S-layer protein of *Dhc* strain 195). The complete list of protein identifications in BDI is presented in **Supplementary Table S8**.

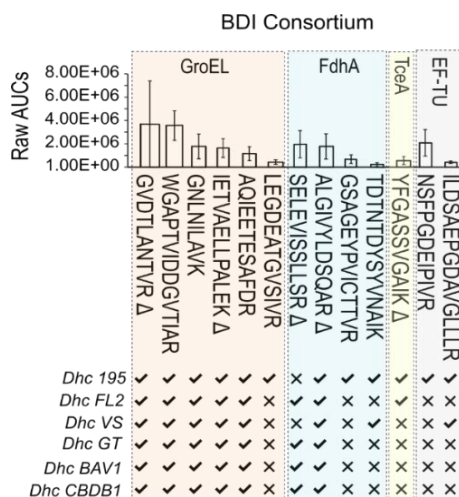


Figure 3.4 LC-MRM-MS *Dhc* biomarker identification in a tryptic digest of the PCE-to-ethene dechlorinating consortium BDI. The Fig. shows the average raw peak area under the curve (AUC) values of the targeted peptides identified in three technical replicate LC-MRM-MS runs. Error bars are the standard error of the mean. Peptides marked with Δ were identified with supporting evidence from spiked-in unlabeled standards. The inserts below the graph show the specificities of the peptides, determined in-silico, to the proteomes of the six most common isolates of *Dhc* bacteria.

Global proteomics analyses also revealed the absence of the BvcA enzyme in BDI, which agreed with the targeted results. This information was also corroborated with prior qPCR experiments showing that *Dhc* strain BAV1 carrying the *bvcA* gene was lost from consortium BDI after transfer with PCE.²⁷⁶ The lack of *Dhc* bacteria expressing BvcA in BDI seems to be compensated, however, with bacteria expressing VcrA (*i.e.*, strains GT and VS). VcrA was not targeted in the MRM assay due to the lack of pure cultures of strains GT and VS, but expression levels of this enzyme were confirmed by global proteomics in consortium BDI, where it may play a role in the dechlorination of *cis*-DCE to ethene. The involvement of microorganisms expressing VcrA, was also supported by the targeted detection of FdhA peptides matching to the proteomes of *Dhc* strains VS and GT.

3.3.4 Application of the selected biomarker panel for targeted proteomics analyses of chloroethene contaminated groundwater.

Seven groundwater samples collected from various international sites impacted with chlorinated ethenes were analyzed by targeted proteomics. Amongst the identified contaminants were TCE, *cis*-DCE and VC. These compounds are substrates and intermediates of the anaerobic reductive dechlorination reactions carried out by *Dhc* bacteria that ultimately yield ethene as the end product. Ethene was detected in these samples at various concentrations. The detection and concentrations of these chemicals provide some level of information about the degree of dechlorination in each sample and are tabulated for each groundwater sample **in Supplementary Table S6**).

qPCR measurements performed on groundwater samples M17, M18, 97, 116, and 29 (33NA4 samples for DNA extraction were not available) showed average total bacterial 16S rRNA gene copies/mL values ranging from $2.6 \times 10^7 \pm 1.4 \times 10^6$ in sample 116 to 9.8×10^5 copies/mL $\pm 4.9 \times 10^5$ in sample M18 (**Figure 3.5**). qPCR measurements of 16S rRNA genes of relevant organohalide respirators (*Dhc*, *Dehalobacter*, and *Dehalogenimonas*) demonstrated the presence of *Dhc* bacteria in all samples, with the highest abundance of *Dhc* 16S rRNA genes quantified in sample M17. At the M17 sampling location, *Dhc* represented ~20% of the total bacterial 16S rRNA genes ($2.0 \times 10^5 \pm 1.7 \times 10^3$ copies/mL). According to empirical information from bioremediation site regulators, they have found that values higher than 1×10^5 copies/mL of organohalide respirators such as *Dhc* are needed for observable dechlorination to occur. Thus, the qPCR data above suggested that M17 has appropriate *Dhc* cellular abundance for dechlorination, whereas the other samples has cellular abundances that appear to be below this minimum threshold.

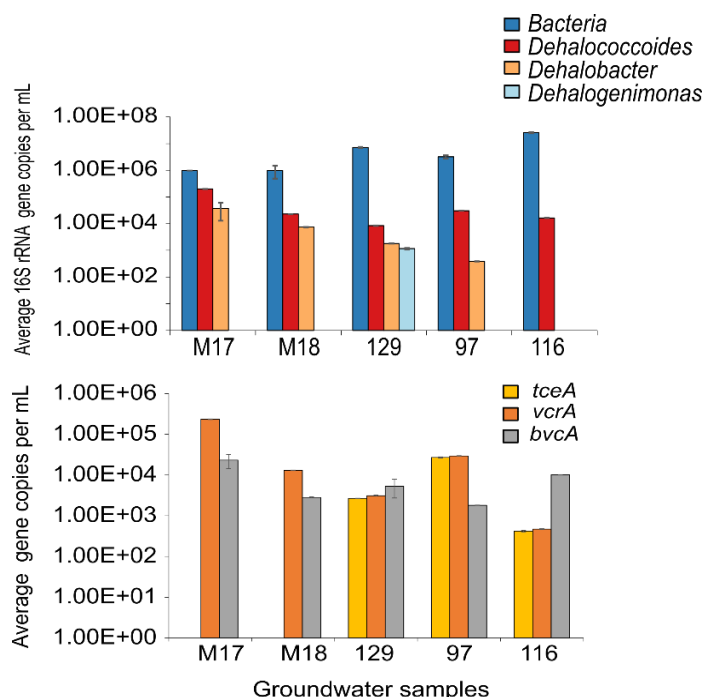


Figure 3.5 qPCR measurements of bacterial, *Dhc*, *Dehalobacter* and *Dehalogenimonas* 16S rRNA gene copy numbers. Gene copy numbers of *tceA* and *bvcA* are also shown. Values are given on a log scale and each bar represents one DNA extraction quantified in triplicate. DNA-based analyses were not performed for sample 33NA4 due to limited availability. *tceA* genes were detected but not quantifiable in samples M17 and M18.

As discussed previously, the identification of *Dhc* genes does not necessarily indicate that *Dhc* is actively dechlorinating TCE or any other chlorinated ethene. Amongst the reasons for this observation are the lack of correlation between dechlorination activity and the abundance of *Dhc* 16S rRNA genes and the variable translation rates of RDase transcripts observed in pure and mixed cultures.^{286, 287} Additionally, in groundwater samples, *Dhc* microorganisms may be present but not contributing significantly to dechlorination processes due to inhibitory mechanisms (*i.e.*, the presence of perfluoroalkyl acids³¹⁹) or competition with other organohalide-respiring bacteria having more favorable chances of growth. Due to these factors, the identification of *Dhc* protein biomarkers of dechlorination would provide more definitive information about whether *Dhc* active involvement in the dechlorination processes in these samples.

Analysis of the groundwater samples included in this study by targeted proteomics, identified *Dhc* biomarker proteins and peptides only in groundwater samples M17 and 33NA4 (**Figure 3.6**). A few of the other samples were viscous and consisted of black oily, sticky material that complicated filtering in Sterivex cartridges and potentially limited DNA and protein extraction

and subsequent measurement. This may be an issue in general sampling at some sites but was beyond the scope of this manuscript.

GroEL proteins were observed in both M17 and 33NA4 samples and were identified by peptides that are highly conserved across the proteomes of multiple *Dhc* strains, including those of the six isolates (**Figures 3.6A and 3.6B**). Besides detection of GroEL in both M17 and 33NA4 samples, targeted peptides from the housekeeping EF-TU and structural S-layer biomarkers were also detected in groundwater M17 (**Figure 3.6B**). For example, the EF-TU peptide ILDTAEPGDAVGLLLR, which differs by a single threonine residue compared to the peptide identified in consortium BDI, and is present in multiple *Dhc* strains, demonstrated the utility of targeted proteomics to differentiate single amino acid changes in the sequences of the analytes. The additional detection of the S-layer peptide AGIIDVPATADDAK in sample M17, which is found in four *Dhc* proteomes, including those of strains GT and FL2, also suggested that specific *Dhc* strains were present in this sample.

Evidence of dechlorination activity was obtained by the detection of two and three FdhA peptides in samples M17 and 33NA4, respectively (**Figures 3.6A and 3.6B**). Common between both samples was the detection of the FdhA peptides ALGIVYLD SQAR and SELEVISSLLSR, which can be found in 25 and 19 *Dhc* proteomes, respectively, of the 31 *Dhc* proteomes available in UniProt (as of July 2018). Peptide ALGIVYLD SQAR has been selected as MRM target for absolute protein abundance quantification in published reports examining pure and mixed *Dhc* cultures,^{33, 103} which also points to its high conservation amongst *Dhc* strains and robust characteristics for mass spectrometric analyses. In addition to the ALGIVYLD SQAR and SELEVISSLLSR peptides, the detection of the FdhA peptide TDTNDYSYVNAIK in groundwater sample 33NA4 suggested that organisms representing *Dhc* strains 195, KBTCE2, CG4 and KBTCE3, were involved in active dechlorination.

Supporting the FdhA observations in samples M17 and 33NA4 and hence, the potential of active dechlorination, we also identified a TceA peptide in sample 33NA4 and a BvcA peptide in M17. For instance, the TceA peptide YFGASSVGAIK in sample 33NA4 (**Figure 3.6B**) suggested the involvement of *Dhc* strains expressing the *tceA* RDase (e.g., strains 195 and FL2) in the dechlorination reactions leading to the transformation of TCE to VC and ethene. Similarly, the BvcA peptide STVAATPVFNSFFR in sample M17 (**Figure 3.6A**), pointed to active transformation reactions of *cis*-DCE to ethene by strain BAV-type *Dhc*.

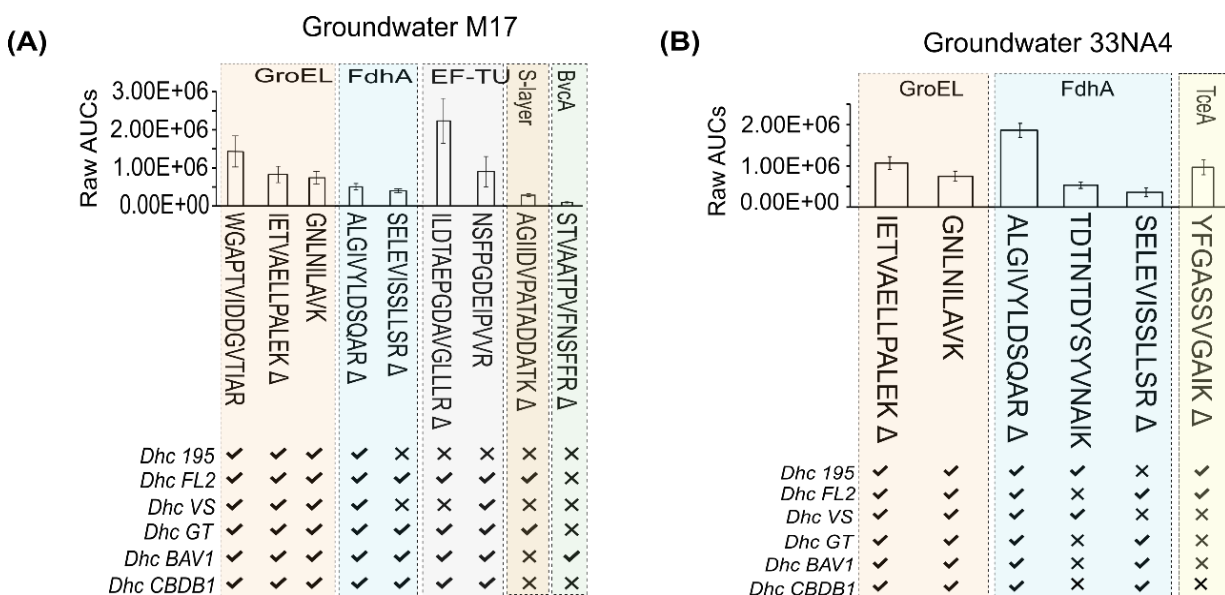


Figure 3.6 LC-MRM-MS *Dhc* biomarker identification in a tryptic digest of groundwater samples M17 and 33NA4. Average raw peak area under the curve (AUC) values of the targeted peptides identified in LC-MRM-MS runs of tryptic digests from groundwater samples (A) M17 and (B) 33NA4, respectively. Error bars are the standard error of the mean (n= 3 technical replicates). Peptides marked with Δ were identified with supporting evidence from spiked-in unlabeled standards. The inserts below each graph show the specificities of the peptides, determined in-silico, to the proteomes of the six most common isolates of *Dhc* bacteria.

The data provided by LC-MRM-MS thus contrasted with the initial qPCR information, in which *Dhc* 16S rRNA genes were detected in all groundwater samples, but peptides of the targeted proteins were not identified in four of them (M18, 97, 116, or 129). This suggested that either the targeted proteins were not expressed in these samples, the proteins were of too low abundance to be detected by targeted proteomics, or the enzymatic digestion of the proteins in a sample could have produced a different set of peptides to the ones targeted. To provide insight into these issues, high-mass-accuracy and high-mass-resolution global proteomics data was also collected. For this purpose, the proteomes of six *Dhc* isolates and other bacteria that have been isolated from aquifers or sediment material contaminated with organic chlorinated compounds were combined into a database for MS spectra search (**Supplementary Table S7**).

Global proteomics revealed that the samples having the highest numbers of *Dhc* protein identifications were samples M17 (125 groups) and 33NA4 (38 groups), in which peptides from *Dhc* biomarkers were also detected by LC-MRM-MS. Indeed, the *Dhc* dechlorination biomarkers BvcA for sample M17, TceA for 33NA4, and FdhA for both, were also identified in the global proteomics datasets (**Supplementary Table S8**). The detection of TceA in sample 33NA4 and

the absence of *Dhgm* proteins by global analyses suggested that the YFGASSVGAIK peptide detected before by targeted proteomics had a *Dhc* origin. We also observed that except for S-layer proteins that were identified by a different set of peptides to the ones targeted in samples 129 and 116, all the other *Dhc* biomarkers were not detected by means of global proteomics analyses in samples 129, 116 and 97 (**Supplementary Table S8**) which largely agreed with the targeted proteomics results. In groundwater M18, instead, *Dhc* GroEL, EF-TU and S-layer proteins were identified but with a different set of peptides. The low numbers of *Dhc* protein groups detected in samples M18, 129, 116 and 97, which included proteins that are not directly involved in mediating dechlorination processes, in combination with the aforementioned *Dhc* 16S rRNA gene data, suggested that *Dhc* cells were present but not actively dechlorinating in these samples or expressing levels of proteins that fall below the detection limits of the proteomics approach.

3.4 Conclusions.

This work demonstrates that the identification of *Dhc* biomarker proteins in contaminated groundwater through targeted proteomics is feasible. Although the approach presented here requires further optimization to provide absolute protein abundance metrics (*i.e.*, molar amounts), the panel of proteins and peptides selected should be useful for further development of a robust quantitative assay.

Successful implementation of targeted proteomics for *Dhc* containing groundwater, in comparison to pure or mixed anaerobic bacterial cultures, requires knowledge of the specificity of the peptides selected from *Dhc* biomarkers in a broader microbiological context. The *in-silico* peptidome analyses conducted in this study suggested that a panel of *Dhc* specific and semi-specific peptides (albeit, found in other bacteria with dechlorination capabilities) from proteins relevant to dechlorination activities (FdhA and RDases), should be used in concert to provide a more accurate identification of *Dhc* in samples.

Regarding this last point, environmental studies utilizing targeted proteomics to monitor the presence and infer the activity of *Dhc* microorganisms in contaminated groundwater, need to define the goal of their research – *i.e.*, is it important to know the presence of active dechlorination in general, or is it also necessary to resolve down to the species/strain level what microbe (*i.e.*, *Dhc* strain) is most active? These are distinct questions – in many cases, the former question can take precedence at sites impacted with specific chlorinated pollutants. Either way, the information provided by targeted proteomics in combination with data contributed by other well-established technologies like qPCR, can offer a more complete view of key microbes and their activities

contributing to contaminant detoxification. Results from gene-centric qPCR and proteomics will provide better guidance on bioremediation decisions, assist remediation project managers to efficiently manage remediation, and provide regulators with a relevant line of evidence that contaminant attenuation is occurring.

Contrary to regular targeted proteomics workflows, this project highlighted the use of semi-specific peptides to target dechlorination biomarkers and how it was still possible to employ them to obtain useful information regarding the presence of biomarkers in a sample. Of course, this is also related to the question under investigation, but this feature made us wonder of the capability of targeted proteomics methodologies to provide information on groups of functionally related proteins, instead of individual analytes, by making use of shared peptides between them. Provided that these groups of proteins can inform about certain biological processes that are necessary for the stable functioning of a system, or that they can inform about some aspect of the underlying biology within it, then, the first step would be to evaluate how feasible is to select shared peptides amongst these proteins. This question was evaluated in the next chapter of this dissertation.

CHAPTER 4

***In-silico* assessment of the selection of a minimal set of unique peptides for glycoside-hydrolases families for the development of a targeted proteomics methodology: Demonstrations with a global non-system specific and a system-specific approach**

Authors: Villalobos Solis, M.I., Chirania, P. and Hettich, R.L.

Authors contributions: Manuel Ivan Villalobos Solis (M.I.V.S.) planned the study alongside Payal Chirania (P.C.) and Robert L. Hettich (R.L.H.). P.C. developed all Phyton scripts to parse and condense metadata. M.I.V.S. analyzed the data, consulted references, and reported findings. R.L.H provided reviews to the study.

4.1 Introduction.

Lignocellulosic biomass that is the residual product from activities like agriculture, forestry, food-processing, industry and wastewater treatment, consists of ~ 75% of polysaccharides that can be used to produce biofuels (*i.e.*, ethanol) and biogas (*i.e.*, mixture of CO₂, methane and other trace gases).^{320, 321} Lignocellulosic biomass does not compete directly with food or feed crops for biofuel production and high biomass can be obtained with low inputs of energy, water, fertilizers, and pesticides.^{320, 322} However, the extraction of the energetic content from lignocellulosic biomass is a challenge. Lignocellulosic materials are made primarily of cellulose (35-50%), hemicellulose (15-35%), the non-polysaccharide aromatic polymer lignin (10-35%), and to a lesser extent of pectin, proteins, ash, salt and minerals.³²³⁻³²⁵ These components create a highly resistant and recalcitrant structure that is difficult to degrade chemically and enzymatically.

One approach to release the energetic content of lignocellulose consists in employing the native metabolic capabilities of aerobic and anaerobic microorganisms like fungi and bacteria to deconstruct lignocellulose into its fundamental components. Biological delignification and depolymerization of polysaccharides found in lignocellulose have been gaining momentum over the past decade due to advantages like mild reaction conditions, less energy demands, and less waste generation compared to chemical pretreatment/enzymatic hydrolysis.³²⁶⁻³²⁸

Particularly for the release of free sugars from polymers buried in lignocellulose, research has focused on the different types of carbohydrate active enzymes (CAZymes) that aerobic and anaerobic bacteria use to hydrolyze lignocellulose biomass.³²⁹ From the CAZymes, the group that are primarily involved in the breakdown of polysaccharides found in lignocellulose are bacterial

free or cell-associated glycoside hydrolases (GHs).³³⁰ GHs are enzymes that hydrolyze the glycosidic linkage between two carbohydrates or between a carbohydrate and a non-carbohydrate group.^{331, 332} The enzymatic activities of bacterial GHs in lignocellulosic environments are broadly classified in:^{331, 333, 334} (a) endoglucanases, that hydrolyze the internal glycosidic bonds along a cellulose chain; (b) exoglucanases, that act first on reducing and nonreducing ends of a cellulose chain and then hydrolyze the rest of the substrate in a processive manner; and (c) β -glucosidase, that hydrolyze only terminal, non-reducing glucose units from soluble cellodextrin and cellobiose.

The ability of bacterial communities to degrade cellulose/hemicellulose through the expression of GHs and other CAZymes has been employed in biochemical transformation technologies for bioenergy production. One of such technologies is the production of biogas from lignocellulose containing wastes (*i.e.*, woody biomass and municipal solid waste) in anaerobic digesters (ADs). In these types of bioreactors, the metabolic capacity of hydrolytic bacteria is synergistically combined with the activity from other “satellite” acidogenic, acetogenic, and methanogenic microbes.³³⁵⁻³³⁸ Hydrolytic bacteria in these consortia use CAZymes to depolymerize polysaccharides into simple sugars that fermentative bacteria use to transform to organic acids. The acidogenic bacteria then convert these acids into H₂, CO₂, and acetate, which in turn methanogenic archaea transform to methane and CO₂.^{339, 340} In return for the provision of substrate, satellite microbes may provide essential nutrients and vitamins that the cellulolytic degrading bacteria can use, or they could provide protective agents against the inhibitory effects of certain products released during the metabolic cross-talk.^{329, 337}

The success of ADs to use complex polysaccharides as source for biogas depends on the activity of cellulose-hydrolyzing bacteria.³⁴¹ Therefore, it has been recommended that anaerobic digesters start with an inoculum containing such type of bacteria.³³⁶ However, even with a starting bacterial community, physical, chemical and biological variables can impact the hydrolytic performance of a bioreactor.³⁴⁰ For example, when highly lignified or high crystalline cellulose substrates are used as feedstock in an AD, hydrolysis can underperform.³³⁹ Microbial composition in ADs is also known to be impacted by operational parameters like pH, temperature and the presence of other substances like ammonium nitrogen.^{342, 343} Therefore, it is important to understand what type of cellulose-degrading microorganisms survive in bioreactor environments and what is known about their metabolism.³³⁶

Communities of microbes living in ADs have been studied with the application of -omics measurements. Metagenomics studies have provided insights in community structure and

metabolic potential,^{339, 341-344} while metatranscriptomics and metaproteomics investigations have linked genetic information to function.³⁴⁵⁻³⁴⁸ Metagenomics studies have described the predominance of the bacterial *Firmicutes*, *Bacteroidetes* and *Proteobacteria* phyla in biogas communities and have highlighted their high polysaccharide hydrolytic potentials.³⁴⁷⁻³⁴⁹ Importantly, these studies have also demonstrated that the relative abundance of bacterial and archaeal phyla depends on process conditions.³⁵⁰ Likewise, metatranscriptomics and metaproteomics data have helped put into perspective the metabolic capacity of microbial communities. Specially, metaproteomics has been useful in observing protein patterns in ADs communities in response to environmental and operational parameters. For example, glycoside hydrolases with multiple enzymatic functions like endoglucanases, beta-galactosidases, alpha-L-fucosidases, alpha-amylases, xylanases and others have been characterized and their expression assigned to specific bacterial genera and even species.^{347-349, 351, 352}

There is no doubt that our molecular understanding of metabolic potential and function within ADs for biogas production has been greatly improved due to the combined application of omics approaches; however, they are time-consuming and labor-intensive to perform and it is difficult to envision them as routine methods to monitor the molecular-level performance of an AD. Data derived from omics techniques could be employed to develop faster and more convenient assays to inform of the molecular changes occurring in a bioreactor.

Indeed, metagenome information has been used to develop RT-qPCR biomarker assays targeting the V6-V8 regions of the 16S rRNA gene of specific bacterial taxa that can inform about the potential of these microorganisms to contribute to hydrolytic and acidogenesis processes in ADs.³⁵³ In this sense, metaproteomics has also the potential to help identify sets of proteins and their shifts in expression patterns that could be used to monitor the metabolic performance of ADs microbial communities or that could be employed as predictive indicators of process failure.^{354, 355} For example, variations in the abundance levels of methyl-coenzyme M reductase, which is a key enzyme in all methanogenic pathways and is highly expressed in biogas reactor systems,^{347, 355, 356} has been proposed as a predictive biomarker of acidification conditions.³⁵⁷

Hydrolysis of polysaccharides is an essential metabolic step in the production of biogas. Several metaproteomic studies have identified and quantified abundance levels of GHs relative to changes in the operational parameters of anaerobic digesters. In this sense, groups of GHs can act as biomarkers of the ability of a system to degrade polysaccharides. Thus, it is plausible to think that monitoring only the presence of these enzymes and their abundance changes in anaerobic digesters could provide of faster decision-making capabilities to adjust operational

parameters concerning the disposition of a digester or the microbial communities thriving within it to degrade polysaccharides. Therefore, instead of investigating the expression of every protein in an anaerobic digester, unique peptides representing groups of GHs could be detected by targeted proteomics techniques faster and with more sensitivity than by executing complete metaproteomics experiments.

A key element to any targeted proteomics study is the selection of a set of peptides that can uniquely identify proteins of interest in a biological background without being redundant to others. GHs are an extensive group of proteins classified into more than 100 families based on amino acid similarities and 3D structure.³⁵⁸⁻³⁶⁰ Because of the diversity of families and number of proteins belonging to them, in this project we specifically elaborate on the challenges associated in selecting unique peptides *in-silico* not for individual GH enzymes but for selected families of GHs. This process was demonstrated using two approaches.

The non-system specific approach involved the use of a curated database from <http://www.cazy.org/> which is the largest repository that comprises genomic, structural and biochemical information on GHs found in all kingdoms of life, as well as of others CAZymes involved in the biosynthesis and modification of carbohydrates and glycoconjugates.³⁶⁰ In contrast, a system-specific approach used a published dataset of 1401 high-quality and medium-high quality metagenome-assembled genomes (MAGs) published by Campanaro *et al.*, 2019 as part of the biogas microbiome project (<https://biogasmicrobiome.env.dtu.dk/>).³⁶¹ This dataset is the most complete repository of genomes in anaerobic digestors up to date and provides with a realistic community of microorganisms (and their proteomes) living in these environments.

4.2 Materials & Methods.

4.2.3 CAZyDB.07312019 from dbCAN2 meta site.

For the non-system approach of peptide selection, the database CAZyDB.07312019.fa stored in the automated Carbohydrate-active enzyme ANnotation (dbCAN2) meta server (<http://bcbl.unl.edu/dbCAN2/>)³⁶² was used. This is a pre-computed file with GenBank proteins annotated as CAZymes by the CAZy database. CAZyDB.07312019.fa has a total of 1,365,566 non-duplicate protein sequences including glycoside transferases (GTs), glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), auxiliary activities (AAs) and

carbohydrate-binding modules (CBMs). The same protein can have more than one annotation (*i.e.*, Protein X can have a GH 5 domain and contain several CBM modules).

4.2.4 Unique tryptic peptides from GH proteins compared to backgrounds of other CAZymes.

Protein sequences belonging to distinct GH families were removed from CAZyDB.07312019.fa. The remaining sequences were used to create independent CAZy protein backgrounds for each family. Databases of proteins in a GH family were clustered at 100% sequence identity with USearch v10.0.240²⁹⁶ (command `cluster_fast`). Target sequences and backgrounds were then loaded to the software Skyline²⁶⁰ to select unique tryptic peptides (6-25 amino acids in length, without Met or Cys residues, and excluding the first 24 N-terminal amino acids of each protein) against the backgrounds.

4.2.5 Generation of a minimum list of unique tryptic peptides for GH families.

Lists of unique peptides against backgrounds and the proteins they mapped to were ran through an *in house* developed Python 3 script to select the minimum number of peptides between proteins in a GH family considering each protein is captured only once. **Figure 4.1A** shows an example of how the script works starting with a list of seven proteins each containing different proportions of six peptides which can be shared or unique. The script first assembles groups of peptides and proteins like those shown in **Figure 4.1B** and then orders them based on groups capturing the greatest number of proteins. In this example, peptide 1, which is the most shared peptide amongst all considered proteins, is found in a total of four proteins. This group of 4 proteins that have peptide 1 is then compared to protein groups captured by other peptides. Based on this comparison, if a protein that has peptide 1 is found in a group with fewer number of proteins, the protein is removed from it. In this case, proteins are removed from peptides 2 and 3, while peptide 4 loses all its proteins and is removed from further analysis. This cycle repeats but now with the second largest group of proteins sharing a peptide which in this case is peptide 5. Following this example, the final minimum list of peptides will have peptide 1 capturing four proteins (A, D, E, F), peptide 5 capturing two proteins (C, G) and either peptides 3 or 6 capturing one protein (B) as shown in **Figure 4.1C**.

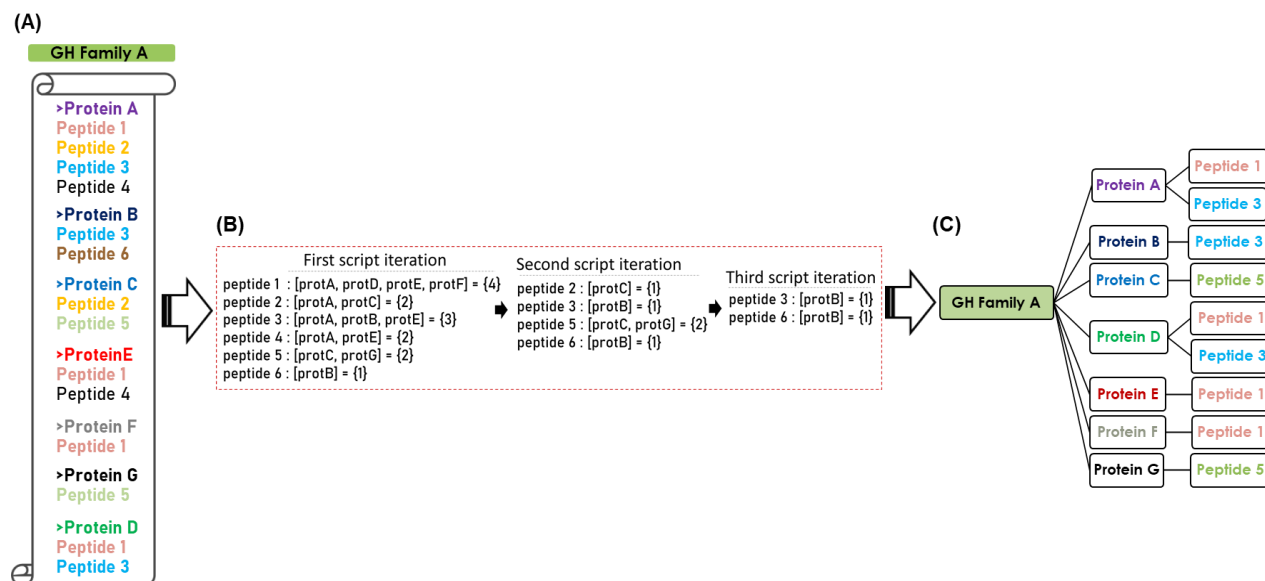


Figure 4.1 Example showing the process to select the minimum number of unique tryptic peptides and their associated number of protein seeds in different GH families. Descriptions of (A), (B) and (C) are presented in the text above this figure.

This bioinformatic approach ensures that proteins in a GH family are selected based on shared peptides that meet the selection criteria and after they are compared to peptides generated from background proteins. Even though this will not be ideal to target specific proteins sequences within a family (*i.e.*, a GH3 protein versus another GH3 protein) it ensures that the final peptide list is not overwhelmingly large and impractical for targeted proteomics applications.

4.2.6 Super kingdom taxonomy of groups of proteins having unique peptides selected from the CAZy database.

Taxonomical information at the superkingdom level of the protein sequences captured by the minimal lists of peptides obtained for different GH families in CAZyDB.07312019.fa was investigated. For this purpose, GH protein accessions and the taxa they belong to were extracted from the CAZy database (<http://www.cazy.org/>) using the *cazy-parser* tool.³⁶³ Scientific names of organisms were then used to retrieve taxa ID and lineage information from NCBI using the R package *taxonomizr* (<https://cran.r-project.org/web/packages/taxonomizr/index.html>).

4.2.7 Re-processing of 1401 MAGs in dbCAN2.

Predicted genes and coding sequences (CDS) data from 1401 bacterial and archaeal high-quality (HQ) [Completeness > 90%, Contamination < 5%] and medium-high quality (MHQ) [90% > Completeness ≥70%; 5% < Contamination < 10%] MAGs reported in Campanaro *et al.*, 2019³⁶¹ were kindly provided by the first author of the study. The biogas microbiome project was a collaborative effort in which 134 published datasets (~0.9 Tbp sequence data) derived from a wide range of different biogas reactor systems (full scale biogas plants and laboratory-scale bioreactors) fed with complex carbohydrates, proteins and lipids, were re-analyzed by means of comprehensive metagenome-centric analyses. The provided CDS were annotated using Prodigal v2.6.2 ran in normal mode. According to the Campanaro *et al.*, 2019 study these predicted proteins were used to predict CAZymes with the Carbohydrate-active enzyme ANnotation (dbCAN2) meta server (<http://bcb.unl.edu/dbCAN2/>).³⁶² Since the protein fasta files from Prodigal lacked annotation information, CAZy annotation was performed again in order to identify individual CAZymes in each MAG for further *in-silico* analyses of unique peptides. The dbCAN2 searches in this chapter were performed using the HMMER³⁶⁴, DIAMOND³⁶⁵ and Hotpep³⁶⁶ tools. Proteins annotated by ≥ 2 tools were only considered to define CAZymes. HMMER annotations took priority over DIAMOND and Hotpep tools. In case where no HMMER annotation was obtained, common annotations between DIAMOND and Hotpep were only considered, otherwise, they were discarded.

4.2.8 Selection of unique tryptic peptides from GH families in the biogas microbiome MAGS and generation of a minimum list of peptides.

Protein sequences from GH families with the highest representation across all MAGs were removed from their original proteome files. The remaining protein sequences in each MAG were used to create independent protein backgrounds for selection of unique peptides. The first 24 N-terminal amino acids from the targeted GH sequences were removed using an *in house* developed script. Targets and backgrounds were then tryptic digested *in-silico* using the prot2pept command of the Unipept³⁶⁷ command line interface (CLI) (<https://unipept.ugent.be/clidocs/prot2pept>). Tryptic peptides of 6-25 amino acids in length, and without Met or Cys residues were filtered from the resultant target peptidomes. Peptidomes from targets and backgrounds were then compared and unique peptides mapping only to targeted GH sequences selected. Lists of unique peptides against backgrounds and the proteins they mapped to were ran through the *in house* developed Python3 script to select the minimum number of peptides representing all proteins in a GH family.

4.3 Results & Discussion.

4.3.1 A minimum set of unique peptides distinguishing between GH families using sequence information from CAZyDB.

Due to the availability of online resources, and because our goal was to target families of GHs and not individual proteins, we tested an approach of *in-silico* peptide selection that retained the least number of peptides while capturing all the sequences contained in a given GH family after peptide selection and peptidome comparisons (see Materials & Methods section). In this sense, a reduced list of unique peptide sequences could, in a broad scale, differentiate between GH family groups (*i.e.*, GH family 3 vs GH family 10) in a targeted proteomics setup.

For this purpose, a curated protein fasta file from the CAZy database was downloaded. This fasta file included a total of 1,365,566 non-duplicated protein sequences including all annotated GH proteins to date as well as other groups of CAZymes. From this file, we first decided to focus on members of GH families with known and diverse enzymatic activities and mechanisms acting on the hydrolysis of celluloses and hemicelluloses (**Table 4.1**). These GH families also covered a wide distribution of organisms in all super kingdoms of life (archaea, bacteria, eukaryotes, and viruses). In addition, the number of protein sequences per GH families available in the database varied widely giving us a chance to better explore the minimum peptide set list approach. Apart from these reasons, these types of CAZy families have been commonly observed in different metaproteomics experiments that have analyzed diverse fungal and bacterial communities directly extracted from the environment or grown in laboratory settings.³⁶⁸⁻³⁷⁶

After comparison of the set of peptides derived from clustered protein sequences (seeds) of each GH family database versus those derived from the sequences of the remaining CAZymes minus the subset of proteins under analysis, we found that > 700 peptides per family are able to capture most of the protein seeds, while at the same time maintaining specificity to each family (**Figure 4.2A**). This suggests that while there is some degree of tryptic peptide redundancy within proteins in GH families, several protein seeds can be identified by single unique peptides. In fact, 26%, 27%, 37%, and 76% of the unique peptides in the GH1, GH3, GH9, and GH51, families were peptides matching to unique proteins (*i.e.*, one peptide → one protein), respectively, whereas every protein in the GH10 family was represented by a single peptide.

Table 4.1 GH families selected for testing the *in-silico* development of a minimum list of unique peptides able to differentiate between groups of enzymes using sequence information from CAZyDB.

GH family	Enzymatic activities	# of protein seeds in CAZyDB
1	β -glucosidase, β -galactosidase, 6-phospho- β -glucosidase and 6-phospho- β -galactosidase, β -mannosidase, β -D-fucosidase and β -glucuronidase	14367
3	Exo-acting β -D-glucosidases, α -L-arabinofuranosidases, β -D-xylopyranosidases, <i>N</i> -acetyl- β -D-glucosaminidases and <i>N</i> -acetyl- β -D-glucosaminide phosphorylases	16148
9	Primarily endo-glucanases and a few endo-xyloglucanases, mixed-linkage endo-glucanases and exo-beta-glucosaminidases	2548
10	Most of the enzymes are endo-beta-1,4-xylanases A few show endo-beta-1,3-xylanase activity.	3195
51	Most of the enzymes are L-arabinofuranosidase but few members are also classified as β -1,4-endoglucanases	2418

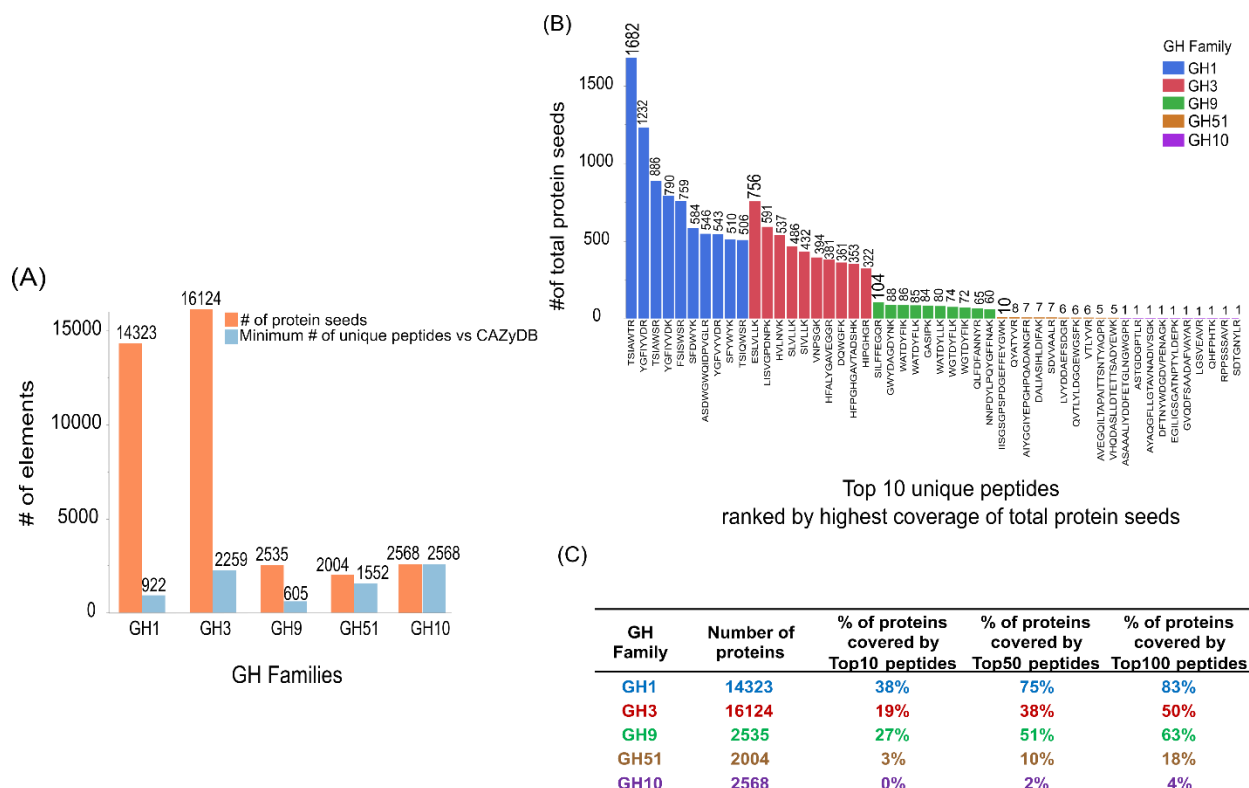


Figure 4.2 Minimum number of unique tryptic peptides and their associated number of protein seeds in different GH families. (A) Total number of unique peptides selected for each GH family after comparison to other proteins in CAZyDB and the protein seeds in which they are found. (B) Top 10 tryptic peptides (from blue bars in A) ranked by the highest number of protein seeds coverage in each family. (C) Percentages of total proteins in GH families covered by top 10, 50 and 100 peptides ranked by protein coverage.

The maximum number of total proteins that a peptide was able to capture in each of the targeted GH families varied between 1682 peptides for the GH1 family to 1 peptide for the GH10 family (**Figure 4.2B**). When categorizing peptides by % of proteins covered, we found that 100 peptides in the GH1, 3 and 9 families, covered more than 50% of all protein seeds, whereas for GH51 and 10 families, the lack of shared peptides meant that more than 100 peptides were necessary to cover at least 50% of the proteins in these groups (**Figure 4.2C**).

Unique peptides did not map to any conserved regions in GHs. For example, when GH1 proteins were analyzed with the ScanProSite tool,³⁷⁷ two conserved regions identified were an N-terminal signature and an active site motif (**Figure 4.3**). The *in-silico* approach to generate peptides in this study did not consider any of the first 24 N-terminal amino acids in proteins due to the potential presence of cleavable N-terminal signal peptides, so it was not unexpected that none of the selected peptides mapped to these regions. However, the unique peptides selected did not map to active sites either. For the cases shown in **Figure 4.3**, the active site of several of these proteins was represented by the sequence LFIVENGFG which was not part of any of the unique peptides identified. We believe that the variability of the adjacent amino acids next to active sites in GHs, as well as the number and type of potential amino acids that can be part of their active sites, may explain the lack of tryptic peptides covering this region in our approach. Besides, tryptic peptides that could be produced near active site regions of GHs did not meet adequate properties for their selection after *in-silico* digestion of proteins (*i.e.*, lengths > 25 amino acids) or even if they were present, they did not capture as many proteins as other peptides.

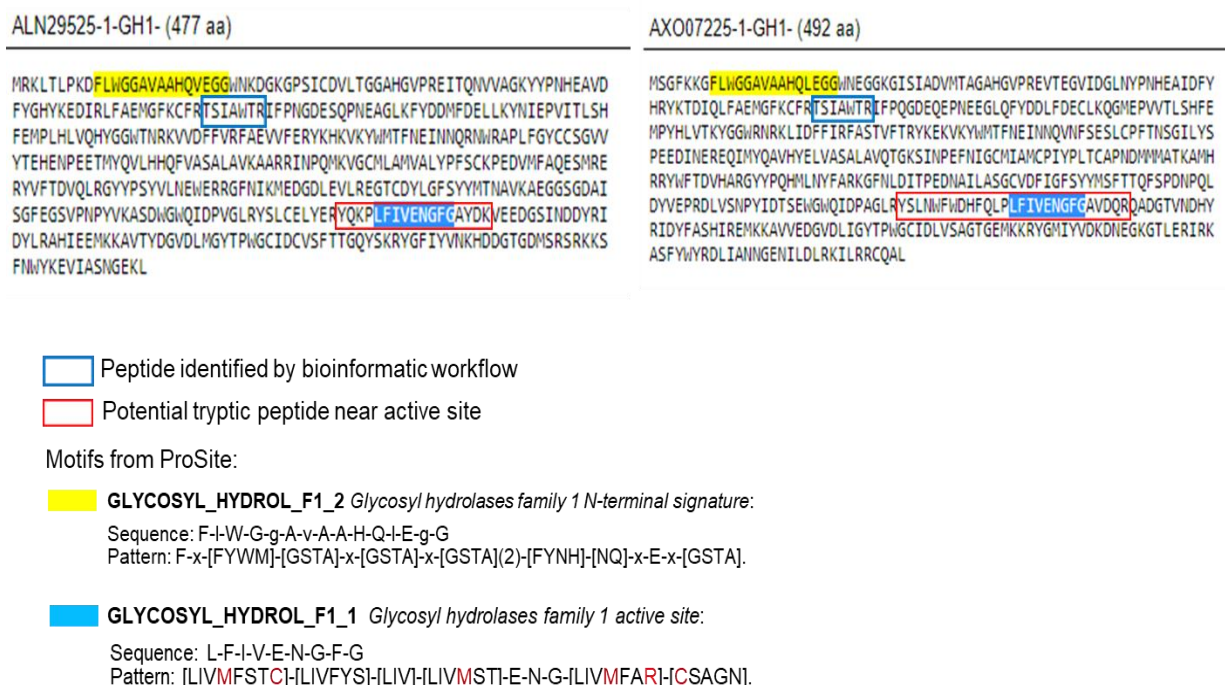


Figure 4.3 Examples of GH1 protein sequences analyzed with ProSite Scan. The unique tryptic peptide TSIWTR (in blue lined box) was amongst the 922 minimum unique peptides that covered 1682 GH1 proteins but did not map to any of the conserved sites on these sequences.

