University of Tennessee, Knoxville
TRACE: Tennessee Research and Creative Exchange

Doctoral Dissertations | Graduate School

12-2022

Metaproteomics as a systems-biology approach to characterize the microbiome functionality and interactions in the human gut

Samantha L. Peters
University of Tennessee, Knoxville, speter29@vols.utk.edu

Follow this and additional works at: https://trace.tennessee.edu/utk_graddiss

Recommended Citation

Peters, Samantha L., "Metaproteomics as a systems-biology approach to characterize the microbiome functionality and interactions in the human gut. " PhD diss., University of Tennessee, 2022.
https://trace.tennessee.edu/utk_graddiss/7691

This Dissertation is brought to you for free and open access by the Graduate School at TRACE: Tennessee Research and Creative Exchange. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of TRACE: Tennessee Research and Creative Exchange. For more information, please contact trace@utk.edu.

To the Graduate Council:

I am submitting herewith a dissertation written by Samantha L. Peters entitled "Metaproteomics as a systems-biology approach to characterize the microbiome functionality and interactions in the human gut." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Robert L. Hettich, Major Professor

We have read this dissertation and recommend its acceptance:

Gladys Alexandre, Melissa Cregger, Jessy Labbé, Margaret Staton

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

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

Metaproteomics as a systems-biology approach to characterize the microbiome functionality and interactions in the human gut

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Samantha Peters

December 2022

Copyright © 2022 by Samantha Peters
All rights reserved.

ACKNOWLEDGMENTS

I would like to thank everyone who made the completion of this Ph.D. dissertation a reality. First and foremost, I would like to thank my research advisor, Robert Hettich for his unwavering guidance. He taught me to embrace science, especially when it is challenging, and that sometimes the long shots are the most rewarding, with the philosophy of "If you want to do cutting edge science, you should expect to find a few arrows in your back".

I would like to thank my mentors and co-workers at SomaLogic who taught me that "DNA isn't destiny" and introduced me to the wonderful world of proteomics. Their guidance and support ignited my passion to pursue interdisciplinary collaborative research and changed the trajectory of my scientific career. I would also like to thank the entire GST family under the fearless leadership of Dr. Albrecht von Arnim for providing the opportunities and environment to pursue genome-based research.

I would like to thank my committee members: Dr. Gladys Alexandre, Dr. Melissa Cregger, Dr. Jesse Labbé, and Dr. Margaret Staton for their effort and valuable mentorship on this dissertation. Their expertise in omics research and open minds helped me view my research from different perspectives and broaden the scope of the work.

I would also like to give thanks to the members of the Hettich lab past and present, for their guidance and comradery throughout this journey. In particular, I would thank Dr. Richard Giannone and Dr. Paul Abraham for their patience in teaching me the technical aspects of MS and for the brainstorming sessions that made much of this research possible.

I would like to thank the many collaborators I have gained along the way who were enthusiastic co-conspirators for the projects presented in this dissertation. They have taught me that good science takes time, but great science also takes teamwork.

Last, but not least, I would like to thank my family and friends who always encouraged me to bet on myself and to find my passion in life. Without their encouragement to embrace my research curiosity and their support through the ups and downs of my journey, I may never have made it down this path in life.

ABSTRACT

Microbial communities are composed of bacteria, archaea, microbial eukaryotes, and viruses. Organisms in these communities assist with critical functions across diverse environments. In host-associated microbiomes, such as the human gut microbiome, microbiota carry out activities that modulate the host immune system and provide metabolic benefits to the host. Due to their diverse and important roles in ecosystem processes, many questions exist about microbial community establishment, partitioning of function, and interactions between microbiota with each other and the surrounding environment. High-throughput meta-omics technologies are powerful methods to address the complexity of microbial communities and give unprecedented insights into the potential and active functions of microbiota within these communities. Among these meta-omics methods, metaproteomics can identify and quantify thousands of proteins from a single environmental sample and can be used to directly measure the active function of individual members of the community. This dissertation combines LC-MS/MS-based metaproteomics with other meta-omics approaches to study interkingdom community interactions and functions in environmental samples, with a focus on the human gut microbiome using tractable models. The projects presented here show (1) the application of critical parameters in LC-MS instrumentation and informatics approaches dictate the measurement depth and quality for complex microbiome samples, (2) interspecies competition and cooperation shape relative community composition and are the driving forces behind community utilization of fiber-derived glycans, (3) interactions with the host immune system and functional partitioning among community members facilitates establishment and persistence in the gut environment, and (4) uncultivated gut bacteriophages can use genetic codes different than their bacterial hosts as a regulatory mechanism during infection. In total, this dissertation makes a major step forward by showing that carefully designed metaproteomic measurements can explain the mechanisms of microbiome interactions and functionality.

TABLE OF CONTENTS

Chapter 1 Introduction to mass spectrometry-based metaproteomic characterization of interkingdom interactions in environmental samples. 1

1.1 The composition of microbial communities..... 1
1.2 Characterizing microbiome functionality using meta-omics technologies..... 3
1.3 Mass spectrometry-based metaproteomics among omics approaches..... 7
1.3.1 Bottom-up vs. middle-down vs. top-down proteomics. 8
1.3.2. Acquisition methods: DDA vs. DIA vs. targeted proteomics..... 11
1.4 Scope of the dissertation..... 13
Chapter 2 - Fundamentals of LC-MS/MS analyses for microbial communities.... 16

2.1 The general workflow for metaproteome measurements..... 16
2.2 Sample preparation. 16
2.2.1 Sample selection and collection..... 18
2.2.2 Cellular lysis and protein extraction..... 20
2.2.3 Protein denaturation, clean-up, and digestion. 22
2.3 Analytical measurements. 26
2.3.1 Liquid Chromatography. 26
2.3.1.1 One-dimensional liquid chromatography. 27
2.3.1.2 Multidimensional liquid chromatography. 29
2.3.1.3 Column loading..... 32
2.3.2 Mass Spectrometry. 33
2.3.2.1 Analytical figures of merit. 33
2.3.2.2 Ionization sources. 36
2.3.2.3 Mass analyzers and detectors. 39
2.3.2.4 Characteristics of MS/MS fragmentation. 42
2.4 Downstream analysis of MS/MS data in bottom-up proteomics..... 43
2.4.1 Peptide identification using database search approaches. 43
2.4.1.1 Constructing protein sequence databases for metaproteomics. .. 45
2.4.1.2 Evaluation of False Discovery Rates. 47
2.4.2 Peptide identification using *de novo* sequencing approaches..... 48
2.4.3 Post-search processing of metaproteomics datasets. 51
2.5 Summary..... 54
Chapter 3 - Technical considerations and optimization of sample preparation techniques for various complex environmental matrices. 55

3.1 Optimizing elements of sample preparation optimization for metaproteomes. 56

3.2 Extracting proteins from complex environmental matrices with limited microbial biomass. 57
3.2.1 Challenges with protein extraction from soil matrices. 57
3.2.2 Results. 61
3.2.2.1 Adaptation of the protein aggregation capture method to remove co-extracted humic substances from soil samples. 61
3.2.2.2 Optimizing conditions for pressure-assisted protein extraction. . 66
3.2.2.3 Evaluation of mechanical and chemical extraction methods in topsoil. 70
3.2.2.4 Optimizing and evaluating protein extraction in permafrost-derived soils. 74
3.2.3 Discussion. 76
3.2.4 Methods. 77
3.2.4.1 Soil selection. 77
3.2.4.2 PAC adaption. 77
3.2.4.3 Extractant compatibility. 78
3.2.4.4 Optimizing conditions for pressure-assisted extraction. 78
3.2.4.5 Comparison of mechanical extraction methods. 79
3.2.4.6 Spiking of Soil with Known Bacterial Isolates. 79
3.2.4.7 Protein extraction and digestion. 80
3.2.4.8 Peptide measurement by LC-MS/MS and data analysis. 80
3.3 Protein clean-up and quantification in the presence of co-extracted interfering molecules. 81
3.4 Separation and fractionation of complex peptide mixtures before MS. 84
Chapter 4 - Challenges of metaproteomic datasets and informatics approaches. . 91
4.1 Introduction to adapting proteomic approaches for metaproteomic datasets. 92
4.2 Evaluating the impact of database search strategies on peptide identification rates. 92
4.3 Identification of peptides from chimeric spectra. 97
4.4 Peptide quantitation and protein inference. 106
4.4.1 Match-between-run (MBR) for peptide quantification. 106
4.4.2 Evaluating false transfer rates associated with MBR. 110
4.4.3 Benchmarking the accuracy of common protein inference strategies for metaproteomic datasets. 111
4.5 Conclusions. 118
Chapter 5 - Interrogation of bacterial cooperation and competition through diet manipulation of human gut communities in gnotobiotic animals. 119

5.1 Introduction. **120**
5.2 Controlled investigation of microbial community dynamics in gnotobiotic animals. **121**
5.3 Characterizing the prevalence of protein post-translational modifications that form in the manufacturing of processed foods. **122**
5.3.1 Project scope. **122**
5.3.2 Results and discussion. **123**
5.4.3 Methods. **127**
5.4 Integration of metaproteomics with other techniques to assess community degradation of dietary fibers **129**
5.4.1 Project scope. **129**
5.4.2 Results and discussion. **130**
5.4.3 Methods. **140**
5.5 Microbiota Functional Activity Biosensors (MFABs). **141**
5.5.1 Project scope. **141**
5.5.2 Results and discussion. **143**
5.5.3 Methods. **146**
Chapter 6 - Probing the diversity of interkingdom interactions in the preterm gut. **147**
6.1 Introduction to the preterm gut environment. **147**
6.2 Multi-omics characterization of the establishment of the preterm infant gut microbiome in the context of eukaryotic membership **149**
6.2.1 Introduction. **149**
6.2.2.1 Metrics and trends of the metaproteomic measurements. *149*
6.2.2.2 Longitudinal metaproteomic characterization simultaneously reveals the presence and functions of bacteria and eukaryotes in the gut microbiomes of preterm infants. *151*
6.2.2.3 Proteomic analysis of one infant with evidence of Candida parapsilosis colonization. *157*
6.2.3 Methods. **161**
6.3 Understanding how the developing human preterm infant gut microbiome responds to antibiotic administration and host immune system-induced metal bactericidal control. **164**
6.3.1 Introduction. **164**
6.3.2 Results. **168**
6.3.2.1 Antibiotic resistance mechanisms help selected microbes overcome susceptibility to antibiotics and persist in the environment. *172*

6.3.2.2 Host sequestration of manganese and zinc does not impact microbes with enhanced import capabilities. 179
6.3.3 Discussion..... 186
6.3.4 Methods..... 193
6.3.4.1 Sample selection, preparation, and measurement. 193
6.3.4.2 Database searching and construction of protein datasets..... 193
6.3.4.3 Functional annotation of proteins. 194
6.3.4.4 Data analysis. 195
Chapter 7 - Validation that human microbiome phages use alternative genetic coding..... 196
7.1 Introduction..... 196
7.1.1 The role of bacteriophages in the environment..... 196
7.1.2 Uncultivated gut bacteriophages with alternative genetic code. 197
7.2 Results and Discussion..... 198
7.2.1 Selective enrichment of virus-like particles from human feces. 198
7.2.2 Confirmation of genetic code 15 usage by gut bacteriophage..... 204
7.3 Conclusions..... 211
7.4 Methods..... 212
7.4.1 Sample selection. 212
7.4.2 Genome predictions and phage genome curation. 212
7.4.3 Sample preparation for LC-MS/MS. 213
7.4.4 LC-MS-MS. 214
7.4.5 Proteomics data analysis. 215
Chapter 8 - Overview and perspectives on MS-based metaproteomics. 217
8.1 Conclusions..... 217
8.2 Remaining challenges and future outlook of metaproteomics..... 222
References 226
VITA..... 254

LIST OF TABLES

Table 3-1 Organisms used for soil spike-in experiments ……..…65

Table 4-1 Comparison of database workflows...…..…….…..…....100

Table 6-1 Fungal metabolic modules identified in GOMIXER analysis……155

LIST OF FIGURES

Figure 1-1 Multi-omics approaches for studying microbiomes enabling comprehensive characterization of microbial ecosystems. 5
Figure 1-2 Bottom-up vs. middle-down vs. top-down proteomic strategies. 9
Figure 2-1 General workflow for bottom-up LC-MS/MS proteomics. 17
Figure 2-2 Chromatographic peak resolution. 28
Figure 2-3 MudPIT LC column set-up. 31
Figure 2-4 Components of a mass spectrometer. All MS instrumentation contains components for ion generation, sorting, and detection. 34
Figure 2-5 Analytical figure of merit. 35
Figure 2-6 Schematic representation of the electrospray ionization (ESI) process. 37
Figure 2-7 Schematic of the Q Exactive Plus mass spectrometer. 41
Figure 2-8 Example of fragmentation ions generated from a peptide. 44
Figure 2-9 Experiment-wide false discovery rates (FDR) evaluation with a target-decoy strategy. 49
Figure 3-1 Aggregation of proteins and humic acids on magnetic bead microparticles. 63
Figure 3-2 Charge state distribution of precursor ions detected by LC-MS/MS. 67
Figure 3-3 Schematic illustration of the sample preparation workflow used to assess protein extraction methods for the analysis of soil metaproteomes. 71
Figure 3-4 The extracted volume of topsoil samples prior to protein clean-up and digestion. 72
Figure 3-5 Evaluation of LC-MS/MS results for optimized protein extraction in permafrost soils. 75
Figure 3-6 Impact of protein quantification method on LC-MS/MS peptide identification and quantification. 83
Figure 3-7 Comparison of 1D-RP-LC measurements of SIHUMIx mock community. 86
Figure 3-8 Functional profiles of identified Clostridium ramosum proteins. 87

Figure 3-9 Identified PSMs, peptides, and proteins in a human fecal sample using different LC separation approaches..... 89
Figure 4-1 PSMs, peptides analytes, and proteins identified from the SIHUMIx mixture by three different search algorithm and PSM validation tools..... 94
Figure 4-2 120-minute gradient unfractionated SDS_PAGE gel band identified proteins..... 96
Figure 4-3 Co-isolation and co-fragmentation of precursor ions in LC-MS/MS bottom-up proteomics experiments..... 98
Figure 4-4 Distribution of peptide sequence lengths (number of amino acids) of peptides identified in each database search workflow..... 101
Figure 4-5 Comparison of database search strategy identifications of a gnotobiotic mouse fecal sample..... 103
Figure 4-6 Community-level comparison of database search strategy identifications of a gnotobiotic mouse fecal sample..... 104
Figure 4-7 MS1 XICs of precursor ion m/z=906.41830..... 108
Figure 4-8 Percent of data matrix filled using MBR implementation. 109
Figure 4-9 *B. cellulosilyticus* peptides utilizing different implementations of MBR. 112
Figure 4-10 Hitchhiking proteins. 114
Figure 4-12 Theoretical tryptic peptidome of *Bacteroides* species in a gnotobiotic mouse model community..... 115
Figure 4-13 Relative summed organismal protein abundance of *Bacteroides* members in the samples lacking *B. cellulosilyticus* based on three quantitation methods..... 117
Figure 5-1 Mass spectrometry analysis of the whey protein isolate..... 124
Figure 5-2 β-lactoglobulin modified peptides..... 126
Figure 5-3 Proteomic and INSeq analyses of fecal samples collected on day 6..... 132
Figure 5-4 Deliberate Manipulation of Community Composition Demonstrates Interspecies Competition for Pea Fiber Arabinan..... 134

Figure 5-5 Detecting Acclimation to the Presence of a Potential Competitor Using Proteomics. 136
Figure 5-6 Alleviation of Competition between Arabinoxylan-Consuming *Bacteroides*. 138
Figure 5-7 The effects of supplementing the HiSF-LoFV control diet with unfractionated pea fiber, PFABN, or SBABN on PUL gene expression..... 144
Figure 6-1 Principal component analysis of all metaproteomic measurements..... 150
Figure 6-2 Taxonomic distribution of quantified proteins across all measurements for infant 06 (A) and infant 74 (B)..... 152
Figure 6-3 Tri-plot representation of short-chain fatty acid production..... 154
Figure 6-4 KO terms related to lipopolysaccharide biosynthesis in infant 74. 156
Figure 6-5 Organismal relative abundance based on summed protein abundances. 158
Figure 6-6 Global KEGG map of metabolism..... 160
Figure 6-7 Fecal samples collected for metaproteomic measurements for all infants in the study..... 169
Figure 6-8 Distribution of protein groups identified in each sample for each protein source (A) and the percentage of human proteins quantified in each Reactome annotation category (B). 170
Figure 6-9 Distribution of antibiotic resistance orthologs (AROs) found in each sample for all 17 infants in the study. 174
Figure 6-10 Relative abundance of organisms at each time point for samples collected for infant 61. 176
Figure 6-11 Relative abundance of organisms at each time point for samples collected for infant 39. 180
Figure 6-12 Functional β-diversity by proteins source for the 12 infants included in the analysis based on maternal antibiotic administration information..... 183
Figure 6-13 Metal homeostasis mechanisms of microbiota and corresponding host immune responses. 189
Figure 7-1 Genome curation and variation. 199

Figure 7-2 Workflow for phage enrichment from human feces. 202
Figure 7-3 Extracted-ion chromatograms of m/z 522.8003 ion. 203
Figure 7-4 Proteomic detection of alternatively coded proteins from two phage genomes. 205
Figure 7-5 Protein sequence coverage map of alternative code phage tail-related protein. 207
Figure 7-6 Example MS/MS spectra of alternative coding tryptic peptides. 209
Figure 7-7 MS/MS spectrum of de novo sequence tag of an alternative coded tryptic peptide (A). 210

Chapter 1 Introduction to mass spectrometry-based metaproteomic characterization of interkingdom interactions in environmental samples.

1.1 The composition of microbial communities.

Microorganisms constitute the largest population group on Earth, with studies predicting the existence of as many as one trillion microbial species[1]. These communities comprise a significant portion of the environment, including the host-associated environments such as the human body and rhizospheres, soils, and oceans. Microbial activities underlie environmental, industrial, and health-related processes including biogeochemical cycling and the homeostasis of health and disease[2]. In fact, they are considered essential for almost all biogeochemical cycling processes[3]. Bacteria and archaea contain most of the total nitrogen and phosphorus stored in living organisms, along with almost half of the stored carbon[4]. Therefore, microorganisms drive a variety of ecosystem processes including decomposition and the catalysis of important transformations in the carbon, nitrogen, phosphorus, and sulfur cycles. Despite this importance in ecosystem processes, fundamental questions exist about how these microbes assemble, interact, and function in communities. These questions remain in large part due to the fact that the majority of microorganisms cannot be isolated or cultured, with estimates that around 25% of microorganisms on Earth belong to phyla with no cultured relatives[5]. Consequently, the mechanistic study of microbial communities is challenging, and our current understanding of the roles of microorganisms in the environment is primarily based on observations from a minority of culturable organisms or from data generated from culture-independent studies[6]. Therefore, uncultured microbes which might dominate environments may have undiscovered functions and relationships with other community members that are important to the ecosystem homeostasis. Since the 1990s, advances in large-scale DNA sequencing technologies have greatly expanded the known diversity of

microbial communities and other complementary omics approaches have emerged to comprehensively study the composition, structure, and function of microbial communities in their natural habitat.

Microbial communities are composed of bacteria, archaea, microeukaryotes, and viruses. The term "microbiome" refers to the entire habitat, which includes the microorganisms, their genomes, higher order eukaryotes (for host-associated microbiomes), and the surrounding environment conditions including extracellular metabolites and bidirectionally coupled abiotic physicochemical processes (physical transport processes and abiotic chemical reactions)[7,8]. This definition is derived from the term "biome", referring to the biotic and abiotic factors of a given environment, which can be described as a reasonably well-defined habitat that has distinct bio-physio-chemical properties[9]. Microbiome research relies on systems biology approaches, where the complex biological system is studied through large-scale quantification of biomolecules including DNA, RNA, proteins, and metabolite. This research area has emerged as an interdisciplinary field, with ties to agriculture, bioeconomy, biotechnology, food science, mathematics, plant pathology and human medicine[9].

The impetus for systems biology-driven microbiome research is that we must be able to study biological life within biomes. In contrast to *in vitro* settings, which do not fully mimic actual environmental settings, we must be able to examine biological life in the context of the natural environmental complexity. This means the measurements used to examine microbial communities must be able to examine multiple kingdoms of life that compose the community, deal with abiotic interferences present in the sample, and the dynamic range of both population and function that exist within a given biome and may not be culturable in a laboratory setting.

Across all these disciplines, there are several driving science questions related to microbial about how they assemble function and interact with each other and their hosts. Among these are how does the assembly of communities occur and how is the assembly process different across environments and microbial communities of various complexity? Early research to determine the mechanisms of assembly primarily

focused on determining which taxa are the initial colonizers of a community and how community composition changes during the development of the community. However, as several large-scale efforts to characterize microbial community compositions across various environments, such as the human microbiome project (HMP)[10], the Earth Microbiome Project[11], and the TerraGenome project[12], given the sheer diversity and variation in community complexity, it is becoming clear that factors beyond taxonomy, such as structure and functionality, may be the driving forces of colonization and establishment.

Other research foci of interest include studying how communities age and persist in spite of environmental stress and if this persistence is enabled by niche ecologies. In addition, there is an ongoing interest to determine how organisms sense a complex environment and decide on an optimal response to that environment to dynamically regulate community behavior and function. In terms of community function and interaction with each other, there have been significant efforts over the past years to investigate bacterial quorum sensing, cooperation and competition between community members, protection from external perturbation, and how communities interact with their surrounding environment through processes such as biogeochemical cycling. For microbial communities that are associated with hosts, such as the mammalian gut and plant rhizosphere communities, significant interest is given towards determining host-microbe interactions, the synergistic balance between host and microbiota in terms of composition and functional activities, to determine how the composition and function of the microbial community ultimately impact host health and disease.

1.2 Characterizing microbiome functionality using meta-omics technologies.

Microbiome research is driven by advances in technology, and various omics technologies have emerged over the past few decades with an aim to measure each of the facets (DNA, RNA, proteins, metabolites), encompassed in the central dogma of molecular biology, which describes the flow of information encoded in DNA to

downstream expression products. Among the omics technologies available to study microbial communities, each technology has its benefits and limitations, and there is no perfect or universal method. Integrating multiple omics technologies to study any given microbiome results in a more complete view of the biological system as a whole and can reduce biases that arise from each technology. **Figure 1-1** outlines some of the primary omics technologies and their science drivers used in microbiome research.

To answer the question "Who is there?", marker gene amplification and sequencing, such as16S rRNA for bacteria and archaea and internal transcribed spacer (ITS) for fungi, is a commonly used method that can provide extensive profiles of microbial communities[13]. This is a fast, low-cost, and extensively used method used to obtain information about the composition of the community and can be amenable for low biomass samples. However, as with all technologies, it is prone to inherent bias that arises during amplification[14] and it is often restricted to genus-level resolution[13,14]. Culturomics, the high throughput isolation of microbes under a diverse set of culture conditions paired with identification by matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF-MS), has gained popularity as a culture-dependent approach to identify bacteria in a microbiome sample[15]. While not applied as often as culture-independent techniques, the complementarity between 16S rRNA sequencing and culturomics for the identification has been demonstrated, as a 2012 study found only a 15% overlap in bacterial identifications between the two techniques when applied to human stool samples[16]. In addition, while this approach is fairly labor and resource intensive, it can result in the cultivation of previously uncultured bacterial lineages for downstream mechanistic study[6].