4.3.2 Unique tryptic peptides selected for group of GHs in CAZyDB can separate groups of proteins based on their superkingdom origins.

CAZyDB contains protein sequences from different domains of life including bacteria, archaea, eukaryotes, and viruses. As there may be more sequence similarity between proteins belonging to the same superkingdom, we decided to look at the number of unique peptides selected for each GH family in CAZyDB under a taxonomical perspective. Interestingly, peptides identified in each family provided a broad level of taxonomical resolution regarding groups of proteins belonging to organisms in the same superkingdom. For example, 9 of the top 10 peptides ranked by protein coverage for GH1 family were specific to bacterial GH proteins while the remaining peptide FSISWSR covered a higher percentage of GH sequences of eukaryotic origin (**Figure 4.4A**). A similar case was observed with protein sequences belonging to the GH51 family. In contrast, 9 out of the top 10 tryptic peptides in the GH9 family covered on their majority eukaryotic GH proteins, while peptide NNPDYLPQYGFFNAK was only part of bacterial GH proteins.

Based on these observations we then investigated the total amount of peptides necessary to cover proteins according to their taxonomical origins. As observed in **Figure 4.4B**, a smaller number of peptides in each of the targeted GH families can indeed provide a level of taxonomical resolution compared to the original numbers selected when broadly categorizing families of GHs. For example, if considering only bacterial GH9 proteins, the total amount of peptides needed to cover all of them is reduced to approximately half of the original number. Although the number of peptides is still high to cover bacterial proteins in certain families like in GH10, GH3, and GH51, this observation demonstrates how one can adapt the use of unique peptides for GH families to provide different levels of information in a biological system.



(B)

GH Family	Total # of peptides selected	Peptides to cover all bacterial proteins	Peptides to cover all archaea proteins	Peptides to cover all eukaryotic proteins	Peptides to cover all viral proteins	Peptides to cover all NA proteins
GH1	922	707	38	255	1	45
GH10	2568	1926	19	364	0	259
GH3	2259	2079	81	298	1	112
GH51	1552	1380	21	136	1	16
GH9	605	361	6	162	0	137

Figure 4.4 Peptide classification by taxonomy using sequence information from CAZyDB. (A) Stacked bars showing the classification of groups of proteins captured by peptides based on their taxonomical origins at the superkingdom level. **(B)** Number of peptides from the ones selected to cover proteins by superkingdoms. NA, lacks annotation at the superkingdom level.

4.3.3 Narrowing down the number of unique peptides by constraining the proteome space.

The results and observations presented from the determination of a minimum set of peptides from sequences extracted directly from CAZyDB taught us that even though protein sequences belonging to these same family of GHs do not share higher amounts of tryptic peptides between them (>700 in each case), it was encouraging to assess, that at least for some families like GH1, GH3 and GH9, 50% or more of the proteins selected in each family after peptide evaluation and comparison to background proteomes could still be covered with 100 peptides, which is a more manageable number of analytes. It was also interesting to uncover that tryptic peptides can provide some taxonomical resolution compared to others within same families which could be useful to target proteins of specific origins.

Large-scale targeted proteomics studies have used between 300-600 peptides to identify and quantify proteins of interest in a sample.^{378, 379} Although some of the numbers presented before fall within this range, to increase the chances of developing a more manageable minimal list of unique peptides for groups of GH families, the proteome space for consideration could be reduced. Besides, all the selected peptides derived from the CAZyDB analysis still need to be evaluated for uniqueness when put into a biological context. In other words, these peptides need to be compared against the peptidome of other organisms found in a sample of interest. Thus, instead of determining which other proteins in databases could also yield the tryptic peptides selected here, we turned our attention to defined and realistic microbial communities like the ones found in anaerobic digesters used for biogas production. For this purpose, data available from the biogas microbiome project was considered by applying all the bioinformatic analysis conducted so far.³⁶¹

4.3.4 Taxonomic diversity of the 1401 high-quality (HQ) and medium-high-quality (MHQ) MAGs in the biogas microbiome.

Hydrolysis of polysaccharides is an important step in the anaerobic digestion food chain happening in anaerobic digesters (ADs) using different lignocellulosic feedstock like animal manure and crop silage.^{336, 380} In the original paper by Campanero *et al.*, 2019, the coding sequences of 1401 HQ and MHQ MAGs were predicted with Prodigal v2.6.2 and then CAZymes annotated using the dbCAN2 software. Although more in-depth functional analyses of CAZymes present in the MAGs were conducted in the Campanero *et al.*, 2019 publication, such as enrichment of functional CAZyme classes to hypothesize relevant participation of microbial phyla in the

degradation of complex carbohydrates; here, we decided to focus on all these MAGs to evaluate the feasibility of selecting unique peptides for GH families in a comprehensive background of microorganisms representative of communities thriving within a biogas reactor environment.

In total, 96.1% of the MAGs provided were of bacterial origin and the remaining 3.9% were of archaeal origin (**Figure 4.5**). These percentages are common in microbial communities in biogas production. The highest number of HQ and MHQ bacterial MAGs belonged to the *Firmicutes* phylum. Amongst the bacterial phyla, several had been associated with the degradation of polysaccharides in ADs fed with lignocellulosic biomass including for example, members in the *Firmicutes*, *Bacteroidetes*, *Fibrobacter*, *Spirochaetes* and *Thermotoga*^{336, 381}. Other phyla are less known to be involved in the hydrolysis of polysaccharides, but the study by Campanero *et al.*, 2019 hypothesized, based on functional enrichment, that MAGs belonging to *Candidatus Hydrogenedentes*, the *Armatimonadetes*, *Lentisphaerae* and *Planctomycetes* are also potentially involved in this process. In fact, one metaproteomics study that characterized the microbial community of an industrial biogas reactor fed with food waste as dominant feedstock operating at thermophilic temperatures (60°C) and elevated levels of free ammonia found that *Planctomycetes* were metabolically active in the hydrolysis of carbohydrates.³⁵²

The biogas microbiome also consists of other populations of satellite microorganisms that act in concert with hydrolytic bacteria to produce methane as the end product in a digester³³⁷. Groups of bacteria in the anaerobic digestion of lignocellulose material are linked to each other as the products produced from one group serve as substrates for another. Besides hydrolysis, other metabolic mechanisms include acidogenesis, acetogenesis and methanogenesis^{335, 340}. MAGs from these other satellite microbes were also identified in Campanero *et al.*, 2019. For example, several MAGs belonged to methanogenic *Archaea*. The HQ and MHQ Archaeal MAGs found in Campanero *et al.*, 2019 were classified in five phyla, including the broad group of *Euryarchaeota*, which contained genera like *Methanobacterium*, *Methanosarcina*, *Methanoculleus* and *Methanocorpusculum*, from which several species have been isolated before.³⁸² Apart from methanogenesis, substrates like acetate, carbon dioxide and hydrogen can be used by groups of bacteria to produce alternative products, and several MAGs from these study also possessed the genetic potential to fix carbon in pathways including the phosphate acetyltransferase-acetate kinase pathway, the reductive acetyl-CoA pathway and the reductive pentose phosphate cycle.

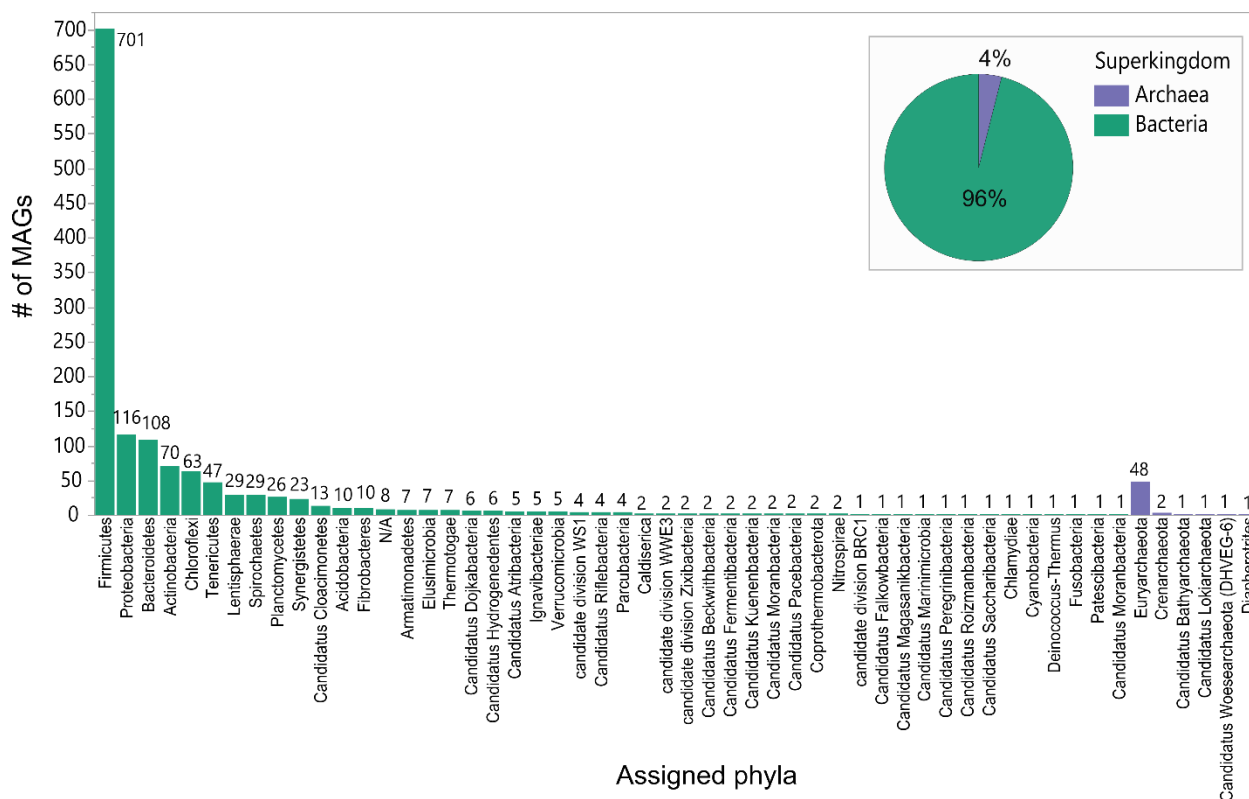


Figure 4.5 Number of HQ & MHQ MAGs identified in the biogas microbiome project. MAGs were assigned to different phyla classifications based on tiered taxonomical assignment strategy described in the original paper by Campanaro *et al.*, 2019. The inset shows the total percentages of MAGs per superkingdom. N/A- MAGs not assigned at the phylum level.

4.3.5 Distribution of GHs and other CAZymes in the proteomes of the 1401 MAGs.

The number of proteins predicted in each MAG varied considerably (**Figure 4.6**). For bacterial sequences, the *Planctomycetes* phylum had the largest median number of 4061 proteins and contained the largest bacterial proteome with 9288 predicted proteins assigned to a MAG from the *Rhodopirellula* genus. The smallest bacterial proteomes were from four MAGs in the *Parcubacteria* phylum with a median of 574 proteins. In terms of archaea the singly identified *Candidatus Lokiarchaeota* MAG had the highest number of proteins (5276) while the singly *Candidatus Woesearchaeota* (DHVEG-6) MAG had the least number of proteins (715).

Due to the distinct metabolic specialization of different microbes during the anaerobic digestion of lignocellulose material, the numbers of CAZymes found in their proteomes were expected to be variable. In total, 62,627 CAZymes were annotated across 1399 MAGs.

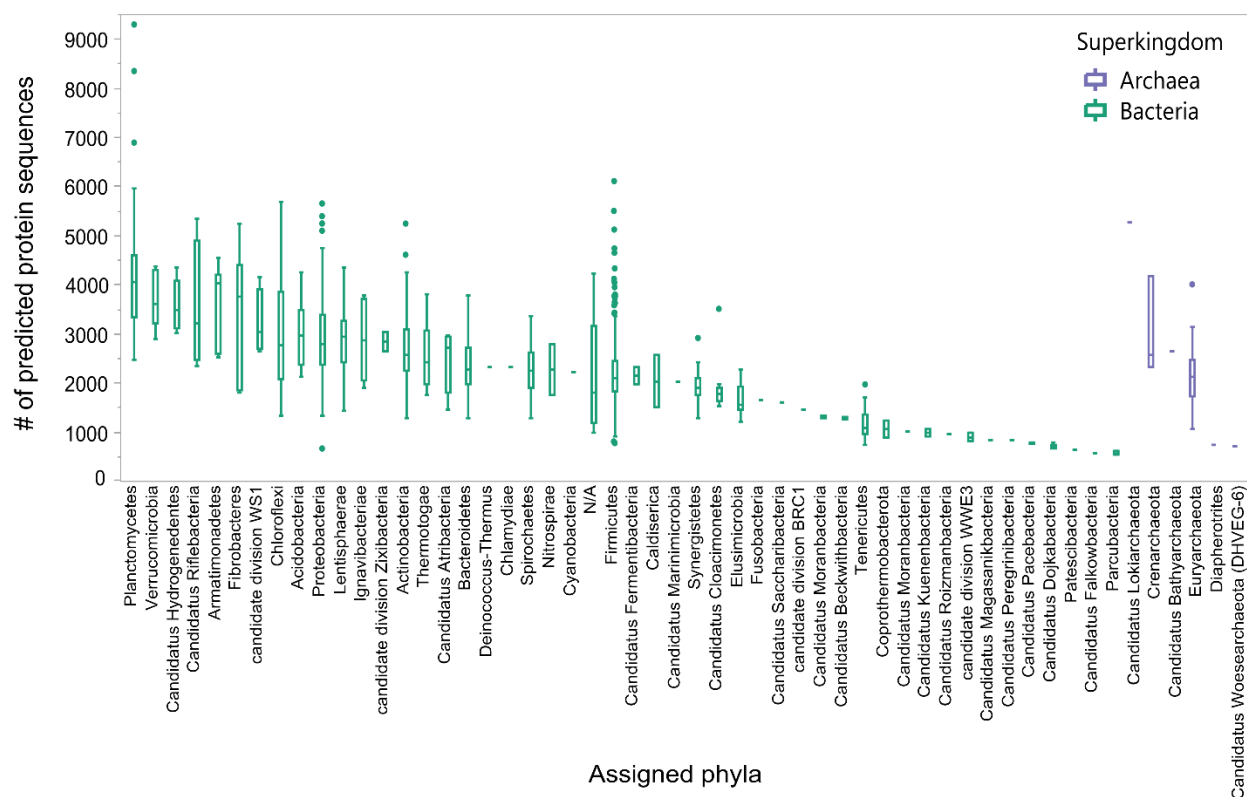


Figure 4.6 Distribution of the sizes of the predicted proteomes from the HQ & MHQ quality MAGs from the biogas microbiome project. The provided CDS were annotated using Prodigal v2.6.2

CAZyme annotation included all the different categories of enzymes described in the CAZy database apart from others like cohesin and S-layer homology domains which are structural components of bacterial cellulosomes.³⁸³ The boxplots presented in **Figure 4.7** show the % of CAZymes in the different proteomes grouped at the phylum level. The median percentage of CAZymes found per phylum was below 4%, which is consistent with the general abundance range of 1-3% of these enzymes from the total gene content of all living organisms and with the >3% of the gene content of organisms with specialized functions like the breakdown of complex carbohydrates found in lignocellulose.³⁸⁴⁻³⁸⁶ The bacterial phylum having MAGs with highest median percentages of CAZymes in their proteomes was that of *Bacteroidetes*, while *Caldiserica* had the lowest median percentages.

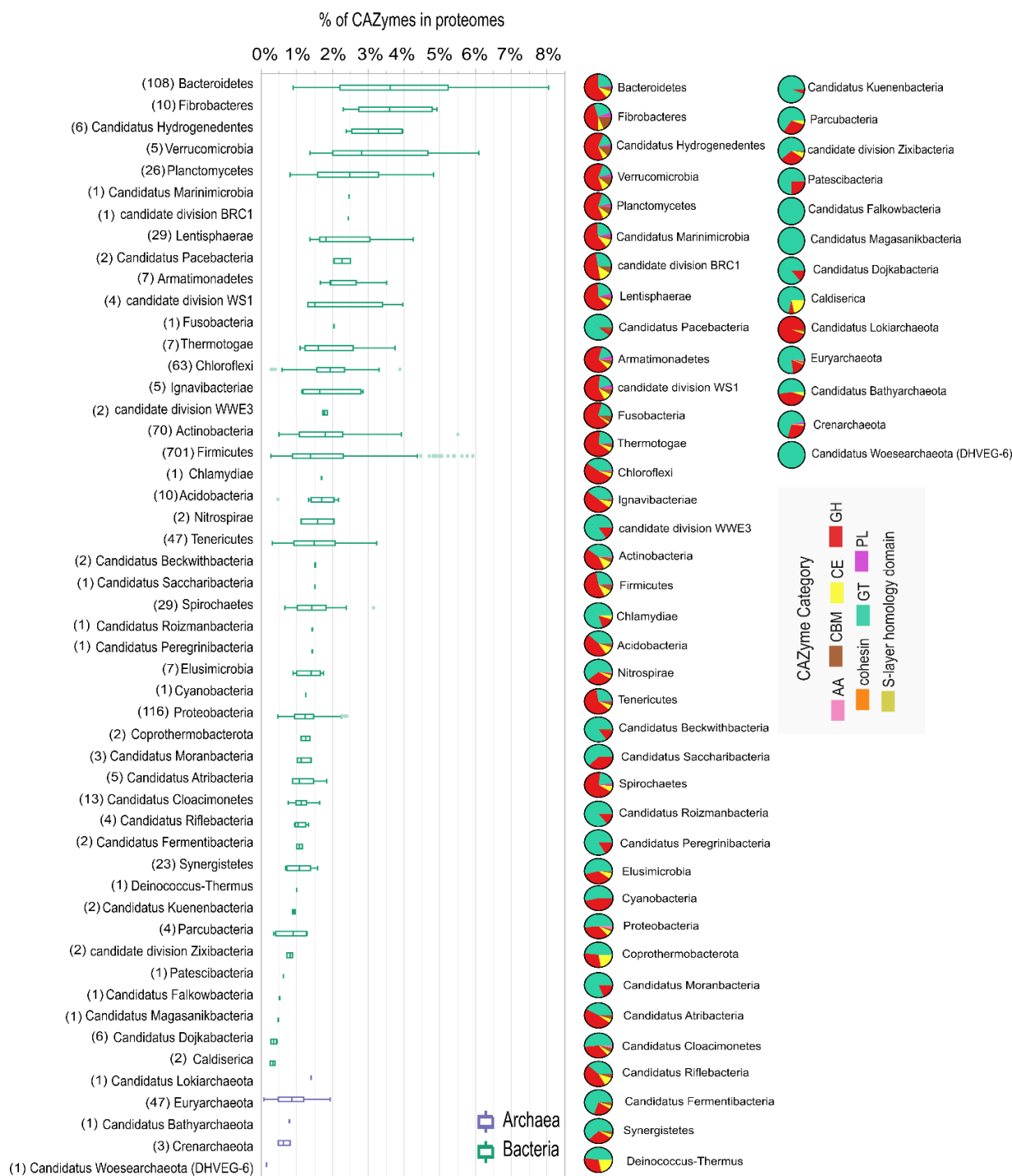


Figure 4.7 CAZymes annotated in the proteomes of different phyla. Box plots (left) show the percentage of CAZymes annotated in the proteomes of different bacterial and archaeal phyla using dbCAN2. The number of annotated MAGs per phylum are shown in parenthesis. Pie charts (right) show the relative fraction of different CAZyme classes, which include AAs (enzymes of the auxiliary activities), CBMs (carbohydrate binding modules), CEs (carbohydrate esterases), GHs (glycoside hydrolases), GTs (glycosyltransferases), and PLs (polysaccharide lyases). Some proteins were also annotated with cohesin and S-layer homology domains, which are involved in cellulosomes structure and formation. MAGs lacking annotations at the phylum level are not shown.

Not surprisingly, MAGs from bacterial phyla like *Bacteroidetes*, *Fibrobacteres*, *Verrucomicrobia*, and *Planctomycetes* which are known to degrade various complex carbohydrates from plant/algae material rich environments, were amongst the top phyla having the highest percentages of CAZymes in their proteomes. However, other less studied taxa, like MAGs assigned to *Candidatus Hydrogenedentes*, *Candidatus Marinimicrobia*, and *candidate division BRC1* also ranked high. The pie charts in **Figure 4.7** show that all these phyla have higher fractions of carbohydrate degrading CAZymes like glycoside hydrolases (GHs), carbohydrate esterases (CEs), and polysaccharide lyases (PLs). More than 50% of the identified CAZymes in these phyla were GH proteins, although this situation was similar in other 11 bacterial phyla proteomes including *Fusobacteria*, *Thermotogae*, and *Firmicutes*.

CAZymes were also annotated in the proteomes of archaeal MAGs with median percentages by phylum below 2%. A large fraction of the annotated CAZymes in archaeal phyla were glycosyl transferases (GTs), which are involved in the transfer of sugar moieties to specific acceptor molecules. Archaea members are known to contain several genes expressing GTs in part due to their intricate protein N-glycosylation mechanisms. In fact, GT2 and GT4 families predominate in Archaea³⁸⁷ and this was observed in members of the Euryarchaeota phyla, in which the 47 MAGs proteomes annotated with CAZymes contained on average 10 more times GT2 and GT4 proteins compared to the numbers of other GT families. Members from the archaea are known to use a wider variety of sugar subunits for N-glycosylation compared to eukaryotes and bacteria. It is hypothesized that N-glycosylation in archaea may contribute to their ability to survive or adapt to harsh environments, and some research has implicated it in cell structural support and protein assembly and function.³⁸⁸

The only distinctive case of CAZymes found in archaeal phyla was that of the single MAG assigned to *Candidatus Lokiarchaeota*, where 94% of identified CAZymes were GHs and the remaining ones were annotated as CBM and CE modules. However, metatranscriptome analyses have revealed the expression of ORFs with similarity to glycoside hydrolases in members of the *Candidatus Lokiarchaeota* group, and anaerobic utilization of carbohydrates has been described as one of their metabolic capacities.³⁸⁹

4.3.6 A unique set of peptides for GH families in in the biogas microbiome.

Proteins from GH families with the highest numbers of representatives across every MAG in the biogas microbiome were submitted to the bioinformatics pipeline of before to find the minimum set of unique tryptic peptides on each of them (**Table 4.2**). Of note, we just wanted to show examples of the selection of unique peptides in a realistic biological system, but one would ideally select families of GH proteins that could act as indicators of the stable hydrolytic functioning of a bioreactor or that can tell something of the underlying biology within it.

As observed in **Figure 4.8A** the number of peptides per each of the five selected families were still above 700 but in this case we did not observe equal number of proteins to peptides as was the case with GH family 10 when using the CAZyDB data. This information demonstrates that the number of shared tryptic peptides amongst groups of GHs is still not as high as expected even within more related organisms living in a specialized environment. The maximum numbers of proteins covered by a peptide were 68, 47, 51, 29 and 13 in GH families 13, 3, 2, 43, and 23, respectively (**Figure 4.8B**). The top 10 peptides ranked by coverage of proteins in each GH family were able to cover between 7%-10% of the total proteins considered for analyses (**Figure 4.8C**). Different than in the previous case of using all the information related to CAZyDB, 200 peptides in GH families 1,3,and 9 are able to cover ~ 50% of all proteins in each of them while more than 200 peptides are necessary to cover 50% or more of the total proteins in GH families with more specialized polysaccharides degrading functions such as GH51 and GH10.

Redundancy of tryptic peptides amongst families of GHs expressed by microorganisms found in specialized systems have not been studied before; however, due to the number of peptides we obtained for each family, we suspected that there is enough sequence variability between different groups of microorganisms that peptides could again provide some level of information regarding the taxonomical origins of the proteins they are derived from. By determining which peptides are specific to certain phyla, reduced number of peptides could in theory be used to target GH families from specific microbes.

Table 4.2 GH families selected from the biogas microbiome data for testing the *in-silico* development of a minimum list of unique peptides able to differentiate between groups of enzymes.

GH family	Enzymatic activities	# of protein seeds across every MAG
13	Some enzymatic activities include: α -amylase, oligo-1,6-glucosidase, α -glucosidase, pullulanase, cyclomaltodextrinase, maltotetraose-forming α -amylase, isoamylase, dextran glucosidase, trehalose-6-phosphate hydrolase, among others acting on complex polysaccharides	4024
2	Most common activities include β -galactosidases, β -glucuronidases, β -mannosidases, exo- β -glucosaminidases and, in plants, a mannosylglycoprotein endo- β -mannosidase	2182
3	Exo-acting β -D-glucosidases, α -L-arabinofuranosidases, β -D-xylopyranosidases, <i>N</i> -acetyl- β -D-glucosaminidases and <i>N</i> -acetyl- β -D-glucosaminide phosphorylases	2134
43	The major activities reported are α -L-arabinofuranosidases, endo- α -L-arabinanases (or endo-processive arabinanases) and β -D-xylosidases	1465
23	GHs in this family are lytic transglycosylases of both bacterial and bacteriophage origin, and family G lysozymes of eukaryotic origin. Both of these enzymes are active on peptidoglycan, but only the lysozymes are active on chitin and chitooligosaccharides.	1090

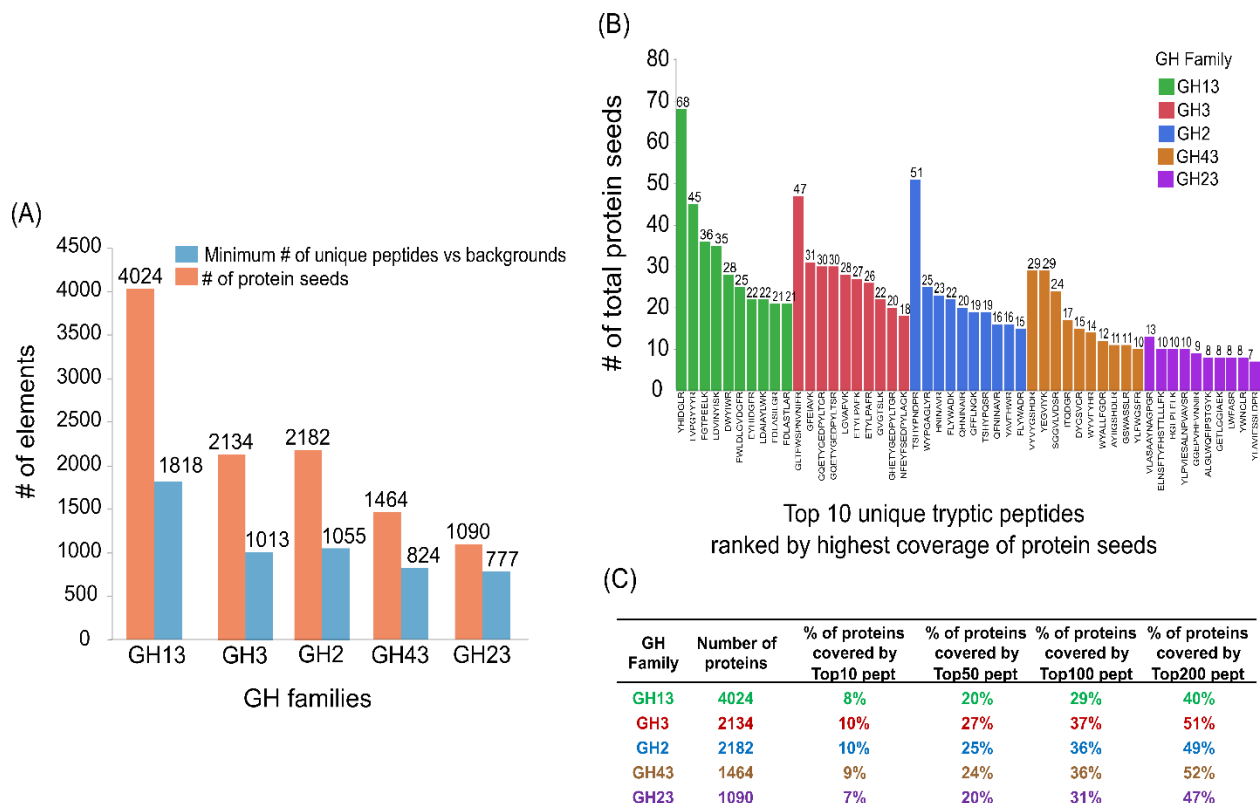


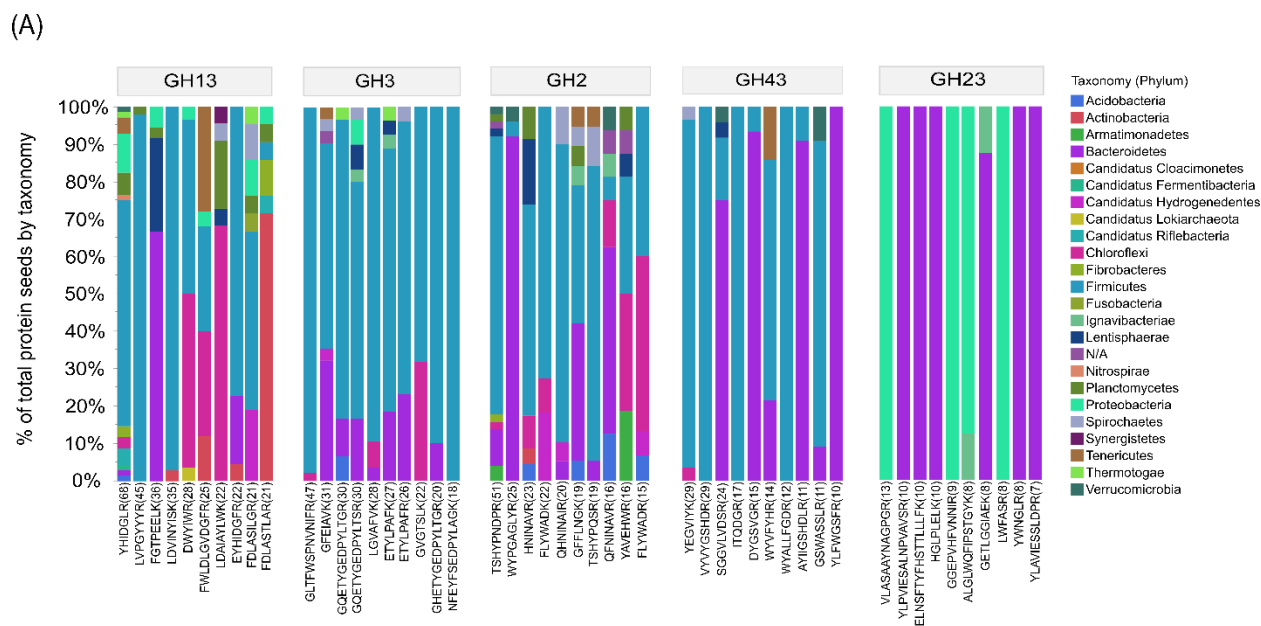
Figure 4.8 Minimum number of unique tryptic peptides and their associated number of protein seeds in different GH families. (A) Total number of unique peptides selected for each GH family after comparison to other proteins in CAZyDB and the protein seeds in which they are found. (B) Top 10 tryptic peptides (from blue bars in A) ranked by the highest number of protein seeds coverage in each family. (C) Percentages of total proteins in GH families covered by top 10, 50 and 100 peptides ranked by protein coverage.

4.3.7 Unique tryptic peptides selected for group of GHs in the gas microbiome can separate groups of proteins based on their taxonomical origins.

In biogas reactors and plants, the hydrolytic ability of anaerobic bacteria of transforming polysaccharides into low molecular weight intermediates that can be used by other microbes during the anaerobic digestion food chain is a key element to the success to these types of systems.³³⁶ The biogas microbiome data provided phylum-level information of all MAGs that were re-annotated using dbCAN2. Hence, we decided to use this data to categorize the peptides we selected and the proteins they mapped to based on their phylum-level origins.

Figure 4.9A shows the top 10 peptides ranked by coverage of number of proteins that were specific to individual phyla contributing to proteins in each GH family. As a first observation, we noticed that *Firmicutes* and *Bacteroidetes* proteins in families GH13, GH3, GH2, and GH43 were amongst the ones with most representation captured by the top 10 peptides in each dataset. In terms of peptides found only in proteins of one phylum, in family GH43, for example, peptides ITQDGR, VYVYGSHDR, WYALLFGDR were identified only in proteins from MAGs assigned to the *Firmicutes* while peptide YLFWGSFR was specific to *Bacteroidetes* proteins. In family GH23, many more peptides covered proteins from single phyla like was the case of five peptides mapping exclusively to proteins from *Bacteroidetes*, three peptides found only in proteins originated from *Proteobacteria*, and one peptide that mapped only to *Firmicutes* proteins.

From the total number of peptides selected for each GH family, we also calculated how many of them were necessary to cover all proteins related to phyla with high hydrolytic potential in anaerobic environments (**Figure 4.9B**). As observed, between 195-854 peptides are necessary to cover all *Firmicutes* proteins in each of the analyzed GH families while the numbers were less for other phyla. These numbers consider shared proteins amongst phyla so it is expected that the numbers could be less if one is indeed only looking to target proteins from a specific phylum and not any other one. These findings are important from a microbiological point of view, as members of the hydrolytic *Firmicutes* and *Bacteroidetes* phyla are the most commonly found in biogas plants,³³⁶ so it may be possible to target GH proteins that are only derived from these phyla.



(B)

GH Family	Total # of peptides selected	# of peptides to cover all Firmicutes proteins	# of peptides to cover all Bacteroidetes proteins	# of peptides to cover all Proteobacteria proteins	# of peptides to cover all Fibrobacteres proteins
GH13	1818	854	191	187	41
GH3	1013	195	223	90	9
GH2	1055	498	248	13	9
GH43	824	381	283	19	34
GH23	777	253	95	238	12

Figure 4.9 Peptide classification by taxonomy using sequence information from the biogas microbiome MAGs. (A) Stacked bars showing the classification of groups of proteins captured by peptides based on their taxonomical origins at the superkingdom level. (B) Number of peptides from the ones selected to cover proteins by superkingdoms. NA, lacks annotation at the superkingdom level.

4.3.8 Unique tryptic peptides selected for group of GHs in the gas microbiome can separate groups of proteins based on their enzymatic specificity.

Several GH families are populated with enzymes having different substrate specificities.³⁹⁰ Such substrate specificity is expressed in the enzymatic commission numbers (EC) given to an enzyme.³⁹¹ For example, enzymes in family GH3 are known to have dual or broad substrate specificities with respect to monosaccharide residues, linkage position and chain length of the substrate.³⁹² Depending on the type of lignocellulose material fed to an anaerobic digester, sets of GHs with more or less specialized enzymatic activities and substrate affinities may become more important for the successful degradation of complex polysaccharides. This reason led us to explore the idea if peptides from selected families of GHs could also provide resolution of groups of proteins within the same family with different enzymatic activities. To check whether this was true, we employed the Ghost Koala annotation pipeline³⁹³ to retrieve EC numbers for all the GHs captured by different peptides in the biogas microbiome.

Interestingly, we observed that the unique peptides we selected grouped proteins within GH families based on different EC numbers. In **Figure 4.10**, we can observe, for example, that out of the 1055 peptides originally selected for family GH2, 443 peptides are only found in GH2 proteins with beta-galactosidase activity (EC 3.2.1.23) while 196 are specific to GH2 proteins that have beta-mannosidase activity (EC.3.2.1.31). Even in the GH13 family, which is known to contain ~30 different enzymatic specificities³⁹⁴, discrete groupings of peptides and proteins according to EC numbers were observed. For example, to target GH13 proteins with Amylo-(1,4 to 1,6)transglucosidase (EC 2.4.1.18) activity, 247 peptides are necessary, while proteins that were annotated as Cyclomaltodextrinases (EC 3.2.1.54), Glucan 1,4-alpha-maltohydrolases (3.2.1.133) and Neopullulanases (3.2.1.135) can be differentiated from other GH13 proteins by 189 peptides.

This differentiation given by the peptides selected here could be useful to target specific groups of GH proteins by substrate affinity in a bioreactor and opens the possibility of monitoring enzymes, independent of their families, but grouped under several EC numbers.

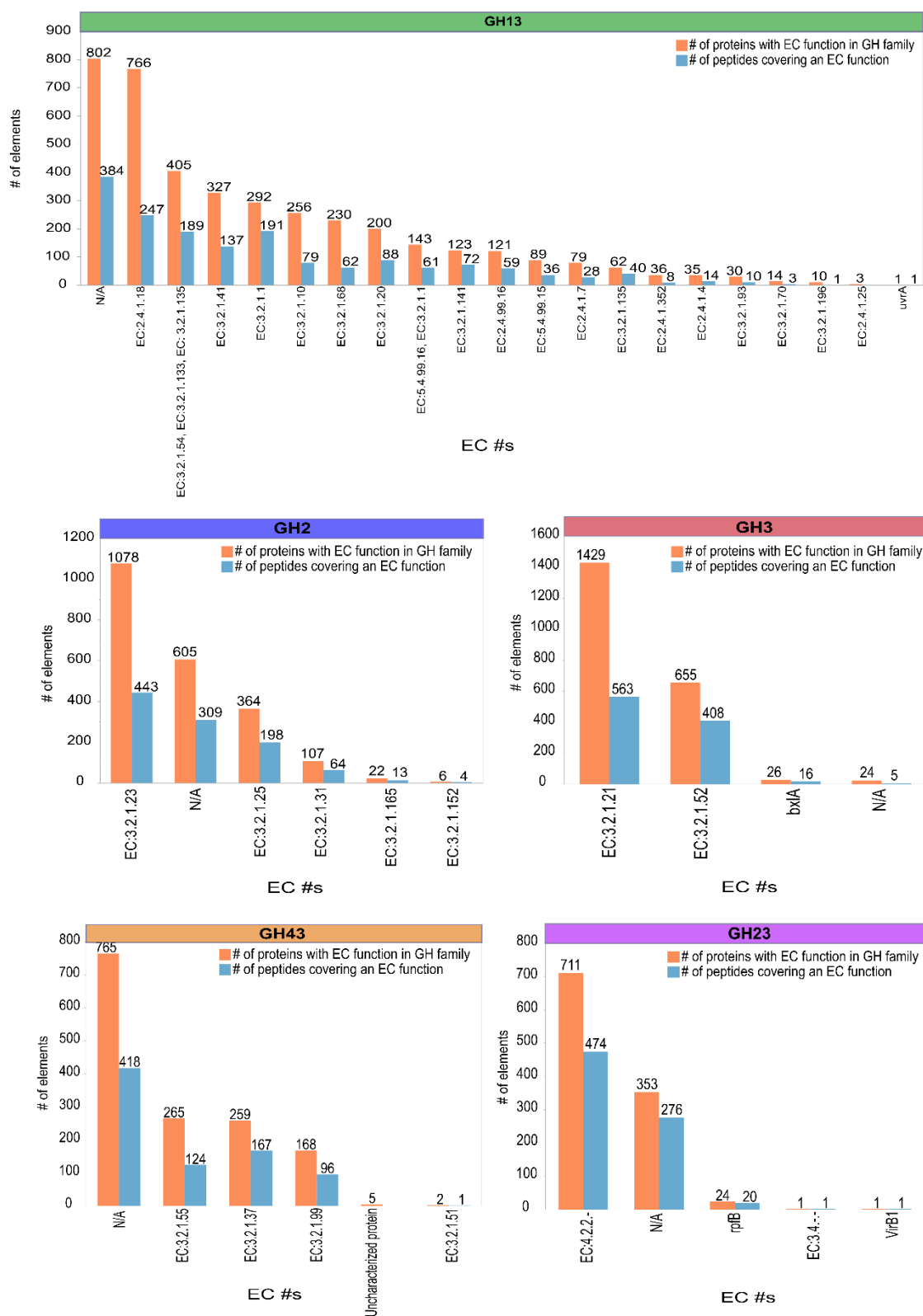


Figure 4.10 Functional classification of groups of proteins captured by numbers of unique peptides. Functional annotation of proteins captured by unique peptides was done with Ghost Koala. N/A= lacks annotation.

4.4 Conclusions.

In this project we have explored the *in-silico* selection of a minimum set of unique peptides targeting groups of proteins in GH families. Contrary to most common targeted proteomics workflows, we decided to make use of shared peptides within proteins in families of GHs to capture all proteins that passed our peptide selection criteria as well as comparisons to background peptidomes. During our analyses we found that the amount of tryptic peptides specific to GH families either using sequence information derived from CAZyDB or the biogas microbiome project were relatively high (>700 in each case); however, these peptides can be adapted to provide different degrees of taxonomical information or even EC number distinction for the case of peptides identified from the biogas data. The high number of tryptic peptides identified for GH proteins was also useful to determine that it may be possible to find unique peptides for individual GH proteins even within the same family (*i.e.*, a GH3 protein versus another GH3).

As we mentioned before, large-scale targeted proteomics studies have employed between 300-600 peptides to target different proteins, however, in less biologically complex systems when compared to biogas reactors or plants.^{378, 379} Thus, reducing the number of peptides to more manageable amounts (<100), if possible, will be a future goal of this work. Other *in-silico* strategies to try out with the purpose of reducing the number of shared peptides amongst proteins can include the digestion of protein targets and background databases with enzymes other than trypsin that could for example, exploit the sequence similarities of active site regions found in several GH families. Of note, after the initial *in-silico* determination of a set of peptides, these analytes need to be tested experimentally to select the ones that can provide adequate signals in a mass spectrometer. This process further reduces the list of initial peptide candidates albeit at the expense of losing some of the proteins of interest.

Targeted proteomics promises a way to identify and quantify proteins in anaerobic digesters that can serve as indicators of the hydrolytic capacity of the system. As of now, several other techniques that measure biodegradable organics present in the sludge fraction of a bioreactor (*i.e.*, oxygen content, C/N ratios measurement) are employed to evaluate the performance of the anaerobic digestion process in faster ways, but these metrics lack molecular-level resolution. In terms of protein abundance, the hydrolytic capacity of anaerobic digesters has been assessed by isolating active enzymes from different sample fractions and conducting *in-vitro* substrate-degradation assays to characterize their enzymatic activity, but these do not reveal sequence-level identities of the CAZymes that are actively participating in the process and neither of their microbial origins.^{332, 395} Thus, information derived from metaproteomics studies and

adapted to develop targeted proteomics assays could get us closer to the high-throughput identification of functional biomarkers for biogas reactors and plants.

The development of targeted proteomics assays relies on lists of protein targets that are often selected based on experimental data obtained in global proteomics experiments. Common workflows for designing targeted proteomics studies also use peptide information collected from global approaches to choose those analytes that can provide not only analytical specificity to proteins of interest, but that are also able to provide intense signals and adequate fragmentation profiles by tandem mass spectrometry. These peptides then become protein surrogates and primary carriers of quantitative information. However, in typical global proteomics experiments some of these peptides may be hidden in the spectral data that is collected. One reason of why this happens is due to the prevalent occurrence of chimeric spectra in LC-MS/MS based global proteomics runs, which causes peptide information to be lost during database searches. In the next chapter the problem of chimeric spectra occurrence in LC-MS/MS global proteomics using spectral data-dependent acquisition was explored under different experimental conditions to evaluate the opportunity of recovering missing peptide identification data from these types of experiments.

CHAPTER 5

An updated perspective on deconvoluting chimeric MS/MS spectra by liquid chromatography and precursor isolation and their subsequent assignment by a multi-peptide-spectrum match algorithm

Text and figures were adapted from the following published journal article:

.....

Villalobos Solis, M.I., Giannone, R.J., Hettich, R.L. and Abraham, P.E. (2019) Exploiting the Dynamic Relationship between Peptide Separation Quality and Peptide Coisolation in a Multiple Peptide Matches-per-Spectrum Approach Offers a Strategy To Optimize Bottom-Up Proteomics Throughput and Depth. *Analytical Chemistry*, **91**, 7273-7279.

Authors contributions: Manuel Ivan Villalobos Solis (M.I.V.S.) planned the study alongside Paul E. Abraham (P.E.A.) and Richard J. Giannone (R.J.G.). M.I.V.S. executed all the experiments and collected all the spectral data. M.I.V.S. and P.E.A. analyzed the results and wrote the main manuscript. R.J.G. and Robert L. Hettich provided critical reviews and proof-read the manuscript. All authors reviewed and approved the final version of the manuscript.

Notes: Supplemental figures and tables mentioned in text are available at [10.1021/acs.analchem.9b00819](https://doi.org/10.1021/acs.analchem.9b00819) and their numbering reflects that of the online material.

.....

5.1 Introduction.

Despite advances in mass accuracy, resolving power, and scan speeds in mass spectrometry instrumentation, one of the remaining challenges of any high-throughput bottom-up proteomics experiment is that only a fraction of the collected tandem mass spectra (MS/MS) can be assigned to peptide sequences with high confidence (usually $\leq 60\%$).¹⁰⁷ Although several reasons contribute to this effect,¹⁰⁹ one that has been under scrutiny by the proteomics community is the occurrence of chimeric spectra (also known as mixture or co-fragmented spectra).

Chimeric spectra are the result of the co-isolation and co-fragmentation of two or more peptide precursor ions with similar m/z and retention time. The complex nature of the samples commonly analyzed in bottom-up proteomics (>100,000 detectable peptide species) and the m/z isolation widths typically used in data-dependent acquisition (DDA) experiments (2-4 m/z)^{107, 109} can result in chimeric spectra representing 50% of the total MS/MS data collected.¹⁰⁹ Recent investigations employing tryptic digests of Henrietta Lacks' 'Immortal' Cell Lines (HeLa) cells demonstrated that even in experiments with a narrow isolation width of 2 m/z , 39% of the total MS/MS spectra collected is chimeric.¹¹⁸

The negative effects of chimeric spectra in proteomics studies have been well-studied. For example, in database-driven peptide and protein identifications, the presence of chimeras deteriorates the search scores of true peptide assignments given by algorithms such as MASCOT

and SEQUEST.¹⁰⁹ In addition, chimeric spectra reduces the accuracy of quantitative isobaric tagging-based quantification methods such as iTRAQ or TMT, in which the contribution of reporter ion intensities from co-fragmented peptides, causes the under-estimation of protein/peptide abundance differences (a phenomenon termed as “ratio compression”).¹¹⁰⁻¹¹² Due to these issues, experimental and computational approaches that minimize the negative impact of chimeric spectra in bottom up experiments have been developed.

Reducing the complexity of samples prior to MS analysis has the advantage of minimizing the chance of co-eluting peptides.^{396, 397} By exploiting independent physicochemical properties of peptides, protocols coupling orthogonal chromatographic separations before MS detection have shown improved separation resolutions and increased peptide and protein identifications, albeit at the cost of increased analysis times.³⁹⁸⁻⁴⁰⁰ With the introduction of ultra-high-pressure chromatography (UHPLC), a new era of high-quality one-dimensional separations has resurfaced. The use of chromatographic pumps that can tolerate up to 10,000 psi and stationary phases with particles sizes of < 2 μ m diameter,^{200, 205} not only has afforded narrower peptide elution profiles and increased ion sensitivities, but the improved column capacities also reduces co-elution of peptides and hence, the occurrence of chimeric spectra.¹⁰⁷ This mode of operation has enabled “single-shot” proteomics studies, in which complex proteomes are analyzed in-depth with the aid of 1D reverse phase chromatographic columns of 50 cm or more in length and effective LC gradients times ranging from 300-500 mins.^{152, 200, 204, 401} Although 1D-UHPLC (including “one-shot” separations) and 2D-HPLC based separations reduce the occurrence of chimeric spectra, even under the best chromatographic separations, the co-elution of thousands of peptides is still unavoidable.¹⁰⁷

Computationally, database searching algorithms aiming to deconvolute chimeric spectra collected in DDA experiments have been developed.^{115, 402-405} These algorithms make use of the multiple peptides-per-spectrum-match approach (mPSM) that, in comparison to most commonly employed search strategies using a single-peptide match-per-spectrum approach (sPSM), try to assign more than one peptide per MS/MS spectrum. A newly developed computational workflow called CharmeRT demonstrated substantial improvement in peptide identifications when compared to other methods.¹¹⁸ This was achieved by implementing a second search strategy coupled with a highly accurate retention time prediction algorithm method. The second search option of CharmeRT is integrated into the database search engine MS Amanda, with validation of multiple peptide assignments to a given MS/MS spectrum performed by Elutator, a new tool built upon the foundations of Percolator that incorporates retention time (RT) prediction in FDR calculations. The impact of having RT prediction in FDR evaluations of first and second searches

translated into 25%-62% more peptide identifications in runs of HeLa tryptic digests compared to more frequently used database search workflows.

Further advancements in LC and MS technology will continue to improve the limit of detection for peptide sequencing in complex mixtures; however, these alone are unlikely to solve the problem with chimeric spectra. We contend that a concomitant evaluation of LC configurations and mPSM search algorithms is required to further increase the number of detectable peptides. Therefore, our study was designed to systematically evaluate several LC peptide separation techniques, ranging from short HPLC gradients to UHPLC and orthogonal separations, as well as precursor isolation *m/z* windows, to better understand the effects of each on the subsequent assignment of MS/MS spectra by a traditional sPSM search strategy compared to a mPSM one, specifically CharmerT.

5.2 Experimental procedures.

5.2.1 Standard samples for LC-MS/MS analyses.

Commercial Pierce HeLa Protein Digest Standards were purchased from Thermo Fisher Scientific (20 µg total amounts). The lyophilized peptides were resuspended in 40 µL of water with 0.1% formic acid as per the manufacturer instructions. Stock solutions of ~0.5 µg/µL were frozen at -80°C and used as needed. Volumes equivalent to 2 µg of HeLa protein digest were analyzed each time using different LC setups and gradients as described below.

5.2.2 1D-LC-MS/MS runs of HeLa tryptic digest standards.

1D-LC-MS/MS runs were carried out using a Proxeon EASY-nLC 1200™ system (Thermo Fisher Scientific) interfaced with a Q Exactive Plus (Thermo Fisher Scientific) mass spectrometer equipped with a nano-electrospray source. For HPLC measurements, peptides were loaded in mobile phase A (0.1% formic acid, 2% acetonitrile) each time onto a trap column (150 mm x 100 µm ID) packed in-house with ~10 cm of 5 µm Kinetex C₁₈ resin (Phenomenex). Peptide separation was conducted on an analytical column (250 mm x 75 µm ID) packed in-house with 5 µm particle size Kinetex C₁₈ resin using a linear gradient from 2 to 22% of mobile phase B (0.1% formic acid, 80% acetonitrile) at a flow rate of 300 nL/min over 90, 210, or 240 min depending on the experiment. Each gradient was followed by an increase to 35% B within 5 minutes, a 5 mins hold in 35% B, and afterwards a decrease to 2% B in 5 minutes.

To test the effects of smaller particle diameter size in number of identifications, the same LC gradients were used in an UHPLC setup, where the trap and analytical front columns were packed with 1.7 μm particle size Kinetex C₁₈ resin. In addition, HeLa peptides were separated with an analytical 500 mm x 75 μm ID column packed with 1.7 μm Kinetex C₁₈ which was placed in a column heater (Sonation GmbH) at a temperature of 60°C. The linear gradient for this configuration was from 2 to 22% solvent B over 500 mins at a flow rate of 250 nL/min.

Optimization of relevant MS parameters in the Q Exactive instrument were performed and were found to agree well with the Q Exactive benchmarking study.⁴⁰⁶ In brief, mass spectra were acquired with the Q Exactive Plus instrument in a top 10 data-dependent acquisition setup. Peptide precursor MS spectra was collected within 300 to 1500 m/z with automatic gain control (AGC) target value of 3×10^6 at a resolution of 70,000 with a maximum injection time (IT) of 25 ms. Precursor ions with charge states ≥ 2 and ≤ 5 and intensity threshold of 1.6×10^5 were selected for higher-energy C-trap collision dissociation (HCD) with a normalized collision energy of 27 eV. Peptide precursor ions collected from the 1D HPLC and UHPLC gradient runs were isolated using a 1.6 m/z isolation width; whereas for the best gradients under HPLC or UHPLC conditions (see Results & Discussions), precursor m/z isolation widths of 0.8 and 3.0 m/z were additionally employed in order to test the effects of co-isolation interference. Fragment ion spectra were always acquired at a resolution of 17,500 at m/z 200 with an AGC target value of 1×10^5 and maximum IT of 50 ms. Dynamic exclusion was set to 20 s to avoid repeated sequencing of peptides. All runs were conducted in triplicate.

5.2.3 2D-LC-MS/MS runs of HeLa tryptic digest standards.

2D LC-MS/MS runs were performed using a Vanquish UHPLC interfaced with a Q Exactive Plus mass spectrometer (Thermo Fisher Scientific) outfitted with a 100 μm ID triphasic precolumn (RP-SCX-RP) coupled to a 250 mm x 75 μm ID nanospray emitter packed with 250 mm of 5 μm Kinetex C₁₈ RP resin.⁴⁰⁷ For each sample, 2 μg of HeLa peptides were loaded to the precolumn by direct flow (2 $\mu\text{L}/\text{min}$) then separated and analyzed across two successive salt cuts of ammonium acetate (35 mM and 500 mM), with each cut followed by a 210 min, split-flow (300 nL/min) organic gradient, wash, and re-equilibration: 0% to 2% solvent B over 2 min; 2% to 22% solvent B over 208 min; 22% to 50% solvent B over 10 min; 50% to 0% solvent B over 10 min, hold at 0% solvent B for 15 min. Peptides were loaded with mobile phase A onto an in-house assembled 100 μm ID fritted precolumn packed with 10 cm of 5 μm particle size Luna strong-cation exchange resin (SCX; Phenomenex). Mass spectra from the eluting peptides was collected

using the same MS and MS/MS parameter settings on the Q Exactive Plus instrument as in the 1D-LC-MS/MS runs.

5.2.4 PSMs, peptide, and protein identifications by database search.

All MS/MS spectra collected were processed in Proteome Discoverer v.2.2. (PD) with MS Amanda v.2.2 and Percolator. Spectral data were searched against the most up-to-date human reference proteome database from UniProt (ID. UP000005640) to which common laboratory contaminants were appended. The following parameters were set up in MS Amanda to derive fully-tryptic peptides: MS1 tolerance = 5 ppm; MS2 tolerance = 0.02 Da; missed cleavages = 2; Carbamidomethyl (C, +57.021 Da) as static modification; oxidation (M, +15.995 Da) and carbamylation (n-terminus, +43.006 Da) as dynamic modifications. The percolator FDR threshold was set to 1% at the PSM and peptide level. In addition, MS/MS spectral data were searched with MS Amanda in which a second search option was enabled and Elutator v2.2 (the CharmerT workflow). Parameters applied for MS Amanda second search were as described in the original CharmerT publication, with the exception that a maximum of 3 additional precursors per PSMs were searched (referred as second searches).¹¹⁸ The Elutator FDR threshold was set to 1% at the PSM and peptide level.

5.2.5 Assessment of the performance of each 1D and 2D LC-MS/MS runs.

The following identification parameters were considered to assess the performance of each 1D and 2D LC-MS/MS runs: number of protein groups, number of modification-specific peptides with charge (referred to as peptide analytes), number of peptides without modification and charge (referred to as peptide sequences) and number of peptide-spectrum matches (PSMs). In addition, we considered the precursor isolation interference percentage calculated by Proteome Discoverer, as a measure of chimerism in the spectra collected:

$$\text{Eq.1 \% isolation interference} = 100 \times \left[1 - \left(\frac{\text{precursor intensity in isolation window}}{\text{total intensity in isolation window}} \right) \right]$$

The full width at half-maximum (FWHM) region was calculated for each LC configuration using the FeatureFinderMetabo node⁴⁰⁸ of OpenMS⁴⁰⁹ using default parameters except for the expected chromatographic peak width (in seconds) setting, which was optimized for each configuration. All spectral data collected in this study was deposited at the ProteomeXchange Consortium via the MASSIVE repository. The project accession is PXD012635 and reviewers can access the data under the username reviewer64304@ebi.ac.uk and password qd5I9bhm.

5.3 Results & Discussions.

5.3.1 Performance metrics across a range of nanoLC peptide separation techniques.

Two common LC peptide separation techniques are HPLC and UHPLC. The former employs analytical columns packed with stationary phases having particle sizes $> 2\ \mu\text{m}$ and flow rates that operate below 450 bar,⁴¹⁰ while the latter uses stationary phases with particle sizes $\leq 2\ \mu\text{m}$ that drive the backpressure of the LC system to 600-1300 bar.²⁰⁵ Our approach started with investigation of the performance of $5\ \mu\text{m}$ C18 packed HPLC and $1.7\ \mu\text{m}$ C18 packed UHPLC columns (250 mm x $75\ \mu\text{m}$ ID) employing three different linear gradient lengths from 2 to 22% solvent B over 90 min, 210 min and 240 min at a constant flow rate of 300 nL/min.

Not surprisingly, the 1D-UHPLC setups outperformed 1D-HPLC in average number of identifications across the different gradients tested, with the best performance achieved in the longer analysis time. For example, in the 240 min gradient, the average numbers of PSMs, peptide sequences, peptide analytes and protein groups were 18%, 56%, 58% and 42% higher than the values obtained for HPLC (**Figures S-1A-D**). These data are undoubtedly explained by UHPLC providing narrower FWHM and boosting the sensitivity of the analysis with increased ion intensities (**Figure S-2**).²⁰⁴ In addition, the overall improved chromatographic resolution provided by UHPLC setups identified $>90\%$ of all the identified HPLC peptides and further yielded between 38-44% new peptide sequences and 25-32% new protein groups not found by their HPLC counterparts (**Figure S-1E**).

The comparison across gradient time lengths on the same LC setups demonstrated that the increase in linear gradient times was more beneficial for UHPLC. For both UHPLC and HPLC, the change from 90 to 210 min increased the average number of PSMs by more than 90% per platform, with at least 10% increases in the numbers of peptide analytes, peptide sequences and protein groups. When the gradient lengths were adjusted from 210 min to 240 min, minor returns in the number of identifications were still observed for the UHPLC setup, with a maximum gain of 12% in the number of PSMs and less than 5% for the remaining identifications. However, for HPLC only an increase of 6% PSMs was observed at 240 min, but the numbers of the other identifications were reduced between 2%-4%, thereby suggesting longer gradients for the HPLC column results in diminishing returns. In general, the effect observed in the 240 min HPLC can be explained by the peak dilution phenomena, in which the use of long gradients with large particle sizes causes a decrease in peak heights which is accompanied by a loss of sensitivity and resolution.^{411, 412}

One way to combat peak dilution is to use extra-long 1D UHPLC columns (> 30 mm) or to fractionate the sample via 2D HPLC. While long 1D-UHPLC analytical columns have been shown to offer in-depth proteome analyses with comparable identification metrics to 2D-HPLC chromatography, these different modes of enhanced LC likely lead to differing degrees of co-isolation. Therefore, both 2D-HPLC and long 1D-UHPLC configurations were employed to assess their overall influence on chimeric MS/MS spectra. As such, a 2D-HPLC separation strategy was evaluated where peptides were first loaded onto a triphasic precolumn and analyzed over two salt cuts, each followed by a 210 min organic gradient, for a total measurement time of 420 minutes. For the long 1D-UHPLC analytical column, a 500 mm long 1D-UHPLC column was evaluated in which separation occurred over a 500 min organic gradient. As observed in **Figures 5.1A-D**, the percentage increases in the average number of identifications were significant for both the 2D-HPLC and the 1D-UHPLC 500 min runs when compared to the results from the 250 mm HPLC and UHPLC configurations tested.

5.3.2 Comparison of precursor isolation interferences across a range of nanoLC peptide separation techniques.

An obvious benefit provided by enhanced LC separation is a reduction in the number of co-eluting peptides. As fewer peptides with similar m/z ratios co-elute, the amount and degree of interfering precursor ions during isolation is expected to decrease. To evaluate this premise across the tested LC configurations, the degree of isolation interference imparted by co-eluting peptides was computed.

Given the already established relationship between precursor ion abundance and the degree of isolation interference in MS/MS spectra,¹⁰⁹ the median isolation interference percentages of identified PSMs in each LC-MS/MS run were plotted against different ranges of precursor ion abundance (**Figure 5.1E**). As expected, higher precursor abundances have lower median isolation interferences in each LC configuration. In general, the majority of PSM precursor abundances ranged from $1e07$ - $1e10$. In this range, higher median isolation interference was observed in the 1D-HPLC setups relative to the 1D-UHPLC setups, while the lowest isolation interferences were found in the 500 min 1D-HPLC runs and 2D-HPLC. Between these more specialized LC setups, precursors identified in the 2D-HPLC runs had a slightly lower median isolation interference percentage. Plotting the 2D-HPLC values for the independent fractions of peptides collected at each salt pulse revealed similar median isolation interference percentages across both fractions, suggesting they have similar peptide complexity (**Figure S-3**).

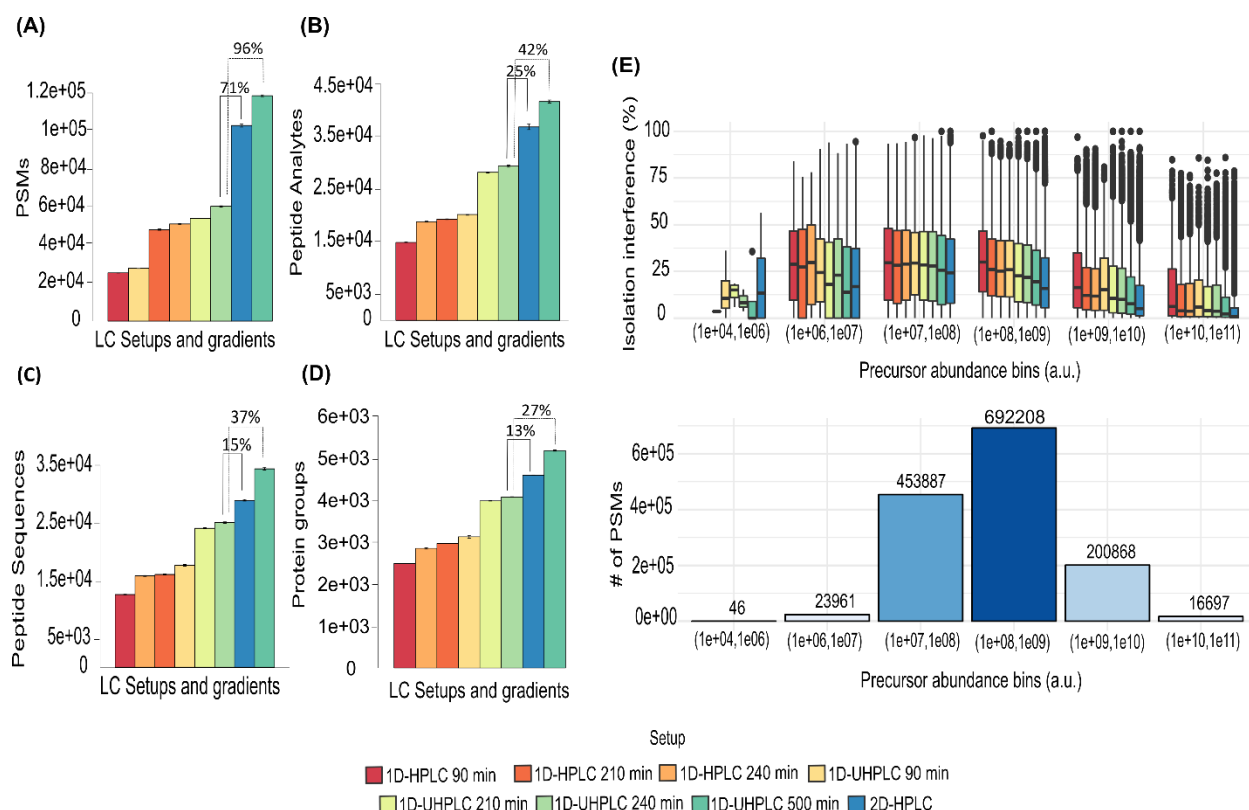


Figure 5.1 Results from the LC-MS/MS analyses of 2µg of HeLa digest using different LC configurations. (A-D) Comparison of the average number of identifications between runs employing different gradient times under HPLC (250 mm x 75 µm column ID, 5 µm C18 particle size) and UHPLC setups (250 mm x 75 µm column ID, 1.7 µm C18 particle size). The alternative 1D-UHPLC 500min (500 mm x 75 µm column ID, 1.7 µm C18 particle size) and 2D-HPLC (250 mm x 75 µm column ID, 5 µm C18 particle size) setups are also shown. Error bars are the standard error of the mean (n= 3 technical replicates). Percentage increase in the average number of identifications of 2D-HPLC and UHPLC 500 min compared to 1D-UHPLC 240min runs are shown in each graph as examples (solid and dashed lines, respectively). **(E)** Boxplots showing the median isolation interference of all identified PSMs precursors across a range of abundances in all initial LC configurations tested. A distribution of the total number of PSMs precursors per abundance ranges is also shown in blue bars. All spectral data were searched with Amanda-Percolator.

5.3.3 Evaluation of the CharmeRT mPSM search algorithm across different LC peptide separation techniques.

Dorfer *et al.* (2018) benchmarked CharmeRT against other widely used search algorithms, including MS Amanda, and demonstrated improved measurement depth. However, an in-depth evaluation of the performance gains across a broad range of peptide separation techniques has not been reported. Therefore, our objective was to compare the CharmeRT workflow (mPSM) against its foundational algorithms, MS Amanda and Percolator (sPSM), across the various LC configurations implemented. As expected, the application of CharmeRT increased the average numbers of identified PSMs between 31%-63% for the HPLC runs, 56%-58% for the UHPLC runs, and up to 64% and 26% for the more specialized 2D-HPLC and 500 min 1D-UHPLC runs, respectively. These numbers translated to increases in the number of peptide sequences identified overall with gains of 32%-55% per HPLC run, 45%-41% per UHPLC, 51% per 2D-HPLC and 22% per 500 min 1D-UHPLC (**Figure 5.2**). The percentage gains in the average number of peptide sequences did not translate into comparable percentage of increases in the average number of protein groups; however, this observation is expected to vary across different peptide mixtures. Nearly all peptides and protein groups identified with MS Amanda-Percolator were also found with CharmeRT (**Figure S-4**). Moreover, most peptides identified from the second search mapped to protein groups identified in the first search (**Figure S-5**). Overall, these observations agreed well with results reported in the original CharmeRT publication.

A major goal of this study was to better understand how varying the quality of peptide separation influences the performance of the CharmeRT search strategy and to determine whether one could achieve similar depth using either a UHPLC with MS Amanda-Percolator or an HPLC configuration with CharmeRT. A comparison of the average number of PSMs, peptide analytes, and peptide sequences between these two approaches revealed that HPLC combined with CharmeRT provided substantially more identifications than the conventional UHPLC Amanda-Percolator search (**Figure 5.2**). More importantly, comparable performance was observed between the HPLC 210 min CharmeRT and UHPLC 240 min Amanda-Percolator using the shorter analytical column (250 mm x 75 μ m ID). Intriguingly, this observation suggests that peptide separation of a complex mixture on a HPLC column with shorter gradient times in combination with a mPSM search workflow can achieve identification metrics comparable to results given by a UHPLC column and a sPSM search strategy.

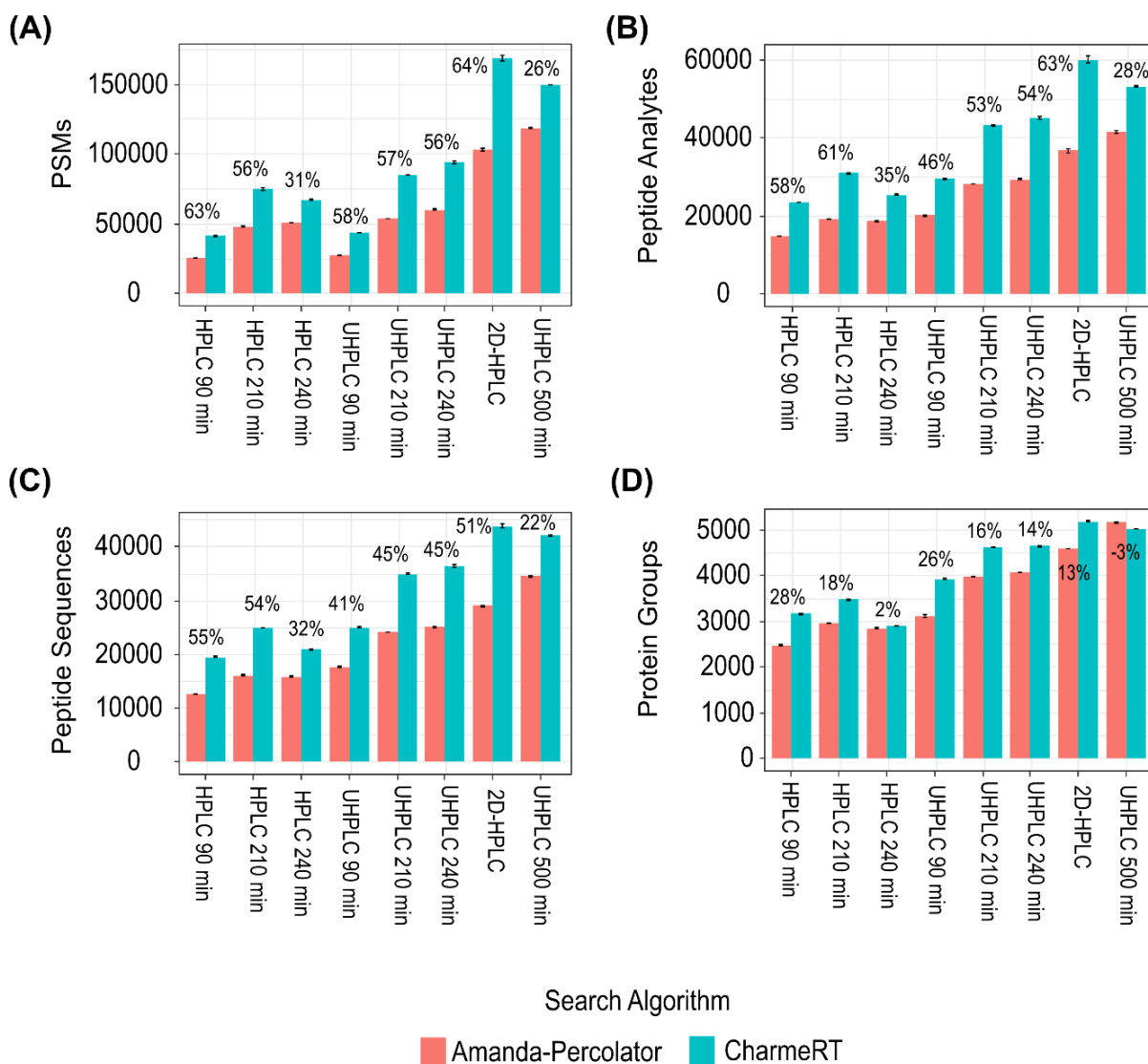


Figure 5.2 Results from the LC-MS/MS analyses of 2µg of HeLa digest using different LC configurations and MS/MS spectra search algorithms. (A-D) Bar charts depicting the average numbers of identifications obtained from the spectral data collected for each LC setup tested (x-axis) and searched with Amanda-Percolator or the CharmerRT workflow. Percentage increase or decrease of the CharmerRT results compared to the ones obtained with Amanda-Percolator are shown above or close to each pair of bars per LC setup. Error bars are the standard error of the mean ($n=3$ technical replicates).

Next, the degree of overlap between the newly identified peptides provided by either UHPLC or the application of CharmerT was evaluated. More specifically, the newly identified peptide sequences and protein groups identified from HPLC runs searched with CharmerT were compared to the results obtained from their UHPLC counterparts searched with MS Amanda-Percolator (**Figure S-6**). These comparisons demonstrate that the newly identified peptides for both approaches were mostly complementary, with only a small overlap (~10%).

When comparing the CharmerT performance between 2D-HPLC and the 1D-UHPLC 500 min gradient runs (500 mm x 75 μ m ID), the 2D-HPLC measurements had the highest percentage gains in terms of average number of PSMs, peptide analytes, peptide sequences and protein groups (64%, 63%, 51%, 13%, respectively) (**Figure 5.2**), despite 2D-HPLC having the least interference (**Figure 5.1E**). To explore this further, the total number of PSMs derived from the 2D-HPLC and 1D-UHPLC 500 min setups at different ranges of isolation interference were binned and quantified based on their CharmerT search depth (**Figure 5.3, isolation width 1.6 m/z**). Interestingly, across every isolation interference bin, the 2D-HPLC configuration afforded 2-4x more CharmerT gains relative to the 1D-UHPLC 500 min configuration. Additionally, a greater search depth (*i.e.*, a higher percentage of scans having 2-3 additional precursors derived from second searches) was observed at higher isolation interference ranges in 2D-HPLC. After exploring accompanying data related to data quality and PSM scoring, it's not immediately obvious what metrics explain this phenomenon. However, a likely explanation is that the 2D separation scheme, which reduces the complexity of the loaded peptide mixture for each salt pulse, retains the CharmerT performance gains like the 1D-HPLC 210 min separation scheme, whereas the 1D-UHPLC 500 min CharmerT performance gains suffers from dilution of lower abundant secondary precursors.

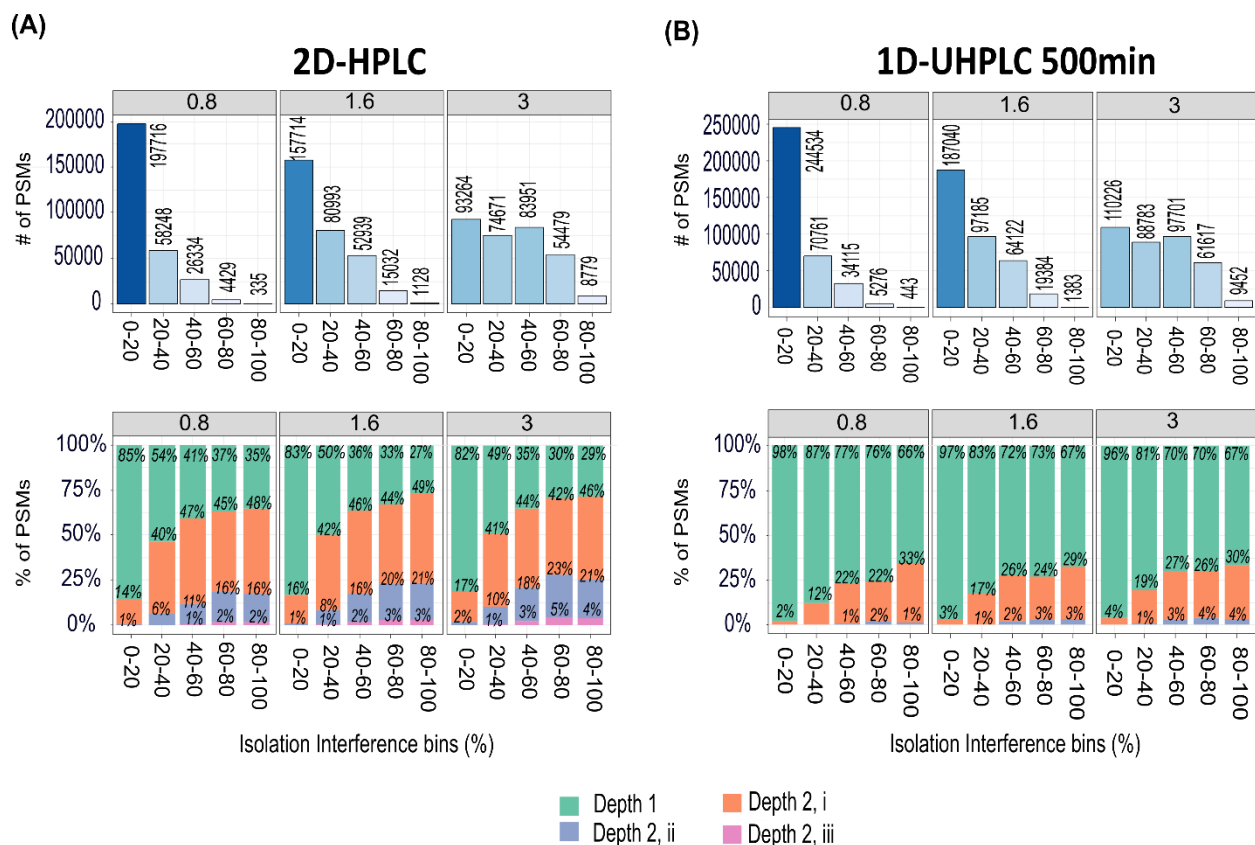


Figure 5.3 Search depth achieved by CharmERT in the spectral data collected through specialized LC modes of operation. Results from (A) 2D-HPLC and (B) 1D-UHPLC 500 mins runs at different isolation windows. Histograms of the number of PSMs per isolation interference range and identified in the first search of CharmERT are shown. The colored stacked bars below each histogram represent the percentage (above each stacked bar) of PSMs identified in the first search of CharmERT from which no additional PSMs derived from second search were identified (green color, Depth 1); from which just one additional PSM was identified (orange color, Depth 2, i); from which two additional PSMs were identified (green color, Depth 2, ii); from which three additional PSMs were identified (pink color, Depth 2, iii).

5.3.4 Evaluation of spectra collected from LC peptide separation techniques with different isolation widths and searched with CharmerRT.

Widely acknowledged, precursor isolation widths can impact the performance in LC-MS/MS measurements: narrower isolation widths lead to loss in signal and wider isolation widths results in more precursor co-isolation/co-fragmentation. As previously demonstrated by Dorfer *et al.*, (2018) applying wider isolation widths during data acquisition improves the performance gains of the CharmerRT workflow. Therefore, our approach was to evaluate different isolation widths across the different LC configurations applied in this study.

To this end, three isolation widths (0.8, 1.6 and 3.0 m/z) that encompass the range often employed in LC-MS/MS measurements were applied and performance assessed across the 1D-HPLC 210 min and 1D-UHPLC 240 min setups, which gave the highest numbers of average identifications across each configuration either when the results were searched with Amanda-Percolator or CharmerRT, as well as the 1D-UHPLC 500 min and 2D-HPLC setups.

At each isolation width, the 1D-HPLC 210 min runs had the highest precursor isolation interference medians, followed by the 1D-UHPLC 240 min runs. The 1D-UHPLC 500 min and 2D-HPLC setups presented similar isolation interference values (**Figure 5.4A**). Given the lower signal and reduced potential for co-fragmentation, the narrower isolation width of 0.8 m/z had the fewest identifications and lowest gains by the CharmerRT workflow (**Figures 5.4B and S7**). Similar to the results observed above, CharmerRT gains in peptide and protein group identifications were higher in the HPLC configurations relative the UHPLC setups, particularly the 2D-HPLC which experienced the widest range in peptide identifications across the three different isolation widths.

In agreement with the original CharmerRT publication, our results revealed CharmerRT improvements as isolation widths increased, except for the 1D-UHPLC 500 min configuration. This is counterintuitive, as the increase in precursor isolation interference becomes seemingly more favorable for the CharmerRT workflow (**Figure 5.3, isolation width 3 m/z and Figure S-8**). This novel insight into mPSM expectations implies that long 1D-UHPLC gradient times do not experience the same benefit in the CharmerRT workflow as observed for the 2D-HPLC and other more common HPLC configurations tested in this study.

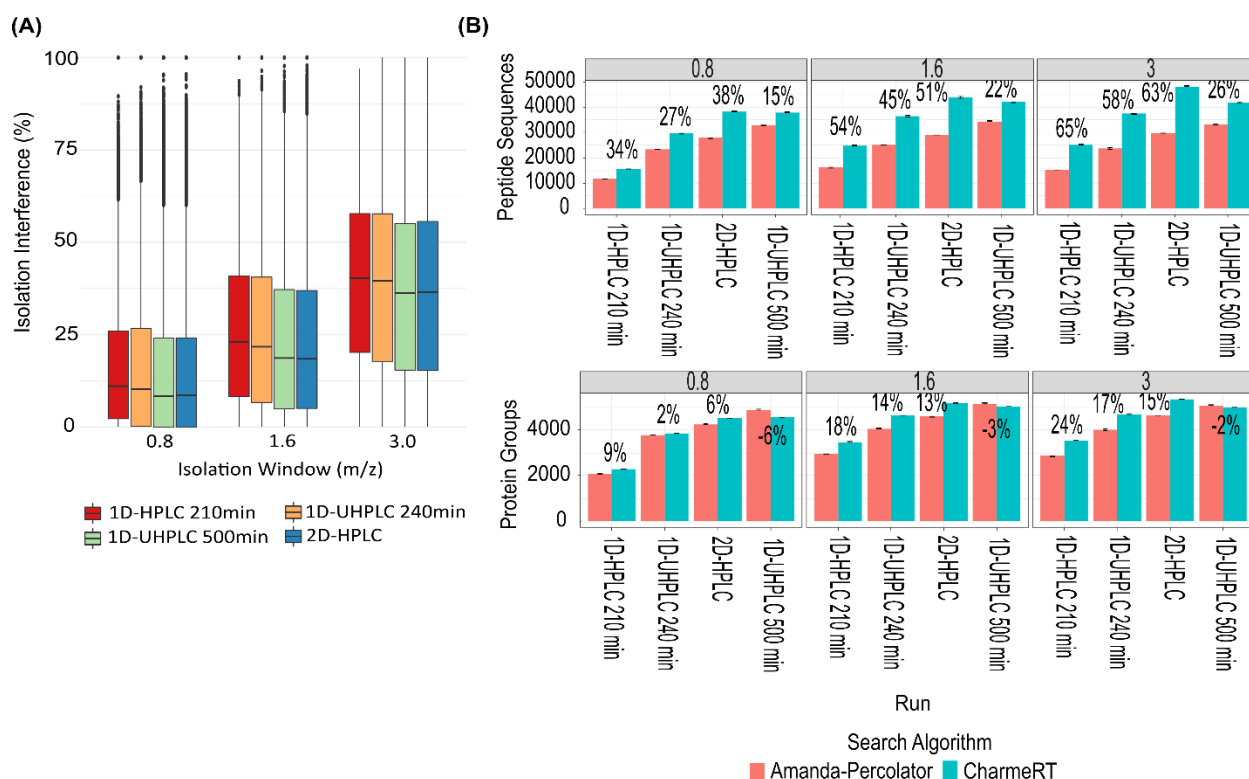


Figure 5.4 Results from the spectral data collected with a range of isolation windows from the best LC configurations. (A) Boxplots showing the median isolation interference of all identified PSMs precursors from the spectral data of the optimized LC gradients collected under a range of isolation windows. (B) Bar charts depicting the average numbers of peptide sequences and protein groups obtained from the spectral data collected for each LC setup (x-axis) under a range of isolation windows. Percentage increase or decrease afforded from the CharmRT results compared to those from Amanda-Percolator are shown above or close to each pair of bars per setup and isolation window. Error bars are the standard error of the mean (n= 3 technical replicates). The average numbers of PSMs and peptides analytes are presented in Figure S-8.

5.4 Conclusions.

By using state-of-the-art LC-MS platforms, proteomes of simple organisms can now be sequenced to near completion in just over an hour.⁴¹³ However, comparable coverage is still not readily achieved in more complex peptide mixtures, like those derived from higher-order eukaryotes or microbial communities. While steady advancements in MS technology and LC will continue to transform the achievable throughput and depth of proteomic analyses, the results presented herein suggest that a better understanding of the dynamic relationship between peptide separation and co-fragmentation can be leveraged by mPSM approaches, such as CharmerT, to enhance saturation in peptide identifications with minimal measurement duration or without advanced LC configurations, *i.e.*, 500 mm long, heated, UHPLC-driven peptide separations. Moreover, this approach coupled to more specialized LC configurations, like 2D-HPLC, offers even greater depth when compared to more traditional single-PSM approaches.

While our results revealed an improvement in peptide separation leads to the expected increases in the number of peptides identified, similar gains can be achieved using shorter gradients coupled with a mPSM approach. This performance differential became less as the peptide separation improved. This observation has substantial implications for LC-MS/MS approaches that use HPLC columns, as it suggests that a new approach towards improving peptide identification rates is to reduce peptide separation while employing a mPSM approach. Moreover, for the 250 mm x 75 μ m ID HPLC and UHPLC columns, a similar depth can be achieved when HPLC measurements employ the CharmerT workflow. Importantly, this suggests that research laboratories separating complex mixtures on a HPLC column with shorter gradient times can achieve comparable identification metrics as if the sample were separated on a UHPLC column.

Many LC-MS proteome specialists view the use of longer columns packed with sub-2 μ m separation particles, as the most direct configuration toward improving peptide identifications.²⁰⁴ Yet, our results offer an alternative perspective, in which the application of CharmerT performed substantially better for a 2D-HPLC strategy when compared to a long 1D-UHPLC column. Therefore, the 2D-HPLC strategy combined with mPSM represents the best path toward improving the number of detectable peptides, enabling a more rapid and comprehensive proteome analysis when compared to the longer UHPLC columns. Moreover, because the demands of UHPLC configurations require LC platforms that withstand very high pressures, which are often unavailable to non-specialists, MS/MS spectral data derived from HPLC separations searched with a mPSM approach could become broadly adopted by the proteomics community.

The occurrence of chimeric spectra in global proteomics experiments using data-dependent acquisition are but one of the reasons behind missing peptide identifications. Other causes are variable peptide ionization efficiencies and ion suppression effects. On the biological side, peptides derived from naturally abundant proteins have better chances of being identified in comparison to the ones produced from less abundant proteins. Related to this point, a vast amount of research now agrees that amongst the molecules in a cell controlling important process such as cell-to-cell signaling, differentiation, growth, and defense are endogenous or native peptides of very low abundance that can be products of protein maturation, turnover or small open reading frames. If global proteomics suffers from all these issues, how can we increase the chances of identifying such part of the proteome that is subjected to dynamic range constraints? This issue was explored in more detail in the following chapter.

CHAPTER 6

LC-MS/MS based discovery and characterization of proteolytic cleavage products (PCPs) produced during the mutualistic symbiotic interactions between the perennial plant *Populus X canescens* and the ectomycorrhizal basidiomycete *Laccaria bicolor*

Text and figures were adapted from the following manuscript submitted and under revision in the Molecular Plant-Microbe Interactions Journal (Revised version submitted on April 3rd, 2020).

Manuel I. Villalobos Solis, Suresh Poudel, Clemence Bonnot, Him K. Shrestha, Robert L. Hettich, Claire Fourrey, Francis Martin, and Paul E. Abraham. (2019) Regulation at the plant-microbe interface: Discovery and characterization of signaling polypeptides using high-performance tandem mass spectrometry

Authors contributions: Manuel Ivan Villalobos Solis (M.I.V.S.) planned the study alongside Suresh Poudel (S.P.) and Paul E. Abraham (P.E.A.). Clemence Bonnot (C.B.) and Claire Fourrey (C.F.) growth and harvested biomass used for proteomics analyses. M.I.V.S. and P.E.A prepared samples from proteomics analyses and collected spectral data. M.I.V.S., S.P. and Him K. Shrestha (H.K.S.) analyzed the proteomics data. M.I.V.S. and P.E.A wrote the manuscript. Robert L. Hettich, C.B, C.F and Francis Martin provided critical reviews and proof-read the manuscript.