To answer the question "What <u>can</u> the microbes do?", metagenomics, the method of sequencing all microbial genomes within a sample, can provide information about the functional genomic capacity or potential of the community. Compared to marker gene sequencing, it results in deeper taxonomic resolution and more detailed genomic information, as it captures all DNA present in the sample, including eukaryotic and viral DNA[13]. However, this method does not discriminate between

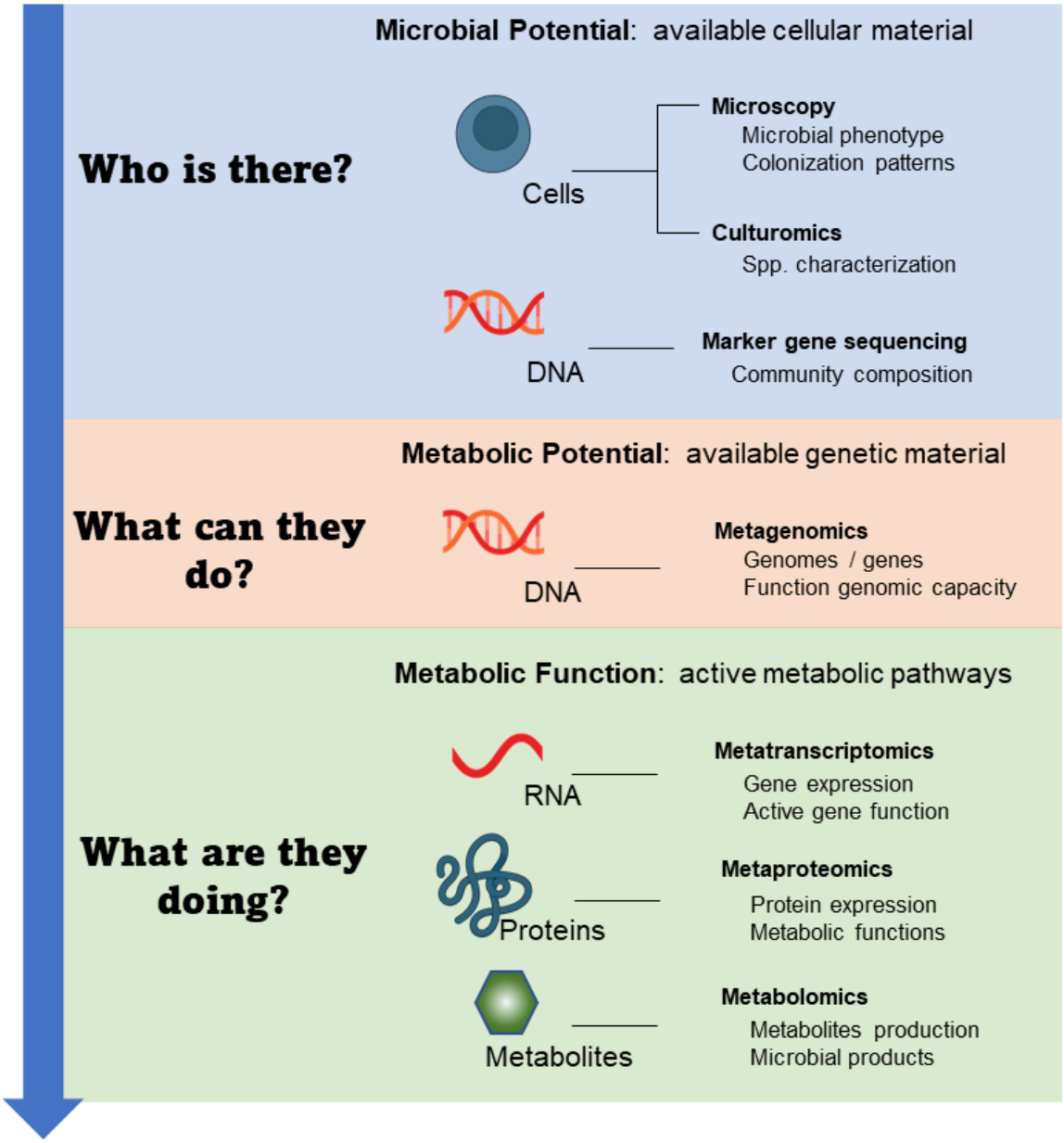

Figure 1-1 Multi-omics approaches for studying microbiomes enabling comprehensive characterization of microbial ecosystems. Microbiome research covers various levels of information (cells, DNA, RNA, protein, and metabolites). Each approach focuses on a different aspect of microbial ecosystems, from determining what microbiota exist in a given habitat (microbial potential) to determining the functional capacity of the community (metabolic potential) to the discovering the active metabolism taking place in the community (metabolic function).

DNA from live, dead, or active organisms and the analysis of the resulting data can be challenging and laborious depending on the complexity of the community.

To answer the question "What are the microbes doing?", three primary functional genomics approaches are used in microbiome research: metatranscriptomics, metaproteomics, and metabolomics. Metatranscriptomics uses RNA sequencing to profile transcription and can provide dynamic information about what genes are being actively transcribed. This method inherently discriminates between dead, dormant, or active microorganisms, but is biased toward organisms with higher transcription rates[13].

Metaproteomics can provide information about what proteins are being translated, and since proteins are major catalytic units in any ecosystem constituting roughly half of the dry mass of a cell[17], they provide a detailed look at the functional activities of cells in a microbiome sample in a spatial context. For example, this approach can discriminate between extracellular, membrane-bound, and cytosolic proteins and therefore functions. One benefit over sequencing-based approaches, metaproteomics can capture information about post-transcriptional regulation processes and post-translational modifications (PTM) of proteins[18,19]. While this approach is technically challenging in terms of sample preparation and does not provide the same depth of measurement as sequencing-based approaches, mass spectrometry has emerged as the predominant method to study protein expression in microbial communities. Metaproteomic analyses require a dynamic range of at least 5,000:1, mass accuracy of <3 parts-per-million (ppm), and mass resolution of 70,000 at m/z 200 to provide the necessary metaproteome depth and coverage. Due to the superior capabilities of MS instrumentation compared to other tools used for protein analysis, including Edman sequencing, gel electrophoresis, antibody-based approaches, protein microarrays, and NMR, this method has emerged over as the protein identification tool of choice for complex microbiome measurement.

Metabolomics can provide information about what metabolites are being produced and consumed in a community. For microbiome research, one primary limitation of metabolomics compared to metatranscriptomics and metaproteomics is

that this approach cannot link the measured entities (metabolites) back to the organism of origin without complementary gene information[20]. In total, to comprehensively study community composition, structure, and function, a combination of omics techniques should be utilized, since no one method captures all information necessary to interpret organism interactions with each other and their environment.

1.3 Mass spectrometry-based metaproteomics among omics approaches.

Metaproteomics was first described as "the large-scale characterization of the entire protein complement of environmental microbiota at a given point in time" by Paul Wilmes and Philip Bonds in an examination of prokaryotic microorganisms from a laboratory-scale activated sludge system that was optimized for enhanced biological phosphorus removal[21]. The operational goal of metaproteomics is to characterize the complete suite of proteins being expressed in a microbial community at any given point in time. In the first demonstration of metaproteomics utilizing 2D-PAGE gel electrophoresis paired with MS to examine protein expression from an activated sludge bacterial consortium, the initial goal of the study was to determine how many spots were observed on the gel compared to how many proteins could be identified. In this study, three proteins could be identified from a gel with over 700 spots, suggesting that the field had a long way to go to characterizing community function and that additional technological advancements, both in metaproteomics and metagenomics, including better metagenome assemblies for identifying the proteins were necessary [21]. In 2005, Ram et al. employed the first demonstrated combined genomics and mass spectrometry-based proteomics approach using multidimensional liquid chromatography to study microorganisms present in an acid mine drainage (AMD) biofilm community and identified over 2000 proteins from the five most abundant members in the community[22]. The application of mass spectrometry opened the doors for the development of the field and since this study, the field has continued to grow[23]. As MS instrumentation have advanced, the dynamic range of instruments is from 5,000-10,000 and mass accuracies are in the parts-per-per-million range. The

measurement depth and accuracy provided by instrumentation developments mean metaproteomics measurements can now confidently identify more than 50,000 non-redundant proteins from complex microbiome samples across a single experimental campaign[24].

In recent years, metaproteomics has been a critical component of microbiome research to advance the field beyond the analysis of the composition and functional potential of the community provided by DNA sequencing methods to better connect the community structure of the gut microbiome to its functional output[25]. Currently, metaproteomics studies have been conducted across a wide array of environments including human microbiomes[26], marine environments[27], soils[28], sediments[29,30], industrial wastewater treatment facilities[31], biogas reactors, and plant-associated microbiomes[32]. Several studies have integrated metaproteomics into multi-omics integrated analyses to provide key information about community function and dynamics that would not be elucidated with a single omics approach, such as the identification of keystone genes in a community[33], examination of the niche breadth of members in a mixed microbial community based on fluctuating environmental conditions[34], and defining interactions between bacteriophages, plasmids, and CRISPR-immunity[35].

1.3.1 Bottom-up vs. middle-down vs. top-down proteomics.

Measurement approaches in MS-based (meta)proteomics research can be conducted and classified as either "bottom-up", "middle-down", or "top-down" (**Figure 1-2**). Bottom-up proteomics is the typical method of choice for large-scale investigations of proteins present in a complex biological sample since the robustness and throughput of this method allow for the identification and quantification of thousands of proteins in a single proteomics experiment[36]. The bottom-up proteomics approach entails proteolytic digestion of proteins into short peptides, typically with sizes of 6-30 amino acids, that can be separated by liquid chromatography (LC) prior to tandem mass spectrometry (MS/MS). While this approach is fairly robust, it does not

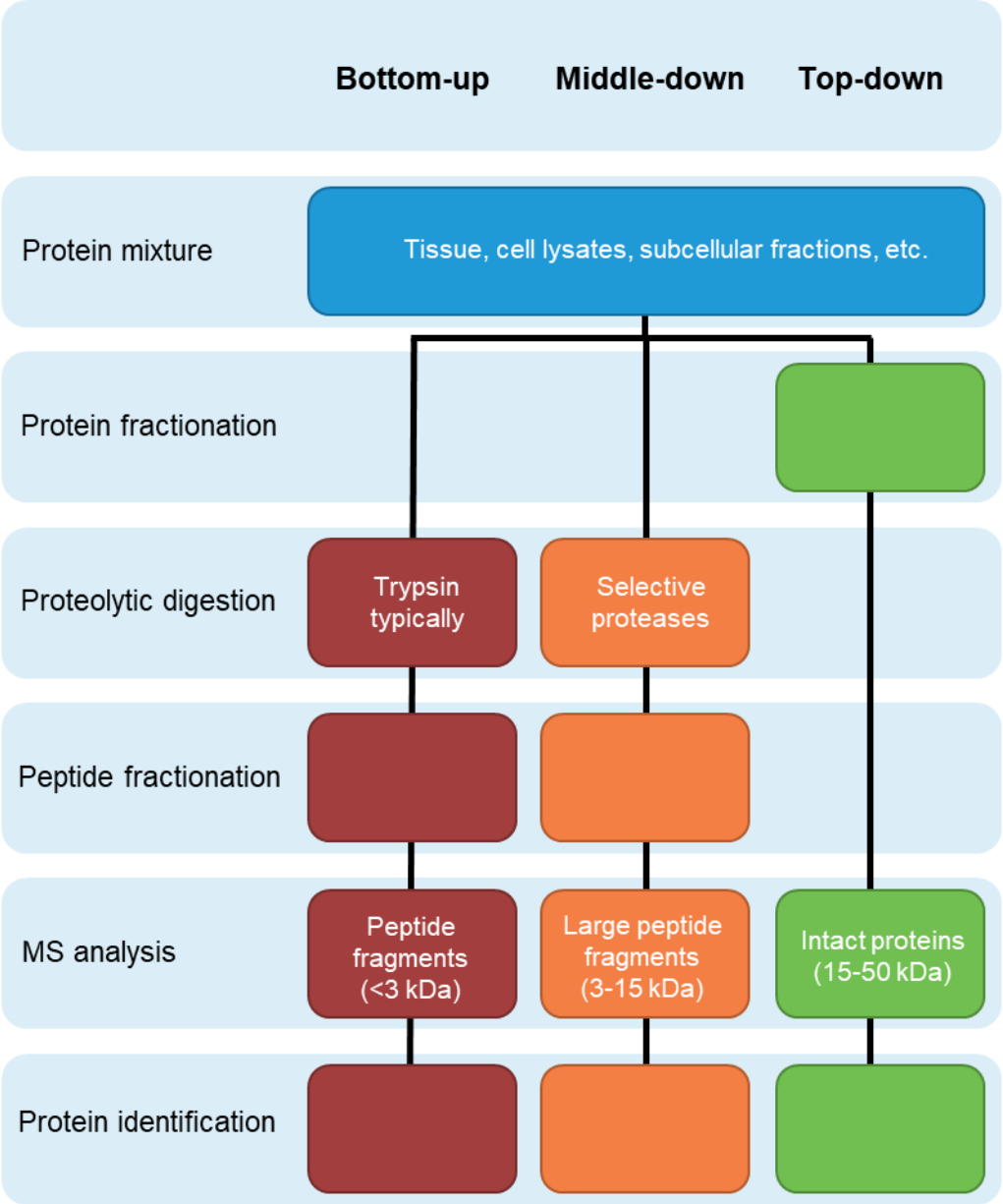

Figure 1-2 Bottom-up vs. middle-down vs. top-down proteomic strategies. The bottom-up and middle-down approaches measure proteolytic peptides digested from proteins that are separated or fractionated prior to MS analysis. The middle-down approach generates longer peptides compared to the bottom-up approach due to restricted proteolysis by selective proteases. Protein identification from both approaches is inferred based on peptide identifications. The top-down approach measures intact proteins which are separated or fractionated before MS analysis.

provide a comprehensive characterization of all peptides in a sample. The properties of enzymatically derived peptides, such as the size of the peptides and the location of specific residues in proteins, lead to both the primary benefits and shortcoming of this approach. For example, trypsin, the most used proteolytic enzyme used for protein digestion, can generate a large number of very short peptides which can be ionized and fragmented easily which is good for LC-MS/MS measurements. However, very short peptide fragments are hard to confidently identify and unambiguously assign to proteins, making protein identification challenging.