Notes: Both M.I.V.S and S.P. are first authors of this study. Supplementary tables are included in the online version of dissertation as downloadable files and their numbering reflect the one followed in this chapter.

6.1 Introduction.

The plant peptidome encompasses a mixed array of low-molecular weight peptides (<10 kDa), or proteolytic cleavage products (PCPs), that are derived from endogenous cellular mechanisms of proteolysis acting in protein turnover,⁴¹⁴ or in the maturation of large protein precursors.^{133, 415} PCPs from protein turnover processes may not have pronounced cellular effects in plants but may act as sources/reservoirs of nitrogen.^{130-132, 416} By contrast, the group of PCPs derived from protein maturation processes have key roles in plant growth and development, reproduction, and stress responses.^{133, 135} These types of bioactive PCPs include well studied families of plant peptide hormones that are classified according to primary sequence characteristics.¹³⁵

Plant PCPs also fulfill diverse roles in the different types of symbiotic interactions with microorganisms in the environment.⁴¹⁷ Although PCPs that are the products of protein turnover events have not been directly implicated in biological functions during symbioses, it is interesting to note that the machinery behind protein degradation appears to be affected in plants undergoing symbiotic relationships with microbes, where the process is deemed essential for cellular

maintenance.⁴¹⁸⁻⁴²⁰ Regarding bioactive PCPs, several of them have been characterized as molecular intermediaries in the responses that plants induce on associated microbiota.⁴²¹⁻⁴²⁴ Amongst the different microbial signals that a plant perceives at any time are microbial bioactive PCPs, which have been shown to alter their hosts cellular behavior in order to survive.⁴²⁵⁻⁴²⁸

The discovery and characterization of PCPs in plants and in plant-microbe interactions have been expedited with the application of experimental mass spectrometry (MS) methodologies that have used biomaterial known to be enriched in selective families of PCPs or genetically altered to overexpress them, followed by purification or enrichment strategies, and in some cases mapping of spectral data to custom-made databases containing known PCP precursors and/or alternate sequence versions of bioactive PCPs.^{415, 429-439}

All these studies have demonstrated the advantage of MS methodologies for the large-scale identification of PCPs in plants and plant-microbial interactions; however, they have also highlighted several technological challenges associated with their application including the relatively natural low abundance of PCPs, as well as undefined proteolytic processing and unusual post-translational modifications that are normal sequence features on them. Furthermore, several of these studies were also tailored specifically for the detection of known PCPs families. Thus, alternative MS approaches that can provide a wider image of known or unknown PCPs involved not only in plants but in plant-microbe interactions are needed.

One of such alternatives is to explore the capabilities of searching MS spectral data with *de novo* database-assisted searches.⁴⁴⁰ In *de novo* database-assisted searches peptide sequence tags derived directly from a MS/MS spectrum are used to better select candidate peptides in a database.^{149, 441} In this way, *de novo* circumvents the low sensitivity and impractical implementation of deriving *in-silico* spectra from database searches with nonspecific cleavage rules,^{441, 442} or of narrowing the scope of identifications by searching spectral data against databases that identify selected groups of PCPs.

Combined with experimental practices that enrich peptides < 10 kDa in samples from tissues and cells,^{138, 139, 142, 144, 145, 429, 435} tandem mass spectral peptide sequencing via *de novo*-assisted database searches could provide a feasible way to comprehensively detect endogenous groups of PCPs produced in tissues of plants and in cells of microbes undergoing symbiosis. In this study, we thus explored the feasibility of the application of 10 kDa MWCO filters to enrich PCPs in biomaterial directly extracted from *Populus X canescens* plants grown *in vitro* with the ectomycorrhizal basidiomycete *Laccaria bicolor*. Interestingly, recent publications on the symbiosis between *Populus* trees and *L. bicolor*, have shown that the establishment and regulation of this interaction is driven by effector-like PCPs derived from small protein

precursors;^{443, 444} however, the comprehensive profiling of all types of PCPs expressed in active ectomycorrhizal *Populus* roots is still missing in literature.

To benchmark the performance of our approach before its application to the *Populus-L. bicolor* system, and evaluate the confidence levels of the *de novo* results, standardized samples containing constant amounts of undigested *Escherichia coli* K12 proteome backgrounds as well as digested Universal Proteome Standard (UPS1) tryptic digests and variable concentrations of known absolute quantification (AQUA) heavy-labeled peptide standards were analyzed. After peptide enrichment using 10 kDa MWCO filters, it was possible to reproducibly recover quantifiable information from UPS1 peptides and AQUA peptide standards.

When applied to the *Populus-L. bicolor* system, our methodology was able to identify PCPs involved in biological processes other than in the plant-fungi interaction, such as PCPs representing the bioactive form of the *Populus* Photosystem II complex 5 kDa (PsbTn) protein in plant leaves. A differential PCP abundance analysis determined a total of 157 PCPs that were significantly more abundant in root tips with established ectomycorrhiza when compared to root tips without established ectomycorrhiza and extramatrical mycelium of *L. bicolor*, thus suggesting roles in the mutualistic interactions between both organisms. Of these, 96 PCPs mapped to 69 *L. bicolor* proteins, 56 PCPs mapped to 64 *Populus* proteins and 5 PCPs were shared between 12 *Populus* and 6 *L. bicolor* proteins, respectively. These findings demonstrate that global profiling of PCPs via LC-MS/MS with *de novo*-assisted database search could significantly advance the comprehensive understanding of the role of these molecules in plant-microbe interactions, and highlights the general applicability of this experimental strategy to acquire direct evidence of PCPs from different plant tissues.

6.2 Experimental procedures.

6.2.1 *E. coli* background proteome sample preparation.

A crude *E. coli* K12 proteome was prepared to provide a sample matrix for known target peptides for benchmarking. Briefly, an *E. coli* cell pellet (OD 0.4) was suspended in lysis buffer (2% sodium deoxycholate (SDC) in 100 mM NH_4HCO_3 (ABC)), disrupted by sonication (30% amplitude, 10 s pulse with 10 s rest, 2 min total pulse time) and boiled for 5 min. Crude protein extract was precleared via centrifugation for 15 min at 14,500 rpm and quantified with a Nanodrops spectrophotometer (Thermo Scientific). The precleared protein extract was then adjusted to 10 mM dithiothreitol (DTT) and incubated under constant shaking for 30 min at 37°C

(600 rpm, Eppendorf Thermomixer) to avoid endogenous peptide cyclization or unwanted protein interactions. Cysteines were then blocked with 30 mM iodoacetamide (IAA) to prevent reformation of disulfide bonds. The *E. coli* proteome sample was stored at -20°C until further use.

6.2.2 Universal Proteomics Standard (UPS1) sample preparation.

One vial of UPS1 (Sigma) was solubilized in cell lysis buffer (2% SDC in 100 mM ABC). Proteins were reduced with 10 mM DTT and incubated at 35°C under constant shaking for 30 min (600 rpm, Eppendorf Thermomixer). Afterwards, proteins were alkylated with 30 mM IAA followed by a 15 min incubation at room temperature in the dark. The UPS1 sample was submitted to a 3 hours trypsin digestion (1:20, enzyme to protein ratio) at room temperature under continuous shaking followed by an overnight trypsin digestion using the same enzyme to protein ratio at 37°C under continuous shaking. The UPS1 tryptic digest was stored at -20°C until further use.

6.2.3 Heavy-labeled peptide standards.

A total of 17 AQUA Ultimate (>97% purity) labeled peptide standards (Thermo Scientific) were combined in a standard solution using 100 mM ABC as dilution buffer to individual concentrations of 10 fmol/μL. The AQUA mix was stored at -20°C until further use.

6.2.4 Preparation of master mixes for benchmarking the performance of a 10 kDa enrichment strategy of PCPs and their identification/quantification via *de novo*-assisted database search.

The master mix for benchmarking the identification of small peptides by the informatics pipeline was prepared on top of Vivaspin 500 μL centrifugal filter units (MWCO 10kDa, Millipore Sigma). To each filter, volumes equivalent to 100 μg of *E. coli* proteome and 0.24 μg of digested UPS1, and volumes equivalent to 1000, 750, 500, 250 and 125 fmol of each standard in the AQUA mix were pipetted ($n=3$ experimental replicate per standard concentration). Additionally, volumes equivalent to 100 μg of *E. coli* proteome were also pipetted into three filters and used as background controls.

All samples were then centrifuged for 30 min at 12,000 g to remove proteins from the *E. coli* background proteome or undigested proteins from the UPS tryptic digest. The flow-through was collected and acidified with 1% formic acid to precipitate SDC. Hydrated ethyl acetate was added to each sample at a 1:1 (v/v) ratio three times to effectively remove SDC. Samples were then placed in a SpeedVac Concentrator (Thermo Fischer Scientific) to remove ethyl acetate. The peptide-enriched flow throughs were desalted on Pierce peptide desalting spin column (Thermo

Scientific) as per the manufacturer instructions. After speed-vac concentration, dry peptide samples were suspended in 10 μ L of 0.1% formic acid solution and placed in autosampler vials.

6.2.5 1D-LC-MS/MS of master mixes samples.

A Proxeon EASY-nLC 1200™ system (Thermo Fisher Scientific) interfaced with a Q Exactive Plus (Thermo Fisher Scientific) mass spectrometer equipped with a nano-electrospray source were used to conduct 1D-LC-MS/MS analyses. Injection volumes of 2 μ L of each master mix ($n=15$) were loaded in mobile phase A (0.1% formic acid, 2% acetonitrile) onto a trap column (150 mm x 100 μ m ID) packed in-house with ~10 cm of 1.7 μ m Kinetex C₁₈ resin (Phenomenex). Peptide separation was performed on an analytical column (300 mm x 75 μ m ID) packed in-house with 1.7 μ m particle size Kinetex C₁₈ resin using a linear gradient from 2 to 22% of mobile phase B (0.1% formic acid, 80% acetonitrile) at a flow rate of 300 nL/min over 90 min.

Mass spectra were acquired with the Q Exactive Plus instrument in a top 10 data-dependent acquisition setup. MS spectra were collected within 300 to 1500 m/z with automatic gain control (AGC) target value of 3×10^6 at a resolution of 70,000 with a maximum injection time (IT) of 25 ms. Precursor ions with charge states ≥ 2 and ≤ 5 and intensity threshold of 1.6×10^5 were isolated using a 1.6 m/z isolation width for higher-energy C-trap collision dissociation (HCD) with a normalized collision energy of 27 eV. MS/MS spectra were acquired at a resolution of 17,500 at m/z 200 with an AGC target value of 1×10^5 and maximum IT of 50 ms. All spectral data collected in this study was deposited at the ProteomeXchange Consortium via the MASSIVE repository. The project identifier is MSV000084471 and the data can be reviewed under the username “MSV000084471_reviewer” and password “Martin”.

6.2.6 Preparation of *Populus X canescens* with or without *Laccaria bicolor* ectomycorrhizae samples.

The hybrid *Populus X canescens* (*Populus tremula* x *Populus alba* line INRA 717-1-B4) clones were micropropagated in vitro and grown in half-strength Murashige and Skoog (MS/2) medium in a growth chamber at 24°C under a 16-h photoperiod. The dikaryotic vegetative mycelia of strain S238N of the ectomycorrhizal fungus *Laccaria bicolor* were maintained on modified Pachlewski agar medium P5 at 25°C in the dark.⁴⁴⁵

For *in vitro* coculture of poplar with *L. bicolor*, we used a sandwich system described before.⁴⁴⁶ The Petri dishes were incubated for 2 weeks vertically in a growth chamber at 20°C under a 16-h photoperiod. From these co-culture plates, we collected extramatrical mycelium,

ectomycorrhizal lateral roots, non-mycorrhizal lateral roots and leaves. Leaves from control plants (no ECM fungi) were harvested on plant grown separately but in the same *in vitro* conditions.

Cuttings of *Populus X canescens* (synonymous *P. tremula* x *P. alba*; INRA clone 717-1B4) of 1 cm diameter and 15 cm were planted in pots containing 5 L of sterile sand and grown in a greenhouse irrigated twice a day with 20 mL of nutrient solution (8 μ M KNO₃, 8 μ M Ca(NO₃)₂·4H₂O, 3 μ M NaH₂PO₄, 3 μ M MgSO₄·7H₂O, 0.2 μ L/L of Kanietra solution (COFAZ, Paris). Xylem sap was extracted by natural root pressure after 7 weeks of growth. To ensure xylem sap flow, poplars were watered with the nutrient solution two hours before sap collection. Using a razor blade, poplars were then decapitated by transversal section of their stem 3 cm above its branching point from the cutting. To avoid contamination of the sap by other tissues of the stem, the cell layers surrounding the stele were removed with care around the section site with a scalpel on a surface of 1 cm length. After cleaning with deionized water and drying with tissues, the exposed stele was attached tightly to 3 cm of Tygon ND 100-65 tubes (Saint-Gobain Performance plastics, Courbevoie, France) of appropriate inner diameter with microporous tape (Urgo, Chenove, France). Flowing xylem sap was collected regularly with a micropipette from the Tygon tubes. Samples were stored on ice during sampling. 150 μ L of sap per poplar was collected over three hours from 16 poplars. The sap from four poplars was pooled to produce one sample before flash freezing the samples in liquid nitrogen.

6.2.7 PCPs extraction from *Populus X canescens* and *Laccaria bicolor* samples.

All samples derived from the ectomycorrhizal and non-ectomycorrhizal interactions of *Populus X canescens* with *L. bicolor* were processed to recover endogenous polypeptides for LC-MS/MS analysis. Briefly, grounded samples were placed in 1 ml lysis buffer (2% SDC in 100 mM ABC), vortexed and then placed in a heat-block for 5 min at 90°C. Afterwards, samples were disrupted by sonication (program: 30% amplitude, 10 s pulse with 10 s rest, 2 min total pulse time) and boiled for an additional 5 min at 90°C. After recovering the supernatant via centrifugation, the samples were adjusted to 10 mM DTT and incubated under constant shaking for 30 min at 37°C (600 rpm, Eppendorf Thermomixer). Cysteines were then blocked with 30 mM iodoacetamide (IAA) to prevent reformation of disulfide bonds.

To separate the low molecular weight fraction of the proteome, samples were transferred into 10 kDa MW cutoff centrifugal filter units and then centrifuged for 30 min at 12,000 xg. The peptide-enriched flow-throughs were collected into fresh Eppendorf tubes and acidified with 1% formic acid to precipitate residual SDC. Samples were then placed in a SpeedVac Concentrator (Thermo Fischer Scientific) to remove ethyl acetate. Peptide samples were desalted on Pierce

peptide desalting spin column (Thermo Scientific) as per the manufacturer instructions. After speed-vac concentration, dry samples were suspended in 10 μ L of 0.1% formic acid solution and placed in autosampler vials.

6.2.8 1D-LC-MS/MS of *Populus X canescens* with or without *Laccaria bicolor* samples.

All samples were analyzed on a Q Exactive Plus mass spectrometer (Thermo Fischer Scientific) coupled with a Proxeon EASY-nLC 1200 liquid chromatography (LC) pump (Thermo Fisher Scientific). Peptides were separated on a 75 μ m inner diameter microcapillary column packed with 50 cm of Kinetex C18 resin (1.7 μ m, 100 Å, Phenomenex) that was heated to 60 °C in a Phoenix S&T NanoLC column heater. For each sample, a 2 μ g aliquot was loaded in buffer A (0.1% formic acid, 2% acetonitrile) and eluted with a linear 90 min gradient of 2 – 20% of buffer B (0.1% formic acid, 80% acetonitrile), followed by an increase in buffer B to 30% for 10 min, another increase to 50% buffer for 10 min and concluding with a 10 min wash at 98% buffer A. The flow rate was kept at 200 nL/min. MS data was acquired with the Thermo Xcalibur software version 4.27.19 using the topN method where N could be up to 10. Target values for the full scan MS spectra were 1×10^6 charges in the 300 – 1,500 m/z range with a maximum injection time of 25 ms. Transient times corresponding to a resolution of 70,000 at m/z 200 were chosen. A 1.6 m/z isolation window and fragmentation of precursor ions was performed by higher-energy C-trap dissociation (HCD) with a normalized collision energy of 30. MS/MS scans were performed at a resolution of 17,500 at m/z 200 with an ion target value of 1×10^6 and a maximum injection time of 50 ms. Dynamic exclusion was set to 20 s to avoid repeated sequencing of peptides.

6.2.9 PSMs, peptide, and protein identifications by PEAKS.

The raw data were interrogated by *de novo*-assisted database searching against the *Populus trichocarpa* v3 reference proteome coupled to the *L. bicolor* UniProt reference proteome (UP000001194) accompanied with common contaminant proteins using PEAKS DB, PEAKS PTM and PEAKS SPIDER in PEAKS X Studio (Bioinformatics Solutions, Waterloo, Canada). The peptide and fragment ion mass tolerances were set to ± 10 ppm and ± 0.02 Da, respectively. The enzyme parameter was set to “no enzyme”. Features associated with chimera scan were enabled. *De novo* ALC score was set at >90%. A false discovery rate of 1% was applied to accept the peptide sequences and a minimum of three peptides were required to identify a protein. For PEAKS DB, carbamidomethylation (+57.02) of cysteine was set as fixed modification and oxidation (+15.99) of methionine was set as a variable modification. PEAKS PTM algorithm was

used to identify other common modifications in PCPs by allowing the search against all possible modifications from the Unimod database. Similarly, PEAKS SPIDER algorithm was used to detect any possible *de novo* sequencing errors and homology peptide mutations.

For differential abundance analysis of PCPs involved in ECM interactions between *Populus* and *L. bicolor*, peptide data from the *Populus* roots with and without ECM, as well as *L. bicolor* extramatrical mycelium sample sets, were filtered to remove all peptides, that were present in only 1/3 replicates. All peptides present in 2/3 biological replicates were considered valid for quantitative analysis. Missing data were imputed by random numbers drawn from a normal distribution (width = 0.3 and downshift = 2.8 using Perseus software (<http://www.perseus-framework.org>)²⁶²). The resulting matrix was subjected to ANOVA ($p < 0.05$) followed by post-hoc Tukey test (Family-Wise Error Rate < 0.05) to assess peptide abundance differences between the different experimental groups. The entire filtering and statistical analyses were done using an in-house developed python script.

6.3 Results & Discussion.

6.3.1 Benchmarking the identification and quantitation of PCPs.

The low nanomolar to mid picomolar concentrations to which bioactive PCPs are thought to be active, and which experimental measurements in plant tissue cultures have found,^{435, 447} prompted us to evaluate the quantitative accuracy and reproducibility of recovering peptides < 10 kDa from a molecular weight-spin column enrichment strategy. For this purpose, we created a standard benchmarking sample that contained invariable amounts of proteolytically digested universal proteomics standard (UPS1 – 48 proteins, 5 pmol each), combined with invariable amounts of an *E. coli* crude protein extract to best mimic small endogenous peptides at relatively low abundances within a complex proteome background. In addition, equimolar solutions of 17 heavy AQUA synthetic peptide (**Table 6.1**) standards in different molar amounts (injection volumes equivalent to 12, 25, 50, 75 or 100 fmol of each standard) were also spiked into the mix of *E. coli* proteins and UPS1 peptides (**Figure 6.1**) for quantitative benchmarking.

Table 6.1 Heavy peptide standards (>97% purity) spiked into constant amounts of background *E. coli* proteome and UPS tryptic digest used to benchmark the *de novo* peptide identification and quantification pipeline. Peptide standards were derived from proteins of *Sphingobium* sp. (Strain NBRC 103272/SYK-6). A 2 μ L injection volume of each benchmark sample was analyzed by LC-MS/MS. The injection volume is equivalent to 12.5, 25, 50, 75 and 100 fmol for each peptide standard. Residues in red are heavy versions of amino acids containing ^{13}C and ^{15}N isotopes.

Heavy Peptide Sequence	Precursor <i>m/z</i>	Precursor Charge
IALAGAGAFGEK	556.8131	2
AEQAAEVAAK	498.2660	2
YDDLVTGK	459.7365	2
LVDDAPK	383.2152	2
SVAPLVADYADR	643.8343	2
GLVIDAGVR	455.2732	2
AIYAQGTVK	479.7760	2
VFTAQR	366.2073	2
ADESAYLDEWNLTAAK	951.4539	2
LFSTDGK	388.2074	2
GGYFLAEDLPADTATR	853.9166	2
GLVAPTGDETR	563.2923	2
ETIEANQLR	542.2870	2
GIDPTTGHYFDDTK	525.5786	3
YIDALDISDQER	724.3505	2
EFDLNFIDK	574.7893	2
LISDPEELSK	569.8077	2

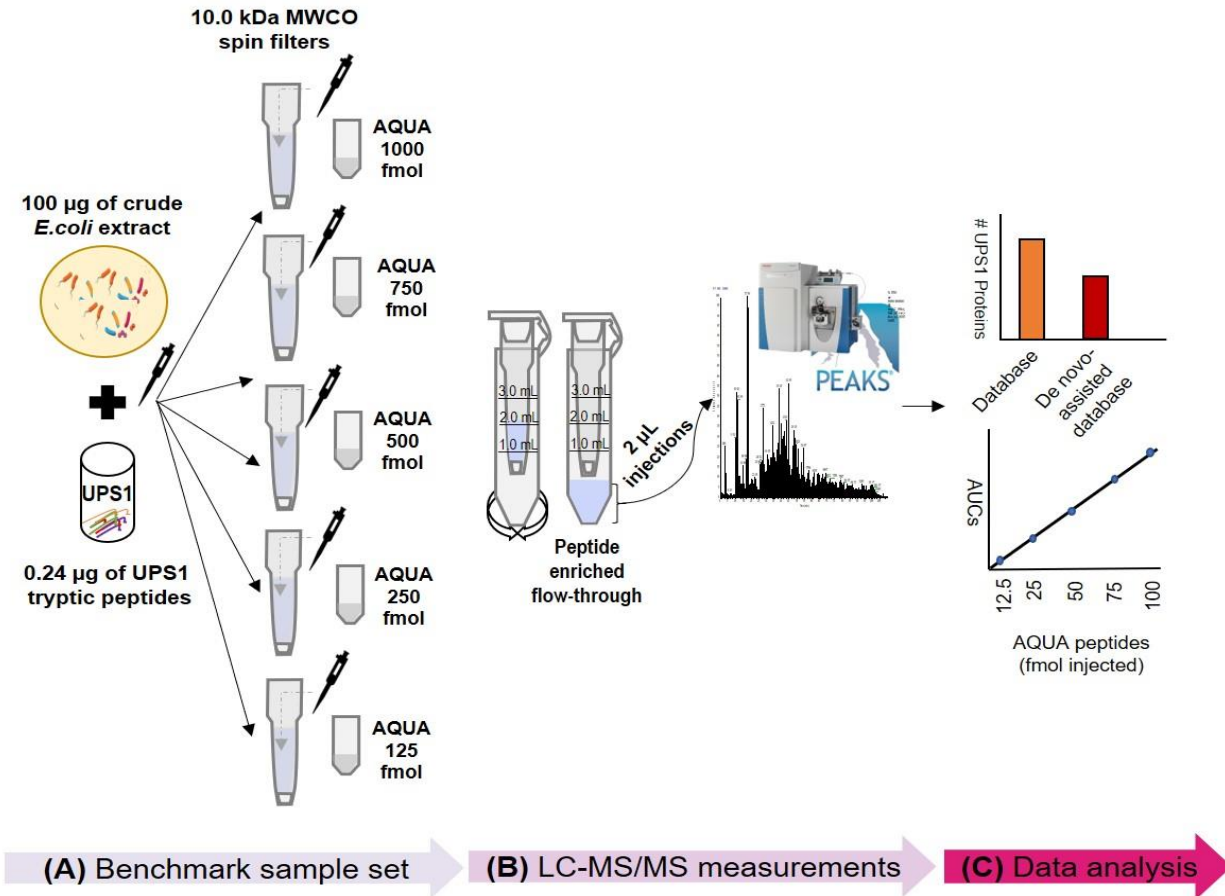


Figure 6.1 Computational and experimental strategies followed for the identification and quantification of small endogenous polypeptides. (A) Samples consisting of 0.24 µg of tryptic digested UPS, 100 µg of *E. coli* protein and different fmol amounts of heavy labeled AQUA peptides were prepared in triplicates on top of 10kDa spin filters. (B) After centrifugation, the flow-throughs were collected and 2 µL of each sample were analyzed via LC-MS/MS using a Q Exactive Plus instrument ($n = 3$ experimental replicates). (C) UPS1 and AQUA peptides were identified by PEAKS *de-novo* sequencing of MS/MS spectral data. The number of UPS peptide-spectrum matches (PSMs) and proteins were compared to values obtained by the *de novo*-assisted database search strategy performed with PEAKS. In addition, MS1 peak areas of the spiked AQUA standards were compared to values obtained with Skyline.

The UPS1 peptides and AQUA standards were identified by *de novo* peptide sequencing using PEAKS.⁴⁴⁸ The PEAKS software is recognized as a standard for automated peptide *de novo* sequencing, which also introduced the capability of incorporating *de novo* results into its database search algorithm to improve the sensitivity and accuracy of the identifications.¹⁴⁹ In particular, *de novo*-assisted database search is useful for novel PCP identification as the exact mechanisms of how these molecules originate is not well-understood and biasing the results towards the use of an experimental protease and a database digested with specific *in-silico* rules would hinder the characterization of PCP.

Overall, the *de novo*-assisted database peptide sequencing reliably identified a similar number of peptide-spectral matches (PSMs) across the experimental benchmark dataset (**Figure 6.2A**) and identified over 55% of the UPS1 proteins in each sample (**Figure 6.2B**), despite the *E. coli* crude lysate representing ~99% of the collected peptide biomass during enrichment (based on BCA assay). In addition, all the 17 AQUA peptides spiked into samples were identified.

The calculated peak area for each AQUA peptide provided by PEAKS was also compared to manually curated data obtained with Skyline, which is another popular software that provides an alternative to extract MS1-level abundances of a relatively low number of peptides.⁴⁴⁹ In this way, we evaluated PEAKS capabilities of providing quantitative values that could be used for relative quantification of novel PCPs identified in complex biological matrices.

As observed in **Figure 6.2C**, for the spiked AQUA peptide standards, excellent agreement between the Skyline MS1 extracted AUCs and PEAKS AUCs were observed ($R^2 > 0.98$). In terms of individual peptide responses (**Figure 6.2D**), the lowest R^2 value observed was for peptide ADESAYLDEWNLTPAAK ($R^2 = 0.808$) while the best linear response at increasing amounts of injected standards was that from peptide EFDLNFIDK ($R^2 = 0.936$). Overall, the measured changes in linear responses demonstrated the ability to accurately recover peptide quantitative information from a complex proteome background after filtration through a 10 kDa membrane filter, data acquisition via state-of-the-art tandem mass spectrometry, and post-processing using PEAKS.

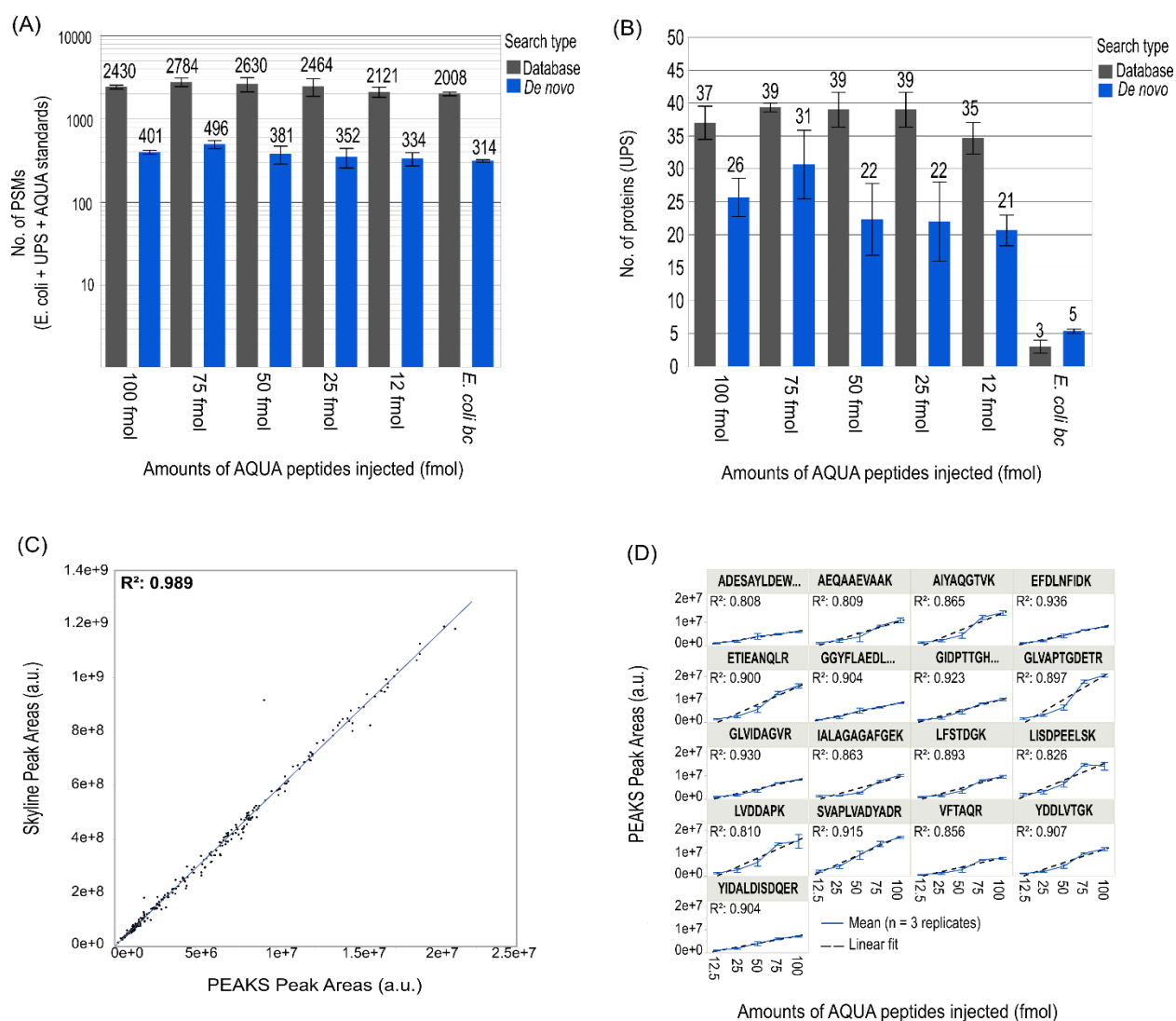


Figure 6.2 Performance of 10 kDa membrane filtering and PEAKS software for peptide identification. For this study we identified peptides from the UPS standards by *de novo* analysis and extracted quantitative information from the AQUA peptide standards spiked into *E. coli* and UPS backgrounds after 10kDa membrane filtering 15 samples x 17 peptides = 255 data points). (A) The number of total identified PSMs for the benchmarking samples and (B) the total number of identified UPS1 proteins are provided. (C) Scatter plot of peptide peak areas reported by PEAKS compared to those identified by automatic and manual validation in Skyline. (D) Linear fit regression analyses of AQUA standards areas under the curve responses according to their molar amounts spiked into benchmarking sample sets. Coefficient of determination values (R^2) are shown in each graph.

6.3.2 Identification metrics of PCPs from *Populus* and its ectomycorrhizal symbiont *L. bicolor*.

The 10 kDa enrichment strategy with the *de novo* sequencing assisted MS/MS database search was applied to *Populus* tissues (*i.e.*, xylem sap, roots, leaves) with or without established *L. bicolor* ectomycorrhizae (ECM) as well as extramatrical *L. bicolor* mycelium (Lb EXM) (**Figure 6.3A**). At a false-discovery rate (FDR) of < 1% at the peptide level, the sample types varied in the total numbers of identified PCPs as well as their similarity to each other (**Figure 6.3B**, horizontal bars in UpSet plot). Overall, the measured PCPs had similar peptide lengths (median length of 21 amino acids) and number of basic residues (average of 1) (**Figure 6.4**) as to what is generally reported in conventional peptide sequencing studies.⁴⁵⁰

Across all datasets, 4564 PCPs were identified from which 1653 and 2859 mapped exclusively and respectively to *Populus* and *L. bicolor* proteins, while 52 PCPs were shared amongst a number of proteins highly conserved between both organisms such as elongation factors and histones (**Supplementary Table S6.1**, see PCPs and Proteins information). Amongst all sample types, the highest numbers of unique PCPs were detected in Lb EXM and mycorrhizal roots (**Figure 6.3B**, vertical bars in UpSet plot). PCPs identifications detected across multiple samples were also observed. The datasets with the greatest number of overlapping PCPs were those of leaves with and without ECM, which is to be expected given the ECM interaction largely affects roots tissues. Additionally, we noticed a significant number of PCPs in common between the Lb EXM samples and the roots samples with and without ECM.

Bioactive PCPs can be occupied by post-translational chemical modifications that influence structure and their biological activity.¹³⁵ For example, the CLE and CEP families have hydroxyprolines that are glycosylated with L-arabinose.^{145, 431, 435, 451} Of the total number of PCPs identified in each dataset, 19-33% contained post-translational modifications or amino acid substitutions as found by the PEAKS PTM and PEAKS Spider modules of the *de novo* assisted database searches, respectively. With the PEAKS software considering 313 natural modifications reported in the Unimod database, we observed PTMs that are likely introduced during sample preparation as well as those that are presumed to occur *in vivo* and have biological significance, like methylation and phosphorylation (**Figure 6.5**).

In the context of sample sets, PCPs matched to a maximum number of 522 proteins (Lb EXM samples) and a minimum number of 87 proteins (xylem sap samples) in the database (**Figure 6.3C**, horizontal bars in UpSet plot). Like the peptide-level observation, we observed substantial overlap in the proteins detected in the Lb EXM, poplar roots with established ECM and roots without ECM (**Figure 6.3C**, vertical bars in UpSet plot).

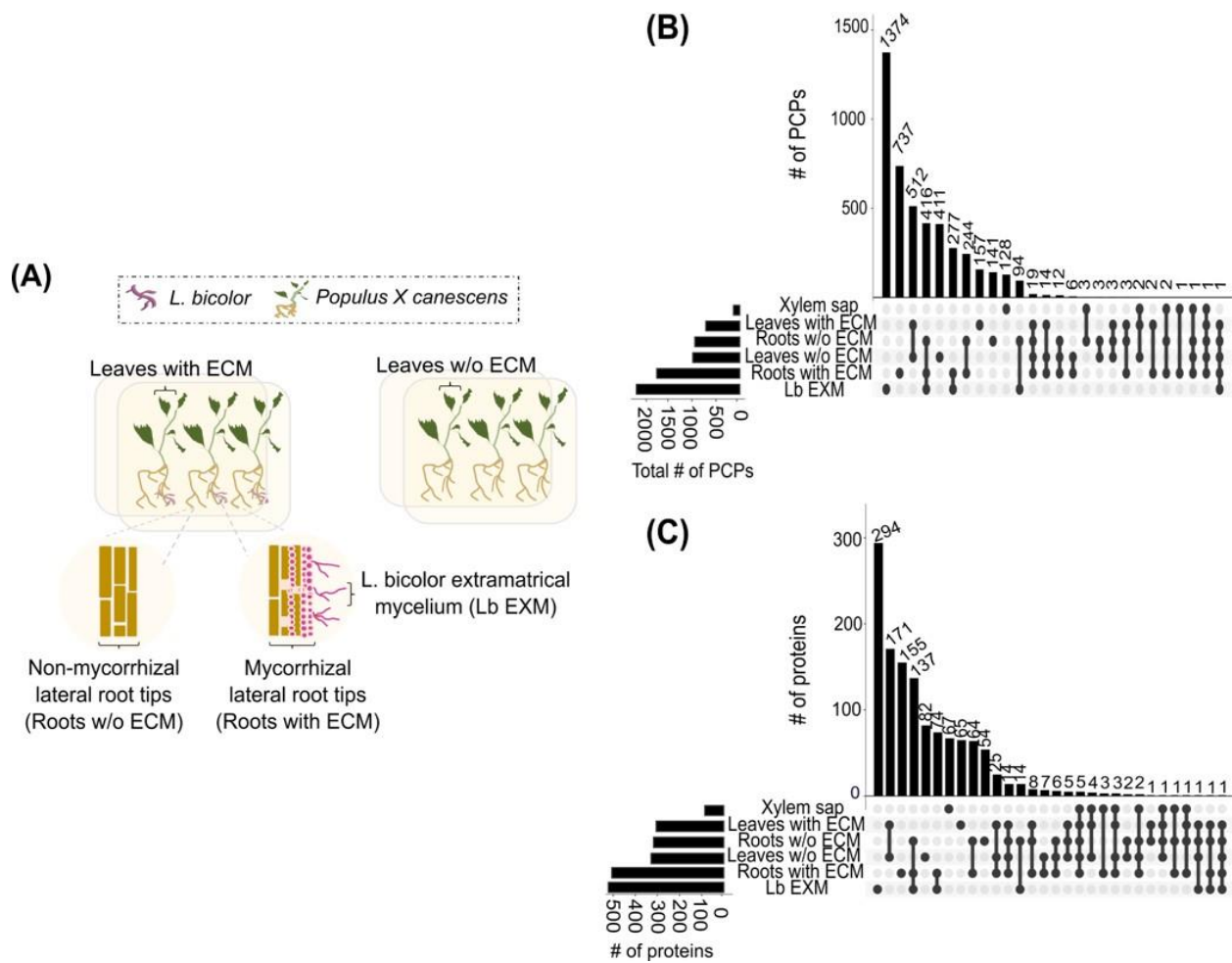


Figure 6.3 Proteomics metrics of the *Populus X canadensis* and *Laccaria bicolor* PMI system. (A) Illustration of the in-vitro sandwich type system used to harvest different *P. X canadensis* extracts and *L. bicolor* extramatrical mycelium. Xylem sap was obtained from plants growing in pots of sterile sand and grown in a greenhouse and is not depicted in this figure. (B) The UpSetR plot depicts the intersection between PCPs identified per experiment type. Horizontal bars indicate the total number of PCPs detected in each experiment (unique + shared); vertical bars depict the number of jointly identified PCPs. (C) The UpSetR plot depicts the intersection between proteins to which PCPs mapped per experiment type. Horizontal bars indicate the total number of proteins detected in each experiment; vertical bars depict the number of jointly identified proteins.

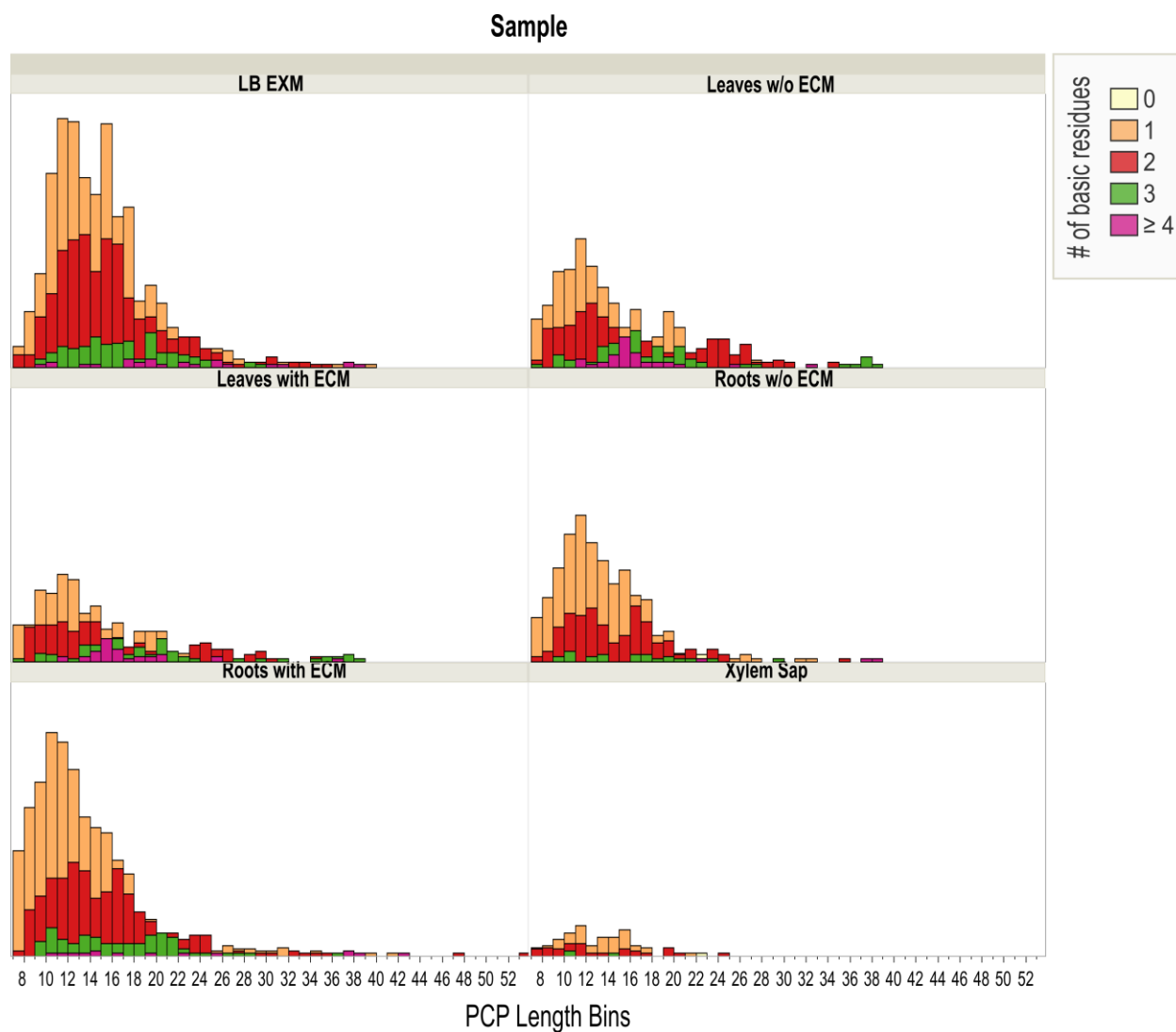


Figure 6.4 Length and basic residue distribution of all identified PCPs (unique & shared) across datasets. Colors overlaid on each histogram represent the number of basic amino acids (Lys, Arg or His) found in each PCP. Lb EXM – *L. bicolor* extramatrical mycelium. ECM – Ectomycorrhiza.