The top-down MS approach is an alternative strategy for protein analysis since large, intact proteins can be analyzed without prior proteolysis. The primary benefit of this approach is the ability to look at 100% sequence coverage for the characterization of all existing proteoforms of a gene that result in structural differences due to variations in the gene sequence at the DNA level, alternative splicing events at the RNA level, and any possible post-translational modifications (PTMs) at the protein level[36]. In addition, it is not subject to the same protein inference problem that hinders bottom-up proteomics approaches. However, while this approach seems the more straightforward compared to bottom-up proteomics, there are technical challenges associated with this approach for the uniform detection and identification of multiple proteins present in a biological sample. These challenges include but are not limited to, the difficulty of separating intact proteins in a mixture by LC, the challenge of ionizing intact proteins, the high instrument resolution and sensitivity required, and the difficulty of confidently matching proteoforms to the observed MS/MS fragmentation information[37].

A third approach, known as middle-down proteomics, has emerged as an alternative to bottom-up and top-down strategies, as it benefits from some of the advantages of each approach while minimizing some of their shortcomings. In contrast to bottom-up proteomics which is optimized for peptides sizes <3 kDa, or top-down proteomics which analyzes proteins or protein fragments 15-50 kDa in size, middle-down proteomics targets protein fragments up to 15 kDa in size[36]. In this approach, proteins are subjected to chemical or enzymatic proteolysis to generate

larger peptide sequences up to 150 amino acids in length[38]. While the complexity of the resulting peptide mixture in a biological sample is reduced compared to bottom-up approaches, the protein sequence coverage achievable increases due to the fragmentation of longer peptides. In addition, similar to top-down approaches, the probability of the presence of multiple PTMs or single-point mutations that arise from splicing variants is increased due to the creation of longer peptide fragments and this valuable PTM information of multiply modified peptides would not necessarily be observed in bottom-up proteomics. However, while several technological advancements have been made for top-down and middle-down approaches due to the advent of new instrumentation, due to the simplicity and robustness of bottom-up proteomics methods, this is still the dominant method used to characterize proteomes. The bottom-up approach was used for all metaproteomic measurements conducted in studies presented in this dissertation.

1.3.2. Acquisition methods: DDA vs. DIA vs. targeted proteomics.

One of the challenges in measuring very complex samples by LC-MS/MS is that thousands of peptide ions with different m/z values and varying abundances can be scanned and measured in a single sample. Even with fast, high-performance MS instrumentation, not all of these analytes can be fragmented and analyzed for peptide sequence information within a single measurement campaign. Therefore, for the analysis of complex samples, intelligent measurement approaches are needed to operate the mass spectrometer in a way that maximizes the efficiency of generating peptide sequence information for protein identification. There are three approaches to generating bottom-up proteomics data: data-dependent acquisition (DDA), data-independent acquisition (DIA), and targeted proteomics using parallel reaction monitoring (PRM)[39]. All three methods exhibit exquisite protein specificity compared to classical antibody-based methods such as western blotting and immunoprecipitation. Global, or shotgun, proteomics using DDA is the preferred approach of many researchers for discovery-based proteomics to provide unbiased and

comprehensive coverage of the proteome since *a priori* knowledge of the proteins present in a biological sample is unnecessary and all proteins can be interrogated in a single LC-MS/MS measurement.

For the DDA-based approach, the mass spectra of all precursor ion species that co-elute at a particular retention time in the LC elution gradient are recorded at the MS1 level. For each full scan, a predetermined number of the most abundant precursor ions, typically the top ten or top twenty most abundant ion species, are subjected to fragmentation for the acquisition of fragmentation (MS/MS or MS2) spectra used for peptide identification. This process is generally referred to as top-N DDA. As this approach preferentially fragments high abundance ions, lower abundance precursor ions will not be sampled unless repeated fragmentation of high abundance ions is prevented. Dynamic exclusion is an instrumental parameter utilized in DDA which temporarily excludes previously fragmented precursor ions from additional fragmentation events over a predetermined time window, thus increasing the detection of low abundance peptide analytes and overall proteome coverage[40]. However, even with instrumentation capabilities, such as dynamic exclusion, in very complex biological samples, such as metaproteomes, several low abundance peptides can co-elute with high abundance peptides and may not be detected unless the MS measurements are paired with enhanced LC peptide separation techniques. Despite this limitation, DDA is the most common data acquisition method used for bottom-up metaproteomics experiments. DIA and targeted proteomics have gained some popularity over recent years for metaproteomics as high-resolution instrumentation with fast scan speeds, such as Orbitrap and TIMS-TOF, have become available[27,41].

In contrast to DDA, where precursor ions are fragmented sequentially, in DIA-based methods, entire m/z ranges, typically spanning 25m/z units, of precursor ions are simultaneously fragmented[39]. Peptide fragmentation information is interpreted from the resulting multiplexed MS/MS spectra by comparing to known fragment spectra from very large spectral libraries or by generating pseudo-MS/MS spectra directly from the DIA data and conducting traditional database searches[42]. Spectral libraries are generated using in-depth DDA measurements of the same samples

measured in DIA, therefore only peptides previously discovered in the DDA measurements and incorporated into the spectral library can be identified by DIA. One of the primary pitfalls of the analysis of DIA data is that the spectral library might not represent the actual diversity of the measured samples[41]. In addition, DIA is currently limited to a dynamic range of 4-5 orders of magnitude[39], which is not sufficient to comprehensively characterize the entire protein content of a complex biological sample, as even the human plasma proteome has an estimated dynamic range of 9-13 orders of magnitude[43]. As this approach is still in its infancy for metaproteomics due to analytical limitations, it is still not widely adopted.

While both DDA and DIA approaches aim to comprehensively characterize the entire protein content in a biological sample, targeted bottom-up proteomics focused only on the measurement of a select set of peptides for predetermined and known proteins of interest. In this approach, proteotypic peptides are selectively isolated and fragmented over their chromatographic elution time. The targeted proteomics approach is significantly faster and more sensitive compared to discovery-based methods[44]. Despite these advantages over discovery-based methods to analyze key proteins of interest, in microbial communities with inherent functional redundancy and therefore many shared peptides, special consideration must be given to select peptides that can identify or quantify specific functions or taxon rather than specific proteins[44]. Therefore, unless there is *a priori* knowledge of what specific peptides may be present in the community based on a matched metagenome and if those peptides are distinguishable enough to answer specific research questions targeted approaches may provide limited value to metaproteomic studies. For all of the metaproteomic studies presented in this dissertation, DDA-based bottom-up proteomics approaches are used.

1.4 Scope of the dissertation.

This dissertation focuses on developing and exploiting mass spectrometry-based metaproteomics to study the functionality and dynamics of environmental

microbiomes, with an emphasis on the human gut environment. In particular, metaproteomic measurements of fecal sample were used to explore several questions related to interkingdom interactions that drive community structure and function. For example, this dissertation explored the reciprocal relationship between host-associated microbiota and the host immune system for the establishment and development of microbial communities and host health, the interactions of microbiota with each other to establish functional partitioning among community members and for the consumption and utilization of dietary resources, and examining the genomic underpinning of phage infections of their bacterial hosts. As LC-MS/MS-based proteomics is at the core of these metaproteomics investigations, Chapter 2 provides broad explanations of the experimental methods used for investigations covered in Chapters 3-7.

To approach fundamental questions about the functions and activities of microbial communities, a portion of the dissertation work focused on the development and optimizations in sample preparation, MS measurement metrics, and new bioinformatics approaches for complex microbiome samples. Chapter 3 highlights advances in sample preparation and LC-MS/MS measurements in a variety of different environmental matrices including soils, filters from the collection of ocean water, and laboratory-derived defined bacterial communities. Chapter 4 provides examples of bioinformatics advances for the application of bottom-up proteomics data analysis packages and software for metaproteomic datasets of varying complexity. The advancements in Chapters 3 and 4 were implemented in the research projects presented in Chapters 5-7.

Chapter 5 is an exploration of controlled human-derived gut communities in gnotobiotic animals and the interaction of microbiota with host dietary compounds in those systems. This chapter presents three projects where metaproteomics was integrated with other complementary techniques used to study the gut microbiome, including omics technologies such as various sequencing techniques and metabolomics, to provide biological insights into community degradation and utilization of dietary compounds. In addition, this chapter also includes an analytical

study utilizing the bacterial consortium from one of the gnotobiotic models presented in this chapter to help better define the limits of metaproteomic depth of measurement.

In contrast to Chapter 5, where gut microbiome functionality is evaluated using defined microbial communities, Chapters 6 and 7 explore inter-kingdom interactions using unmanipulated human gut microbiome samples. Chapter 6 presents two studies of the preterm infant gut environment, which is a tractable *in vivo* model to study colonization dynamics and host-microbe interactions. The first study explores early-life community colonization and establishment in the context of microeukaryotic membership. The second study is an in-depth longitudinal analysis looking at bacterial functions and persistence over the first several weeks of life in relation to host immune responses. Chapter 7 is a proteogenomic analysis of uncultivated bacteriophages in the human gut that utilize different genetic codes than their bacterial hosts during their infection process.

Finally, Chapter 8 presents a broad perspective of emerging technological advances in the field of metaproteomics that are under development. The chapter also describes how these advancements may apply to LC-MS/MS-based metaproteomics studies in coming years to look at microbial community function and how they will help shape microbiome research moving forward.

Chapter 2 - Fundamentals of LC-MS/MS analyses for microbial communities.

2.1 The general workflow for metaproteome measurements.

The MS-based proteomics approach is fundamentally a collection of methodologies involving sample preparation, measurement, and data analysis that each must be considered and optimized for addressing particular research questions as each methodology has particular strengths and weaknesses. Therefore, before embarking on any MS-based proteomics experiment, consideration should be given to the type of samples being measured, instrumentation capabilities, and analysis strategy best suited to the biological research questions being addressed in the study. All bottom-up proteomics experiments follow the same general workflow can be adapted at each step. The shotgun bottom-up proteomics approach via one- or two-dimensional liquid chromatography coupled with nano-electrospray tandem mass spectrometry was employed for all metaproteome experiments described in this dissertation **(Figure 2-1).** The overall workflow was designed to address two primary challenges that exist for metaproteomics measurement considerations in terms of sample preparation: (1) how can we get microbial proteomes out of complex environmental matrices, and (2) how do we make this method compatible with MS. This general workflow can be adapted for the analysis of proteins from a wide variety of samples, ranging from isolated bacterial cells to complex environmental matrices.