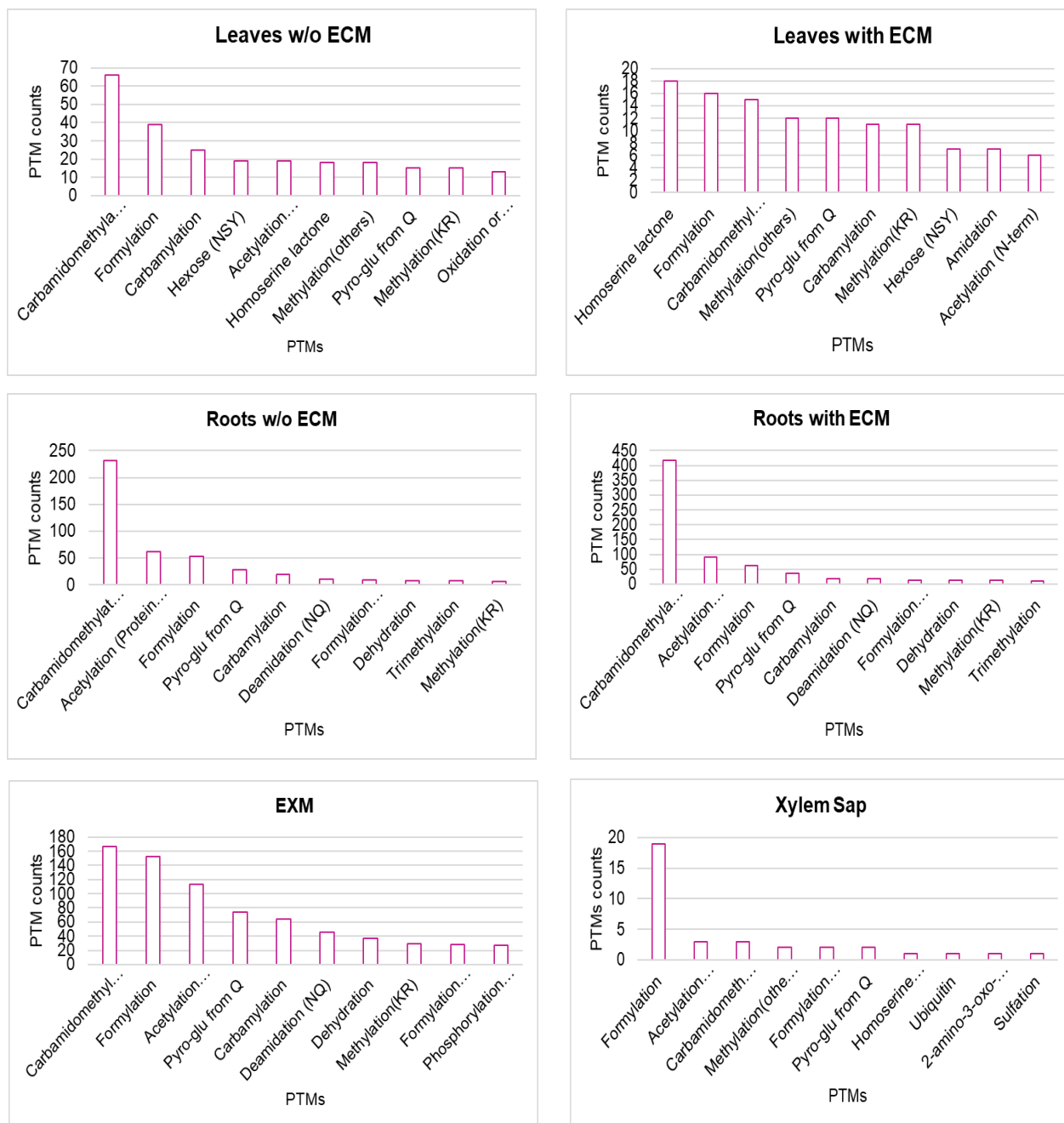


Figure 6.5 Counts of the top 10 most represented PTMs in PCPs identified in different sample types. All PCPs identified by the PEAKS PTM module in each dataset were considered to plot the counts of PTMs observed (Note: The same peptide can have more than one PTM).

6.3.3 *De novo* database-assisted searches enable the identification of non-functional and functional PCPs in the peptidome of different sample types.

PCPs derived from the action of proteolytic mechanisms during protein turnover and during protein maturation contribute to the total pool of native peptides in a cell. In plants, protein turnover is a fundamental component in plant development.^{133, 414, 452} As such, PCPs derived from the degradation of functional proteins can represent a significant fraction of a cell peptidome, but many of their functions are not well understood.⁴¹⁵ On the contrary, several families of bioactive PCPs in plants are derived from protein precursors during certain conditions, like in the presence of microbial pathogens or of heavy metals, and many of their functions have been characterized. Nevertheless, methodologies that can globally profile the expression of both types of PCPs directly in plant tissues and in their microbial partners are still lacking, but our experimental approach can potentially fill this void.

In our datasets, many PCPs mapped to housekeeping proteins like ribosomal subunits, elongation factors and RuBisCO subunits (**Supplementary Table S6.1**). The role of these PCPs in ECM samples or samples without ECM were not further investigated or directly compared, but it is interesting to point that a recent study reported an increase of protein turnover in leaves of *Populus × canescens* plants inoculated *in vitro* with the ECM forming fungi *Paxillus involutus* in comparison to control plants.⁴²⁰ Similar observations have also been made for arbuscular mycorrhizal symbiotic systems grown in lab.^{418, 419} Hence, our MS methodology, in combination with others like traditional proteomics, could be useful for more in-depth investigations on the plant proteolytic “degradome”^{133, 414} by directly analyzing PCPs abundance changes from plant tissues experiencing plant-microbial interactions compared to controls.

Apart from detecting members of the degradome our approach was able to capture previously reported endogenous mechanisms of protein maturation happening in different sample types. For example, we were able to identify PCPs that mapped to the N- and C-terminus regions of histone H2BK4 of *L. bicolor*, suggesting a histone clipping mechanism,^{453, 454} (**Figure 6.6A**) and PCPs in *Populus* leaves that mapped to the proposed mature form of protein PsbTn (~ 3 kDa), which is the smallest subunit of the Photosystem II complex (**Figure 6.6B**).⁴⁵⁵ In addition, we observed several uncharacterized PCPs that may point to unknown mechanisms of protein maturation; for example, N-terminus PCPs mapping specifically to the cysteine-rich domain (CFEM) of a glycosylphosphatidylinositol (GPI) anchored protein of *L. bicolor*, the former being a domain that is common only to fungal extracellular membrane proteins (**Figure 6.6C**).⁴⁵⁶

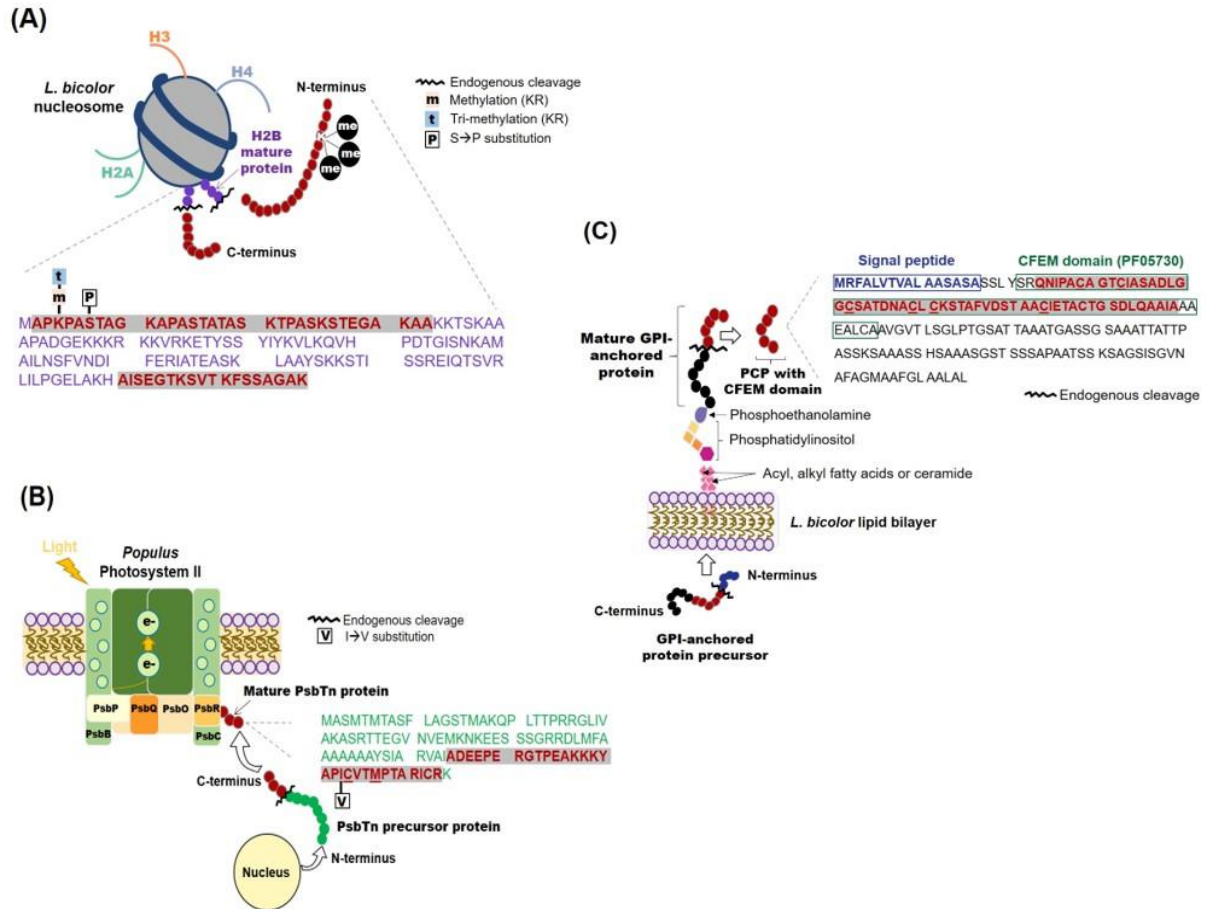


Figure 6.6 Visual representations of detected PCPs (highlighted in grey and in red letters) produced as results of known and hypothesized protein maturation mechanisms in *Populus* and *L. bicolor* samples. (A) PCPs mapped to the N-terminus and C-terminus regions of *L. bicolor* histone H2BK4 in *L. bicolor* extramatrical mycelium (Lb EXM) as well as root with and without established ECM. (B) The mature form of protein PsbTn from the Photosystem II complex of *Populus* was identified in leaves with and without ECM. (C) Identified PCPs from a CFEM-domain containing protein with a predicted GPI-anchored site. PCPs spanned most of the CFEM domain. Peptide mapping data is from Lb EXM.

6.3.4 Comparative quantitative analysis of PCPs defines a set of significant analytes in ectomycorrhiza symbiosis between *Populus X canescens* and *L. bicolor*.

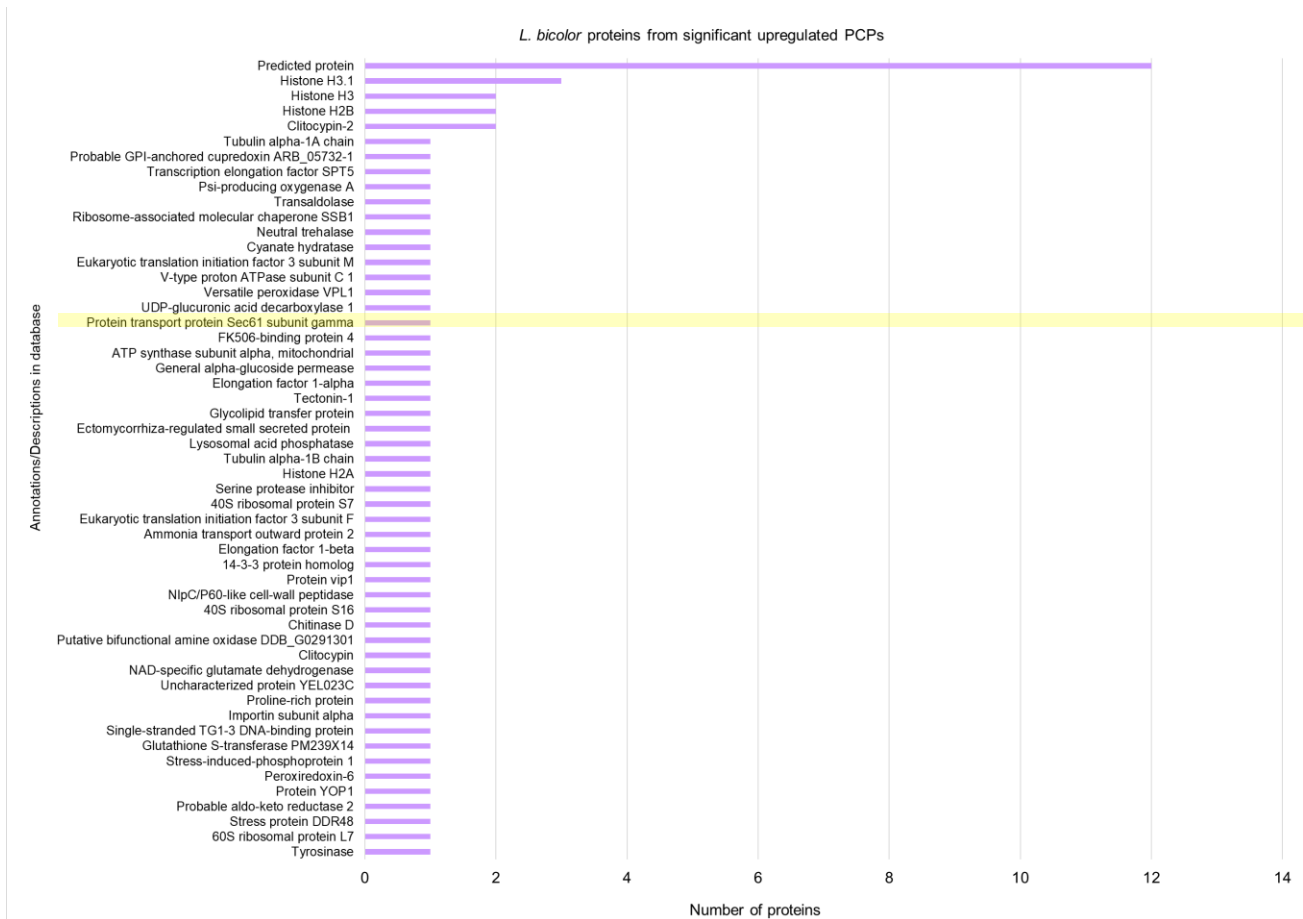
A differential expression analysis of PCPs and the proteins they mapped to in root samples with or without ECM and Lb EXM was useful to narrow the dataset into groups of PCPs with biological significance in the symbiotic associations between plant and fungus. Overall, 398 PCPs were differentially abundant (ANOVA p-value < 0.05; **Supplementary Table S6.2**) between roots with and without established ECM and Lb EXM. To further delineate the relevancy of PCPs to Lb symbiosis with *Populus* roots, TukeyHSD post-hoc test was performed to quantify abundance differences between relevant sample groups. Based on this analysis, a total of 157 PCPs were determined to be significant in the mutualistic interactions between *Populus* and *L. bicolor*. Of the 157 PCPs, 96 mapped to 69 *L. bicolor* proteins, 56 mapped to 64 *Populus* proteins and 5 were shared between 12 *Populus* and 6 *L. bicolor* proteins, respectively.

Like other large-scale ‘omic’ analyses, measured differences between relative abundances of analytes can be explained by numerous phenomena. Regarding PCPs, quantitative differences can arise from several levels of regulation (e.g. the amount of protein expression, the rate of protein translation, protein and peptide turnover rates, post-translational modifications, etc.) and therefore, while we highlight several intriguing observations below and their likely implications, the mechanism(s) regulating these differences are largely unknown.

6.3.4.1 *L. bicolor* endogenous peptides associated with symbiosis with *Populus*.

Of the 157 PCPs determined to be significant in the mutualistic interactions between *Populus* and *L. bicolor* 96 PCPs mapped to 69 *L. bicolor* proteins (**Figure 6.7A**). Amongst the differentially abundant *L. bicolor* PCPs in root with established ECM compared to Lb EXM were those belonging to a variant of the cysteine-rich C-terminus of the *L. bicolor* Mycorrhiza-Induced Small Secreted protein MiSSP7.6, that in our database is annotated as MiSSP7.7 and has 97.3% pairwise identity to MiSPP7.6 (**Figure 6.7B**). The SignalP-5.0⁴⁵⁷ and Phobius⁴⁵⁸ webserver agreed on the presence of a signal peptide on the first 21 amino acid residues of MiSSP7.7. It has been previously reported that MiSSP7.6 is secreted by *L. bicolor* and later imported into *Populus* cells as a mature peptide of 54 amino acids. The matured MiSSP7.6 PCP localizes in the nuclei of plant cells where it interacts with Zinc-Finger transcription factors belonging to the Myb/SANT-like DNA-binding domain protein family.⁴²⁷ The C-terminal PCP detected in this study may point to a plausible mechanism of protein maturation.

(A)



(B)

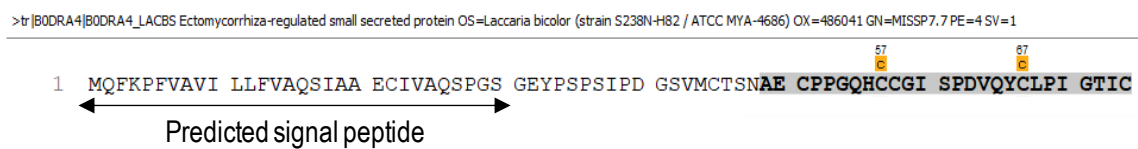


Figure 6.7 *L. bicolor* proteins from significant upregulated peptides associated with ECM. (A) A total of 96 PCPs mapped to 69 *L. bicolor* proteins, amongst them, a MiSSP7.7 protein (97.3%identity to MiSSP7.6 reported in literature, highlighted in yellow). **(B)** The region to which identified PCPs mapped to MiSSP7.6 is highlighted in grey color. A predicted signal peptide in the protein is also shown.

In addition, we identified 4 putative small secreted proteins of < 250 amino acids (B0FD3, B0D9T2, B0DP29 and B0DXP4)⁴⁵⁹ lacking functional annotation. As research on the symbiosis between *Populus* and *L. bicolor* has been uncovering novel SSPs derived from both organisms necessary for the establishment and regulation of the symbiotic association of *L. bicolor* and Poplar,^{427, 428, 443, 444} this dataset contributes additional novel protein candidates associated with ectomycorrhizal symbiosis.

Another group of expressed PCPs in roots with ECM that mapped to several *L. bicolor* proteins involved in the detoxification of reactive oxygen species (ROS). This observation is likely related to the type of stress responses that mycorrhizal fungi can induce to plants which include the generation of ROS species like hydrogen peroxide, superoxide anion, and hydroxyl radicals.⁴⁶⁰ Amongst those proteins, we identified a tyrosinase (B0DMA1) that provides resistance against reactive oxygen species and phenolic compounds during early symbiosis,⁴⁶¹ a cysteine peroxiredoxin (B0CY32) and a manganese peroxidase (B0DVT9).

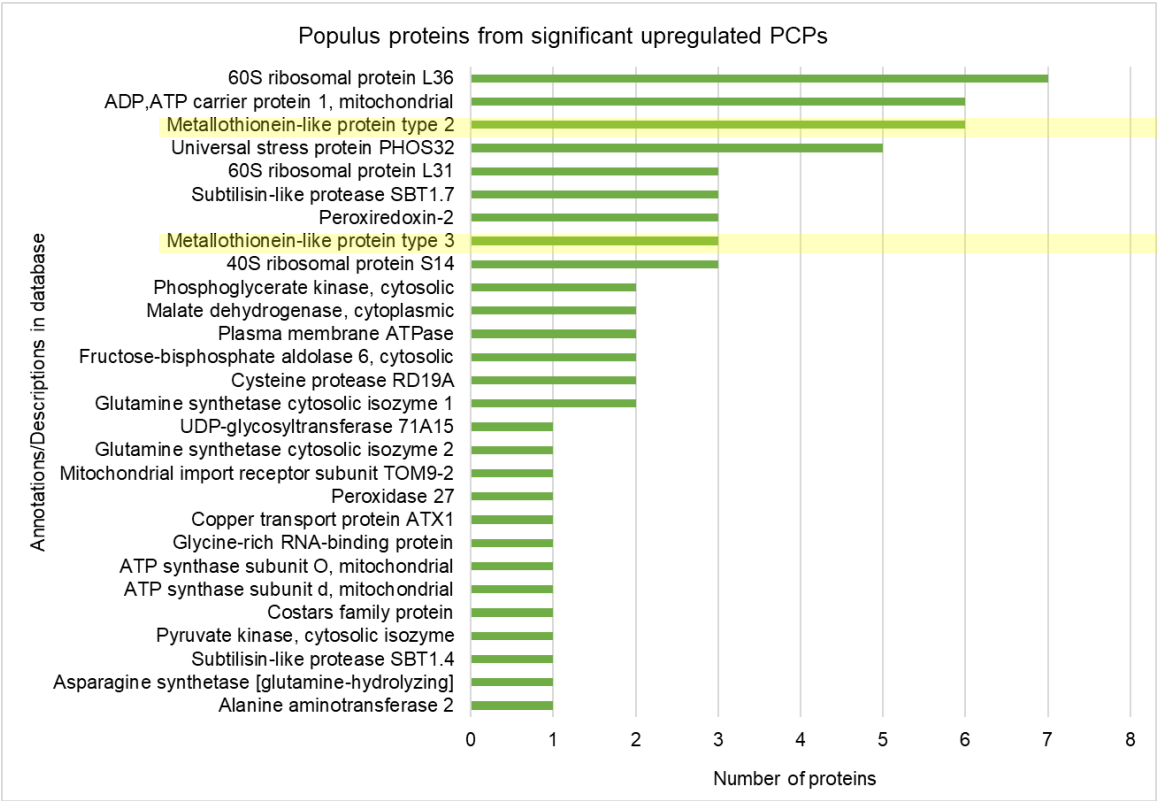
Interestingly, amongst the differently expressed *L. bicolor* proteins we also identified PCPs that mapped to a lectin-related tectonin 1 protein (B0D1K0). In general, fungal lectins are best known to be part of the defense mechanisms against soil microbes and predators that act by targeting glycoepitopes in the cells surfaces of the latter;⁴⁶² however, fungal lectins that recognize host cell glycans in ectomycorrhiza, albeit, not specifically in *Populus* – *L. bicolor* ECM, have also been hypothesized to help mediate the adhesion of partners to hosts.⁴⁶³ Interestingly, a study investigating the molecular mechanisms behind the physical interactions of *Laccaria bicolor* S238N hyphae and its mycorrhiza-helper bacteria *Pseudomonas fluorescens* BBc6R, found upregulation of the *L. bicolor* tectonin 2 gene, where it was suggested to have a role in cell-to-cell recognition.⁴⁶⁴ Thus, it is plausible to hypothesize that the tectonin 1 found here has a similar function to tectonin 2, but with Poplar root cells instead, and opens an interesting experimental opportunity for its functional validation.

6.3.4.2 *Populus X canescens* endogenous peptides associated with symbiosis with *L. bicolor*.

Overall, the majority of the 56 *Populus* PCPs found to be significant in ectomycorrhizal root compared to Lb EXM mapped to proteins related to stress response mechanisms (**Figure 6.8A**), with the metallothionein (MT) protein family being highly represented (**Figure 6.8B**). During endomycorrhizal symbiosis, metallothionein proteins are generally hypothesized to provide protection against oxidative stress including the one caused by heavy metals.^{465, 466} Interestingly, the differentially abundant polypeptides map exclusively to the N-terminus and C-terminus of these proteins, which are both cysteine-rich regions. An intriguing explanation for the specific mapping to these regions is that *Populus* MTs experience post-translational cleavage, a largely overlooked regulatory mechanism for this protein family. Cleavage of MT by proteases has been shown to be largely dependent on the loading of the protein with heavy metals⁴⁶⁷ and occurs among the large stretch of amino acid residues separating the cysteine-rich domains.⁴⁶⁸ A study on a plant MT expressed in *E. coli* showed that the linker region was sensitive to the bacterial proteolytic system;⁴⁶⁹ however, whether this process occurs in nature or its functional importance in symbiosis has not yet been shown.

Additionally, PCPs mapped to *Populus* house-keeping proteins involved in protein expression machinery (*i.e.*, subunits of ribosomal proteins) and/or components of active cell signaling/communication systems (*i.e.*, ADP/ATP carrier proteins and metal transport proteins). Similar observations have been reported at the transcript level in active ECM symbiosis between *Pisolithus tinctorius* and *Eucalyptus globulus* where the authors concluded that this was expected due to intense bilateral metabolic transfers and high assimilative activity.⁴⁷⁰ Of note, although we cannot accurately pinpoint to specific mechanisms of how the PCPs observed from these proteins are being produced endogenously, there is no indication in literature that such molecules are further processed to mature PCPs and thus, we can hypothesize that mechanisms of protein turnover or degradation are the most plausible explanations behind their identification by our experimental approach. Complementary proteomics analyses measuring the relative levels of protein precursors from which these PCPs are derived from can confirm the natural degradation of these proteins and/or provide a different functional perspective from pools of PCPs, similar to what others have found before.⁴¹⁵

(A)



(B)

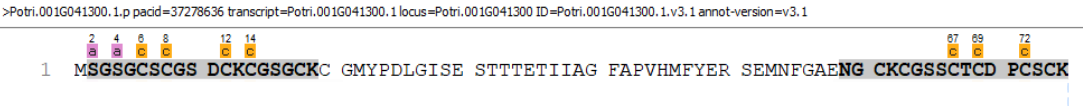


Figure 6.8 *Populus X canescens* proteins from significant upregulated PCPs associated with ECM. (A) A total of 56 PCPs mapped to 64 *Populus* proteins, amongst them, proteins related to stress protection like metallothioneins (highlighted in yellow). Other PCPs mapped to *Populus* house-keeping proteins involved in protein expression machinery and/or signaling systems. (B) Example of one of the six metallothionein-like type 2 proteins identified (ID. Potri.001G041300.1). All identified PCPs from this protein mapped to its N- and C-terminus regions (highlighted in grey).

6.3.4.5 Differentially abundant PCPs shared between *L. bicolor* proteins and *Populus X canescens* proteins.

The five differentially expressed and shared PCPs between *Populus* and *L. bicolor* mapped mainly to functional domains of polyubiquitin and ubiquitin proteins which are well conserved proteins across life.⁴⁷¹ Ubiquitins can be found as monomers and polymers and have distinct roles in intracellular protein degradation via the proteasome pathway.^{472, 473} Several studies have demonstrated high and significant expression of ubiquitin genes in the establishment of ectomycorrhizal symbiotic associations. For example, in the *Eucalyptus globulus-Pisolithus tinctorius* ectomycorrhiza, upregulation of plant derived transcripts encoding components of the ubiquitin pathway were reported, where it was hypothesized that these proteins help mediating the downregulation of plant protein synthesis in colonizing roots.⁴⁷⁰ Another study analyzing the transcriptional responses in the ECM interaction between the basidiomycete *Paxillus involutus* and its host birch *Betula pendula* compared to extramatrical mycelium, shown a polyubiquitin that was upregulated by 4.2-fold in ECM.⁴⁷¹ In addition, the generation of reactive oxygen species like hydrogen peroxide is known to induce the expression of polyubiquitin in cells as a mechanism of detoxification⁴⁶⁰ and in this study we saw significant expression of fungi ROS-related enzymes in root with ECM.

6.4 Conclusions.

Over the last couple of years an increasing body of evidence at the transcript level has demonstrated the expression of bioactive PCPs in biotrophic plant-fungal associations. However, direct evidence of the products of genes and transcripts believe to be involved, or participating in ECM establishment and/or regulation, is still lacking in literature, but can certainly provide a new layer of evidence in the functional understanding of these molecules. As certain characteristics typical of endogenous PCPs in any biological system can make their identification laborious and difficult (*i.e.*, unknown cleavage specificity, the presence of multiple PTMs, and relatively low natural abundances), experimental approaches that can comprehensively identify them with sensitivity and specificity are most welcome.

In this study, we systematically investigated the use of a 10 kDa membrane filtering strategy to enrich and quantify *Populus* and *L. bicolor* PCPs expressed in active mycorrhizal lateral root tips and identified novel peptide sequences using LC-MS/MS with de novo-assisted database search strategy. Particularly, de novo-assisted database searches offer a way to detect

functional products of genes and transcripts without biasing the identification towards the use of specific proteases or cleavage mechanisms like in traditional large-scale protein studies.

The differential expression analysis of PCPs conducted throughout this study enumerated PCPs that have a role in the symbiotic interactions between both *Populus* and *L. bicolor* and provide theories of their biosynthetic origins. For example, we found evidence of a *L. bicolor* PCPs mapping to the C-terminal region of a MiSSP7.6 protein, or of PCPs that mapped to N-terminus and C-terminus regions of *Populus* metallothionein proteins that are related to stress responses happening during the formation of ECM. These cases highlighted the utility of the LC-MS/MS methodology to capture potential mechanisms of endogenous protein maturation. Interestingly several PCPs mapped to small *L. bicolor* proteins of unknown function from which four were predicted to be secreted.

Housekeeping proteins supporting the crosstalk and metabolic adjustments between plant and fungi such as transporters and ribosomal proteins, were also detected. It was hypothesized that the identification of PCPs, and their PTMs, mapping to these types of proteins are related to turnover/degradation proteolytic mechanisms which are supposed to be upregulated in AM and ECM symbioses.⁴¹⁸⁻⁴²⁰ Their detection may prove useful alongside relative measurements of protein precursors to determine if such proteins are being subjected to natural turnover or if these PCPs serve another role in the symbiotic system under study. Overall, this study provides experimental support for a quantitative strategy to investigate the global expression of PCPs in PMI systems allowing to identify those with potential roles in their associations.

CHAPTER 7

Outlook, challenges, and conclusions

7.1 The potential of targeted proteomics to monitor environmentally relevant microorganisms in bioremediation.

Anthropogenic pollutants are found worldwide. The expansion of human population and its related activities such as industrial production processes and agriculture has led to an increase in the amounts of toxic organic and inorganic compounds that are released in the environment. To partially alleviate this global problem, active research on naturally occurring microorganisms at contaminated sites that can degrade, transform, or accumulate a wide range of hazardous compounds such as hydrocarbons, heavy metals, pesticides and chlorinated compounds⁸⁸ has sparked interest into harnessing their intrinsic metabolic capabilities as a sustainable way to degrade and detoxify contaminants in the environment. These strategies, known as monitored natural attenuation (MNA) strategies,⁹⁷ differ to others that employ genetically modified or non-native species, and require a deep understanding on microbial physiology and the cellular processes behind contaminant degradation.

As with other areas of microbiological research, the application of 'omics' technologies to this field has not only greatly improved our fundamental understanding of the biology of these microorganisms under controlled conditions; but they have also allowed their direct study in contaminated environments, which has provided information on complex microbial dynamics happening *in situ*.⁸⁷ The latter has been particularly useful because unlike laboratory-scale simulations, the application of MNA strategies takes place across environments with different geochemical gradients, and diverse geophysical and hydrological complexities⁴⁷⁴ typically involving the presence of more than one contaminant and of different strains of microbes with specialized metabolic capacities.

In particular, the application of quantitative proteomics techniques for bacterial-mediated bioremediation research provides critical insights into important cellular activities with temporal and spatial resolution. Different than genes and/or transcripts, proteins are the direct catalysts behind the variety of metabolic responses exerted by bacteria to contaminants in the environment. Proteome analysis of these environmentally relevant group of microorganisms, either growing under laboratory-controlled conditions or directly from environmental samples, can be used to identify and compare the suite of expressed proteins from a particular sample in a particular state, in order to elucidate metabolic pathways or identify group of proteins that can potentially be used as biomarkers that register the desirable degradation response to a contaminant in the

environment. Measuring proteins in bacteria provides direct and relevant information that can be used to show that certain bioprocesses are occurring *in situ* and/or evaluate a certain set of conditions that favor the degradation of a contaminant or class of contaminants.

As a case example, genomics, transcriptomics and proteomics research on the anaerobe *Dehalococcoides mccartyi* (*Dhc*) bacteria has provided enough gene and protein level evidence over the years to assess the metabolic capabilities of *Dhc* bacteria to transform chlorinated ethenes to benign ethene via the expression of reductive dehalogenase enzymes (RDases). Early proteomics studies, for example, focused on the identification of RDase enzymes on cultures of *Dhc* bacteria amended with different organohalides as electron acceptors and in varying degrees of concentrations, which allowed to gain a deeper understanding of the overall physiological activity of the organism. Proteomics research on *Dhc* bacteria has also enabled a refined validation of transcriptome results, as it has been observed in some cases that transcripts are not always translated into functional proteins. Improvements in sample preparation and MS technologies have enabled a deeper *Dhc* proteome coverage with percentages in different *Dhc* strains ranging from 60-72% up to the ~ 9% that was reported in earlier studies.

The application of proteomics techniques to *Dhc* research has sparked the interest of the research community of employing proteomics techniques for more specialized applications like *in situ* monitoring of bacteria in contaminated groundwater as well as calculations of protein activities. Indeed, directly applying proteomics to groundwater could provide information into the actual functional activities of *Dhc* associated with each sample and complement state-of-the-art gene-targeted approaches that are commonly used to monitor *Dhc* cell abundances *in situ*. These two goals are also appealing to site manager regulators that look to know if *Dhc* is present in a site, at what levels, and if the bacteria are performing the desirable pollutant degrading activity. One mode of MS data collection that can be used to answer these questions is targeted proteomics via MRM-MS or PRM-MS. Monitoring specific *Dhc* RDases in groundwater with sensitivity and in a faster way than with global proteome measurements, can in theory allow the determination of *Dhc* bacteria in a contaminated site while at the same time providing relevant information for the potential of enzymatically driven dechlorination. Furthermore, the shift from relative to absolute quantitation can be used to calculate enzymatic degradation rates.

Few studies have explored the use of targeted proteomics for the organohalide respiration capabilities of *Dhc* bacteria, but the systems under investigation have been restricted to anaerobic mixed cultures grown in the laboratory. The results presented in those studies have provided levels of targeted protein copy numbers per cell and even enzymatic kinetic rates. However, transitioning this knowledge from simpler cultures to groundwater, where multiple metabolic

processes mediated by taxonomically diverse bacteria are simultaneously happening is not trivial. For instance, peptides selected in targeted studies that act as surrogates to monitor the expression and calculate absolute abundance levels of a protein, can be not specific when analyzing an environmental sample. Such a case was demonstrated with the study presented in Chapter 3 of this dissertation, where the specificity of each selected peptides had to be compared *in-silico* to the proteomes of other known type of bacteria living in groundwater or even on a broader taxonomical context.

Importantly, the natural dynamic ranges of proteins expressed in an environmental sample as well as other interferences can impact the detection of *Dhc* RDases in these types of samples, and even when targeted proteomics provides sensitivities that are 1-2 orders of magnitude higher than global measurements, targeted peptides and proteins may still not be detected. This demonstrates the need of identifying peptides by this method as a first step to plan and execute their absolute quantitation, which was the main goal of Chapter 3. It is expected that the list of peptides selected in this study can then be useful for the development of a quantitative study, as several of these peptides were detected in different groundwater matrices and their specificities checked. Further studies could evaluate the dynamic ranges of these peptides and provide estimates of the number of protein copies per volume of groundwater.

The application of targeted proteomics technology for protein analysis in the field of environmental biotechnology is still in its infancy but when combined with adequate functional genomic/proteomic analysis of degradative bacteria can potentially yield information of genes/proteins acting as bioindicators of important metabolic process in the environment. A recent example includes the quantitative measurement of nitrogen regulatory proteins, with peptides capable of distinguishing proteins derived from either *Prochlorococcus* and *Synechococcus* cyanobacteria living in sea samples collected from the Central Pacific Ocean.⁴⁷⁵ Interestingly, the applications of targeted proteomics are not only limited to bacteria, but recent efforts have also been applied to higher organisms. For example, in the sentinel amphipod *Gammarus fossarum*, a targeted proteomics method using MRM-MS was recently developed for the simultaneous quantification of 38 peptides reporting for 25 proteins, whose changes in abundance act as sensitive indicators for the presence of contaminants in river water.⁴⁷⁶

These are just some examples of studies demonstrating the use of targeted proteomics strategies to monitor protein biomarkers in the environment, and the advantage that it provides but many more are yet to come in the next years.

7.2 Measurement of groups of proteins by targeted proteomics to monitor metabolic processes in anaerobic bioreactors for biogas production.

Biogas is one of the products obtained from the microbial-mediated anaerobic degradation of organic materials with high lignocellulose content like food, municipal, and agricultural wastes. Production of biogas in anaerobic digesters holds great potential for the generation of electricity and vehicle fuel. The application of molecular “omics” techniques to anaerobic digesters has now enabled a more detailed understanding of the microbial metabolic processes occurring during the production of biogas. Particularly, metaproteomics has linked genetic information to the functional characterization of microbial communities thriving in anaerobic environments and has provided information on regulatory processes necessary for enzyme production. Amongst other advantages of metaproteomics is also the potential to identify proteins, and/or their temporal shifts in abundances, in response to environmental and operational parameters which could be used to monitor the performance of an anaerobic digester, for example, of either of its stable or unstable operation.

In fact, several publications in the field have mentioned an interest in the use of proteins that are highly expressed during the process of biogas formation as functional biomarkers that could act as predictive biomarkers of perturbations in a system.^{354, 355, 477} One of those examples are variations in the abundance levels of the methyl-coenzyme M reductase, a key enzyme in all methanogenic pathways from which decrease in abundance has been correlated to acidification conditions in biogas plants.³⁵⁷ Besides proteins in methanogenesis, other proteins of potential interest could be glycoside hydrolases (GHs) which are important for the deconstruction of polysaccharides buried in lignocellulosic material. Hydrolysis of sugar polymers is considered a bottleneck in the anaerobic digestion process because it can either underperform or outperform other metabolic steps based on initial conditions, *i.e.*, the use of highly lignified substrate sources. Although metaproteomics approaches to reveal protein content and abundance in anaerobic digesters can in theory improve process efficiency and stability, the amount of time and effort required behind these types of experiments limits them to be performed routinely and in as many samples as possible. However, targeted proteomics experiments present an interesting alternative to discovery-based proteomics measurements of proteins as they allow for flexible, quantitative, and relatively routine measurement of hundreds of peptides within a shorter time frame.⁴⁷⁸ Compared to immunoassays, targeted proteomics also features faster assays development and deployment and in principle can distinguish similar proteoforms such as isoforms and post-translational modified proteins.⁴⁷⁹

In Chapter 4, the initial steps in the development of a targeted proteomics assay to detect GHs in anaerobic digesters that could provide information of the hydrolytic performance of the latter, were presented. However, different than in the development of other targeted proteomics assays, which are defined by the selectivity of peptides to specific proteins, it was decided to show the selection of unique peptides capturing entire families of GHs. Some of these peptides, although shared amongst individual members in a family, were specific to GH families meaning that they cannot be produced from other proteins in the background nor found across different families. Provided that sufficient data acquisition has suggested the expression of groups of GHs at determined ratios to be indicative of the hydrolytic capacity of a digester, then monitoring groups of proteins without necessarily knowing the signal contribution of specific proteins in a group could provide of a faster assessment of the hydrolytic capacity of a reactor.

It was observed that several of the selected peptides mapped exclusively to GHs belonging to different microbes but also to GHs with different enzymatic substrate affinities within same families of enzymes. The latter characteristic may be even more important for anaerobic digesters as the starting substrates can vary widely in their lignocellulosic content. Chapter 4 is just the beginning of the development of a targeted proteomics assay and more work is required to get a more manageable number of peptide candidates capturing groups of GHs to test experimentally in a mass spectrometer. As proteins give the most direct evidence of the functionality of active organisms within microbial communities, targeted proteomics is an interesting alternative to metaproteomics and immunoassays towards the detection of proteins biomarkers in selected metabolic routes in anaerobic digesters.

7.3 Getting more out of DDA-collected spectra in bottom-up proteomics experiments.

The interpretation of the thousands or hundreds of thousands of peptide sequences in a protein database and their assignment to experimental spectra is a key challenge of MS-based bottom-up proteomics experiments. After a peptide is fragmented in the mass spectrometer, a series of characteristic set of ions, such as *y*- and *b*-ions, are used by a search engine to compute the probability that an experimental spectrum represents an amino acid sequence in a database. These peptide assignments are often achieved through scoring of the experimental spectrum relative to a theoretical spectrum of a candidate peptide.⁴⁸⁰ By setting up a score significance threshold, peptides that are above the threshold are accepted, while peptides with a lower score are rejected. Thus, the success that any given search engine has of finding and assigning the correct peptide to a mass spectrum depends of how many fragment ions are characteristic for a particular peptide sequence and how many other peptides are present in the database.⁴⁸¹ Therefore, when two or more different peptides of similar masses co-elute within the same time window in a data-dependent acquisition procedure (see section 1.1.2), the fragment ions of these “interfering” peptides, add to the uncertainty of peptide identification.⁴⁸²

Chimeric spectra (see section 1.4.1) are unavoidable in LC-MS/MS bottom up proteomics experiments that analyze complex biological mixtures. As mentioned before in Chapter 4, several studies have reported that the percentage of chimeric spectra collected in bottom-up proteomics experiments analyzing single cell isolates can be as high as 50% of the total data,¹⁰⁹ with newer studies using state-of-the art mass spectrometer, putting this percentage close to 39%.¹¹⁸ Thus, it is safe to assume that even greater percentages are expected in experiments analyzing samples that are the target of metaproteome studies. Interestingly, one study reported that even when every possible peptide precursor ion could be targeted for fragmentation in a LC-MS/MS experiment via DDA, cofragmentation of precursor ions would be unavoidable as many peptides coelute from the liquid chromatography part of the experiment.¹⁰⁷ The same study also suggested two ways of solving this issue: Computationally and/or by improving the capabilities of MS technology by affording high resolution precursor selection.

Computationally, and in order to correctly assign multiple peptides to the concurrent product ion spectra, conventional database algorithms have been improved over time and some of them were already described in section 1.4.2. The project presented in Chapter 3, evaluated the performance of the CharmERT computational workflow to deconvolute chimeric spectra under different LC setups and isolation windows commonly employed in bottom-up proteomics research and proposed that when combined with database engines using the multiple peptides-per-

spectrum-match approach (mPSM), like CharmERT, HPLC configurations can still provide comparable identification metrics to UHPLC configurations that overall improve chromatographic peptide resolution and lower peptide coisolation, but that are difficult and costly to implement.

However, besides specialized software that is tailored to deconvolute chimeric spectra in proteomics data, other interesting alternatives are coming to light. For example, two research groups^{483, 484} have reported substantial improvements in the accurate prediction of theoretical fragmentation spectra that is used in database search by using deep learning, a branch of machine learning algorithms.⁴⁸⁵ In brief, these algorithms are trained on large dataset of experimental fragmentation spectra, but different than other methods, they can more accurately predict fragment ion intensities, by taking into consideration additional factors like collision energy and fragmentation modes. These accurate reference spectra are then used for matching against experimental spectra from either DDA or DIA experiments.⁴⁸⁰ These studies reported small increases in the number of peptide identifications at 1% FDR, but both algorithms provided higher identification gains when searching large databases, such as those that would be typical used for metaproteomics.⁴⁸⁰ One of those studies⁴⁸³ concluded by stating that one potential use of deep learning algorithms could be in deconvoluting chimeric spectra, perhaps by better differentiating clusters of “contaminating” fragment spectra based on similarities to training datasets, or by combining information from predicted retention time profiles of peptide sequences that can co-elute under different chromatography conditions.

Besides the use of deep learning algorithms, developments in LC-MS/MS instrumentation can also help in the task of increasing measurement depths to what has already been collected. For example, newer types of mass spectrometers, like the Orbitrap Fusion Lumos Tribrid⁴⁸⁶ instrument provides resolving power up to 500,000 FWH at m/z 200 and scan rates up to 20 Hz, and it would be interesting to compare how the improved mass resolution helps in the process of selecting precursor ions that are targeted for MS/MS fragmentation. In the same way different modes of LC operation are being explored to counter, in part, the effects of coelution in bottom-up proteomics experiments. For example, one group explored a multiple solvent approach of resolving peptides from a HeLa cell digest consisting of consecutive and separate gradient runs using acetonitrile, methanol, and acetone, to reduce the effect of coelution and ion suppression. Their strategy resulted in almost a 30% increase in the number of peptide identifications compared to a simple technical triplicate of the same digest but using a typical single acetonitrile solvent gradient.⁴⁸⁷ Another group explored temperature manipulation at different stages of a two LC column setup and suggested that loading peptides at subambient temperature (0 °C) could improve the characterization of hydrophilic peptides, which do not resolve quite well in proteomics

experiments, by providing narrower peak shapes; but that also heating of the main analytical column could help improving chromatographic resolution by reducing peak widths and facilitate peptide elution.⁴⁸⁸

Due to the nature of the samples analyzed by bottom-up global LC-MS/MS proteomics experiments, and of the data-dependent spectral acquisition mode used typically on them, some level of peptide information often escapes these types of analyses in the form of chimeric spectra. The occurrence of chimeric spectra lowers the rate of peptide spectral matches that are assigned by common database search algorithms, but they also present an opportunity to increase measurement depth. Although for low complexity samples, such as single cell isolates, the depth of measurement achievable today be enough to answer the questions under investigation, studies analyzing higher complexity samples, could benefit from the application of bioinformatic software or experimental tools that can recover more information from chimeric spectra, or lower their chances of occurrence, respectively. As database search algorithms get improved and LC-MS/MS instrumentation cheaper, deconvoluting chimeric spectra from any given LC-MS/MS bottom-up proteomics dataset could become a routinely step of any proteomics pipeline.

7.4 Advantages of LC-MS/MS peptidomics to uncover bioactive peptides in different systems.

The discovery of peptides of less than 100 amino acids that exist within cells and tissues of different organisms across prokaryotes and eukaryotes have provided new insights into diversity of molecules that govern the functionality of biological systems. These molecules are involved in important biological roles including translational control, cell-to-cell communication, development and protection mechanisms.⁴⁸⁹⁻⁴⁹² While gene-based methods and targeted protein-based methods like western blots have been the main tools used for the discovery and characterization of endogenous peptides, the application of LC-MS/MS based methodologies has provided comprehensive evidence of their expression and stability. LC-MS/MS peptidomics workflows aim to detect and identify endogenous peptides in a biological system with molecular weights ranging between 0.5-15 kDa.¹³⁷ Although peptidomics has borrowed from analytical LC-MS/MS proteomics strategies,⁴⁹³ there are fundamental differences and considerations that need to be addressed when planning an experiment of the latter type.

One of those considerations is the distinct biosynthetic origins of endogenous peptides. In diverse organisms, these origins may not only include protein maturation mechanisms or natural turnover but also small, and in many cases, unannotated open reading frames. Each of these origins reflect the acting of frequently unknown proteolytic mechanisms; post-translational modifications, and the co-existence of different versions of peptides derived from the same gene or precursor in a sample. Thus, to capture the biological origins of these peptides, LC-MS/MS based peptidomics strategies may not necessarily require the enzymatic digestion of the sample¹⁴⁷ (*i.e.*, the use of trypsin), and in some cases, this may be even detrimental for their identifications, due to the limited number of cleavage sites available on them.

The issue gets further complicated when deciding how to search spectral data derived from endogenous peptides. Typical LC-MS/MS based proteomics employ databases that are pre-digested *in-silico* to generate a list of theoretical peptides and spectra that is matched to experimental data; however, without the constraints of sites for proteases in peptidomics data analysis, database searches can become a real challenge. Using multiple proteases for *in-silico* digestions not only increases the search space and computational resources needed to assign peptide sequences to experimental spectra, but, false positive rates can also reach very high levels compared to traditional proteomics.¹⁴⁷ Other studies, especially in plants, have used custom-made databases including known variants of families of endogenous peptides, but this limits the scopes of identifications or works better with those analytes that have been studied with more detail. In addition, peptidomics studies may also rely on single identifications, as peptides

can originate from small protein precursors or, as mentioned before, be directly translated from sORFs that, in addition, may not be represented in protein databases.⁴⁹⁴ Besides these issues, others, like the relative natural abundance of peptides in a proteome (compared for example to housekeeping proteins), their variable sizes (2-100 amino acids) and properties, can make their identification by MS troublesome.

Several of these issues were taken into consideration in the development of the research project presented in Chapter 6 of this proposal. Here, a method of identifying natural occurring peptides or proteolytic cleavage products (PCPs) relevant in the establishment of ectomycorrhizal symbiotic associations between the basidiomycete *Laccaria bicolor* and the perennial plant *Populus X canescens*, through LC-MS/MS was developed. This dual system presents a good opportunity to test the capabilities of a LC-MS/MS based peptidomics approach to provide experimental evidence of the expression of native peptides at a global scale. An increasing body of gene-based bioinformatic research has shown the presence of small secreted proteins (≤ 250 amino acids) in the genomes of *Populus* and *L. bicolor*,^{495, 496} and experimental data has demonstrated different levels of gene/transcript-level upregulation during their mutualistic interactions.⁴⁹⁷⁻⁵⁰¹ Moreover, there is specific protein-level evidence of the mechanisms of some small secreted proteins from *L. bicolor*, such as the mature form of the MiSSP7.2 protein (68 amino acids) that acts by stabilizing the *Populus* jasmonic acid signaling repressor PtJAZ6,⁴⁹⁷ or more recently of four *Populus* secreted peptides and small proteins (ranging from 69 to 263 amino acids) that localize to the nucleus of *L. bicolor*, with two of them impacting hyphal growth and morphology.⁴⁴⁴

As many of these small secreted proteins in *Populus* and *L. bicolor* do not have characterized mechanisms of expression, it was decided not to use any protease for their identification via LC-MS/MS so not to bias the analysis towards specific groups of peptides or hinder their identification due to limited cleavage sites. Furthermore, an enrichment protocol using 10 kDa filter membranes was used to deplete high abundant molecular weight proteins and/or their proteolytic residues that could interfere with their identification due to electrospray signal suppression, for example. The workaround to a regular database search was found in *de novo*-assisted database searches, a capability implemented in the proteomics software PEAKS.¹⁴⁹ As the name implies, this mode of database search derives peptides sequence information from experimental spectral data without relying on reference theoretical spectra generated from the *in-silico* digestion of a proteome database. However, in contrast to stand-alone *de novo* sequencing, the sequence tags identified *de novo* are used to identify proteins in a database.

The application of this peptidomic approach to *Populus* tissues, including leaves and roots, as well as roots with established *L. bicolor* ectomycorrhiza identified 1660 and 2870 *Populus* and *L. bicolor* unique PCPs, respectively, having a median length of 21 amino acids. These PCPs mapped to diverse proteins in the database and were able to capture previously reported endogenous mechanisms of protein maturation. One example was the mature form of protein PsbTn (~ 3 kDa) in *Populus* leaves, which is the smallest subunit of the Photosystem II complex. Besides qualitative information, comparison of relative abundance values of PCPs demonstrated upregulation of 157 analytes in root tips with established ectomycorrhiza when compared to root tips without established ectomycorrhiza and free-living mycelium of *L. bicolor*. These PCPs mapped to 64 *Populus* proteins and 69 *L. bicolor* proteins in our database, with several of them previously implicated in biologically relevant associations between plant and fungus. *L. bicolor*, for example, expressed proteins involved in the detoxification of reactive oxygen species which are known to be induced during the formatting of mycorrhiza. Interestingly, the peptidome method was also able to identify small secreted proteins, four of them without functional annotations and another annotated as a MiSSP 7.6, from which it was plausible to hypothesize a new mechanism of protein maturation as the only peptides that mapped to it were exclusive to its C-terminus. Although no predicted small secreted proteins from *Poplar* were identified in this study, the PCPs identified mapped to stress related proteins like metallothioneins, an observation that has been also captured in gene-based studies.

Endogenous peptides have been regarded as the “rising stars in the proteome”.⁵⁰² A recent issue of the journal Proteomics highlighted the functional importance that these molecules have in diverse organisms across different domains of life. With advances in RNA-seq and better bioinformatic approaches for *in-silico* prediction, new regions in the genomes of even “well-known” model organisms like *E. coli* and *Arabidopsis thaliana* with peptide encoding potential are being discovered.^{59, 503} However, as the same Proteomics issue also mentioned, is the integration of different sources of evidence that needs to become mainstream to detect and assess the coding potential of such regions. Peptides derived from the maturation of annotated proteins, but with unknown endogenous proteolytic mechanisms are also not exempt from this view. This is where the power of LC-MS/MS peptidomics comes handy. The ability to detect and quantify the expression of peptides under different conditions, not only can corroborate data derived from gene-based techniques, but it can also, in a high-throughput fashion provide evidence of novel molecules that can be candidates of further experimentation. With more technical advances in LC-MS technology, but also in computational algorithms for peptide sequencing such as with

better *de novo* spectra assignments,⁵⁰⁴ LC-MS/MS peptidomics will become a powerhouse for the study of these “long overlooked, small... but mighty!”⁵⁰² molecules.

7.5 Concluding perspective.

Liquid chromatography interfaced with tandem mass spectrometry has enabled and widened the scope of protein-based knowledge to our understanding of the inner mechanisms of cells. From the early discovery and characterization of a dozen-to-hundreds of proteins to current nearly 70-80% of complete proteome coverage on bacterial isolates, the in depth characterization of the proteome has allowed researchers to understand that the array of proteins in an organism is a dynamic and interconnected entity regulating multiple biological functions. However, as with every other technique used in biology, proteomics by LC-MS/MS is not exempt of challenges. Some of them include the lack of relatively streamlined experimental approaches for the recovery of proteins from different sample types and the amount of initial material necessary to perform these types of analyses when compared to genomics or transcriptomics. Likewise, there is no standard sequencing platform and the high costs of state-of-the-art equipment, which keeps getting better over time, has confined most of high-impact proteomic analyses to a small number of laboratories around the world.

Despite these challenges, proteomics nowadays is sufficiently advanced to warrant the characterization of a great variety of biological systems. Modern mass spectrometry techniques have been deployed to analyze unique sample types with biological and clinical importance, from bacterial communities thriving in acid mine drainage to cerebrospinal fluids. Although it may seem with discomfort to more traditional researchers that proteomics enables the free conception of hypotheses, this has facilitated a variety of follow-up experiments that have accelerated biological research. It is the opinion of this author that the era of forming scientists specialized in the lifelong learning of a few dozen of proteins will be even more challenged in the upcoming years as the comprehensive characterization of the proteome becomes a routine and while proteomics remains a rapidly developing and open-ended endeavor.

Besides the broader molecular understanding of organisms contributed by global LC-MS/MS proteomics approaches, the adaptability of mass spectrometry has been demonstrated with other more specialized techniques like targeted proteomics via multiple reaction monitoring mass spectrometry. This technique has provided effective ways to address the identification and absolute quantification of proteins of interest spanning wide dynamic ranges in cells. The non-

scanning nature of MRM proteomics improves sensitivity relative to how data is collected in global proteomics approaches like data-dependent acquisition. In addition, targeted proteomics provides interesting advantages over techniques like Western Blot and ELISA assays including the ability to analyze with specificity and sensitivity multiple protein targets at a time and not needing to use antibodies. As technology gets more financially accessible, and software for data analysis more user friendly, we may see targeted proteomics replace these traditional biomolecular techniques at least in laboratories doing recurrent analyses of hundreds of samples like hospitals, clinics, environmental monitoring labs, and others.

As the reader noticed throughout this dissertation, several challenges of LC-MS/MS bottom up proteomics were tackled using different biological systems (*i.e.*, groundwater, HeLA cells, plant, and fungi extracts). This not only demonstrates the multifaceted characteristic of proteomics but also its robustness as technology. Interestingly, these two attributes are also shaping a new generation of researchers with all-around knowledge, those that can interpret and condense enormous of data no matter of its origins to uncover the underlying biology within it.

List of References

1. Tyers, M.; Mann, M., From genomics to proteomics. *Nature* **2003**, 422, (6928), 193-7.
2. 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.
3. 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.; Lehoczký, 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.; Kasprzyk, 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.; Szustakowski, 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.; Szustakowski, J.; International Human Genome Sequencing, C., Initial sequencing and analysis of the human genome. *Nature* **2001**, 409, (6822), 860-921.
4. Liebler, D. C., *Introduction to proteomics: tools for the new biology*. Springer Science & Business Media: 2001.
5. Twyman, R., *Principles of proteomics*. Taylor & Francis: 2004.
6. Stagljar, I., The power of OMICS. *Biochem Biophys Res Commun* **2016**, 479, (4), 607-609.
7. Wanichthanarak, K.; Fahrman, J. F.; Grapov, D., Genomic, Proteomic, and Metabolomic Data Integration Strategies. *Biomark Insights* **2015**, 10, (Suppl 4), 1-6.
8. Mishra, N. C., *Introduction to proteomics: principles and applications*. Wiley Online Library: 2010; Vol. 146.

9. Spellman, P. T.; Sherlock, G.; Zhang, M. Q.; Iyer, V. R.; Anders, K.; Eisen, M. B.; Brown, P. O.; Botstein, D.; Futcher, B., Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* **1998**, 9, (12), 3273-97.
10. Zhou, J., Microarrays for bacterial detection and microbial community analysis. *Curr Opin Microbiol* **2003**, 6, (3), 288-94.
11. Wang, Z.; Gerstein, M.; Snyder, M., RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **2009**, 10, (1), 57-63.
12. 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.
13. Fenn, J., Electrospray ionization mass spectrometry: how it all began. *Journal of biomolecular techniques: JBT* **2002**, 13, (3), 101.
14. Hunt, D. F.; Henderson, R. A.; Shabanowitz, J.; Sakaguchi, K.; Michel, H.; Sevilir, N.; Cox, A. L.; Appella, E.; Engelhard, V. H., Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **1992**, 255, (5049), 1261-1263.
15. Mann, M., The Rise of Mass Spectrometry and the Fall of Edman Degradation. *Clin Chem* **2016**, 62, (1), 293-4.
16. 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.
17. Mann, M., Origins of mass spectrometry-based proteomics. *Nat Rev Mol Cell Biol* **2016**, 17, (11), 678.
18. Abraham, P. E.; Garcia, B. J.; Gunter, L. E.; Jawdy, S. S.; Engle, N.; Yang, X.; Jacobson, D. A.; Hettich, R. L.; Tuskan, G. A.; Tschaplinski, T. J., Quantitative proteome profile of water deficit stress responses in eastern cottonwood (*Populus deltoides*) leaves. *Plos One* **2018**, 13, (2), e0190019.
19. Poudel, S.; Giannone, R. J.; Basen, M.; Nookaew, I.; Poole, F. L.; Kelly, R. M.; Adams, M. W.; Hettich, R. L., The diversity and specificity of the extracellular proteome in the cellulolytic bacterium *Caldicellulosiruptor bescii* is driven by the nature of the cellulosic growth substrate. *Biotechnology for biofuels* **2018**, 11, (1), 80.
20. Yates, J. R., 3rd, Proteomics of Communities: Metaproteomics. *J Proteome Res* **2019**, 18, (6), 2359.
21. Li, G.-W.; Xie, X. S., Central dogma at the single-molecule level in living cells. *Nature* **2011**, 475, (7356), 308.
22. Cravatt, B. F.; Simon, G. M.; Yates III, J. R., The biological impact of mass-spectrometry-based proteomics. *Nature* **2007**, 450, (7172), 991.
23. Han, X.; Aslanian, A.; Yates, J. R., 3rd, Mass spectrometry for proteomics. *Curr Opin Chem Biol* **2008**, 12, (5), 483-90.
24. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. *Nature* **2003**, 422, (6928), 198-207.
25. Restrepo-Perez, L.; Joo, C.; Dekker, C., Paving the way to single-molecule protein sequencing. *Nat Nanotechnol* **2018**, 13, (9), 786-796.
26. Alberts, B.; Johnson, A.; Lewis, J.; Walter, P.; Raff, M.; Roberts, K., *Molecular Biology of the Cell 4th Edition: International Student Edition*. Routledge: 2002.
27. Lodish, H.; Berk, A.; Kaiser, C. A.; Krieger, M.; Scott, M. P.; Bretscher, A.; Ploegh, H.; Matsudaira, P., *Molecular cell biology*. Macmillan: 2008.
28. Gillet, L. C.; Leitner, A.; Aebersold, R., Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annu Rev Anal Chem (Palo Alto Calif)* **2016**, 9, (1), 449-72.
29. Vogel, C.; Marcotte, E. M., Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews genetics* **2012**, 13, (4), 227.

30. Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R., Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology* **1999**, *19*, (3), 1720-1730.
31. Tian, Q.; Stepaniants, S. B.; Mao, M.; Weng, L.; Feetham, M. C.; Doyle, M. J.; Yi, E. C.; Dai, H.; Thorsson, V.; Eng, J.; Goodlett, D.; Berger, J. P.; Gunter, B.; Linseley, P. S.; Stoughton, R. B.; Aebersold, R.; Collins, S. J.; Hanlon, W. A.; Hood, L. E., Integrated genomic and proteomic analyses of gene expression in Mammalian cells. *Mol Cell Proteomics* **2004**, *3*, (10), 960-9.
32. Liu, F.; Lu, J.; Hu, W.; Wang, S.-Y.; Cui, S.-J.; Chi, M.; Yan, Q.; Wang, X.-R.; Song, H.-D.; Xu, X.-N., New perspectives on host-parasite interplay by comparative transcriptomic and proteomic analyses of *Schistosoma japonicum*. *PLoS Pathogens* **2006**, *2*, (4), e29.
33. Rowe, A. R.; Heavner, G. L.; Mansfeldt, C. B.; Werner, J. J.; Richardson, R. E., Relating chloroethene respiration rates in *Dehalococcoides* to protein and mRNA biomarkers. *Environ. Sci. Technol.* **2012**, *46*, (17), 9388-9397.
34. Nie, L.; Wu, G.; Culley, D. E.; Scholten, J. C.; Zhang, W., Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. *Crit Rev Biotechnol* **2007**, *27*, (2), 63-75.
35. Zhang, W.; Li, F.; Nie, L., Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology* **2010**, *156*, (2), 287-301.
36. Cravatt, B. F.; Simon, G. M.; Yates, J. R., 3rd, The biological impact of mass-spectrometry-based proteomics. *Nature* **2007**, *450*, (7172), 991-1000.
37. Aebersold, R.; Mann, M., Mass-spectrometric exploration of proteome structure and function. *Nature* **2016**, *537*, (7620), 347-55.
38. Domon, B.; Aebersold, R., Mass spectrometry and protein analysis. *Science* **2006**, *312*, (5771), 212-7.
39. 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.
40. Lipton, M. S.; Paša-Tolić, L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarithes, H., Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proceedings of the National Academy of Sciences* **2002**, *99*, (17), 11049-11054.
41. Walther, T. C.; Mann, M., Mass spectrometry-based proteomics in cell biology. *J Cell Biol* **2010**, *190*, (4), 491-500.
42. Petricoin III, E. F.; Ardekani, A. M.; Hitt, B. A.; Levine, P. J.; Fusaro, V. A.; Steinberg, S. M.; Mills, G. B.; Simone, C.; Fishman, D. A.; Kohn, E. C., Use of proteomic patterns in serum to identify ovarian cancer. *The lancet* **2002**, *359*, (9306), 572-577.
43. America, A. H.; Cordewener, J. H., Comparative LC-MS: a landscape of peaks and valleys. *Proteomics* **2008**, *8*, (4), 731-49.
44. Schiess, R.; Wollscheid, B.; Aebersold, R., Targeted proteomic strategy for clinical biomarker discovery. *Molecular oncology* **2009**, *3*, (1), 33-44.
45. Dunham, W. H.; Mullin, M.; Gingras, A. C., Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* **2012**, *12*, (10), 1576-90.
46. Ori, A.; Iskar, M.; Buczak, K.; Kastiris, P.; Parca, L.; Andres-Pons, A.; Singer, S.; Bork, P.; Beck, M., Spatiotemporal variation of mammalian protein complex stoichiometries. *Genome Biol* **2016**, *17*, 47.
47. Leitner, A.; Faini, M.; Stengel, F.; Aebersold, R., Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends Biochem Sci* **2016**, *41*, (1), 20-32.
48. Liu, F.; Heck, A. J., Interrogating the architecture of protein assemblies and protein interaction networks by cross-linking mass spectrometry. *Curr Opin Struct Biol* **2015**, *35*, 100-8.

49. Joachimiak, L. A.; Walzthoeni, T.; Liu, C. W.; Aebersold, R.; Frydman, J., The structural basis of substrate recognition by the eukaryotic chaperonin TRiC/CCT. *Cell* **2014**, *159*, (5), 1042-1055.
50. Politis, A.; Stengel, F.; Hall, Z.; Hernandez, H.; Leitner, A.; Walzthoeni, T.; Robinson, C. V.; Aebersold, R., A mass spectrometry-based hybrid method for structural modeling of protein complexes. *Nat Methods* **2014**, *11*, (4), 403-406.
51. Mann, M.; Jensen, O. N., Proteomic analysis of post-translational modifications. *Nat Biotechnol* **2003**, *21*, (3), 255-61.
52. Aebersold, R.; Goodlett, D. R., Mass spectrometry in proteomics. *Chem Rev* **2001**, *101*, (2), 269-95.
53. Wang, Y.; Hou, Y.; Qiu, J.; Li, Z.; Zhao, J.; Tong, X.; Zhang, J., A Quantitative Acetylomic Analysis of Early Seed Development in Rice (*Oryza sativa* L.). *Int J Mol Sci* **2017**, *18*, (7).
54. 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., Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011**, *480*, (7376), 254.
55. Toby, T. K.; Fornelli, L.; Kelleher, N. L., Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu Rev Anal Chem (Palo Alto Calif)* **2016**, *9*, (1), 499-519.
56. Compton, P. D.; Zamdborg, L.; Thomas, P. M.; Kelleher, N. L., On the scalability and requirements of whole protein mass spectrometry. *Anal Chem* **2011**, *83*, (17), 6868-74.
57. Schaffer, L. V.; Millikin, R. J.; Miller, R. M.; Anderson, L. C.; Fellers, R. T.; Ge, Y.; Kelleher, N. L.; LeDuc, R. D.; Liu, X.; Payne, S. H.; Sun, L.; Thomas, P. M.; Tucholski, T.; Wang, Z.; Wu, S.; Wu, Z.; Yu, D.; Shortreed, M. R.; Smith, L. M., Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics* **2019**, *19*, (10), e1800361.
58. Shushkova, N. A.; Vavilov, N. E.; Novikova, S. E.; Farafonova, T. E.; Tikhonova, O. V.; Liao, P. C.; Zgodá, V. G., [Quantitative proteomics of human blood exosomes]. *Biomed Khim* **2018**, *64*, (6), 496-504.
59. VanOrsdel, C. E.; Kelly, J. P.; Burke, B. N.; Lein, C. D.; Oufiero, C. E.; Sanchez, J. F.; Wimmers, L. E.; Hearn, D. J.; Abuikhdair, F. J.; Barnhart, K. R.; Duley, M. L.; Ernst, S. E. G.; Kenerson, B. A.; Serafin, A. J.; Hemm, M. R., Identifying New Small Proteins in *Escherichia coli*. *Proteomics* **2018**, *18*, (10), e1700064.
60. Zhong, C.; Li, J.; Mao, L.; Liu, M.; Zhu, X.; Li, W.; Sun, M.; Ji, X.; Xiao, F.; Yang, L.; Zhang, W.; Liao, Z., Proteomics analysis reveals heat shock proteins involved in caprine parainfluenza virus type 3 infection. *BMC Vet Res* **2019**, *15*, (1), 151.
61. Nesvizhskii, A. I.; Aebersold, R., Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **2005**, *4*, (10), 1419-40.
62. Niessen, M. A. W.; Falck, D., Introduction to Mass Spectrometry, a Tutorial. In *Analyzing Biomolecular Interactions by Mass Spectrometry*, pp 1-54.
63. Cosme, J.; Tchernyshyov, I.; Gramolini, A., Basic Concepts in Mass Spectrometry and Protein Quantitation. In *Manual of Cardiovascular Proteomics*, Agnetti, G.; Lindsey, M. L.; Foster, D. B., Eds. Springer International Publishing: Cham, 2016; pp 15-32.
64. Di Silvestre, D.; Brambilla, F.; Agnetti, G.; Mauri, P., Bottom-Up Proteomics. In *Manual of Cardiovascular Proteomics*, Agnetti, G.; Lindsey, M. L.; Foster, D. B., Eds. Springer International Publishing: Cham, 2016; pp 155-185.
65. 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. *Molecular & cellular proteomics* **2005**, *4*, (10), 1487-1502.
66. Picotti, P.; Aebersold, R., Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Methods* **2012**, *9*, (6), 555-66.
67. Marx, V., Targeted proteomics. *Nat Methods* **2013**, *10*, (1), 19-22.
68. Lange, V.; Picotti, P.; Domon, B.; Aebersold, R., Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular systems biology* **2008**, *4*, 222.

69. Otto, A.; Bernhardt, J.; Hecker, M.; Völker, U.; Becher, D., Proteomics: From relative to absolute quantification for systems biology approaches. In *Methods in Microbiology*, Elsevier: 2012; Vol. 39, pp 81-106.
70. Kirkpatrick, D. S.; Gerber, S. A.; Gygi, S. P., The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods* **2005**, *35*, (3), 265-73.
71. Peterson, A. C.; Russell, J. D.; Bailey, D. J.; Westphall, M. S.; Coon, J. J., Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol Cell Proteomics* **2012**, *11*, (11), 1475-88.
72. Wohlbrand, L.; Trautwein, K.; Rabus, R., Proteomic tools for environmental microbiology - a roadmap from sample preparation to protein identification and quantification. *Proteomics* **2013**, *13*, (18-19), 2700-30.
73. Kito, K.; Ito, T., Mass spectrometry-based approaches toward absolute quantitative proteomics. *Curr Genomics* **2008**, *9*, (4), 263-74.
74. Otto, A.; Becher, D.; Schmidt, F., Quantitative proteomics in the field of microbiology. *Proteomics* **2014**, *14*, (4-5), 547-65.
75. 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.
76. Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* **2007**, *389*, (4), 1017-31.
77. Ankney, J. A.; Muneer, A.; Chen, X., Relative and Absolute Quantitation in Mass Spectrometry-Based Proteomics. *Annu Rev Anal Chem (Palo Alto Calif)* **2018**, *11*, (1), 49-77.
78. Liu, H.; Sadygov, R. G.; Yates, J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **2004**, *76*, (14), 4193-201.
79. Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M., Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* **2005**, *4*, (9), 1265-72.
80. Ahrne, E.; Molzahn, L.; Glatter, T.; Schmidt, A., Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics* **2013**, *13*, (17), 2567-78.
81. Trudgian, D. C.; Ridlova, G.; Fischer, R.; Mackeen, M. M.; Ternette, N.; Acuto, O.; Kessler, B. M.; Thomas, B., Comparative evaluation of label-free SINQ normalized spectral index quantitation in the central proteomics facilities pipeline. *Proteomics* **2011**, *11*, (14), 2790-7.
82. Zybaylov, B.; Mosley, A. L.; Sardi, M. E.; Coleman, M. K.; Florens, L.; Washburn, M. P., Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J Proteome Res* **2006**, *5*, (9), 2339-47.
83. Dowle, A. A.; Wilson, J.; Thomas, J. R., Comparing the Diagnostic Classification Accuracy of iTRAQ, Peak-Area, Spectral-Counting, and emPAI Methods for Relative Quantification in Expression Proteomics. *J Proteome Res* **2016**, *15*, (10), 3550-3562.
84. Wienkoop, S.; Weckwerth, W., Relative and absolute quantitative shotgun proteomics: targeting low-abundance proteins in *Arabidopsis thaliana*. *J Exp Bot* **2006**, *57*, (7), 1529-35.
85. Lopez-Barea, J.; Gomez-Ariza, J. L., Environmental proteomics and metallomics. *Proteomics* **2006**, *6 Suppl 1*, S51-62.
86. Schneider, T.; Riedel, K., Environmental proteomics: analysis of structure and function of microbial communities. *Proteomics* **2010**, *10*, (4), 785-98.
87. Zhao, B.; Poh, C. L., Insights into environmental bioremediation by microorganisms through functional genomics and proteomics. *Proteomics* **2008**, *8*, (4), 874-881.
88. Hivrale, A. U.; Pawar, P. K.; Rane, N. R.; Govindwar, S. P., Application of genomics and proteomics in bioremediation. In *Toxicity and waste management using bioremediation*, IGI Global: 2016; pp 97-112.

89. Maron, P. A.; Ranjard, L.; Mougél, C.; Lemanceau, P., Metaproteomics: a new approach for studying functional microbial ecology. *Microb Ecol* **2007**, *53*, (3), 486-93.
90. Jayapal, K. P.; Philp, R. J.; Kok, Y. J.; Yap, M. G.; Sherman, D. H.; Griffin, T. J.; Hu, W. S., Uncovering genes with divergent mRNA-protein dynamics in *Streptomyces coelicolor*. *PLoS One* **2008**, *3*, (5), e2097.
91. Chovanec, P.; Basu, P.; Stolz, J. F., Application of Proteomics in Bioremediation. In *Microbial Metal and Metalloid Metabolism*, American Society of Microbiology: 2011; pp 247-259.
92. Nesatyy, V. J.; Suter, M. J., Proteomics for the analysis of environmental stress responses in organisms. *Environ Sci Technol* **2007**, *41*, (20), 6891-900.
93. Gouveia, D.; Almunia, C.; Cogne, Y.; Pible, O.; Degli-Esposti, D.; Salvador, A.; Cristobal, S.; Sheehan, D.; Chaumot, A.; Geffard, O., Ecotoxicoproteomics: A decade of progress in our understanding of anthropogenic impact on the environment. *Journal of proteomics* **2019**, *198*, 66-77.
94. Segura, A.; Godoy, P.; van Dillewijn, P.; Hurtado, A.; Arroyo, N.; Santacruz, S.; Ramos, J. L., Proteomic analysis reveals the participation of energy- and stress-related proteins in the response of *Pseudomonas putida* DOT-T1E to toluene. *J Bacteriol* **2005**, *187*, (17), 5937-45.
95. Wilkins, M. J.; Callister, S. J.; Miletto, M.; Williams, K. H.; Nicora, C. D.; Lovley, D. R.; Long, P. E.; Lipton, M. S., Development of a biomarker for *Geobacter* activity and strain composition; proteogenomic analysis of the citrate synthase protein during bioremediation of U(VI). *Microb Biotechnol* **2011**, *4*, (1), 55-63.
96. Singh, O. V.; Nagaraj, N. S., Transcriptomics, proteomics and interactomics: unique approaches to track the insights of bioremediation. *Brief Funct Genomic Proteomic* **2006**, *4*, (4), 355-62.
97. Koenigsberg, S. S.; Hazen, T. C.; Peacock, A. D., Environmental biotechnology: a bioremediation perspective. *Remediation Journal: The Journal of Environmental Cleanup Costs, Technologies & Techniques* **2005**, *15*, (4), 5-25.
98. Qian, C.; Hettich, R. L., Optimized Extraction Method To Remove Humic Acid Interferences from Soil Samples Prior to Microbial Proteome Measurements. *J Proteome Res* **2017**, *16*, (7), 2537-2546.
99. Stergachis, A. B.; MacLean, B.; Lee, K.; Stamatoyannopoulos, J. A.; MacCoss, M. J., Rapid empirical discovery of optimal peptides for targeted proteomics. *Nat Methods* **2011**, *8*, (12), 1041-3.
100. Picotti, P.; Bodenmiller, B.; Mueller, L. N.; Domon, B.; Aebersold, R., Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. *Cell* **2009**, *138*, (4), 795-806.
101. Whiteaker, J. R.; Lin, C.; Kennedy, J.; Hou, L.; Trute, M.; Sokal, I.; Yan, P.; Schoenherr, R. M.; Zhao, L.; Voytovich, U. J.; Kelly-Spratt, K. S.; Krasnoselsky, A.; Gafken, P. R.; Hogan, J. M.; Jones, L. A.; Wang, P.; Amon, L.; Chodosh, L. A.; Nelson, P. S.; McIntosh, M. W.; Kemp, C. J.; Paulovich, A. G., A targeted proteomics-based pipeline for verification of biomarkers in plasma. *Nat Biotechnol* **2011**, *29*, (7), 625-34.
102. Kuhn, E.; Wu, J.; Karl, J.; Liao, H.; Zolg, W.; Guild, B., Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and ¹³C-labeled peptide standards. *Proteomics* **2004**, *4*, (4), 1175-1186.
103. Werner, J. J.; Ptak, A. C.; Rahm, B. G.; Zhang, S.; Richardson, R. E., Absolute quantification of Dehalococcoides proteins: enzyme bioindicators of chlorinated ethene dehalorespiration. *Environ. Microbiol.* **2009**, *11*, (10), 2687-2697.
104. Morris, R. M.; Fung, J. M.; Rahm, B. G.; Zhang, S.; Freedman, D. L.; Zinder, S. H.; Richardson, R. E., Comparative proteomics of *Dehalococcoides* spp. reveals strain-specific peptides associated with activity. *Appl. Environ. Microbiol.* **2007**, *73*, (1), 320-326.
105. Morris, R. M.; Sowell, S.; Barofsky, D.; Zinder, S.; Richardson, R., Transcription and mass-spectroscopic proteomic studies of electron transport oxidoreductases in *Dehalococcoides ethenogenes*. *Environ. Microbiol.* **2006**, *8*, (9), 1499-1509.

106. 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.
107. Michalski, A.; Cox, J.; Mann, M., More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC-MS/MS. *Journal of Proteome Research* **2011**, 10, (4), 1785-1793.
108. Wang, J.; Bourne, P. E.; Bandeira, N., MixGF: spectral probabilities for mixture spectra from more than one peptide. *Molecular & Cellular Proteomics* **2014**, 13, (12), 3688-3697.
109. Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M., Quantifying the Impact of Chimera MS/MS Spectra on Peptide Identification in Large-Scale Proteomics Studies. *Journal of Proteome Research* **2010**, 9, (8), 4152-4160.
110. Karp, N. A.; Huber, W.; Sadowski, P. G.; Charles, P. D.; Hester, S. V.; Lilley, K. S., Addressing Accuracy and Precision Issues in iTRAQ Quantitation. *Molecular & Cellular Proteomics* **2010**, 9, (9), 1885-1897.
111. Savitski, M. M.; Mathieson, T.; Zinn, N.; Sweetman, G.; Doce, C.; Becher, I.; Pachl, F.; Kuster, B.; Bantscheff, M., Measuring and Managing Ratio Compression for Accurate iTRAQ/TMT Quantification. *Journal of Proteome Research* **2013**, 12, (8), 3586-3598.
112. Evans, C.; Noirel, J.; Ow, S. Y.; Salim, M.; Pereira-Medrano, A. G.; Couto, N.; Pandhal, J.; Smith, D.; Pham, T. K.; Karunakaran, E.; Zou, X.; Biggs, C. A.; Wright, P. C., An insight into iTRAQ: where do we stand now? *Analytical and Bioanalytical Chemistry* **2012**, 404, (4), 1011-1027.
113. Granholm, V.; Kall, L., Quality assessments of peptide-spectrum matches in shotgun proteomics. *Proteomics* **2011**, 11, (6), 1086-93.
114. Klammer, A. A.; Park, C. Y.; Noble, W. S., Statistical calibration of the SEQUEST XCorr function. *J Proteome Res* **2009**, 8, (4), 2106-13.
115. Zhang, N.; Li, X. J.; Ye, M.; Pan, S.; Schwikowski, B.; Aebersold, R., ProbiDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* **2005**, 5, (16), 4096-106.
116. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *Journal of Proteome Research* **2011**, 10, (4), 1794-1805.
117. Zhang, B.; Pirmoradian, M.; Chernobrovkin, A.; Zubarev, R. A., DeMix workflow for efficient identification of cofragmented peptides in high resolution data-dependent tandem mass spectrometry. *Mol Cell Proteomics* **2014**, 13, (11), 3211-23.
118. Dorfer, V.; Maltsev, S.; Winkler, S.; Mechtler, K., CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. *Journal of Proteome Research* **2018**, 17, (8), 2581-2589.
119. Modell, J. W.; Hopkins, A. C.; Laub, M. T., A DNA damage checkpoint in *Caulobacter crescentus* inhibits cell division through a direct interaction with FtsW (vol 25, pg 1328, 2011). *Gene Dev* **2011**, 25, (15), 1662-1662.
120. Karimova, G.; Davi, M.; Ladant, D., The beta-Lactam Resistance Protein Blr, a Small Membrane Polypeptide, Is a Component of the *Escherichia coli* Cell Division Machinery. *J Bacteriol* **2012**, 194, (20), 5576-5588.
121. Wen, J.; Lease, K. A.; Walker, J. C., DVL, a novel class of small polypeptides: overexpression alters *Arabidopsis* development. *Plant J* **2004**, 37, (5), 668-77.
122. Pauli, A.; Norris, M. L.; Valen, E.; Chew, G. L.; Gagnon, J. A.; Zimmerman, S.; Mitchell, A.; Ma, J.; Dubrulle, J.; Reyon, D.; Tsai, S. Q.; Joung, J. K.; Saghatelian, A.; Schier, A. F., Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors. *Science* **2014**, 343, (6172), 746-+.

123. Dong, X.; Wang, D.; Liu, P.; Li, C.; Zhao, Q.; Zhu, D.; Yu, J., Zm908p11, encoded by a short open reading frame (sORF) gene, functions in pollen tube growth as a profilin ligand in maize. *J Exp Bot* **2013**, *64*, (8), 2359-72.
124. Doyle, J.; Brinkworth, C. S.; Wegener, K. L.; Carver, J. A.; Llewellyn, L. E.; Olver, I. N.; Bowie, J. H.; Wabnitz, P. A.; Tyler, M. J., nNOS inhibition, antimicrobial and anticancer activity of the amphibian skin peptide, citropin 1.1 and synthetic modifications - The solution structure of a modified citropin 1.1. *Eur J Biochem* **2003**, *270*, (6), 1141-1153.
125. Boohaker, R. J.; Lee, M. W.; Vishnubhotla, P.; Perez, J. M.; Khaled, A. R., The Use of Therapeutic Peptides to Target and to Kill Cancer Cells. *Curr Med Chem* **2012**, *19*, (22), 3794-3804.
126. Khozani, R. S.; Shahbazzadeh, D.; Harzandi, N.; Feizabadi, M. M.; Bagheri, K. P., Kinetics Study of Antimicrobial Peptide, Melittin, in Simultaneous Biofilm Degradation and Eradication of Potent Biofilm Producing MDR *Pseudomonas aeruginosa* Isolates. *Int J Pept Res Ther* **2019**, *25*, (1), 329-338.
127. De Coninck, B.; Carron, D.; Tavormina, P.; Willem, L.; Craik, D. J.; Vos, C.; Thevissen, K.; Mathys, J.; Cammue, B. P. A., Mining the genome of *Arabidopsis thaliana* as a basis for the identification of novel bioactive peptides involved in oxidative stress tolerance. *Journal of Experimental Botany* **2013**, *64*, (17), 5297-5307.
128. Wolf-Levy, H.; Javitt, A.; Eisenberg-Lerner, A.; Kacen, A.; Ulman, A.; Sheban, D.; Dassa, B.; Fishbain-Yoskovitz, V.; Carmona-Rivera, C.; Kramer, M. P.; Nudel, N.; Regev, I.; Zahavi, L.; Elinger, D.; Kaplan, M. J.; Morgenstern, D.; Levin, Y.; Merbl, Y., Revealing the cellular degradome by mass spectrometry analysis of proteasome-cleaved peptides. *Nat Biotechnol* **2018**.
129. Chen, Y. L.; Fan, K. T.; Hung, S. C.; Chen, Y. R., The role of peptides cleaved from protein precursors in eliciting plant stress reactions. *New Phytol* **2020**, *225*, (6), 2267-2282.
130. Huffaker, R. C., Proteolytic activity during senescence of plants. *New Phytol* **1990**, *116*, 199-231.
131. Müntz, K., Proteases and proteolytic cleavage of storage proteins in developing and germinating dicotyledonous seeds. *Journal of Experimental Botany* **1996**, *47*, (5), 605-622.
132. Muntz, K.; Belozersky, M. A.; Dunaevsky, Y. E.; Schlereth, A.; Tiedemann, J., Stored proteinases and the initiation of storage protein mobilization in seeds during germination and seedling growth. *J Exp Bot* **2001**, *52*, (362), 1741-52.
133. Farrokhi, N.; Whitelegge, J. P.; Brusslan, J. A., Plant peptides and peptidomics. *Plant Biotechnol J* **2008**, *6*, (2), 105-34.
134. Fricker, L. D.; Lim, J.; Pan, H.; Che, F. Y., Peptidomics: identification and quantification of endogenous peptides in neuroendocrine tissues. *Mass Spectrom Rev* **2006**, *25*, (2), 327-344.
135. Tavormina, P.; De Coninck, B.; Nikonorova, N.; De Smet, I.; Cammue, B. P., The Plant Peptidome: An Expanding Repertoire of Structural Features and Biological Functions. *Plant Cell* **2015**, *27*, (8), 2095-118.
136. Skold, K.; Svensson, M.; Kaplan, A.; Bjorkesten, L.; Astrom, J.; Andren, P. E., A neuroproteomic approach to targeting neuropeptides in the brain. *Proteomics* **2002**, *2*, (4), 447-54.
137. Schulz-Knappe, P.; Zucht, H. D.; Heine, G.; Jurgens, M.; Hess, R.; Schrader, M., Peptidomics: the comprehensive analysis of peptides in complex biological mixtures. *Comb Chem High Throughput Screen* **2001**, *4*, (2), 207-17.
138. Kay, R.; Barton, C.; Ratcliffe, L.; Matharoo-Ball, B.; Brown, P.; Roberts, J.; Teale, P.; Creaser, C., Enrichment of low molecular weight serum proteins using acetonitrile precipitation for mass spectrometry based proteomic analysis. *Rapid Commun Mass Sp* **2008**, *22*, (20), 3255-3260.
139. Kawashima, Y.; Fukutomi, T.; Tomonaga, T.; Takahashi, H.; Nomura, F.; Maeda, T.; Kodaera, Y., High-yield peptide-extraction method for the discovery of subnanomolar biomarkers from small serum samples. *J Proteome Res* **2010**, *9*, (4), 1694-705.

140. Gupta, R.; Wang, Y. M.; Agrawal, G. K.; Rakwal, R.; Jo, I. H.; Bang, K. H.; Kim, S. T., Time to dig deep into the plant proteome: a hunt for low-abundance proteins. *Front Plant Sci* **2015**, 6.
141. Gygi, S. P.; Corthals, G. L.; Zhang, Y.; Rochon, Y.; Aebersold, R., Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A* **2000**, 97, (17), 9390-5.
142. Cassidy, L.; Prasse, D.; Linke, D.; Schmitz, R. A.; Tholey, A., Combination of Bottom-up 2D-LC-MS and Semi-top-down GelFree-LC-MS Enhances Coverage of Proteome and Low Molecular Weight Short Open Reading Frame Encoded Peptides of the Archaeon *Methanosarcina mazei*. *J Proteome Res* **2016**, 15, (10), 3773-3783.
143. Cassidy, L.; Kaulich, P. T.; Tholey, A., Depletion of High-Molecular-Mass Proteins for the Identification of Small Proteins and Short Open Reading Frame Encoded Peptides in Cellular Proteomes. *J Proteome Res* **2019**.
144. Slavoff, S. A.; Mitchell, A. J.; Schwaid, A. G.; Cabili, M. N.; Ma, J.; Levin, J. Z.; Karger, A. D.; Budnik, B. A.; Rinn, J. L.; Saghatelian, A., Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* **2013**, 9, (1), 59-64.
145. Ohyama, K.; Ogawa, M.; Matsubayashi, Y., Identification of a biologically active, small, secreted peptide in *Arabidopsis* by in silico gene screening, followed by LC-MS-based structure analysis. *Plant J* **2008**, 55, (1), 152-60.
146. Möhring, T.; Kellmann, M.; Jürgens, M.; Schrader, M., Top-down identification of endogenous peptides up to 9 kDa in cerebrospinal fluid and brain tissue by nanoelectrospray quadrupole time-of-flight tandem mass spectrometry. *Journal of mass spectrometry* **2005**, 40, (2), 214-226.
147. Schrader, M.; Schulz-Knappe, P.; Fricker, L. D., Historical perspective of peptidomics. *EuPA Open Proteomics* **2014**, 3, 171-182.
148. Menschaert, G.; Vandekerckhove, T. T.; Baggerman, G.; Schoofs, L.; Luyten, W.; Van Criekeing, W., Peptidomics coming of age: a review of contributions from a bioinformatics angle. *J Proteome Res* **2010**, 9, (5), 2051-61.
149. Zhang, J.; Xin, L.; Shan, B.; Chen, W.; Xie, M.; Yuen, D.; Zhang, W.; Zhang, Z.; Lajoie, G. A.; Ma, B., PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* **2012**, 11, (4), M111 010587.
150. Tsiatsiani, L.; Heck, A. J., Proteomics beyond trypsin. *FEBS J* **2015**, 282, (14), 2612-26.
151. Burkhardt, J. M.; Schumbrutski, C.; Wortelkamp, S.; Sickmann, A.; Zahedi, R. P., Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *J Proteomics* **2012**, 75, (4), 1454-62.
152. Cristobal, A.; Hennrich, M. L.; Giansanti, P.; Goerdal, S. S.; Heck, A. J.; Mohammed, S., In-house construction of a UHPLC system enabling the identification of over 4000 protein groups in a single analysis. *Analyst* **2012**, 137, (15), 3541-8.
153. Lo, I.; Deneff, 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), 537-41.
154. Secher, A.; Kelstrup, C. D.; Conde-Frieboes, K. W.; Pyke, C.; Raun, K.; Wulff, B. S.; Olsen, J. V., Analytic framework for peptidomics applied to large-scale neuropeptide identification. *Nat Commun* **2016**, 7, 11436.
155. Canas, B.; Pineiro, C.; Calvo, E.; Lopez-Ferrer, D.; Gallardo, J. M., Trends in sample preparation for classical and second generation proteomics. *J Chromatogr A* **2007**, 1153, (1-2), 235-58.
156. 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.