2.2 Sample preparation.

The first step in the metaproteomics workflow is the experimental design to appropriately address the posed science question(s), including the selection and collection of samples. This often involves careful consideration of a several key metrics. After collection, a series of steps in the workflow are conducted related to the liberation

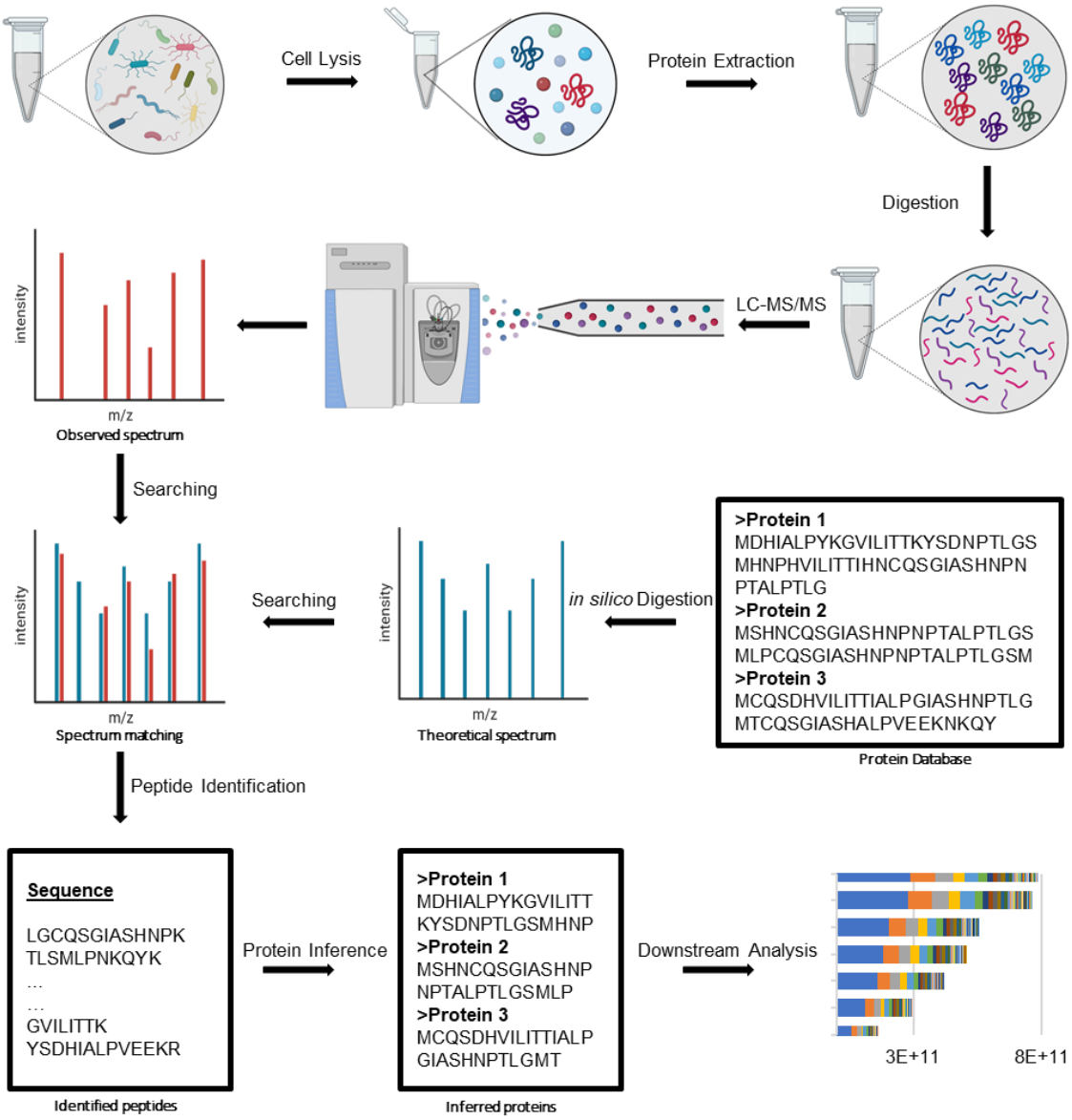

Figure 2-1 General workflow for bottom-up LC-MS/MS proteomics. A typical bottom-up proteomics experiment follows six basic steps. (1) Cells are lysed. (2) Proteins are extracted from the resulting lysate and matrix. (3) Extracted proteins are digested into peptides using proteolytic enzymes. (4) Digested peptides are analyzed by LC-MS/MS. (5) Collected MS/MS spectra matched against protein sequence databases to identify peptides. (6) Proteins are inferred from the identified peptides before downstream data analysis.

and collection of all of the proteins present in the sample, including a combination of homogenization and cellular lysis techniques to break up the sample matrix and release proteins from cells. Proteins are then extracted from the environmental matrix and subjected to clean-up procedures to remove non-proteinaceous material and prepare the samples for enzymatic digestion. Digested peptides are separated by high-performance liquid chromatography (HPLC) before MS analysis. In the mass spectrometer, the mass to charge (m/z) of peptide analytes are measured in a full mass spectrum (MS1). Peptides with the most abundant peaks from each MS1 spectrum are subjected to fragmentation in order to generate tandem mass spectra (MS/MS or MS2) used for amino acid sequence identification by various downstream informatic workflows. Identified peptides are scored for quality and assembled into proteins. Proteins are quantified based on the spectral counts or MS1 intensities of their corresponding peptides. Finally, the proteins can be functionally annotated and analyzed. Several variations of the general workflow have emerged for metaproteomics analyses, providing the experimental flexibility necessary for diverse environmental matrices. Since each of the steps in the workflow brings specific challenges and benefits, and can potentially influence the outcomes of the analyses, each needs to be evaluated when determining the specific approach necessary for an individual study. Specific details and considerations for each step in the workflow are discussed below.

2.2.1 Sample selection and collection.

For all research studies, the initial step of determining the appropriate experimental, methodological, and statistical design to generate meaningful data is critical for accurately addressing the specific scientific questions of the study. When designing a study, consideration must be given to the appropriate controls and metadata of the study to create study designs that isolate and interrogate the specific research variables of interest.

Cross-sectional studies can be used to find differences in microbiomes

between different study groups across a population (for example, between different sampling environments or between healthy and diseased individuals). However, as microbiomes vary greatly between individuals/site and can be influenced by various environmental factors, differences may arise due to confounding factors other than the research focus of the study, which may obscure patterns in the data if those confounders are not controlled and accounted for in the experimental design[13]. For example, clinical human microbiome research has inherent confounders such as the patient age, gender, diet, medications, and other lifestyle factors that may potentially mask differences between microbiota between study groups. Since poor study design and metadata collection can negatively influence experimental results, when selecting patients or samples to use for the study, implementing design approaches such as stratification-based designs can help resolve differences caused by confounding effects. Longitudinal studies can be designed with sampling frequencies selected that are appropriate to capture the temporal and developmental dynamics of the community as an alternative approach to cross-sectional studies that capture one moment in time. A key benefit of longitudinal analyses is the ability to capture the functional activities of core and transient microbiota in strongly dynamic systems [9].

Methodological and technical variations should be accounted for when selecting samples for the study. The extraction method used must be considered to determine whether or not it will be biased in the types of microbes or proteins extracted and whether or not there will be sufficient microbial yield from the amount of sample collected. Any potential interferences present in the environmental matrix that may negatively impact measurements must be considered. Appropriate controls, including methodological and reagent blanks, can be incorporated into the sample preparation process to evaluate experimental bias and variability. If replicates are necessary for the study, consideration should be given to what types of replicates are necessary (biological or technical) and constitutes a biological replicate for the system. As most microbial communities exist in highly dynamic systems with both biotic and abiotic factors, spatial and temporal characteristics must be considered when determining what constitutes a biological replicate. Finally, when collecting materials, one must

decide if paired experiments will be conducted using other omics technologies and whether there is enough available material that can be collected for a matched sample for all experiments. If paired experiments will be conducted appropriate sample handling techniques must be implemented for each technique. For example, the proteome is dynamic compared to the genome, and therefore sample integrity is based on rapid freezing to avoid proteome changes due to improper sample handling.

2.2.2 Cellular lysis and protein extraction.

Following sample collection, microbial cells are lysed and proteins are extracted from the environmental matrix using chemical or mechanical methods, or a combination of both methods. The choice of a particular lysis approach depends on the sample type, downstream methodology, and user preference. Chemical methods include lysis using detergents, such as sodium dodecyl sulfate (SDS) or Triton X-100, to destabilize cell membrane structure and break lipid-protein interactions. Mechanical disruption includes freeze/thaw treatments, sonication, bead-beating, or homogenization. As microbiome samples may contain different cell types (Gram-negative and Gram-positive bacteria, archaea, and fungi), each of which have different cell wall structures, and therefore differences in ease of lysis. In addition, environmental samples may have non-cellular abiotic components in the matrix which hinder lysis. Therefore, for large-scale metaproteomics, a combination of chemical and mechanical cell lysis techniques are often used to increase protein yields from high complexity microbiome samples composed of microbiota with different optimal lysis conditions[45,46].

For chemical lysis, the detergent sodium dodecyl sulfate (SDS) is widely used in protein extraction methods due to increased efficiency of this detergent for whole-cell protein solubilization. The choice of detergent is for cell lysis and protein extraction is critical, as it impacts downstream sample preparation procedures. Some detergents can interfere with proteolytic digestion, and can also interfere with LC-MS/MS measurements by clogging the LC columns and damaging the mass

spectrometry instrumentation, or suppression ion signal[47]. Since SDS can be effectively removed from the sample prior to proteolytic digestion, this detergent is frequently selected for bottom-up proteomics experiments. Several proteomics publications have demonstrated that including SDS in a protein extraction protocol increases the peptide identification rates from a wide range of proteins, including peptides that are found in membrane-associated proteins[48], even when applied to a variety of microbial systems[49–51].

For mechanical-based cell disruption, a key factor in determining the method applied is the number of samples in the experiment, the amount of microbial biomass in the sample and the presence of abiotic components in the samples. For complex environmental microbiome samples in which there is a limited amount of material, sonication procedures are effective, but this method is prone to cross-contamination or sample loss associated with direct contact of sonication probes with the samples. In addition, for some sample types that contain a lot of large debris, such as soils, to use sonication for cell lysis, additional pre-lysis procedures are necessary to avoid damaging the sonication probe. Bead-beating is commonly employed in metaproteomics experiments, and the composition, quantity, and size of beads used in the process is dictated by the origin of the sample and the amount of biomass. This approach has been demonstrated to increase protein yields of Gram-positive bacteria and some yeast species without impacting protein recoveries from Gram-negative bacteria compared to other extraction methods[52,53].

In this dissertation, chemical extraction via SDS was combined with mechanical disruption via bead-beating for most of the metaproteome studies presented. SDS was paired with homogenization or sonication of the soil and ocean metaproteome samples in Chapter 2 and the uncultivated gut bacteriophage fecal samples presented in Chapter 7.

2.2.3 Protein denaturation, clean-up, and digestion.

For successful downstream protein digestion, protein denaturation procedures are often performed concurrently with the cell lysis and protein extraction process as many of the reagents and procedures are the same. Protein denaturation and reduction steps are often included in workflows to generate and maintain protein structures that which are amenable to enzymatic digestion by exposing the proteolytic cleavage sites throughout the protein. In addition to physical protein denaturation processes such as heat treatment or sonication, chemical additives are often included in the lysis buffers that help with protein solubilization and prevent protein renaturation. Three classes of chemical additives are common in bottom-up proteomics: (1) chaotropes, (2) detergents, and (3) reducing agents. Chaotropes, such as urea prevent protein aggregation by disrupting intra- and inter-molecular hydrophobic interactions and hydrogen bonds[54]. Detergents, such as Triton-X and SDS, are used to increase the solubility of proteins by preventing hydrophobic interactions [47,54]. Reducing agents, such as dithiothreitol (DTT) reduce disulfide links and are paired with alkylating agents such as iodoacetamide (IAA). IAA is added to the samples following disulfide reduction to protect newly produced free sulfhydryl groups by alkylation[54]. For all work presented in this study, proteins were solubilized using heat treatment with SDS followed by reduction and alkylation by DTT and IAA.

Many of the reagents used during the cell lysis, protein extraction and denaturation process are not compatible with ESI-MS experiments. For example, the presence of incompatible buffers, detergents or salts can interfere with analyte ionization and can introduce adducts into the mass spectrometer that will overwhelm the signal of the peptides[55]. Therefore, following protein denaturation and alkylation, various strategies can be implemented to clean up the extracted proteins to remove these incompatible substances prior to proteolytic digestion and LC-MS/MS measurements. These strategies include, but are not limited to, trichloroacetic acid (TCA) precipitation, filter assisted sample preparation (FASP), chloroform-methanol exchange (CME), and bead-based approaches such as protein aggregation capture

(PAC). For the TCA precipitation method, when TCA is added to the cell lysates, proteins are precipitated out of solution and the resulting protein pellet is washed by acetone to effectively remove SDS, salts, and nucleic acids from the sample[54]. This method has also been shown to help remove other interfering materials present in the sample matric including humic substances in soils[56]. Chloroform-methanol precipitation was first described to isolate and purify lipids[57], but has since been used in preparation methods for the purification of proteins. In essence, methanol, chloroform, and water are added in a 4:1:3 volume ratio and are mixed with the cell lysate solution. Proteins are separated at the water-organic interface from any unwanted material present in the sample, including detergents like SDS. More recently, bead-based methods, such as the PAC method, have emerged for protein preparation. These methods exploit the inherent instability of denatured and precipitated proteins to nonspecifically immobilize on microbeads regardless of the bead surface chemistry[58,59]. This approach can remove MS-incompatible additives such as SDS and is not as limited by the amount of proteinaceous biomass in the sample as other clean-up procedures such as CME. In this dissertation, both the CME method and the PAC method were used for projects presented in Chapters 3-7.

Following protein extraction and clean-up, proteins are subjected to proteolysis to generate peptides which are used for LC-MS/MS analysis. For most bottom-up analyses, specific peptide bonds are cleaved using a proteolytic enzyme (protease) that is suitable for the particular research questions being addressed. Each protease has a distinct specificity to cleave certain amide bonds in the protein primary structure, therefore producing a very specific pool of under controlled digestion conditions[54]. The choice of protease dictates the characteristics of the proteolytic peptides produced, therefore influencing the ability to infer the proteins present in the sample, along with any post-translational modifications of those proteins. The primary goals of protease selection are (1) to produce proteolytic peptides that will be observed via mass spectrometry-based on their biophysical properties such as amino acid length and hydrophobicity, (2) to provide adequate protein sequence coverage, and (3)

generate a sufficient amount of unique peptides in order to infer specific proteins and a large portion of the proteome[60].