157. Anderson, N. L.; Anderson, N. G., The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* **2002**, *1*, (11), 845-67.
158. Gundry, R. L.; White, M. Y.; Murray, C. I.; Kane, L. A.; Fu, Q.; Stanley, B. A.; Van Eyk, J. E., Preparation of proteins and peptides for mass spectrometry analysis in a bottom-up proteomics workflow. *Curr Protoc Mol Biol* **2009**, Chapter 10, Unit10 25.
159. Whitman, W. B.; Coleman, D. C.; Wiebe, W. J., Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **1998**, *95*, (12), 6578-83.
160. Arenella, M.; Giagnoni, L.; Masciandaro, G.; Ceccanti, B.; Nannipieri, P.; Renella, G., Interactions between proteins and humic substances affect protein identification by mass spectrometry. *Biol Fert Soils* **2014**, *50*, (3), 447-454.
161. 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.
162. Qian, C.; Hettich, R. L., Optimized Extraction Method To Remove Humic Acid Interferences from Soil Samples Prior to Microbial Proteome Measurements. *Journal of Proteome Research* **2017**, *16*, (7), 2537-2546.
163. Chandramouli, K.; Qian, P. Y., Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Hum Genomics Proteomics* **2009**, 2009.
164. Nakamura, T.; Myint, K. T.; Oda, Y., Ethylenediaminetetraacetic acid increases identification rate of phosphoproteomics in real biological samples. *J Proteome Res* **2010**, *9*, (3), 1385-91.
165. Hamdan, M. H.; Righetti, P. G., *Proteomics today: protein assessment and biomarkers using mass spectrometry, 2D electrophoresis, and microarray technology*. John Wiley & Sons: 2005; Vol. 18.
166. Zhou, J. Y.; Dann, G. P.; Shi, T. J.; Wang, L.; Gao, X. L.; Su, D. A.; Nicora, C. D.; Shukla, A. K.; Moore, R. J.; Liu, T.; Camp, D. G.; Smith, R. D.; Qian, W. J., Simple Sodium Dodecyl Sulfate-Assisted Sample Preparation Method for LC-MS-Based Proteomics Applications. *Analytical Chemistry* **2012**, *84*, (6), 2862-2867.
167. Swiderek, K. M.; Alpert, A. J.; Heckendorf, A.; Nugent, K.; Patterson, S. D., Structural analysis of proteins and peptides in the presence of detergents: Tricks of the trade. *ABRF News Artic. Methodol* **1997**, *8*, 17-25.
168. Görg, A.; Weiss, W.; Dunn, M. J., Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **2004**, *4*, (12), 3665-3685.
169. Arnold, U.; Ulbrich-Hofmann, R., Quantitative protein precipitation from guanidine hydrochloride-containing solutions by sodium deoxycholate/trichloroacetic acid. *Anal Biochem* **1999**, *271*, (2), 197-9.
170. Peterson, G. L., A simplification of the protein assay method of Lowry et al. which is more generally applicable. *Analytical biochemistry* **1977**, *83*, (2), 346-356.
171. Crowell, A. M. J.; Wall, M. J.; Doucette, A. A., Maximizing recovery of water-soluble proteins through acetone precipitation. *Anal Chim Acta* **2013**, *796*, 48-54.
172. Wessel, D.; Flugge, U. I., A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal Biochem* **1984**, *138*, (1), 141-3.
173. Ferro, M.; Seigneurin-Berny, D.; Rolland, N.; Chapel, A.; Salvi, D.; Garin, J.; Joyard, J., Organic solvent extraction as a versatile procedure to identify hydrophobic chloroplast membrane proteins. *Electrophoresis* **2000**, *21*, (16), 3517-3526.
174. Burgess, R. R., Protein precipitation techniques. *Methods Enzymol* **2009**, *463*, 331-42.
175. 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.
176. Isaacson, T.; Damasceno, C. M.; Saravanan, R. S.; He, Y.; Catala, C.; Saladie, M.; Rose, J. K., Sample extraction techniques for enhanced proteomic analysis of plant tissues. *Nat Protoc* **2006**, *1*, (2), 769-74.

177. Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates, J. R., 3rd, Protein analysis by shotgun/bottom-up proteomics. *Chem Rev* **2013**, *113*, (4), 2343-94.
178. 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.
179. Masuda, T.; Tomita, M.; Ishihama, Y., Phase transfer surfactant-aided trypsin digestion for membrane proteome analysis. *J Proteome Res* **2008**, *7*, (2), 731-40.
180. Yeung, Y. G.; Stanley, E. R., Rapid detergent removal from peptide samples with ethyl acetate for mass spectrometry analysis. *Curr Protoc Protein Sci* **2010**, Chapter 16, Unit 16 12.
181. Nielsen, M. L.; Vermeulen, M.; Bonaldi, T.; Cox, J.; Moroder, L.; Mann, M., Iodoacetamide-induced artifact mimics ubiquitination in mass spectrometry. *Nat Methods* **2008**, *5*, (6), 459-60.
182. Loziuk, P. L.; Wang, J.; Li, Q. Z.; Sederoff, R. R.; Chiang, V. L.; Muddiman, D. C., Understanding the Role of Proteolytic Digestion on Discovery and Targeted Proteomic Measurements Using Liquid Chromatography Tandem Mass Spectrometry and Design of Experiments. *Journal of Proteome Research* **2013**, *12*, (12), 5820-5829.
183. Brownridge, P.; Beynon, R. J., The importance of the digest: proteolysis and absolute quantification in proteomics. *Methods* **2011**, *54*, (4), 351-60.
184. Murphy, J. P.; Pinto, D. M., Targeted proteomic analysis of glycolysis in cancer cells. *J Proteome Res* **2011**, *10*, (2), 604-13.
185. Wisniewski, J. R.; Mann, M., Consecutive proteolytic digestion in an enzyme reactor increases depth of proteomic and phosphoproteomic analysis. *Anal Chem* **2012**, *84*, (6), 2631-7.
186. Norrgran, J.; Williams, T. L.; Woolfitt, A. R.; Solano, M. I.; Pirkle, J. L.; Barr, J. R., Optimization of digestion parameters for protein quantification. *Anal Biochem* **2009**, *393*, (1), 48-55.
187. Hurrell, T.; Segeritz, C. P.; Vallier, L.; Lilley, K. S.; Cromarty, A. D., A proteomic time course through the differentiation of human induced pluripotent stem cells into hepatocyte-like cells. *Sci Rep-Uk* **2019**, *9*.
188. Kelliher, C. M.; Foster, M. W.; Motta, F. C.; Deckard, A.; Soderblom, E. J.; Moseley, M. A.; Haase, S. B., Layers of regulation of cell-cycle gene expression in the budding yeast *Saccharomyces cerevisiae*. *Molecular Biology of the Cell* **2018**, *29*, (22), 2644-2655.
189. Swaney, D. L.; Wenger, C. D.; Coon, J. J., Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics. *Journal of Proteome Research* **2010**, *9*, (3), 1323-1329.
190. Siepen, J. A.; Keevil, E. J.; Knight, D.; Hubbard, S. J., Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. *J Proteome Res* **2007**, *6*, (1), 399-408.
191. Thiede, B.; Lamer, S.; Mattow, J.; Siejak, F.; Dimmler, C.; Rudel, T.; Jungblut, P. R., Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. *Rapid Commun Mass Spectrom* **2000**, *14*, (6), 496-502.
192. Wither, M. J.; Hansen, K. C.; Reisz, J. A., Mass Spectrometry-Based Bottom-Up Proteomics: Sample Preparation, LC-MS/MS Analysis, and Database Query Strategies. *Curr Protoc Protein Sci* **2016**, *86*, 16 4 1-16 4 20.
193. Kulak, N. A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M., Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat Methods* **2014**, *11*, (3), 319-24.
194. Rappsilber, J.; Mann, M.; Ishihama, Y., Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nature Protocols* **2007**, *2*, (8), 1896-1906.
195. Wolters, D. A.; Washburn, M. P.; Yates, J. R., 3rd, An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* **2001**, *73*, (23), 5683-90.

196. Boysen, R. I.; Hearn, M. T., High-performance liquid chromatography of peptides and proteins. In *Amino Acids, Peptides and Proteins in Organic Chemistry: Analysis and Function of Amino Acids and Peptides*, Wiley-VCH Verlag GmbH & Co. KGaA: 2012; pp 167-210.
197. Shen, Y.; Smith, R. D., Advanced nanoscale separations and mass spectrometry for sensitive high-throughput proteomics. *Expert Rev Proteomics* **2005**, 2, (3), 431-47.
198. Mitulovic, G., New HPLC Techniques for Proteomics Analysis: A Short Overview of Latest Developments. *J Liq Chromatogr R T* **2015**, 38, (3), 390-403.
199. Fekete, S.; Schappler, J.; Veuthey, J. L.; Guillaume, D., Current and future trends in UHPLC. *Trac-Trend Anal Chem* **2014**, 63, 2-13.
200. Pirmoradian, M.; Budamgunta, H.; Chingin, K.; Zhang, B.; Astorga-Wells, J.; Zubarev, R. A., Rapid and Deep Human Proteome Analysis by Single-dimension Shotgun Proteomics. *Molecular & Cellular Proteomics* **2013**, 12, (11), 3330-3338.
201. Rodriguez-Aller, M.; Gurny, R.; Veuthey, J. L.; Guillaume, D., Coupling ultra high-pressure liquid chromatography with mass spectrometry: constraints and possible applications. *J Chromatogr A* **2013**, 1292, 2-18.
202. Novakova, L.; Matysova, L.; Solich, P., Advantages of application of UPLC in pharmaceutical analysis. *Talanta* **2006**, 68, (3), 908-18.
203. Guillaume, D.; Nguyen, D. T.; Rudaz, S.; Veuthey, J. L., Method transfer for fast liquid chromatography in pharmaceutical analysis: application to short columns packed with small particle. Part II: gradient experiments. *Eur J Pharm Biopharm* **2008**, 68, (2), 430-40.
204. Shishkova, E.; Hebert, A. S.; Coon, J. J., Now, More Than Ever, Proteomics Needs Better Chromatography. *Cell Syst* **2016**, 3, (4), 321-324.
205. Lesur, A.; Domon, B., Advances in high-resolution accurate mass spectrometry application to targeted proteomics. *Proteomics* **2015**, 15, (5-6), 880-890.
206. Anspach, J. A.; Maloney, T. D.; Colón, L. A., Ultrahigh-pressure liquid chromatography using a 1-mm id column packed with 1.5- μ m porous particles. *Journal of separation science* **2007**, 30, (8), 1207-1213.
207. Nagaraj, N.; Kulak, N. A.; Cox, J.; Neuhauser, N.; Mayr, K.; Hoerning, O.; Vorm, O.; Mann, M., System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap. *Molecular & Cellular Proteomics* **2012**, 11, (3).
208. Fournier, M. L.; Gilmore, J. M.; Martin-Brown, S. A.; Washburn, M. P., Multidimensional separations-based shotgun proteomics. *Chem Rev* **2007**, 107, (8), 3654-86.
209. Giddings, J. C., Two-dimensional separations: concept and promise. *Anal Chem* **1984**, 56, (12), 1258A-1260A, 1262A, 1264A passim.
210. Cai, X. M.; Guo, Z. M.; Xue, X. Y.; Xu, J. Y.; Zhang, X. L.; Liang, X. M., Two-dimensional liquid chromatography separation of peptides using reversed-phase/weak cation-exchange mixed-mode column in first dimension. *Journal of Chromatography A* **2012**, 1228, 242-249.
211. McDonald, W. H.; Ohi, R.; Miyamoto, D. T.; Mitchison, T. J.; Yates III, J. R., Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. *International Journal of Mass Spectrometry* **2002**, 219, (1), 245-251.
212. Kebarle, P.; Verkerk, U. H., Electrospray: from ions in solution to ions in the gas phase, what we know now. *Mass Spectrom Rev* **2009**, 28, (6), 898-917.
213. Taylor, G. I., Disintegration of water drops in an electric field. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **1964**, 280, (1382), 383-397.
214. Cloupeau, M.; Prunetfoch, B., Electrohydrodynamic Spraying Functioning Modes - a Critical-Review. *J Aerosol Sci* **1994**, 25, (6), 1021-1036.
215. Tang, K. Q.; Gomez, A., Generation of Monodisperse Water Droplets from Electrosprays in a Corona-Assisted Cone-Jet Mode. *J Colloid Interf Sci* **1995**, 175, (2), 326-332.

216. Rayleigh, L., XX. On the equilibrium of liquid conducting masses charged with electricity. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1882**, *14*, (87), 184-186.
217. Dole, M.; Mack, L. L.; Hines, R. L., Molecular Beams of Macroions. *J Chem Phys* **1968**, *49*, (5), 2240-&.
218. Iribarne, J. V.; Thomson, B. A., On the evaporation of small ions from charged droplets. *J Chem Phys* **1976**, *64*, (6), 2287-2294.
219. Wilm, M.; Mann, M., Analytical properties of the nanoelectrospray ion source. *Anal Chem* **1996**, *68*, (1), 1-8.
220. Murray, K. K.; Boyd, R. K.; Eberlin, M. N.; Langley, G. J.; Li, L.; Naito, Y., Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). *Pure Appl Chem* **2013**, *85*, (7), 1515-1609.
221. Dass, C., *Fundamentals of contemporary mass spectrometry*. John Wiley & Sons: 2007; Vol. 16.
222. Boyd, R. K.; Basic, C.; Bethem, R. A., *Trace quantitative analysis by mass spectrometry*. John Wiley & Sons: 2011.
223. Page, J. S.; Kelly, R. T.; Tang, K.; Smith, R. D., Ionization and transmission efficiency in an electrospray ionization-mass spectrometry interface. *J Am Soc Mass Spectr* **2007**, *18*, (9), 1582-1590.
224. De Hoffmann, E.; Stroobant, V., *Mass spectrometry: principles and applications*. 2007. Great Britain: John Wiley & Sons Ltd.
225. Wiza, J. L., Microchannel plate detectors. *Nucl. Instrum. Methods* **1979**, *162*, (1-3), 587-601.
226. Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S., Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom Rev* **1998**, *17*, (1), 1-35.
227. Miller, P. E.; Denton, M. B., The quadrupole mass filter: basic operating concepts. *Journal of chemical education* **1986**, *63*, (7), 617.
228. Lange, V.; Picotti, P.; Domon, B.; Aebersold, R., Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **2008**, *4*, (1), 222.
229. Tabb, D. L.; Vega-Montoto, L.; Rudnick, P. A.; Variyath, A. M.; Ham, A. J.; Bunk, D. M.; Kilpatrick, L. E.; Billheimer, D. D.; Blackman, R. K.; Cardasis, H. L.; Carr, S. A.; Clauser, K. R.; Jaffe, J. D.; Kowalski, K. A.; Neubert, T. A.; Regnier, F. E.; Schilling, B.; Tegeler, T. J.; Wang, M.; Wang, P.; Whiteaker, J. R.; Zimmerman, L. J.; Fisher, S. J.; Gibson, B. W.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Stein, S. E.; Tempst, P.; Paulovich, A. G.; Liebler, D. C.; Spiegelman, C., Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J Proteome Res* **2010**, *9*, (2), 761-76.
230. Makarov, A., Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem* **2000**, *72*, (6), 1156-62.
231. Olsen, J. V.; Schwartz, J. C.; Griep-Raming, J.; Nielsen, M. L.; Damoc, E.; Denisov, E.; Lange, O.; Remes, P.; Taylor, D.; Splendore, M.; Wouters, E. R.; Senko, M.; Makarov, A.; Mann, M.; Horning, S., A Dual Pressure Linear Ion Trap Orbitrap Instrument with Very High Sequencing Speed. *Molecular & Cellular Proteomics* **2009**, *8*, (12), 2759-2769.
232. Michalski, A.; Damoc, E.; Hauschild, J. P.; Lange, O.; Wiegand, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S., Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Molecular & Cellular Proteomics* **2011**, *10*, (9).
233. Sleno, L.; Volmer, D. A., Ion activation methods for tandem mass spectrometry. *J Mass Spectrom* **2004**, *39*, (10), 1091-112.
234. Roepstorff, P.; Fohlman, J., Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* **1984**, *11*, (11), 601.

235. Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* **2004**, *101*, (26), 9528-33.
236. Mikesch, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E.; Shabanowitz, J.; Hunt, D. F., The utility of ETD mass spectrometry in proteomic analysis. *Biochim Biophys Acta* **2006**, *1764*, (12), 1811-22.
237. Kim, M. S.; Pandey, A., Electron transfer dissociation mass spectrometry in proteomics. *Proteomics* **2012**, *12*, (4-5), 530-42.
238. 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.
239. 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.
240. Craig, R.; Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, (9), 1466-1467.
241. Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X. Y.; Shi, W. Y.; Bryant, S. H., Open mass spectrometry search algorithm. *Journal of Proteome Research* **2004**, *3*, (5), 958-964.
242. Wenger, C. D.; Coon, J. J., A Proteomics Search Algorithm Specifically Designed for High-Resolution Tandem Mass Spectra. *Journal of Proteome Research* **2013**, *12*, (3), 1377-1386.
243. Verheggen, K.; Raeder, H.; Berven, F. S.; Martens, L.; Barsnes, H.; Vaudel, M., Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass Spectrom Rev* **2017**.
244. Mayne, J.; Ning, Z. B.; Zhang, X.; Starr, A. E.; Chen, R.; Deeke, S.; Chiang, C. K.; Xu, B.; Wen, M.; Cheng, K.; Seebun, D.; Star, A.; Moore, J. I.; Figeys, D., Bottom-Up Proteomics (2013-2015): Keeping up in the Era of Systems Biology. *Analytical Chemistry* **2016**, *88*, (1), 95-121.
245. Burger, T., Gentle Introduction to the Statistical Foundations of False Discovery Rate in Quantitative Proteomics. *J Proteome Res* **2018**, *17*, (1), 12-22.
246. Nesvizhskii, A. I., A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* **2010**, *73*, (11), 2092-123.
247. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007**, *4*, (3), 207-14.
248. Aggarwal, S.; Yadav, A. K., False discovery rate estimation in proteomics. In *Statistical Analysis in Proteomics*, Springer: 2016; pp 119-128.
249. Li, S.; Tang, H., Computational Methods in Mass Spectrometry-Based Proteomics. In *Translational Biomedical Informatics: A Precision Medicine Perspective*, Shen, B.; Tang, H.; Jiang, X., Eds. Springer Singapore: Singapore, 2016; pp 63-89.
250. Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J., Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **2007**, *4*, (11), 923-5.
251. Tabb, D. L., The SEQUEST family tree. *J Am Soc Mass Spectrom* **2015**, *26*, (11), 1814-9.
252. Li, Y. F.; Radivojac, P., Computational approaches to protein inference in shotgun proteomics. *BMC Bioinformatics* **2012**, *13 Suppl 16*, S4.
253. Steen, H.; Mann, M., The ABC's (and XYZ's) of peptide sequencing. *Nature reviews Molecular cell biology* **2004**, *5*, (9), 699.
254. Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A.; Working Group on Publication Guidelines for, P.; Protein Identification, D., The need for guidelines in

publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol Cell Proteomics* **2004**, 3, (6), 531-3.

255. Wilkins, M. R.; Appel, R. D.; Van Eyk, J. E.; Chung, M. C.; 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.

256. 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.

257. Dorfer, V.; Pichler, P.; Stranzl, T.; Stadlmann, J.; Taus, T.; Winkler, S.; Mechtler, K., MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J Proteome Res* **2014**, 13, (8), 3679-84.

258. Domon, B.; Aebersold, R., Options and considerations when selecting a quantitative proteomics strategy. *Nature Biotechnology* **2010**, 28, (7), 710-721.

259. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010**, 26, (7), 966-8.

260. Pino, L. K.; Searle, B. C.; Bollinger, J. G.; Nunn, B.; MacLean, B.; MacCoss, M. J., The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass. Spectrom. Rev.* **2017**.

261. Finney, G. L.; Blackler, A. R.; Hoopmann, M. R.; Canterbury, J. D.; Wu, C. C.; MacCoss, M. J., Label-free comparative analysis of proteomics mixtures using chromatographic alignment of high-resolution mu LC-MS data. *Analytical Chemistry* **2008**, 80, (4), 961-971.

262. 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), 731-740.

263. Karpievitch, Y. V.; Dabney, A. R.; Smith, R. D., Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* **2012**, 13 Suppl 16, S5.

264. Chawade, A.; Alexandersson, E.; Levander, F., Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *J Proteome Res* **2014**, 13, (6), 3114-20.

265. Sandin, M.; Teleman, J.; Malmstrom, J.; Levander, F., Data processing methods and quality control strategies for label-free LC-MS protein quantification. *Bba-Proteins Proteom* **2014**, 1844, (1), 29-41.

266. Webb-Robertson, B. J. M.; Matzke, M. M.; Jacobs, J. M.; Pounds, J. G.; Waters, K. M., A statistical selection strategy for normalization procedures in LC-MS proteomics experiments through dataset-dependent ranking of normalization scaling factors. *Proteomics* **2011**, 11, (24), 4736-4741.

267. Callister, S. J.; Barry, R. C.; Adkins, J. N.; Johnson, E. T.; Qian, W. J.; Webb-Robertson, B. J.; Smith, R. D.; Lipton, M. S., Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res* **2006**, 5, (2), 277-86.

268. Lazar, C.; Gatto, L.; Ferro, M.; Bruley, C.; Burger, T., Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J Proteome Res* **2016**, 15, (4), 1116-25.

269. Webb-Robertson, B. J.; Wiberg, H. K.; Matzke, M. M.; Brown, J. N.; Wang, J.; McDermott, J. E.; Smith, R. D.; Rodland, K. D.; Metz, T. O.; Pounds, J. G.; Waters, K. M., Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res* **2015**, 14, (5), 1993-2001.

270. Streit, M.; Gehlenborg, N., Bar charts and box plots. *Nat Methods* **2014**, 11, (2), 117.

271. Lex, A.; Gehlenborg, N., POINTS OF VIEW Sets and intersections. *Nature Methods* **2014**, *11*, (8), 779-779.
272. Oveland, E.; Muth, T.; Rapp, E.; Martens, L.; Berven, F. S.; Barsnes, H., Viewing the proteome: How to visualize proteomics data? *Proteomics* **2015**, *15*, (8), 1341-1355.
273. Ringner, M., What is principal component analysis? *Nature Biotechnology* **2008**, *26*, (3), 303-304.
274. Polpitiya, A. D.; Qian, W. J.; Jaitly, N.; Petyuk, V. A.; Adkins, J. N.; Camp, D. G.; Anderson, G. A.; Smith, R. D., DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics* **2008**, *24*, (13), 1556-1558.
275. Lee, J.; Im, J.; Kim, U.; Löffler, F. E., A Data Mining Approach to Predict In Situ Detoxification Potential of Chlorinated Ethenes. *Environ. Sci. Technol.* **2016**, *50*, (10), 5181-5188.
276. Löffler, F. E.; Ritalahti, K. M.; Zinder, S. H., Dehalococcoides and reductive dechlorination of chlorinated solvents. In *Bioaugmentation for groundwater remediation*, Springer: 2013; pp 39-88.
277. Fletcher, K. E.; Costanza, J.; Cruz-Garcia, C.; Ramaswamy, N. S.; Pennell, K. D.; Löffler, F. E., Effects of elevated temperature on Dehalococcoides dechlorination performance and DNA and RNA biomarker abundance. *Environ. Sci. Technol.* **2011**, *45*, (2), 712-718.
278. Hug, L. A.; Maphosa, F.; Leys, D.; Löffler, F. E.; Smidt, H.; Edwards, E. A.; Adrian, L., Overview of organohalide-respiring bacteria and a proposal for a classification system for reductive dehalogenases. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2013**, *368*, (1616), 20120322.
279. Löffler, F. E.; Edwards, E. A., Harnessing microbial activities for environmental cleanup. *Curr Opin Biotechnol* **2006**, *17*, (3), 274-284.
280. Ritalahti, K. M.; Amos, B. K.; Sung, Y.; Wu, Q.; Koenigsberg, S. S.; Löffler, F. E., Quantitative PCR targeting 16S rRNA and reductive dehalogenase genes simultaneously monitors multiple Dehalococcoides strains. *Appl. Environ. Microbiol.* **2006**, *72*, (4), 2765-2774.
281. Lendvay, J.; Löffler, F. E.; Dollhopf, M.; Aiello, M.; Daniels, G.; Fathepure, B.; Gebhard, M.; Heine, R.; Helton, R.; Shi, J., Bioreactive barriers: a comparison of bioaugmentation and biostimulation for chlorinated solvent remediation. *Environ. Sci. Technol.* **2003**, *37*, (7), 1422-1431.
282. Rahm, B. G.; Chauhan, S.; Holmes, V. F.; Macbeth, T. W.; Kent Jr, S.; Alvarez-Cohen, L., Molecular characterization of microbial populations at two sites with differing reductive dechlorination abilities. *Biodegradation* **2006**, *17*, (6), 523-534.
283. Major, D. W.; McMaster, M. L.; Cox, E. E.; Edwards, E. A.; Dworatzek, S. M.; Hendrickson, E. R.; Starr, M. G.; Payne, J. A.; Buonamici, L. W., Field demonstration of successful bioaugmentation to achieve dechlorination of tetrachloroethene to ethene. *Environ. Sci. Technol.* **2002**, *36*, (23), 5106-5116.
284. Maphosa, F.; de Vos, W. M.; Smidt, H., Exploiting the ecogenomics toolbox for environmental diagnostics of organohalide-respiring bacteria. *Trends Biotechnol.* **2010**, *28*, (6), 308-316.
285. Rahm, B. G.; Morris, R. M.; Richardson, R. E., Temporal expression of respiratory genes in an enrichment culture containing Dehalococcoides ethenogenes. *Appl. Environ. Microbiol.* **2006**, *72*, (8), 5486-5491.
286. Da Silva, M. L.; Alvarez, P., Exploring the correlation between halo-respirer biomarker concentrations and TCE dechlorination rates. *Journal of Environmental Engineering* **2008**, *134*, (11), 895-901.
287. Heavner, G. L.; Mansfeldt, C. B.; Debs, G. E.; Hellerstedt, S. T.; Rowe, A. R.; Richardson, R. E., Biomarkers' Responses to Reductive Dechlorination Rates and Oxygen Stress in Bioaugmentation Culture KB-1TM. *Microorganisms* **2018**, *6*, (1), 13.
288. Schiffmann, C.; Hansen, R.; Baumann, S.; Kublik, A.; Nielsen, P. H.; Adrian, L.; von Bergen, M.; Jehmlich, N.; Seifert, J., Comparison of targeted peptide quantification assays for

reductive dehalogenases by selective reaction monitoring (SRM) and precursor reaction monitoring (PRM). *Anal. Bioanal. Chem.* **2014**, 406, (1), 283-291.

289. Yan, J.; Simsir, B.; Farmer, A. T.; Bi, M.; Yang, Y.; Campagna, S. R.; Löffler, F. E., The corrinoid cofactor of reductive dehalogenases affects dechlorination rates and extents in organohalide-respiring *Dehalococcoides mccartyi*. *ISME J.* **2016**, 10, (5), 1092-1101.

290. Löffler, F. E.; Yan, J.; Ritalahti, K. M.; Adrian, L.; Edwards, E. A.; Konstantinidis, K. T.; Muller, J. A.; Fullerton, H.; Zinder, S. H.; Spormann, A. M., *Dehalococcoides mccartyi* gen. nov., sp. nov., obligately organohalide-respiring anaerobic bacteria relevant to halogen cycling and bioremediation, belong to a novel bacterial class, *Dehalococcoidia classis nov.*, order *Dehalococcoidales ord. nov.* and family *Dehalococcoidaceae fam. nov.*, within the phylum *Chloroflexi*. *Int. J. Syst. Evol. Microbiol.* **2013**, 65, (2), 625-635.

291. Yang, S.; Giannone, R. J.; Dice, L.; Yang, Z. K.; Engle, N. L.; Tschaplinski, T. J.; Hettich, R. L.; Brown, S. D., *Clostridium thermocellum* ATCC27405 transcriptomic, metabolomic and proteomic profiles after ethanol stress. *BMC genomics* **2012**, 13, (1), 1.

292. 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.

293. Giannone, R. J.; Huber, H.; Karpinets, T.; Heimerl, T.; Küper, 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.

294. Johnson, C. W.; Abraham, P. E.; Linger, J. G.; Khanna, P.; Hettich, R. L.; Beckham, G. T., Eliminating a global regulator of carbon catabolite repression enhances the conversion of aromatic lignin monomers to muconate in *Pseudomonas putida* KT2440. *Metabolic Engineering Communications* **2017**.

295. Chen, Y. Y.; Chambers, M. C.; Li, M.; Ham, A. J.; Turner, J. L.; Zhang, B.; Tabb, D. L., IDPQuantify: combining precursor intensity with spectral counts for protein and peptide quantification. *J Proteome Res* **2013**, 12, (9), 4111-4121.

296. Edgar, R. C., Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**, 26, (19), 2460-2461.

297. Kumar, S.; Stecher, G.; Tamura, K., MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **2016**, 33, (7), 1870-1874.

298. Edgar, R. C., MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **2004**, 5, 113.

299. Jones, D. T.; Taylor, W. R.; Thornton, J. M., The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **1992**, 8, (3), 275-282.

300. Felsenstein, J., Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution* **1985**, 39, (4), 783-791.

301. Surinova, S.; Huttenhain, R.; Chang, C. Y.; Espona, L.; Vitek, O.; Aebersold, R., Automated selected reaction monitoring data analysis workflow for large-scale targeted proteomic studies. *Nat. Protoc.* **2013**, 8, (8), 1602-1619.

302. Mesuere, B.; Jeugt, F.; Devreese, B.; Vandamme, P.; Dawyndt, P., The unique peptidome: Taxon-specific tryptic peptides as biomarkers for targeted metaproteomics. *Proteomics* **2016**, 16, (17), 2313-2318.

303. Pöritz, M.; Goris, T.; Wubet, T.; Tarkka, M. T.; Buscot, F.; Nijenhuis, I.; Lechner, U.; Adrian, L., Genome sequences of two dehalogenation specialists—*Dehalococcoides mccartyi* strains BTF08 and DCMB5 enriched from the highly polluted Bitterfeld region. *FEMS Microbiol. Lett.* **2013**, 343, (2), 101-104.

304. Men, Y.; Feil, H.; VerBerkmoes, N. C.; Shah, M. B.; Johnson, D. R.; Lee, P. K.; West, K. A.; Zinder, S. H.; Andersen, G. L.; Alvarez-Cohen, L., Sustainable syntrophic growth of *Dehalococcoides ethenogenes* strain 195 with *Desulfovibrio vulgaris* Hildenborough and

- Methanobacterium congolense: global transcriptomic and proteomic analyses. *ISME J.* **2012**, 6, (2), 410.
305. Schiffmann, C. L.; Jehmlich, N.; Otto, W.; Hansen, R.; Nielsen, P. H.; Adrian, L.; Seifert, J.; von Bergen, M., Proteome profile and proteogenomics of the organohalide-respiring bacterium *Dehalococcoides mccartyi* strain CBDB1 grown on hexachlorobenzene as electron acceptor. *J. Proteomics* **2014**, 98, 59-64.
306. Adrian, L.; Rahnenfuhrer, J.; Gobom, J.; Holscher, T., Identification of a chlorobenzene reductive dehalogenase in *Dehalococcoides* sp. strain CBDB1. *Appl. Environ. Microbiol.* **2007**, 73, (23), 7717-7724.
307. Schipp, C. J.; Marco-Urrea, E.; Kublik, A.; Seifert, J.; Adrian, L., Organic cofactors in the metabolism of *Dehalococcoides mccartyi* strains. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2013**, 368, (1616), 20120321.
308. Türkowsky, D.; Jehmlich, N.; Diekert, G.; Adrian, L.; von Bergen, M.; Goris, T., An integrative overview of genomic, transcriptomic and proteomic analyses in organohalide respiration research. *FEMS Microbiol. Ecol.* **2018**, 94, (3), fty013.
309. Hartwig, S.; Dragomirova, N.; Kublik, A.; Türkowsky, D.; von Bergen, M.; Lechner, U.; Adrian, L.; Gary Sawers, R., A H₂-oxidizing, 1, 2, 3-trichlorobenzene-reducing multienzyme complex isolated from the obligately organohalide-respiring bacterium *dehalococcoides mccartyi* strain Cbdb1. *Environ. Microbiol. Rep.* **2017**.
310. Zinder, S. H., *Dehalococcoides* has a dehalogenation complex. *Environ. Microbiol.* **2016**, 18, (9), 2773-2775.
311. Kublik, A.; Deobald, D.; Hartwig, S.; Schiffmann, C. L.; Andrades, A.; Bergen, M.; Sawers, R. G.; Adrian, L., Identification of a multi-protein reductive dehalogenase complex in *Dehalococcoides mccartyi* strain CBDB1 suggests a protein-dependent respiratory electron transport chain obviating quinone involvement. *Environ. Microbiol.* **2016**.
312. Taş, N.; Eekert, V.; Miriam, H.; De Vos, W. M.; Smidt, H., The little bacteria that can—diversity, genomics and ecophysiology of ‘*Dehalococcoides*’ spp. in contaminated environments. *Microb. Biotechnol.* **2010**, 3, (4), 389-402.
313. He, J.; Sung, Y.; Krajmalnik-Brown, R.; Ritalahti, K. M.; Löffler, F. E., Isolation and characterization of *Dehalococcoides* sp. strain FL2, a trichloroethene (TCE)- and 1,2-dichloroethene-respiring anaerobe. *Environ. Microbiol.* **2005**, 7, (9), 1442-1450.
314. Sung, Y.; Ritalahti, K. M.; Apkarian, R. P.; Löffler, F. E., Quantitative PCR confirms purity of strain GT, a novel trichloroethene-to-ethene-respiring *Dehalococcoides* isolate. *Appl. Environ. Microbiol.* **2006**, 72, (3), 1980-1987.
315. Yang, Y.; Higgins, S. A.; Yan, J.; Simsir, B.; Chourey, K.; Iyer, R.; Hettich, R. L.; Baldwin, B.; Ogles, D. M.; Löffler, F. E., Grape pomace compost harbors organohalide-respiring *Dehalogenimonas* species with novel reductive dehalogenase genes. *ISME J* **2017**.
316. Yi, S.; Seth, E. C.; Men, Y. J.; Stabler, S. P.; Allen, R. H.; Alvarez-Cohen, L.; Taga, M. E., Versatility in corrinoid salvaging and remodeling pathways supports corrinoid-dependent metabolism in *Dehalococcoides mccartyi*. *Appl. Environ. Microbiol.* **2012**, 78, (21), 7745-7752.
317. Hayer-Hartl, M.; Bracher, A.; Hartl, F. U., The GroEL-GroES Chaperonin Machine: A Nano-Cage for Protein Folding. *Trends Biochem. Sci.* **2016**, 41, (1), 62-76.
318. Yang, Y.; Capiro, N. L.; Marcet, T. F.; Yan, J.; Pennell, K. D.; Löffler, F. E., Organohalide Respiration with Chlorinated Ethenes under Low pH Conditions. *Environmental Science & Technology* **2017**, 51, (15), 8579-8588.
319. Harding-Marjanovic, K.; Yi, S.; Weathers, T.; Sharp, J.; Sedlak, D.; Alvarez-Cohen, L., Effects of aqueous film-forming foam (AFFF) on trichloroethene (TCE) degradation by a *Dehalococcoides* enrichment culture. *Environ. Sci. Technol* **2015**, 50, (7), 3352-3361.
320. Sawatdeenarunat, C.; Surendra, K. C.; Takara, D.; Oechsner, H.; Khanal, S. K., Anaerobic digestion of lignocellulosic biomass: challenges and opportunities. *Bioresour Technol* **2015**, 178, 178-186.

321. Jeihanipour, A.; Bashiri, R., Perspective of biofuels from wastes. In *Lignocellulose-based bioproducts*, Springer: 2015; pp 37-83.
322. McKendry, P., Energy production from biomass (Part 1): Overview of biomass. *Bioresour Technol* **2002**, *83*, (1), 37-46.
323. Zamani, A., Introduction to lignocellulose-based products. In *Lignocellulose-Based Bioproducts*, Springer: 2015; pp 1-36.
324. Frigon, J. C.; Guiot, S. R., Biomethane production from starch and lignocellulosic crops: a comparative review. *Biofuel Bioprod Bior* **2010**, *4*, (4), 447-458.
325. Jørgensen, H.; Pinelo, M., Enzyme recycling in lignocellulosic biorefineries. *Biofuels, Bioproducts and Biorefining* **2017**, *11*, (1), 150-167.
326. Lee, J., Biological conversion of lignocellulosic biomass to ethanol. *J Biotechnol* **1997**, *56*, (1), 1-24.
327. Agbor, V. B.; Cicek, N.; Sparling, R.; Berlin, A.; Levin, D. B., Biomass pretreatment: fundamentals toward application. *Biotechnol Adv* **2011**, *29*, (6), 675-85.
328. Monlau, F.; Barakat, A.; Trably, E.; Dumas, C.; Steyer, J. P.; Carrere, H., Lignocellulosic Materials Into Biohydrogen and Biomethane: Impact of Structural Features and Pretreatment. *Crit Rev Env Sci Tec* **2013**, *43*, (3), 260-322.
329. Alvarez, C.; Reyes-Sosa, F. M.; Diez, B., Enzymatic hydrolysis of biomass from wood. *Microb Biotechnol* **2016**, *9*, (2), 149-56.
330. van der Lelie, D.; Taghavi, S.; McCorkle, S. M.; Li, L. L.; Malfatti, S. A.; Monteleone, D.; Donohoe, B. S.; Ding, S. Y.; Adney, W. S.; Himmel, M. E.; Tringe, S. G., The metagenome of an anaerobic microbial community decomposing poplar wood chips. *Plos One* **2012**, *7*, (5), e36740.
331. Artzi, L.; Bayer, E. A.; Morais, S., Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. *Nat Rev Microbiol* **2017**, *15*, (2), 83-95.
332. Gasch, C.; Hildebrandt, I.; Rebbe, F.; Röske, I., Enzymatic monitoring and control of a two-phase batch digester leaching system with integrated anaerobic filter. *Energy, Sustainability and Society* **2013**, *3*, (1), 10.
333. Batista-Garcia, R. A.; Sanchez-Carbente, M. D.; Talia, P.; Jackson, S. A.; O'Leary, N. D.; Dobson, A. D. W.; Folch-Mallol, J. L., From lignocellulosic metagenomes to lignocellulolytic genes: trends, challenges and future prospects. *Biofuel Bioprod Bior* **2016**, *10*, (6), 864-882.
334. Lynd, L. R.; Weimer, P. J.; van Zyl, W. H.; Pretorius, I. S., Microbial cellulose utilization: fundamentals and biotechnology. *Microbiol Mol Biol Rev* **2002**, *66*, (3), 506-77, table of contents.
335. Adney, W. S.; Rivard, C. J.; Ming, S. A.; Himmel, M. E., Anaerobic digestion of lignocellulosic biomass and wastes. Cellulases and related enzymes. *Appl Biochem Biotechnol* **1991**, *30*, (2), 165-83.
336. Azman, S.; Khadem, A. F.; Van Lier, J. B.; Zeeman, G.; Plugge, C. M., Presence and Role of Anaerobic Hydrolytic Microbes in Conversion of Lignocellulosic Biomass for Biogas Production. *Crit Rev Env Sci Tec* **2015**, *45*, (23), 2523-2564.
337. Bayer, E. A.; Morag, E.; Lamed, R., The cellulosome--a treasure-trove for biotechnology. *Trends Biotechnol* **1994**, *12*, (9), 379-86.
338. Boontian, N., Conditions of the anaerobic digestion of biomass. *International Journal of Biological, Biomolecular, Agricultural, Food and Biotechnological Engineering* **2014**, *8*, (9), 1036.
339. Bertucci, M.; Calusinska, M.; Goux, X.; Rouland-Lefevre, C.; Untereiner, B.; Ferrer, P.; Gerin, P. A.; Delfosse, P., Carbohydrate Hydrolytic Potential and Redundancy of an Anaerobic Digestion Microbiome Exposed to Acidosis, as Uncovered by Metagenomics. *Appl Environ Microbiol* **2019**, *85*, (15).
340. Gerardi, M. H., *The microbiology of anaerobic digesters*. John Wiley & Sons: 2003.
341. Kougias, P. G.; Campanaro, S.; Treu, L.; Tsapekos, P.; Armani, A.; Angelidaki, I., Spatial Distribution and Diverse Metabolic Functions of Lignocellulose-Degrading Uncultured Bacteria as Revealed by Genome-Centric Metagenomics. *Appl Environ Microbiol* **2018**, *84*, (18).

342. Campanaro, S.; Treu, L.; Kougias, P. G.; Luo, G.; Angelidaki, I., Metagenomic binning reveals the functional roles of core abundant microorganisms in twelve full-scale biogas plants. *Water Res* **2018**, *140*, 123-134.
343. Theuerl, S.; Klang, J.; Heiermann, M.; De Vrieze, J., Marker microbiome clusters are determined by operational parameters and specific key taxa combinations in anaerobic digestion. *Bioresour Technol* **2018**, *263*, 128-135.
344. Treu, L.; Kougias, P. G.; Campanaro, S.; Bassani, I.; Angelidaki, I., Deeper insight into the structure of the anaerobic digestion microbial community; the biogas microbiome database is expanded with 157 new genomes. *Bioresour Technol* **2016**, *216*, 260-266.
345. Gullert, S.; Fischer, M. A.; Turaev, D.; Noebauer, B.; Ilmberger, N.; Wemheuer, B.; Alawi, M.; Rattei, T.; Daniel, R.; Schmitz, R. A.; Grundhoff, A.; Streit, W. R., Deep metagenome and metatranscriptome analyses of microbial communities affiliated with an industrial biogas fermenter, a cow rumen, and elephant feces reveal major differences in carbohydrate hydrolysis strategies. *Biotechnol Biofuels* **2016**, *9*, 121.
346. Bremges, A.; Maus, I.; Belmann, P.; Eikmeyer, F.; Winkler, A.; Albersmeier, A.; Puhler, A.; Schluter, A.; Sczyrba, A., Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant. *Gigascience* **2015**, *4*, 33.
347. Hanreich, A.; Schimpf, U.; Zakrzewski, M.; Schluter, A.; Benndorf, D.; Heyer, R.; Rapp, E.; Puhler, A.; Reichl, U.; Klocke, M., Metagenome and metaproteome analyses of microbial communities in mesophilic biogas-producing anaerobic batch fermentations indicate concerted plant carbohydrate degradation. *Syst Appl Microbiol* **2013**, *36*, (5), 330-8.
348. Ortseifen, V.; Stolze, Y.; Maus, I.; Sczyrba, A.; Bremges, A.; Albaum, S. P.; Jaenicke, S.; Fracowiak, J.; Puhler, A.; Schluter, A., An integrated metagenome and -proteome analysis of the microbial community residing in a biogas production plant. *J Biotechnol* **2016**, *231*, 268-279.
349. Abendroth, C.; Simeonov, C.; Pereto, J.; Antunez, O.; Gavidia, R.; Luschnig, O.; Porcar, M., From grass to gas: microbiome dynamics of grass biomass acidification under mesophilic and thermophilic temperatures. *Biotechnol Biofuels* **2017**, *10*, 171.
350. Hassa, J.; Maus, I.; Off, S.; Puhler, A.; Scherer, P.; Klocke, M.; Schluter, A., Metagenome, metatranscriptome, and metaproteome approaches unraveled compositions and functional relationships of microbial communities residing in biogas plants. *Appl Microbiol Biotechnol* **2018**, *102*, (12), 5045-5063.
351. Lu, F.; Bize, A.; Guillot, A.; Monnet, V.; Madigou, C.; Chapleur, O.; Mazeas, L.; He, P.; Bouchez, T., Metaproteomics of cellulose methanisation under thermophilic conditions reveals a surprisingly high proteolytic activity. *ISME J* **2014**, *8*, (1), 88-102.
352. Hagen, L. H.; Frank, J. A.; Zamanzadeh, M.; Eijsink, V. G. H.; Pope, P. B.; Horn, S. J.; Arntzen, M. O., Quantitative Metaproteomics Highlight the Metabolic Contributions of Uncultured Phylotypes in a Thermophilic Anaerobic Digester. *Appl Environ Microbiol* **2017**, *83*, (2).
353. Lebuhn, M.; Hanreich, A.; Klocke, M.; Schluter, A.; Bauer, C.; Perez, C. M., Towards molecular biomarkers for biogas production from lignocellulose-rich substrates. *Anaerobe* **2014**, *29*, 10-21.
354. Vanwonterghem, I.; Jensen, P. D.; Ho, D. P.; Batstone, D. J.; Tyson, G. W., Linking microbial community structure, interactions and function in anaerobic digesters using new molecular techniques. *Current opinion in biotechnology* **2014**, *27*, 55-64.
355. Abram, F.; Enright, A. M.; O'reilly, J.; Botting, C.; Collins, G.; O'flaherty, V., A metaproteomic approach gives functional insights into anaerobic digestion. *Journal of applied microbiology* **2011**, *110*, (6), 1550-1560.
356. Hanreich, A.; Heyer, R.; Benndorf, D.; Rapp, E.; Pioch, M.; Reichl, U.; Klocke, M., Metaproteome analysis to determine the metabolically active part of a thermophilic microbial community producing biogas from agricultural biomass. *Canadian journal of microbiology* **2012**, *58*, (7), 917-922.

357. Heyer, R.; Kohrs, F.; Reichl, U.; Benndorf, D., Metaproteomics of complex microbial communities in biogas plants. *Microb Biotechnol* **2015**, 8, (5), 749-63.
358. Henrissat, B., A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J* **1991**, 280 (Pt 2), 309-16.
359. Henrissat, B.; Bairoch, A., Updating the sequence-based classification of glycosyl hydrolases. *Biochem J* **1996**, 316 (Pt 2), 695-6.
360. Henrissat, B.; Davies, G. J., Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics. *Plant Physiol* **2000**, 124, (4), 1515-9.
361. Campanaro, S.; Treu, L.; Rodriguez-R, L. M.; Kovalovszki, A.; Ziels, R. M.; Maus, I.; Zhu, X.; Kougias, P. G.; Basile, A.; Luo, G., The anaerobic digestion microbiome: a collection of 1600 metagenome-assembled genomes shows high species diversity related to methane production. *bioRxiv* **2019**, 680553.
362. Yin, Y.; Mao, X.; Yang, J.; Chen, X.; Mao, F.; Xu, Y., dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* **2012**, 40, (Web Server issue), W445-51.
363. Honorato, R., CAZy-parser a way to extract information from the Carbohydrate-Active enZymes Database. *Journal of Open Source Software* **2016**, 1, (8), 53.
364. Finn, R. D.; Clements, J.; Eddy, S. R., HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **2011**, 39, (Web Server issue), W29-37.
365. Buchfink, B.; Xie, C.; Huson, D. H., Fast and sensitive protein alignment using DIAMOND. *Nature methods* **2015**, 12, (1), 59.
366. Busk, P. K.; Pilgaard, B.; Lezyk, M. J.; Meyer, A. S.; Lange, L., Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function. *BMC Bioinformatics* **2017**, 18, (1), 214.
367. Gurdeep Singh, R.; Tanca, A.; Palomba, A.; Van der Jeugt, F.; Verschaffelt, P.; Uzzau, S.; Martens, L.; Dawyndt, P.; Mesuere, B., Unipept 4.0: Functional Analysis of Metaproteome Data. *J Proteome Res* **2019**, 18, (2), 606-615.
368. Warnecke, F.; Luginbuhl, P.; Ivanova, N.; Ghassemian, M.; Richardson, T. H.; Stege, J. T.; Cayouette, M.; McHardy, A. C.; Djordjevic, G.; Aboushadi, N.; Sorek, R.; Tringe, S. G.; Podar, M.; Martin, H. G.; Kunitz, V.; Dalevi, D.; Madejska, J.; Kirton, E.; Platt, D.; Szeto, E.; Salamov, A.; Barry, K.; Mikhailova, N.; Kyrpides, N. C.; Matson, E. G.; Ottesen, E. A.; Zhang, X.; Hernandez, M.; Murillo, C.; Acosta, L. G.; Rigoutsos, I.; Tamayo, G.; Green, B. D.; Chang, C.; Rubin, E. M.; Mathur, E. J.; Robertson, D. E.; Hugenholtz, P.; Leadbetter, J. R., Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **2007**, 450, (7169), 560-5.
369. Wongwilaiwalin, S.; Rattanachomsri, U.; Laothanachareon, T.; Eurwilaichitr, L.; Igarashi, Y.; Champreda, V., Analysis of a thermophilic lignocellulose degrading microbial consortium and multi-species lignocellulolytic enzyme system. *Enzyme and Microbial Technology* **2010**, 47, (6), 283-290.
370. Aylward, F. O.; Burnum, K. E.; Scott, J. J.; Suen, G.; Tringe, S. G.; Adams, S. M.; Barry, K. W.; Nicora, C. D.; Piehowski, P. D.; Purvine, S. O., Metagenomic and metaproteomic insights into bacterial communities in leaf-cutter ant fungus gardens. *The ISME journal* **2012**, 6, (9), 1688.
371. D'Haeseleer, P.; Gladden, J. M.; Allgaier, M.; Chain, P. S.; Tringe, S. G.; Malfatti, S. A.; Aldrich, J. T.; Nicora, C. D.; Robinson, E. W.; Pasa-Tolic, L.; Hugenholtz, P.; Simmons, B. A.; Singer, S. W., Proteogenomic analysis of a thermophilic bacterial consortium adapted to deconstruct switchgrass. *Plos One* **2013**, 8, (7), e68465.
372. Jimenez, D. J.; Maruthamuthu, M.; van Elsas, J. D., Metasecretome analysis of a lignocellulolytic microbial consortium grown on wheat straw, xylan and xylose. *Biotechnol Biofuels* **2015**, 8, 199.

373. Zhu, N.; Yang, J.; Ji, L.; Liu, J.; Yang, Y.; Yuan, H., Metagenomic and metaproteomic analyses of a corn stover-adapted microbial consortium EMSD5 reveal its taxonomic and enzymatic basis for degrading lignocellulose. *Biotechnol Biofuels* **2016**, *9*, 243.
374. Speda, J.; Jonsson, B. H.; Carlsson, U.; Karlsson, M., Metaproteomics-guided selection of targeted enzymes for bioprospecting of mixed microbial communities. *Biotechnol Biofuels* **2017**, *10*, 128.
375. Hori, C.; Gaskell, J.; Cullen, D.; Sabat, G.; Stewart, P. E.; Lail, K.; Peng, Y.; Barry, K.; Grigoriev, I. V.; Kohler, A.; Fauchery, L.; Martin, F.; Zeiner, C. A.; Bhatnagar, J. M., Multi-omic Analyses of Extensively Decayed *Pinus contorta* Reveal Expression of a Diverse Array of Lignocellulose-Degrading Enzymes. *Appl Environ Microbiol* **2018**, *84*, (20).
376. Alessi, A. M.; Bird, S. M.; Oates, N. C.; Li, Y.; Dowle, A. A.; Novotny, E. H.; deAzevedo, E. R.; Bennett, J. P.; Polikarpov, I.; Young, J. P. W.; McQueen-Mason, S. J.; Bruce, N. C., Defining functional diversity for lignocellulose degradation in a microbial community using multi-omics studies. *Biotechnol Biofuels* **2018**, *11*, 166.
377. de Castro, E.; Sigrist, C. J.; Gattiker, A.; Bulliard, V.; Langendijk-Genevaux, P. S.; Gasteiger, E.; Bairoch, A.; Hulo, N., ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res* **2006**, *34*, (Web Server issue), W362-5.
378. Colangelo, C. M.; Ivosev, G.; Chung, L.; Abbott, T.; Shifman, M.; Sakaue, F.; Cox, D.; Kitchen, R. R.; Burton, L.; Tate, S. A.; Gulcicek, E.; Bonner, R.; Rinehart, J.; Nairn, A. C.; Williams, K. R., Development of a highly automated and multiplexed targeted proteome pipeline and assay for 112 rat brain synaptic proteins. *Proteomics* **2015**, *15*, (7), 1202-14.
379. Gallien, S.; Kim, S. Y.; Domon, B., Large-Scale Targeted Proteomics Using Internal Standard Triggered-Parallel Reaction Monitoring (IS-PRM). *Molecular & Cellular Proteomics* **2015**, *14*, (6), 1630-1644.
380. Jia, Y.; Ng, S. K.; Lu, H.; Cai, M.; Lee, P. K. H., Genome-centric metatranscriptomes and ecological roles of the active microbial populations during cellulosic biomass anaerobic digestion. *Biotechnol Biofuels* **2018**, *11*, 117.
381. Rui, J.; Li, J.; Zhang, S.; Yan, X.; Wang, Y.; Li, X., The core populations and co-occurrence patterns of prokaryotic communities in household biogas digesters. *Biotechnol Biofuels* **2015**, *8*, 158.
382. Holmes, D.; Smith, J., Biologically produced methane as a renewable energy source. In *Advances in applied microbiology*, Elsevier: 2016; Vol. 97, pp 1-61.
383. Lombard, V.; Golaconda Ramulu, H.; Drula, E.; Coutinho, P. M.; Henrissat, B., The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **2014**, *42*, (Database issue), D490-5.
384. El Kaoutari, A.; Armougom, F.; Gordon, J. I.; Raoult, D.; Henrissat, B., The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat Rev Microbiol* **2013**, *11*, (7), 497-504.
385. Coutinho, P. M.; Stam, M.; Blanc, E.; Henrissat, B., Why are there so many carbohydrate-active enzyme-related genes in plants? *Trends in plant science* **2003**, *8*, (12), 563-565.
386. Huang, L.; Zhang, H.; Wu, P.; Entwistle, S.; Li, X.; Yohe, T.; Yi, H.; Yang, Z.; Yin, Y., dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation. *Nucleic Acids Res* **2018**, *46*, (D1), D516-D521.
387. Magidovich, H.; Eichler, J., Glycosyltransferases and oligosaccharyltransferases in Archaea: putative components of the N-glycosylation pathway in the third domain of life. *FEMS Microbiol Lett* **2009**, *300*, (1), 122-30.
388. Calo, D.; Kaminski, L.; Eichler, J., Protein glycosylation in Archaea: sweet and extreme. *Glycobiology* **2010**, *20*, (9), 1065-76.

389. Orsi, W. D.; Vuillemin, A.; Rodriguez, P.; Coskun, O. K.; Gomez-Saez, G. V.; Lavik, G.; Morholz, V.; Ferdelman, T. G., Metabolic activity analyses demonstrate that Lokiarchaeon exhibits homoacetogenesis in sulfidic marine sediments. *Nat Microbiol* **2020**, 5, (2), 248-255.
390. Henrissat, B.; Davies, G., Structural and sequence-based classification of glycoside hydrolases. *Curr Opin Struct Biol* **1997**, 7, (5), 637-44.
391. Cornish-Bowden, A., Current IUBMB recommendations on enzyme nomenclature and kinetics. *Perspectives in Science* **2014**, 1, (1-6), 74-87.
392. Fincher, G.; Mark, B.; Brumer, H. Glycoside Hydrolase Family 3. <http://www.cazypedia.org/> (May 6, 2020),
393. Kanehisa, M.; Sato, Y.; Morishima, K., BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* **2016**, 428, (4), 726-731.
394. Cantarel, B. L.; Coutinho, P. M.; Rancurel, C.; Bernard, T.; Lombard, V.; Henrissat, B., The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* **2009**, 37, (Database issue), D233-8.
395. Adney, W. S.; Rivard, C. J.; Grohmann, K.; Himmel, M. E., Characterization of Polysaccharidase Activity Optima in the Anaerobic-Digestion of Municipal Solid-Waste. *Biotechnol Lett* **1989**, 11, (3), 207-210.
396. Ow, S. Y.; Salim, M.; Noirel, J.; Evans, C.; Wright, P. C., Minimising iTRAQ ratio compression through understanding LC-MS elution dependence and high-resolution HILIC fractionation. *Proteomics* **2011**, 11, (11), 2341-2346.
397. Yang, Y.; Qiang, X.; Owsiany, K.; Zhang, S.; Thannhauser, T. W.; Li, L., Evaluation of different multidimensional LC-MS/MS pipelines for isobaric tags for relative and absolute quantitation (iTRAQ)-based proteomic analysis of potato tubers in response to cold storage. *Journal of proteome research* **2011**, 10, (10), 4647-4660.
398. Toll, H.; Oberacher, H.; Swart, R.; Huber, C. G., Separation, detection, and identification of peptides by ion-pair reversed-phase high-performance liquid chromatography-electrospray ionization mass spectrometry at high and low pH. *J Chromatogr A* **2005**, 1079, (1-2), 274-86.
399. Manadas, B.; English, J. A.; Wynne, K. J.; Cotter, D. R.; Dunn, M. J., Comparative analysis of OFFGel, strong cation exchange with pH gradient, and RP at high pH for first-dimensional separation of peptides from a membrane-enriched protein fraction. *Proteomics* **2009**, 9, (22), 5194-8.
400. Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C., Orthogonality of separation in two-dimensional liquid chromatography. *Anal Chem* **2005**, 77, (19), 6426-34.
401. Thakur, S. S.; Geiger, T.; Chatterjee, B.; Bandilla, P.; Frohlich, F.; Cox, J.; Mann, M., Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol Cell Proteomics* **2011**, 10, (8), M110 003699.
402. Wang, J.; Bourne, P. E.; Bandeira, N., Peptide identification by database search of mixture tandem mass spectra. *Mol Cell Proteomics* **2011**, 10, (12), M111 010017.
403. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* **2011**, 10, (4), 1794-805.
404. Shteynberg, D.; Mendoza, L.; Hoopmann, M. R.; Sun, Z.; Schmidt, F.; Deutsch, E. W.; Moritz, R. L., reSpect: software for identification of high and low abundance ion species in chimeric tandem mass spectra. *J Am Soc Mass Spectrom* **2015**, 26, (11), 1837-47.
405. Niu, M.; Mao, X.; Ying, W.; Qin, W.; Zhang, Y.; Qian, X., Determination of monoisotopic masses of chimera spectra from high-resolution mass spectrometric data by use of isotopic peak intensity ratio modeling. *Rapid Commun Mass Spectrom* **2012**, 26, (16), 1875-86.
406. Scheltema, R. A.; Hauschild, J. P.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M., The Q Exactive HF, a Benchtop mass spectrometer with a

pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol Cell Proteomics* **2014**, 13, (12), 3698-708.

407. Clarkson, S. M.; Giannone, R. J.; Kridelbaugh, D. M.; Elkins, J. G.; Guss, A. M.; Michener, J. K., Construction and Optimization of a Heterologous Pathway for Protocatechuate Catabolism in *Escherichia coli* Enables Bioconversion of Model Aromatic Compounds. *Appl Environ Microbiol* **2017**, 83, (18).

408. Kenar, E.; Franken, H.; Forcisi, S.; Wörmann, K.; Häring, H.-U.; Lehmann, R.; Schmitt-Kopplin, P.; Zell, A.; Kohlbacher, O., Automated label-free quantification of metabolites from liquid chromatography–mass spectrometry data. *Molecular & cellular proteomics* **2014**, 13, (1), 348-359.

409. Rost, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H. C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmstrom, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O., OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods* **2016**, 13, (9), 741-8.

410. Minakuchi, H.; Nakanishi, K.; Soga, N.; Ishizuka, N.; Tanaka, N., Octadecylsilylated porous silica rods as separation media for reversed-phase liquid chromatography. *Analytical chemistry* **1996**, 68, (19), 3498-3501.

411. Liu, H.; Finch, J. W.; Lavalley, M. J.; Collamati, R. A.; Benevides, C. C.; Gebler, J. C., Effects of column length, particle size, gradient length and flow rate on peak capacity of nano-scale liquid chromatography for peptide separations. *J Chromatogr A* **2007**, 1147, (1), 30-6.

412. Sandra, K.; Moshir, M.; D'hondt, F.; Verleysen, K.; Kas, K.; Sandra, P., Highly efficient peptide separations in proteomics: Part 1. Unidimensional high performance liquid chromatography. *Journal of Chromatography B* **2008**, 866, (1-2), 48-63.

413. Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J., The one hour yeast proteome. *Mol Cell Proteomics* **2014**, 13, (1), 339-47.

414. Schaller, A., A cut above the rest: the regulatory function of plant proteases. *Planta* **2004**, 220, (2), 183-97.

415. Fesenko, I. A.; Arapidi, G. P.; Skripnikov, A. Y.; Alexeev, D. G.; Kostryukova, E. S.; Manolov, A. I.; Altukhov, I. A.; Khazigaleeva, R. A.; Seredina, A. V.; Kovalchuk, S. I.; Ziganshin, R. H.; Zgoda, V. G.; Novikova, S. E.; Semashko, T. A.; Slizhikova, D. K.; Ptushenko, V. V.; Gorbachev, A. Y.; Govorun, V. M.; Ivanov, V. T., Specific pools of endogenous peptides are present in gametophore, protonema, and protoplast cells of the moss *Physcomitrella patens*. *BMC Plant Biol* **2015**, 15, 87.

416. Chen, G. H.; Huang, L. T.; Yap, M. N.; Lee, R. H.; Huang, Y. J.; Cheng, M. C.; Chen, S. C., Molecular characterization of a senescence-associated gene encoding cysteine proteinase and its gene expression during leaf senescence in sweet potato. *Plant Cell Physiol* **2002**, 43, (9), 984-91.

417. Plett, J. M.; Martin, F. M., Know your enemy, embrace your friend: using omics to understand how plants respond differently to pathogenic and mutualistic microorganisms. *Plant J* **2018**, 93, (4), 729-746.

418. Song, F.; Qi, D.; Liu, X.; Kong, X.; Gao, Y.; Zhou, Z.; Wu, Q., Proteomic analysis of symbiotic proteins of *Glomus mosseae* and *Amorpha fruticosa*. *Sci Rep* **2015**, 5, 18031.

419. Takeda, N.; Sato, S.; Asamizu, E.; Tabata, S.; Parniske, M., Apoplastic plant subtilases support arbuscular mycorrhiza development in *Lotus japonicus*. *Plant J* **2009**, 58, (5), 766-77.

420. Szuba, A.; Marczak, L.; Karlinski, L.; Mucha, J.; Tomaszewski, D., Regulation of the leaf proteome by inoculation of *Populus x canescens* with two *Paxillus involutus* isolates differing in root colonization rates. *Mycorrhiza* **2019**, 29, (5), 503-517.

421. Farkas, A.; Maroti, G.; Durgo, H.; Gyorgypal, Z.; Lima, R. M.; Medzihradszky, K. F.; Kereszt, A.; Mergaert, P.; Kondorosi, E., Medicago truncatula symbiotic peptide NCR247

- contributes to bacteroid differentiation through multiple mechanisms. *P Natl Acad Sci USA* **2014**, *111*, (14), 5183-5188.
422. Mergaert, P.; Nikovics, K.; Kelemen, Z.; Maunoury, N.; Vaubert, D.; Kondorosi, A.; Kondorosi, E., A novel family in *Medicago truncatula* consisting of more than 300 nodule-specific genes coding for small, secreted polypeptides with conserved cysteine motifs. *Plant Physiol* **2003**, *132*, (1), 161-73.
423. van Loon, L. C.; Rep, M.; Pieterse, C. M., Significance of inducible defense-related proteins in infected plants. *Annu Rev Phytopathol* **2006**, *44*, 135-62.
424. Stec, B., Plant thionins--the structural perspective. *Cell Mol Life Sci* **2006**, *63*, (12), 1370-85.
425. Mitchum, M. G.; Wang, X.; Wang, J.; Davis, E. L., Role of nematode peptides and other small molecules in plant parasitism. *Annu Rev Phytopathol* **2012**, *50*, 175-95.
426. Miyawaki, K.; Tabata, R.; Sawa, S., Evolutionarily conserved CLE peptide signaling in plant development, symbiosis, and parasitism. *Curr Opin Plant Biol* **2013**, *16*, (5), 598-606.
427. Plett, J. M.; Kemppainen, M.; Kale, S. D.; Kohler, A.; Legue, V.; Brun, A.; Tyler, B. M.; Pardo, A. G.; Martin, F., A secreted effector protein of *Laccaria bicolor* is required for symbiosis development. *Curr Biol* **2011**, *21*, (14), 1197-203.
428. Pellegrin, C.; Daguerre, Y.; Ruytinx, J.; Guinet, F.; Kemppainen, M.; Frey, N. F. D.; Puech-Pages, V.; Hecker, A.; Pardo, A. G.; Martin, F. M.; Veneault-Fourrey, C., *Laccaria bicolor* MiSSP8 is a small-secreted protein decisive for the establishment of the ectomycorrhizal symbiosis. *Environ Microbiol* **2019**.
429. Ohyama, K.; Shinohara, H.; Ogawa-Ohnishi, M.; Matsubayashi, Y., A glycopeptide regulating stem cell fate in *Arabidopsis thaliana*. *Nat Chem Biol* **2009**, *5*, (8), 578-80.
430. Amano, Y.; Tsubouchi, H.; Shinohara, H.; Ogawa, M.; Matsubayashi, Y., Tyrosine-sulfated glycopeptide involved in cellular proliferation and expansion in *Arabidopsis*. *Proc Natl Acad Sci U S A* **2007**, *104*, (46), 18333-8.
431. Shinohara, H.; Matsubayashi, Y., Chemical synthesis of *Arabidopsis* CLV3 glycopeptide reveals the impact of hydroxyproline arabinosylation on peptide conformation and activity. *Plant Cell Physiol* **2013**, *54*, (3), 369-74.
432. Kondo, T.; Sawa, S.; Kinoshita, A.; Mizuno, S.; Kakimoto, T.; Fukuda, H.; Sakagami, Y., A plant peptide encoded by CLV3 identified by in situ MALDI-TOF MS analysis. *Science* **2006**, *313*, (5788), 845-8.
433. Matsuzaki, Y.; Ogawa-Ohnishi, M.; Mori, A.; Matsubayashi, Y., Secreted peptide signals required for maintenance of root stem cell niche in *Arabidopsis*. *Science* **2010**, *329*, (5995), 1065-7.
434. Whitford, R.; Fernandez, A.; Tejos, R.; Perez, A. C.; Kleine-Vehn, J.; Vanneste, S.; Drozdzecki, A.; Leitner, J.; Abas, L.; Aerts, M.; Hoogewijs, K.; Baster, P.; De Groodt, R.; Lin, Y. C.; Storme, V.; Van de Peer, Y.; Beeckman, T.; Madder, A.; Devreese, B.; Luschnig, C.; Friml, J.; Hilson, P., GOLVEN secretory peptides regulate auxin carrier turnover during plant gravitropic responses. *Dev Cell* **2012**, *22*, (3), 678-85.
435. Mohd-Radzman, N. A.; Binos, S.; Truong, T. T.; Imin, N.; Mariani, M.; Djordjevic, M. A., Novel MtCEP1 peptides produced in vivo differentially regulate root development in *Medicago truncatula*. *J Exp Bot* **2015**, *66*, (17), 5289-300.
436. Patel, N.; Mohd-Radzman, N. A.; Corcilus, L.; Crossett, B.; Connolly, A.; Cordwell, S. J.; Ivanovici, A.; Taylor, K.; Williams, J.; Binos, S.; Mariani, M.; Payne, R. J.; Djordjevic, M. A., Diverse Peptide Hormones Affecting Root Growth Identified in the *Medicago truncatula* Secreted Peptidome. *Mol Cell Proteomics* **2018**, *17*, (1), 160-174.
437. Zhang, K.; McKinlay, C.; Hocart, C. H.; Djordjevic, M. A., The *Medicago truncatula* small protein proteome and peptidome. *J Proteome Res* **2006**, *5*, (12), 3355-67.
438. Van de Velde, W.; Zehirov, G.; Szatmari, A.; Debreczeny, M.; Ishihara, H.; Kevei, Z.; Farkas, A.; Mikulass, K.; Nagy, A.; Tiricz, H.; Satiat-Jeunemaitre, B.; Alunni, B.; Bourge, M.;

- Kucho, K.; Abe, M.; Kereszt, A.; Maroti, G.; Uchiumi, T.; Kondorosi, E.; Mergaert, P., Plant peptides govern terminal differentiation of bacteria in symbiosis. *Science* **2010**, 327, (5969), 1122-6.
439. Durgo, H.; Klement, E.; Hunyadi-Gulyas, E.; Szucs, A.; Kereszt, A.; Medzihradszky, K. F.; Kondorosi, E., Identification of nodule-specific cysteine-rich plant peptides in endosymbiotic bacteria. *Proteomics* **2015**, 15, (13), 2291-5.
440. Muth, T.; Hartkopf, F.; Vaudel, M.; Renard, B. Y., A Potential Golden Age to Come-Current Tools, Recent Use Cases, and Future Avenues for De Novo Sequencing in Proteomics. *Proteomics* **2018**, 18, (18), e1700150.
441. Dong, M.; Ye, M.; Cheng, K.; Dong, J.; Zhu, J.; Qin, H.; Bian, Y.; Zou, H., Identification of phosphopeptides with unknown cleavage specificity by a de novo sequencing assisted database search strategy. *Proteomics* **2014**, 14, (21-22), 2410-6.
442. Kapp, E. A.; Schutz, F.; Connolly, L. M.; Chakel, J. A.; Meza, J. E.; Miller, C. A.; Fenyo, D.; Eng, J. K.; Adkins, J. N.; Omenn, G. S.; Simpson, R. J., An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. *Proteomics* **2005**, 5, (13), 3475-3490.
443. Vincent, D.; Kohler, A.; Claverol, S.; Solier, E.; Joets, J.; Gibon, J.; Lebrun, M. H.; Plomion, C.; Martin, F., Secretome of the free-living mycelium from the ectomycorrhizal basidiomycete *Laccaria bicolor*. *J Proteome Res* **2012**, 11, (1), 157-71.
444. Plett, J. M.; Yin, H.; Mewalal, R.; Hu, R.; Li, T.; Ranjan, P.; Jawdy, S.; De Paoli, H. C.; Butler, G.; Burch-Smith, T. M.; Guo, H. B.; Ju Chen, C.; Kohler, A.; Anderson, I. C.; Labbe, J. L.; Martin, F.; Tuskan, G. A.; Yang, X., *Populus trichocarpa* encodes small, effector-like secreted proteins that are highly induced during mutualistic symbiosis. *Sci Rep* **2017**, 7, (1), 382.
445. Deveau, A.; Palin, B.; Delaruelle, C.; Peter, M.; Kohler, A.; Pierrat, J. C.; Sarniguier, A.; Garbaye, J.; Martin, F.; Frey-Klett, P., The mycorrhiza helper *Pseudomonas fluorescens* BBc6R8 has a specific priming effect on the growth, morphology and gene expression of the ectomycorrhizal fungus *Laccaria bicolor* S238N. *New Phytol* **2007**, 175, (4), 743-755.
446. Felten, J.; Kohler, A.; Morin, E.; Bhalerao, R. P.; Palme, K.; Martin, F.; Dittengou, F. A.; Legue, V., The Ectomycorrhizal Fungus *Laccaria bicolor* Stimulates Lateral Root Formation in Poplar and Arabidopsis through Auxin Transport and Signaling. *Plant Physiol* **2009**, 151, (4), 1991-2005.
447. Narvaez-Vasquez, J.; Pearce, G.; Ryan, C. A., The plant cell wall matrix harbors a precursor of defense signaling peptides. *Proc Natl Acad Sci U S A* **2005**, 102, (36), 12974-7.
448. Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* **2003**, 17, (20), 2337-42.
449. Schilling, B.; Rardin, M. J.; MacLean, B. X.; Zawadzka, A. M.; Frewen, B. E.; Cusack, M. P.; Sorensen, D. J.; Bereman, M. S.; Jing, E.; Wu, C. C.; Verdin, E.; Kahn, C. R.; Maccoss, M. J.; Gibson, B. W., Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation. *Mol Cell Proteomics* **2012**, 11, (5), 202-14.
450. Jarnuczak, A. F.; Lee, D. C.; Lawless, C.; Holman, S. W.; Evers, C. E.; Hubbard, S. J., Analysis of Intrinsic Peptide Detectability via Integrated Label-Free and SRM-Based Absolute Quantitative Proteomics. *J Proteome Res* **2016**, 15, (9), 2945-59.
451. Xu, T.-T.; Ren, S.-C.; Song, X.-F.; Liu, C.-M., CLE19 expressed in the embryo regulates both cotyledon establishment and endosperm development in Arabidopsis. *Journal of experimental botany* **2015**, 66, (17), 5217-5227.
452. Nelson, C. J.; Alexova, R.; Jacoby, R. P.; Millar, A. H., Proteins with high turnover rate in barley leaves estimated by proteome analysis combined with in planta isotope labeling. *Plant Physiol* **2014**, 166, (1), 91-108.

453. Yi, S. J.; Kim, K., Histone tail cleavage as a novel epigenetic regulatory mechanism for gene expression. *Bmb Rep* **2018**, *51*, (5), 211-218.
454. Dhaenens, M.; Glibert, P.; Meert, P.; Vossaert, L.; Deforce, D., Histone proteolysis: A proposal for categorization into clipping and degradation. *Bioessays* **2015**, *37*, (1), 70-79.
455. Shi, L. X.; Schroder, W. P., The low molecular mass subunits of the photosynthetic supracomplex, photosystem II. *Biochim Biophys Acta* **2004**, *1608*, (2-3), 75-96.
456. Kulkarni, R. D.; Kelkar, H. S.; Dean, R. A., An eight-cysteine-containing CFEM domain unique to a group of fungal membrane proteins. *Trends Biochem Sci* **2003**, *28*, (3), 118-21.
457. Armenteros, J. J. A.; Tsirigos, K. D.; Sonderby, C. K.; Petersen, T. N.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H., SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology* **2019**, *37*, (4), 420-+.
458. Kall, L.; Krogh, A.; Sonnhammer, E. L. L., Advantages of combined transmembrane topology and signal peptide prediction - the Phobius web server. *Nucleic Acids Res* **2007**, *35*, W429-W432.
459. Pellegrin, C.; Morin, E.; Martin, F. M.; Veneault-Fourrey, C., Comparative Analysis of Secretomes from Ectomycorrhizal Fungi with an Emphasis on Small-Secreted Proteins. *Front Microbiol* **2015**, *6*, 1278.
460. Johansson, T.; Le Quere, A.; Ahren, D.; Soderstrom, B.; Erlandsson, R.; Lundeberg, J.; Uhlen, M.; Tunlid, A., Transcriptional responses of *Paxillus involutus* and *Betula pendula* during formation of ectomycorrhizal root tissue. *Mol Plant Microbe Interact* **2004**, *17*, (2), 202-15.
461. Piñero, S.; Rivera, J.; Romero, D.; Cevallos, M. A.; Martínez, A.; Bolívar, F.; Gosset, G., Tyrosinase from *Rhizobium etli* is involved in nodulation efficiency and symbiosis-associated stress resistance. *Journal of molecular microbiology and biotechnology* **2007**, *13*, (1-3), 35-44.
462. Sommer, R.; Makshakova, O. N.; Wohlschlager, T.; Hutin, S.; Marsh, M.; Titz, A.; Kunzler, M.; Varrot, A., Crystal Structures of Fungal Tectonin in Complex with O-Methylated Glycans Suggest Key Role in Innate Immune Defense. *Structure* **2018**, *26*, (3), 391-402 e4.
463. Giollant, M.; Guillot, J.; Damez, M.; Dusser, M.; Didier, P.; Didier, E., Characterization of a Lectin from *Lactarius-Deterrimus* - Research on the Possible Involvement of the Fungal Lectin in Recognition between Mushroom and Spruce during the Early Stages of Mycorrhizae Formation. *Plant Physiol* **1993**, *101*, (2), 513-522.
464. Deveau, A.; Palin, B.; Delaruelle, C.; Peter, M.; Kohler, A.; Pierrat, J.-C.; Sarniguet, A.; Garbaye, J.; Martin, F.; Frey-Klett, P., The mycorrhiza helper *Pseudomonas fluorescens* BBc6R8 has a specific priming effect on the growth, morphology and gene expression of the ectomycorrhizal fungus *Laccaria bicolor* S238N. *New Phytol* **2007**, *175*, (4), 743-755.
465. Lanfranco, L.; Bolchi, A.; Ros, E. C.; Ottonello, S.; Bonfante, P., Differential expression of a metallothionein gene during the presymbiotic versus the symbiotic phase of an arbuscular mycorrhizal fungus. *Plant Physiol* **2002**, *130*, (1), 58-67.
466. Reddy, M. S.; Prasanna, L.; Marmesse, R.; Fraissinet-Tachet, L., Differential expression of metallothioneins in response to heavy metals and their involvement in metal tolerance in the symbiotic basidiomycete *Laccaria bicolor*. *Microbiology* **2014**, *160*, (10), 2235-2242.
467. Klaassen, C. D.; Choudhuri, S.; McKim Jr, J. M.; Lehman-McKeeman, L. D.; Kershaw, W. C., In vitro and in vivo studies on the degradation of metallothionein. *Environmental health perspectives* **1994**, *102*, (suppl 3), 141-146.
468. Hensbergen, P.; Donker, M.; Hunziker, P.; Van der Schors, R.; Van Straalen, N., Two metal-binding peptides from the insect *Orchesella cincta* (Collembola) as a result of metallothionein cleavage. *Insect biochemistry and molecular biology* **2001**, *31*, (11), 1105-1114.
469. Kille, P.; Winge, D. R.; Harwood, J. L.; Kay, J., A plant metallothionein produced in *E. coli*. *FEBS letters* **1991**, *295*, (1-3), 171-175.
470. Voiblet, C.; Duplessis, S.; Encelot, N.; Martin, F., Identification of symbiosis-regulated genes in *Eucalyptus globulus*-*Pisolithus tinctorius* ectomycorrhiza by differential hybridization of arrayed cDNAs. *The Plant Journal* **2001**, *25*, (2), 181-191.

471. Morel, M.; Jacob, C.; Kohler, A.; Johansson, T.; Martin, F.; Chalot, M.; Brun, A., Identification of genes differentially expressed in extraradical mycelium and ectomycorrhizal roots during *Paxillus involutus*-*Betula pendula* ectomycorrhizal symbiosis. *Appl Environ Microbiol* **2005**, *71*, (1), 382-91.
472. Sultan, A.; Andersen, B.; Svensson, B.; Finnie, C., Exploring the Plant-Microbe Interface by Profiling the Surface-Associated Proteins of Barley Grains. *J Proteome Res* **2016**, *15*, (4), 1151-67.
473. Zeppa, S.; Potenza, L.; Polidori, E.; Guescini, M.; Agostini, D.; Giomaro, G.; Stocchi, V., Cloning and characterisation of a polyubiquitin gene from the ectomycorrhizal fungus *Tuber borchii* Vittad. *Current genetics* **2001**, *40*, (1), 49-53.
474. Chakraborty, R.; Wu, C. H.; Hazen, T. C., Systems biology approach to bioremediation. *Curr Opin Biotechnol* **2012**, *23*, (3), 483-90.
475. Saito, M. A.; Dorsk, A.; Post, A. F.; McIlvin, M. R.; Rappe, M. S.; DiTullio, G. R.; Moran, D. M., Needles in the blue sea: Sub-species specificity in targeted protein biomarker analyses within the vast oceanic microbial metaproteome. *Proteomics* **2015**, *15*, (20), 3521-3531.
476. Gouveia, D.; Chaumot, A.; Charnot, A.; Almunia, C.; Francois, A.; Navarro, L.; Armengaud, J.; Salvador, A.; Geffard, O., Ecotoxico-Proteomics for Aquatic Environmental Monitoring: First in Situ Application of a New Proteomics-Based Multibiomarker Assay Using Caged Amphipods. *Environmental Science & Technology* **2017**, *51*, (22), 13417-13426.
477. Hanreich, A., Phylogenetic and functional analysis of microbial, biogas-producing communities during the hydrolysis of lignocellulosic biomass.
478. Kiyonami, R.; Schoen, A.; Prakash, A.; Peterman, S.; Zabrouskov, V.; Picotti, P.; Aebersold, R.; Huhmer, A.; Domon, B., Increased selectivity, analytical precision, and throughput in targeted proteomics. *Mol Cell Proteomics* **2011**, *10*, (2), M110 002931.
479. Vidova, V.; Spacil, Z., A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition. *Anal Chim Acta* **2017**, *964*, 7-23.
480. Rost, H. L., Deep learning adds an extra dimension to peptide fragmentation. *Nat Methods* **2019**, *16*, (6), 469-470.
481. Kryuchkov, F.; Verano-Braga, T.; Hansen, T. A.; Sprenger, R. R.; Kjeldsen, F., Deconvolution of mixture spectra and increased throughput of peptide identification by utilization of intensified complementary ions formed in tandem mass spectrometry. *Journal of proteome research* **2013**, *12*, (7), 3362-71.
482. Xu, G.; Stupak, J.; Yang, L.; Hu, L.; Guo, B.; Li, J., Deconvolution in mass spectrometry based proteomics. *Rapid Commun Mass Spectrom* **2018**, *32*, (10), 763-774.
483. Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H. C.; Aiche, S.; Kuster, B.; Wilhelm, M., Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat Methods* **2019**, *16*, (6), 509-518.
484. Tiwary, S.; Levy, R.; Gutenbrunner, P.; Salinas Soto, F.; Palaniappan, K. K.; Deming, L.; Berndt, M.; Brant, A.; Cimermancic, P.; Cox, J., High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat Methods* **2019**, *16*, (6), 519-525.
485. Min, S.; Lee, B.; Yoon, S., Deep learning in bioinformatics. *Brief Bioinform* **2017**, *18*, (5), 851-869.
486. Zhang, Y.; Wen, Z.; Washburn, M. P.; Florens, L., Evaluating Chromatographic Approaches for the Quantitative Analysis of a Human Proteome on Orbitrap-Based Mass Spectrometry Systems. *Journal of proteome research* **2019**, *18*, (4), 1857-1869.
487. Budamgunta, H.; Maes, E.; Willems, H.; Menschaert, G.; Schildermans, K.; Kumar, A. A.; Boonen, K.; Baggerman, G., Multiple solvent elution, a method to counter the effects of coelution and ion suppression in LC-MS analysis in bottom up proteomics. *J Chromatogr B Analyt Technol Biomed Life Sci* **2019**, *1124*, 256-264.

488. Young, C.; Podtelejnikov, A. V.; Nielsen, M. L., Improved Reversed Phase Chromatography of Hydrophilic Peptides from Spatial and Temporal Changes in Column Temperature. *Journal of proteome research* **2017**, *16*, (6), 2307-2317.
489. Orr, M. W.; Mao, Y.; Storz, G.; Qian, S. B., Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res* **2019**.
490. Hsu, P. Y.; Benfey, P. N., Small but Mighty: Functional Peptides Encoded by Small ORFs in Plants. *Proteomics* **2018**, *18*, (10), e1700038.
491. Masachis, S.; Segorbe, D.; Turra, D.; Leon-Ruiz, M.; Furst, U.; El Ghalid, M.; Leonard, G.; Lopez-Berges, M. S.; Richards, T. A.; Felix, G.; Di Pietro, A., A fungal pathogen secretes plant alkalinizing peptides to increase infection. *Nat Microbiol* **2016**, *1*, (6), 16043.
492. Saghatelian, A.; Couso, J. P., Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat Chem Biol* **2015**, *11*, (12), 909-16.
493. Baggerman, G.; Verleyen, P.; Clynen, E.; Huybrechts, J.; De Loof, A.; Schoofs, L., Peptidomics. *Journal of Chromatography B* **2004**, *803*, (1), 3-16.
494. Delcourt, V.; Staskevicius, A.; Salzet, M.; Fournier, I.; Roucou, X., Small Proteins Encoded by Unannotated ORFs are Rising Stars of the Proteome, Confirming Shortcomings in Genome Annotations and Current Vision of an mRNA. *Proteomics* **2018**, *18*, (10), e1700058.
495. Yang, X.; Tschaplinski, T. J.; Hurst, G. B.; Jawdy, S.; Abraham, P. E.; Lankford, P. K.; Adams, R. M.; Shah, M. B.; Hettich, R. L.; Lindquist, E., Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome research* **2011**, *21*, (4), 634-641.
496. Garcia, K.; Ané, J.-M., Comparative analysis of secretomes from ectomycorrhizal fungi with an emphasis on small-secreted proteins. *Frontiers in microbiology* **2016**, *7*, 1734.
497. Martin, F.; Kohler, A.; Murat, C.; Veneault-Fourrey, C.; Hibbett, D. S., Unearthing the roots of ectomycorrhizal symbioses. *Nat Rev Microbiol* **2016**, *14*, (12), 760-773.
498. Felten, J.; Kohler, A.; Morin, E.; Bhalerao, R. P.; Palme, K.; Martin, F.; Ditengou, F. A.; Legué, V., The ectomycorrhizal fungus *Laccaria bicolor* stimulates lateral root formation in poplar and *Arabidopsis* through auxin transport and signaling. *Plant Physiol* **2009**, *151*, (4), 1991-2005.
499. Martin, F.; Aerts, A.; Ahrén, D.; Brun, A.; Danchin, E.; Duchaussoy, F.; Gibon, J.; Kohler, A.; Lindquist, E.; Pereda, V., The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* **2008**, *452*, (7183), 88.
500. Courty, P. E.; Labbe, J.; Kohler, A.; Marçais, B.; Bastien, C.; Churin, J. L.; Garbaye, J.; Le Tacon, F., Effect of poplar genotypes on mycorrhizal infection and secreted enzyme activities in mycorrhizal and non-mycorrhizal roots. *J Exp Bot* **2011**, *62*, (1), 249-60.
501. Labbé, J.; Jorge, V.; Kohler, A.; Vion, P.; Marçais, B.; Bastien, C.; Tuskan, G. A.; Martin, F.; Le Tacon, F., Identification of quantitative trait loci affecting ectomycorrhizal symbiosis in an interspecific F1 poplar cross and differential expression of genes in ectomycorrhizas of the two parents: *Populus deltoides* and *Populus trichocarpa*. *Tree Genetics & Genomes* **2011**, *7*, (3), 617-627.
502. Rothnagel, J.; Menschaert, G., Short Open Reading Frames and Their Encoded Peptides. *Proteomics* **2018**, *18*, (10), 1700035.
503. Hanada, K.; Higuchi-Takeuchi, M.; Okamoto, M.; Yoshizumi, T.; Shimizu, M.; Nakaminami, K.; Nishi, R.; Ohashi, C.; Iida, K.; Tanaka, M.; Horii, Y.; Kawashima, M.; Matsui, K.; Toyoda, T.; Shinozaki, K.; Seki, M.; Matsui, M., Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc Natl Acad Sci U S A* **2013**, *110*, (6), 2395-400.
504. Tran, N. H.; Zhang, X.; Xin, L.; Shan, B.; Li, M., De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences* **2017**, *114*, (31), 8247-8252.

VITA

Manuel Ivan Villalobos Solis was born in Tuxtla Gutierrez, Chiapas, Mexico. He holds a bachelor's degree in Biotechnology Engineering from the Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM) in Monterrey, Nuevo Leon, Mexico. He was first introduced to field of biological mass spectrometry while completing master's degree dissertation work under the tutelage of Dr. Mark Lefsrud at McGill University in Montreal, QC, Canada. Following the completion of his Master's degree in 2009, he enrolled in the graduate school of Genome Science and Technology (GST) at the University of Tennessee-Knoxville, where he completed PhD dissertation work focused on mass spectrometry-based proteomics at the Biological Mass Spectrometry & Laser Spectroscopy group at Oak Ridge National Laboratory under the supervision of Dr. Robert L. Hettich.