Some factors to consider when choosing one protease over another for protein digestion are the size of the peptides needed for analysis, the selectivity of the protease, and the type of proteins being analyzed. Generally, smaller peptides sizes and molecular weights that are amenable to mass spectrometric analysis are preferred. However, the peptides should be long enough in sequence to still be able to confidently identify the amino acid sequence from the MS/MS spectrum. Optimal peptide size for bottom-up proteomics analysis is a peptide length of 10-15 amino acids and a molecular weight of 5-30kDa. Peptide size can also influence the confidence of PTM identifications. For example, if the goal is localization of PTMs on specific residues, longer peptide sequences are sufficient for protein identification, but shorter peptide sequences reduce ambiguity among possible PTM localization sites are there are a limited number of possible modification sites. The activity parameters of each protease, including the pH, denaturant tolerance, and temperature, should also be considered. These activity parameters are important because how the protein is prepared prior to and during digestion (buffers used, etc.) has an impact on which protease can used. If upstream sample processing steps cannot be adapted for compatibility with a protease of interest, an alternative protease must be selected. For example, the proteins of interest for the study are membrane proteins, using enzymes that target hydrophobic residues, such as chymotrypsin, pepsin, or thermolysin, is generally the best option. For metaproteomics research, selecting a protease which can generate enough uniquely identifiable is especially important as peptides cannot only be ambiguous between proteins within a single microbial species but also across several species. Therefore, the selected protease should provide high enough coverage to discriminate between organisms with high proteome redundancy.

Trypsin is a serine protease which is highly specific and cuts on the C-terminal side of lysine or arginine residues, but it will miss if proline is on the C-terminal side of either of those residues. This is the most widely used protease in bottom-up proteomics because it has several advantages. First, this enzyme cleaves only after

lysine or arginine residues so it generates peptide sizes of 800-2000 Da, since these particular amino acids occur in proteins approximately every 10-12 residues in the protein sequence[54]. Second, tryptic peptides are easily ionizable and detectable by ESI-MS instrumentation. In the gas phase, the charge is localized the on the C-terminus which favors consistent fragmentation by collision induced dissociation (CID) and higher energy collisional dissociation (HCD) fragmentation methods[61,62]. The fragmentation spectra of tryptic peptides are typically very rich, which means a protein can often be confidently identified with the information produced by only one tryptic peptide[54].

In some cases, an alternative protease to trypsin should be selected depending on the research questions in the study to generate more critical peptides for the identification of specific types of proteins, sequence variants, protein regions, or PTMs of interest. For example, if the research is exclusively focused on membrane proteins, trypsin might not be the best choice of protease. Since there are not as many lysine or arginine residues within transmembrane segments of proteins, trypsin would not the most efficient protease for proteolytic digestion, as it would not be able to cleave in as many locations and would produce longer peptide fragments that are not as amenable to MS interrogation. In this case, a nonspecific protease such pepsin may be the better option even though it has less specificity than trypsin. Pepsin will cleave at phenylalanine, leucine, and glutamic acid residues (but not at valine, alanine, or glycine). It is also tolerant of low pH and low temperatures. Therefore, it will generate a large number of peptides in transmembrane regions of the protein where there are few lysine or arginine residues.

In Chapter 5, one research study is presented where a combination of proteolytic enzymes was used to interrogate hexose PTMs on lysine and arginine residues. In this case, proteins were first subjected to trypsin digestion. Since hexose modifications on lysine or arginine residues block tryptic cleavage, a large portion of the proteins remained undigested or under-digested. To increase protein sequence coverage and provide a complementary set of peptides that contained hexose-modified lysine and arginine residues, these proteins were further digested with chymotrypsin.

All other work presented in Chapters 3-7 of this dissertation exclusively relied on trypsin as the protease of choice.

2.3 Analytical measurements.

Once proteolytic peptides have been generated, it is advantageous to separate them by liquid chromatography prior to entry into the mass spectrometer for MS/MS interrogation. The combination of these two techniques (LC and MS) provides a powerful analytical measurement platform to fractionate, identify, and quantify unknown and known molecules, like peptides and metabolites. The combination also enables elucidation of the chemical and structural properties of those molecules that would not necessarily be possible with one analytical technique alone. Appropriate selection of the types of liquid chromatography and mass spectrometry used for the analysis depends on the type of data required to answer a particular research question. While not an exhaustive review of these analytical techniques, the following section provides details about each of these analytical techniques that are relevant to the metaproteomics research presented in Chapters 3-7.

2.3.1 Liquid Chromatography.

Liquid chromatography (LC) is a widely used analytical technique that is often paired with mass spectrometry, and is used to separate compounds in a sample, such as proteolytic peptides, prior to MS analysis. Peptide separation is important for both peptide identification and quantification for bottom-up proteomics with DDA acquisition. The primary goal of liquid chromatography for bottom-up proteomics is to separate the peptide mixture into a series of chromatographic peaks. Enhanced separation strategies prior to MS interrogation have several benefits for peptide identification, including the reduction of ion suppression and co-eluted peptides[63,64].

Fundamentally, LC-based separation of sample analytes is based on the interaction of the analytes with the LC mobile and stationary phases. Typically, the

extent of separation is related to each analyte's competitive affinity for the LC mobile phase vs. the stationary phase. In theory, with sufficient separation, each peak constitutes a single peptide analyte from the mixture and the resolution between two peaks is a quantitative measure of their separation. Three factors that influence chromatographic resolution are efficiency, retention, and selectivity. Of these three factors, selectivity has the largest impact on resolution and can be defined as a measure of the ability of the LC system to distinguish between eluted analytes. It can be visualized as the distance between two chromatographic peaks. For example, **Figure 2-2** shows examples of what two chromatographic peaks would look like if the separation approach provides poor, moderate, or high resolution. To increase chromatographic resolution, the selectivity can be altered by changing LC conditions related to stationary phase and mobile phase composition. Details about each type of phase and applications for separation of complex peptide mixtures are described below.

2.3.1.1 One-dimensional liquid chromatography.

The most basic design of a LC-MS proteomics workflow separates peptides by a single chromatographic dimension. Analyte interactions with the stationary phase in an analytical column are influenced by the hydrophobicity, polarity, and nature of the stationary phase. Since analyte interactions directly affect the selectivity of the chromatographic separation, these parameters should be considered when selecting an appropriate stationary phase for the separation of peptides. Three other details to consider in determining the degree of separation for the column is the alkyl length of the hydrocarbon attached to the stationary phase silica particles, the solvent choice used with the stationary, and the slope or separation time needed[65].

Reversed phase (RP) is the most common stationary phase used in LC-MS proteomics experiments for the separation of analytes. This separation approach is based on the hydrophobic interactions between the stationary phase and peptides. The

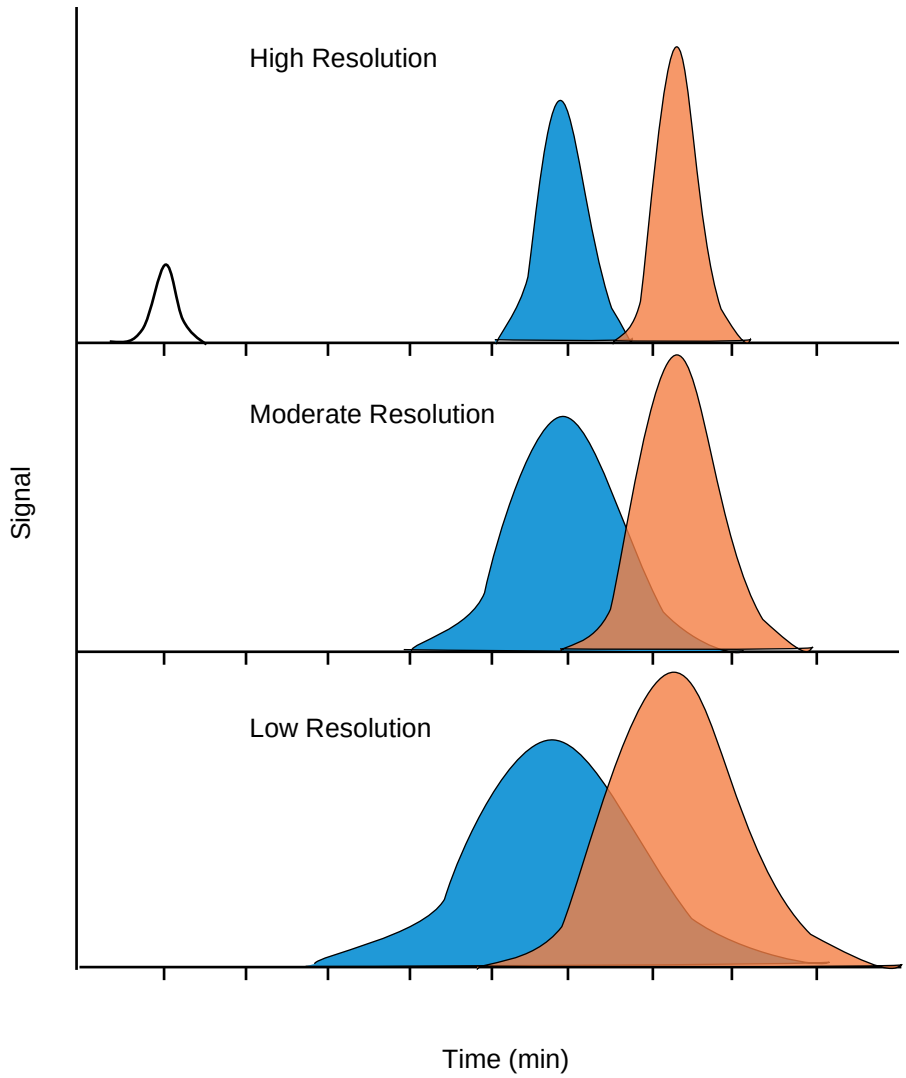

Figure 2-2 Chromatographic peak resolution. The depicted examples show the relationship between the separation and the resolution of two analytes in a mixture. The orange and blue peaks each represent the elution profiles of two peptide analytes. The white peak represents the chromatographic void volume, where unretained analytes are eluted at the beginning of the separation gradient.

bound peptides can then be efficiently concentrated on the column and desalted for better MS performance. Over a user-defined separation period, the organic solvent concentration of the elution buffer is gradually shifted from low organic conditions to low aqueous conditions, and peptides will elute off the column according to the strength of the hydrophobic interactions with the RP resin[66]. The most common RP stationary phase is C_{18} covalently bound to a base silica material, although other alkyl chain lengths are available.

Significant of research has been conducted with the aim of enhancing the quality or efficiency of separation. Separation efficiency can be calculated by the peak capacity of the column, which is the maximum number of analytes that can be separated in a given separation period. Peak capacity is determined by the column length and the particle size of the stationary phase. Therefore, for one-dimensional approaches, increasing peak resolution is often achieved by using longer column lengths and decreasing particle sizes to increase peak capacity or by increasing the separation time. In addition, decreasing column diameter to the micrometer scale to operate in a nanoflow regime (nano-LC) improves sensitivity and ionization efficiency and also is beneficial for samples with limited biomass as it requires less material for column loading compared to larger diameter columns[67].

2.3.1.2 Multidimensional liquid chromatography.

For samples with increasing biological complexity, one-dimensional (1D) separation strategies do not always provide the chromatographic resolution necessary for the analysis of protein mixtures which prompted the development of multidimensional liquid chromatography for proteomics[68]. Multidimensional liquid chromatography approaches combine two or more forms of chromatography for orthogonal separation of peptide mixtures. This combination of LC approaches increases peak capacity, and therefore resolving power, of the separations in order to better fractionate the eluting peptides prior to MS interrogation[68]. Compared to 1D approaches, multidimensional separations provide better resolution of peptides that

differ in charge and hydrophobicity, leading to three primary benefits: (1) minimize ion suppression, (2) improve ionization efficiency, and (3) minimizes under-sampling by simplifying the complexity of peptides entering the MS at a particular point in time[68]. These benefits are especially important for complex metaproteomic samples when DDA acquisition is used since the higher resolution and peak capacity improve data acquisition in a top-N strategy. Improved data acquisition results in a larger dynamic range of the measurement and a better representation of the peptides present in the sample. However, these gains in resolution and dynamic range come at the cost of significantly longer LC run times compared to 1D separation strategies.

There are several multidimensional separation strategies, where separation is achieved using columns that combine RP with other stationary phases such as cation (strong cation exchange (SCX), weak cation exchange (WCX)) or anion exchange columns (strong anion exchange (SAX), weak anion exchange (WAX)) or by RP separations at low or high pH, in order to achieve deeper and orthogonal separation[66,69]. Among the most popular separation approach for complex mixtures is a two-dimensional (2D) separation strategy known as Multidimensional Protein Identification Technology (MudPIT)[70]. The approach exploits two stationary phases, RP and SCX, for orthogonal separation of peptides by both hydrophobicity and charge (**Figure 2-3**). It can be a fully automated, online two-dimensional separation approach where samples are loaded onto back and front columns for in-line clean-up and separation prior to MS/MS to minimize sample loss.

There are three major steps of the approach: (1) sample loading, (2) salt cuts, and (3) LC-MS/MS analysis. In the first step, the sample is loaded onto the RP segment of the back column for washing and desalting with solvent A (low organic content). Then solvent B (high organic content) is flowed over the column and the peptides are eluted from the RP onto the SCX segment of the column for the second step of the process. In the second step, peptides are separated by charge/ionic strength. Step elution of SCX is conducted by flowing increasing amounts of ammonium acetate across the column to displace fractions of peptides from the SCX onto the second RP segment of the back column. The fraction of peptides that has been pushed on to the

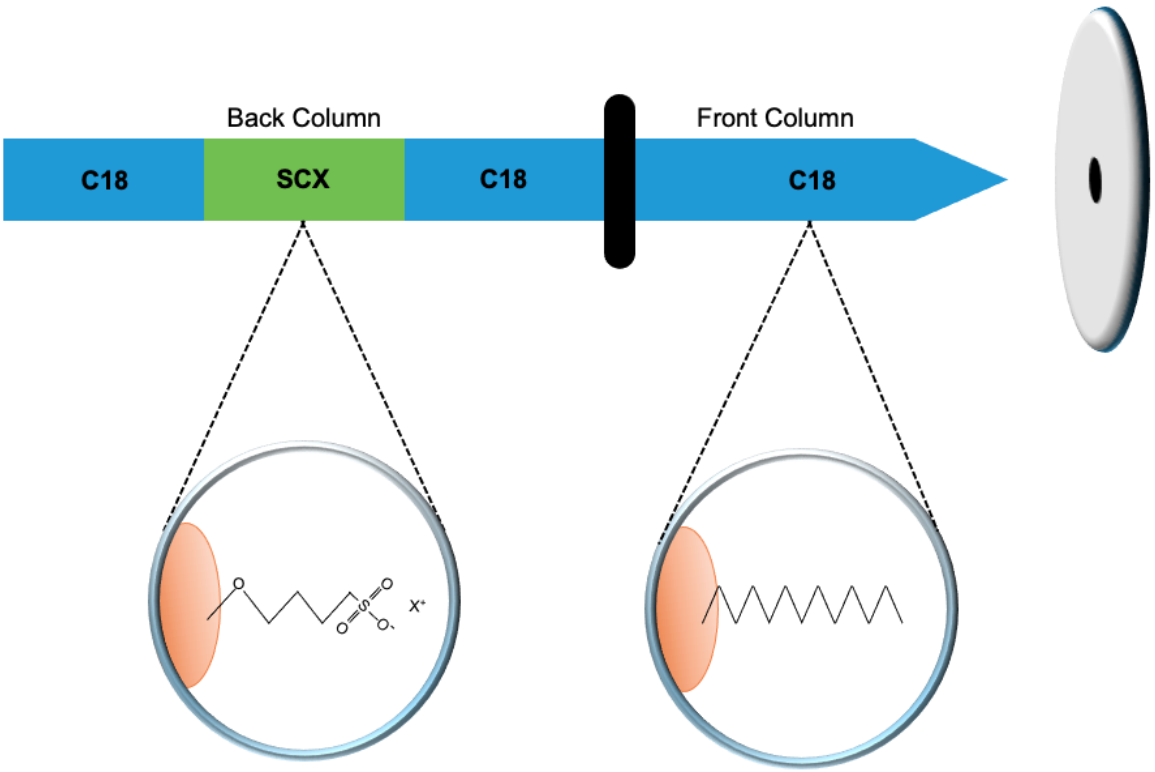

Figure 2-3 MudPIT LC column set-up. A triphasic back column is packed with C-18 resin and SCX resin for in-line clean-up desalting and orthogonal separation. This back-column is paired with a C-18 RP packed analytical column. Following chromatographic separations, peptides elute of the analytical column and are ionized using ESI before introduction into the mass spectrometer.

RP segment is ready for the third step in the approach. In the final step of the process, peptides are separated based on hydrophobicity. In this step, a gradient elution profile is set up where increasing concentrations of solvent B is applied to the column to separate peptides over the front column. Eluting peptides are then directed into the ESI source for ionization and subsequent measurement by MS/MS.

One alternative to the MudPIT approach using RP and SCX to separate peptide mixtures, another two-dimensional strategy exploits the separation of peptides based on pH in combination with RP. Changes in pH of the peptide solution is used to protonate or deprotonate the peptides. Separation of the peptides is based on the protonation level. At low pH, the peptides retain their protons because there are enough protons in the solution and the peptide will be retained on the RP column. At higher pH, the peptides will be deprotonated and will no longer be retained on the RP column.

2.3.1.3 Column loading.

Regardless of the liquid chromatography approached used to measure peptide mixtures, column loading is an important consideration. If insufficient amounts of peptides are loaded onto the column, this can result in ion signals that are below the mass spectrometer's limit of detection. This leads to a reduction in the number of peptide and protein identifications. It also hampers peak area quantification as the low signal-to-noise ratio makes peak integration difficult[71]. Column overloading, where excess amounts of peptides are added to the resin, has also been demonstrated to decrease peptide and protein identification rates [72]. Column overloading results in peak tailing, retention time shifts, high ionization competition, and poorer peptide separation [71]. It can also lead to column clogging resulting in sample loss or increased variability in the sample measurements.

Work presented in Chapter 3 will explore how aspects of liquid chromatography can impact peptide identification for mixtures of varying complexity.

In particular, the impact of different stationary phase separation strategies and column loading will be optimized for LC-MS/MS workflow of metaproteomics samples.

2.3.2 Mass Spectrometry.

In principle, mass spectrometers measure the mass-to-charge ratio (m/z) of gas-phase ions. The basic design of all mass spectrometers includes three primary components: (1) an ionization source, (2) a mass analyzer, and (3) an ion detector **(Figure 2-4).** After sample introduction to the mass spectrometer, the ion source generates gas-phase ions which are separated by a mass analyzer and detected by the ion detector. As there are numerous combinations of MS components and therefore several ways to ionize, analyze, and detect peptides, there is no single MS instrument configuration that is superior to all others. However, there are generally a few types of MS instrumentation that dominate proteomics research due to the intrinsic capabilities of these instruments. Details about each of these mass spectrometer components that are relevant for the work presented throughout the dissertation are discussed in this section.

2.3.2.1 Analytical figures of merit.

To determine the appropriate MS instrumentation required for a research project, it is useful to compare MS components based on some analytical figures of merit. There are several figures of merit that are used for evaluating MS instrumentation, some of which are depicted in **Figure 2-5**. These include mass resolution/resolving power (the ability to differentiate between two adjacent peaks in a mass spectrum, which can be given as the full width at half-height of a single well-resolved peak), mass accuracy (the difference between the measured mass to the calculated mass and stated as the error in terms of parts-per-million (ppm) or a percentage), mass range (the range of molecular masses, from smallest to largest, that can be measured), dynamic range (the measure of the detection range that can be

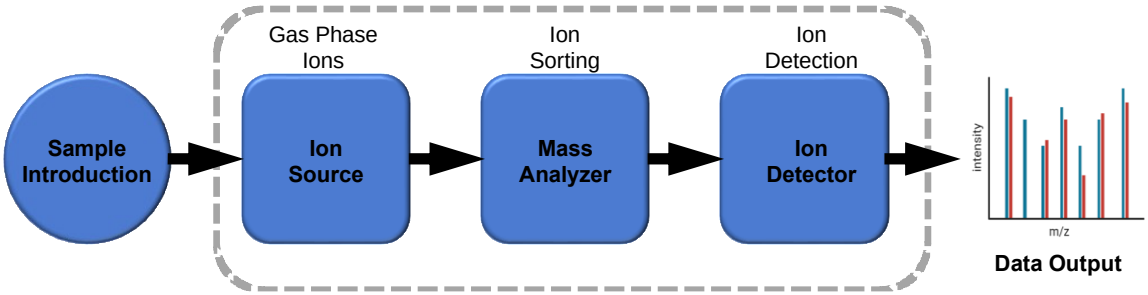

Figure 2-4 Components of a mass spectrometer. All MS instrumentation contains components for ion generation, sorting, and detection.

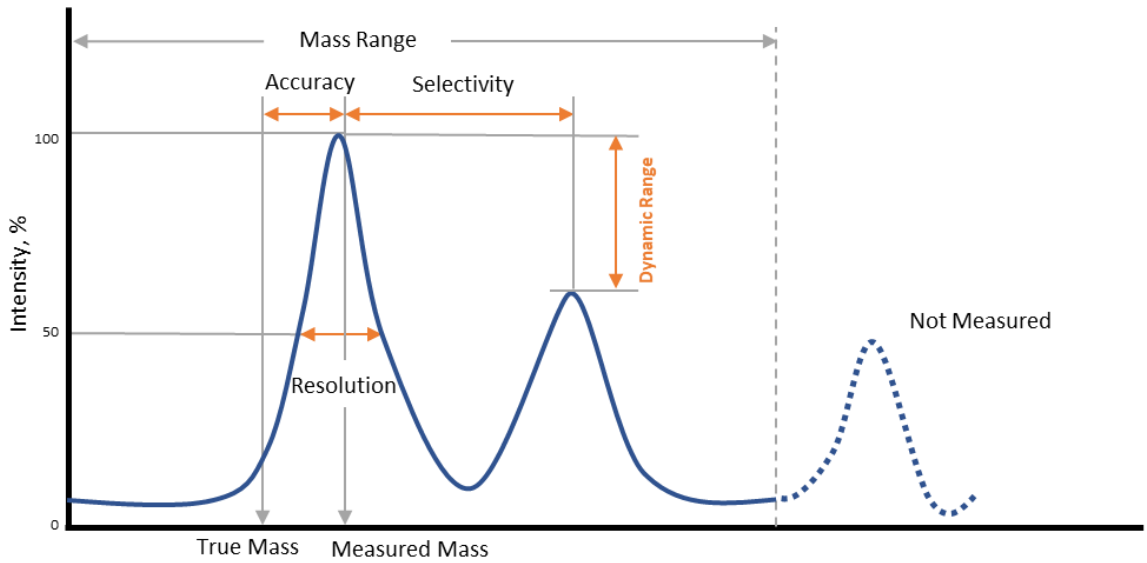

Figure 2-5 Analytical figure of merit. Figure of merit relevant to the selection of MS instrumentation.

detected in a single sample and is calculated as the ratio between the largest and smallest detectable signal), the duty cycle (which is the fraction of time that an instrument is collecting data), the detection limits (defined as the smallest amount of sample that can be detected with a signal to noise ratio of 3:1), and the scan speed (which is the number of spectra that can be collected in a given unit of time). Relevant figures of merit for metaproteomics measurements will be described in this chapter.

2.3.2.2 Ionization sources.

There are several ionization methods used in mass spectrometry, including electron ionization (EI), chemical ionization (CI), Atmospheric Pressure Chemical Ionization (APCI), Electrospray ionization (ESI), and Matrix-Assisted Laser Desorption/Ionization (MALDI). The primary factor for selecting an ionization technique is the type of analyte molecules being measured. Two of the most common methods, ESI [73] and MALDI[74]are soft ionization methods that are applicable for biological samples as they are able to convert larger molecular weight compounds, like peptides and proteins into gas-phase ions. Soft-ionization techniques are preferred for proteomics because hardly any internal energy is transferred to the ions, which minimizes in-source fragmentation[75]. As all work presented in this dissertation utilizes ESI, this method will be described in further detail.

In order to transfer a charged peptide analyte from solution to the gas phase via ESI, there are three basic steps: (1) the production of charged droplets from the high-voltage capillary (emitter) tip when the analyte solution is introduced, (2) solvent evaporation and disintegration of the droplets into very small and highly charged droplets, and (3) the formation of gas-phase ions[76] **(Figure 2-6).** Protonation or deprotonation of peptides is achieved through solvent additives. For example, to generate positively charged peptide ions, acids such as trifluoroacetic acid (TFA) or formic acid (FA) are often added to peptide solvent to lower the pH of the mobile phase, which protonates the acidic side chains of the peptides. To produce charged droplets in positive ion mode, a high-voltage electric field is applied at the emitter tip

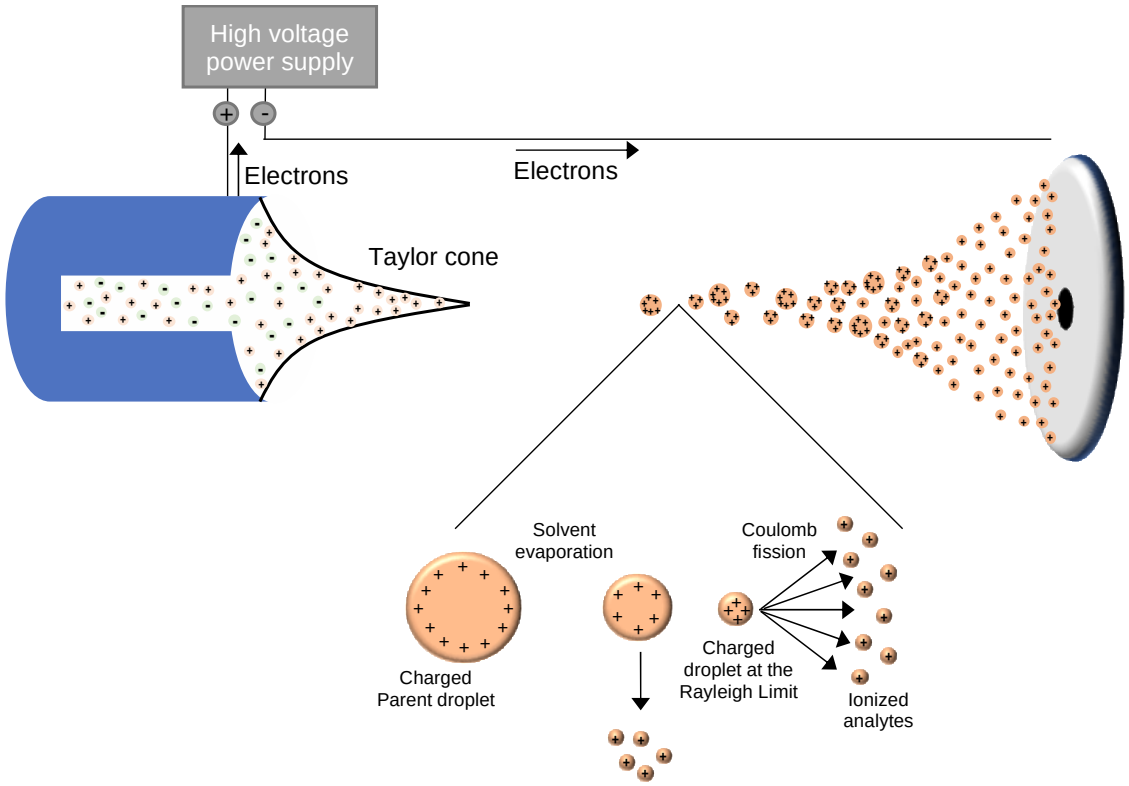

Figure 2-6 Schematic representation of the electrospray ionization (ESI) process.

from the LC set-up, which causes polarization of the solvent near the meniscus of the solution that contains the peptides or other analytes of interest. This leads to an enrichment of positive ions at the meniscus near the tip surface and negative ions away from the meniscus[76], leading to the formation of the "Taylor cone" where distortion of the meniscus creates a cone-shaped spray that points towards a counter electrode on the mass spectrometer. As charged droplets get expelled from the emitter, they travel towards the mass spectrometer and the droplet solvent evaporates. As the solvent of the charged parent droplets evaporates, the droplets disintegrate into smaller offspring droplets

At this point, the gas-phase ions are formed from the highly charged droplets by one of two proposed mechanisms. Researchers debate this concept because there are opposing theories on the exact mechanism by which the final gas-phase ions are produced. In the charge residual model, the charged droplets would only be able to get so small before the surface density of charges would become so high that some of the surface ions would be ejected and pushed directly into the gas phase. In a second mechanism, the highly-charged droplets undergo asymmetric fission to form smaller droplets. In this process, as enough solvent evaporates, the charged droplets reach an unstable state as they reach the Rayleigh limit, where the surface tension is equal to the repulsion forces of the charges[77], and the droplets can no longer sustain the Coulomb force of repulsion. At this point, the droplets undergo Coulomb fission, where a lot of small droplets that are highly charged will break from the larger parent droplet. As the droplets break apart, they don't break apart symmetrically. Each droplet typically loses 2% of the total mass and 15% of the total charge of the parent droplet after each fission event[78]. This fission process continues until single ions, typically multiply charged, are emitted from the small droplets after all of the solvent has evaporated from the droplet. Some of these desolvated ions then enter the enter the mass analyzer.

One major drawback of ESI is that it susceptible to ion suppression. In the presence of nonvolatile solutes, such as ammonium sulfate, changes the efficiency of droplet formation and evaporation, which inhibits formation of gas-phase ions[79]. In

addition, in high concentration samples, the efficiency of ion formation is also decreased by competition between compounds for limited charges or space on the droplet surface[55]. To overcome or minimize ion suppression during ESI, several strategies can be implemented into the proteomics workflow. These include protein clean-up/purifications steps like solid-phase extraction (SPE) or protein precipitation to remove salts from the sample, changing chromatographic conditions by shifting the retention times of eluting analytes away from regions affected by ion suppression or by changing mobile phase additives[75]. Another way to decrease ion suppression is by reducing the amount of sample being introduced to the ion source, through minimizing the volume of sample injected, diluting the sample, or reducing the ESI flow rates. ESI flow rates can be substantially reduced by using nano-electrospray ionization (nano-ESI), which decreases the number of analytes and nonvolatile compounds being introduced into the source and enhances the droplet desolvation process[80,81]. The advent of nano-ESI had greatly improved the sensitivity of ESI, which traditionally has been lower than MALDI, and is the ionization method of choice for peptide analysis of samples with limited amount of material available, which can be common in metaproteomics. Nano-ESI was used for all measurements presented in this dissertation.

2.3.2.3 Mass analyzers and detectors.

After ions have been transferred into the gas phase by ESI and introduced into the mass spectrometer, the mass analyzer separates ions by their m/z before they reach the ion detector where the number of ions that are resolved by the mass analyzer are recorded. All mass analyzers selectively separate ions using either electric or magnetic fields that are static or dynamic. A mass analyzer can be used as a stand-alone system or can be combined into tandem MS (MS/MS) systems depending on research needs[82]. For proteomics workflows that use ESI for ion generation, four types of mass analyzers are generally employed: (1) ion traps (including quadrupole ion traps, linear ion traps, and Orbitraps), (2) quadrupoles, (3) time-of-flight (TOF), and (4) Fourier-

transform ion cyclotron resonance (FTICR) mass analyzers[83]. Hybrid instruments, which have been designed to combine the analytical characteristics and capabilities of different mass analyzers are used in bottom-up MS/MS proteomics for peptide sequence identification.

Among the most common mass analyzers used currently for bottom-up proteomics is the Orbitrap. Data acquisition in an Orbitrap mass analyzer is based on Fourier transformations of image currents of trapped ions[82]. The Orbitrap mass analyzer operates as an electrostatic trap that is composed of three electrodes. It contains two cup-shaped outer electrodes that are facing each other and a central spindle-shaped electrode. When a voltage is applied between the electrodes, populations of ions are trapped and orbit around a central. Since the electric field is linear along the axis, the axial oscillations are maintained with a frequency that is characteristic of their *m/z* values[84]. This ion movement generates an image current in the outer electrodes that is then Fourier transformed into the time domain to produce mass spectra[83].

All research presented in this dissertation was conducted using a Q-Exactive Plus mass spectrometer (Thermo Scientific), which is a hybrid quadrupole-Orbitrap instrument[85]. Inside this mass spectrometer there are five basic components: (1) radio frequency (RF) lens, (2) Advanced Active Beam Guide (AABG), (3) Quadrupole Mass Filter, (4) Orbitrap mass analyzer, and (5) HCD collision cell (**Figure 2-7**). Briefly, under DDA top-N conditions, after ions are generated by ESI and introduced into the MS via an ion transfer tube which is heated to enhance ion solvation, they pass through the RF lens, also known as a shrig or s-lens, which is a stacked-ring RF ion guide. The RF lens focuses the ions into a tight beam, which increases sensitivity. Termed the "advanced axial beam guide" or AABG, ions then pass over an injection flatapole and axial bent flatapole to reduce noise by preventing high-velocity ion clusters and uncharged neutral ion species from entering the quadrupole. After passing through the bent flatapole, ions enter the quadrupole for mass filtering. The quadrupole transmits and filters ions according to their m/z values before they are collected into packets in the C-trap (a "C"-shaped RF-based quadrupole, where

Figure 2-7 Schematic of the Q Exactive Plus mass spectrometer. This MS was used for all proteomic measurements conducted in Chapters 3-7. Image taken from Thermo Scientific product documentation.

ions are cooled using nitrogen gas) and stabilized before being sent to the Orbitrap for MS1 detection[86]. For MS/MS data collection, MS1 precursor ions are passed to the HCD collision cell (higher-energy RF-only collision octapoles) for fragmentation. These fragment ions are then collected into packets in the C-trap and sent tangentially to the Orbitrap for detection. The process is repeated until all the most abundant precursor ions eluting at that time, based upon the top-N user selection, are fragmented and MS/MS spectra are collected.

The Q-Exactive Plus features several characteristics that make it particularly amenable to bottom-up metaproteomic analyses. It includes performance specifications that are necessary for measurement of complex microbiome samples, which were described briefly in Chapter 1. These include a mass range of 50-6,000 m/z, which can sufficiently capture proteolytic peptides and attomole-femtomole sensitivity. It features a mass resolving power up to 140,000 at m/z 200, a dynamic range >5,000:1, and a ~2 ppm mass accuracy. Finally, this instrument is very fast with scan speeds up to 12 Hz and duty cycle times of one second for a top10 method[85]. Therefore, combined with enhanced LC separation strategies, this instrumentation provides the depth, sensitivity, and accuracy needed to sufficiently measure complex microbiome samples.

2.3.2.4 Characteristics of MS/MS fragmentation.

The type and quality of MS/MS spectra generated is based in part by the fragmentation method used. Each MS instrument can perform one or more types of fragmentation type, which is selected based on instrument capability and application. The two most common types of fragmentation in bottom-up proteomics are collisional induced dissociation (CID) and (2) higher energy collisional dissociation (HCD)[62]. Less common types of peptide fragmentation include electron-capture dissociation (ECD) and electron-transfer dissociation (ETD), which are used for specialized proteomics applications like PTM localization and glycopeptide analysis studies[82]. Fragment ions are characterized by where bonds are broken along the peptide

backbone (**Figure 2-8A**). CID and HCD fragmentation types produce series of b- and y-fragment ions, while ETD and ECD produce c- and z-fragment ion structures[61,82,83]. B- and y-ions, which differ based on charge location (**Figure 2-8B**), are useful for determining peptide amino acid sequences. Peptide sequence can be determined manually by associating the m/z difference between fragmentation ions with the masses of individual amino acids. **Figure 2-8C** shows a MS/MS spectrum where the complete b- and y-ion series was observed. While MS/MS spectra can be manually interpreted to identify peptides in a sample, several computational algorithms have been developed for fast, reliable, and automated identification of peptide analytes. These approaches are discussed below.

2.4 Downstream analysis of MS/MS data in bottom-up proteomics.

In this dissertation, two methods were used to analyze collected MS/MS spectra for the identification and quantification of peptides. MS/MS spectra were interrogated by database search approaches for all work presented in Chapters 3-6. In Chapter 7, database searches were complemented by *de novo* peptide sequencing. Both strategies will be described in the following sections, along with specifics regarding the implementation of each approach and methods to analyze the resulting peptide information.

2.4.1 Peptide identification using database search approaches.

In bottom-up proteomics workflows, after MS/MS spectra are collected for the peptides present in the sample, those MS/MS spectra are interrogated to determine the amino acid sequences of the identified peptides. This peptide identification is most commonly performed via database search engines where the acquired experimental MS/MS spectra are matched against theoretical spectra derived from a constructed protein database. These matches, known as peptide spectrum matches (PSMs), are then scored for each spectrum, with the highest scoring PSM matching the theoretical

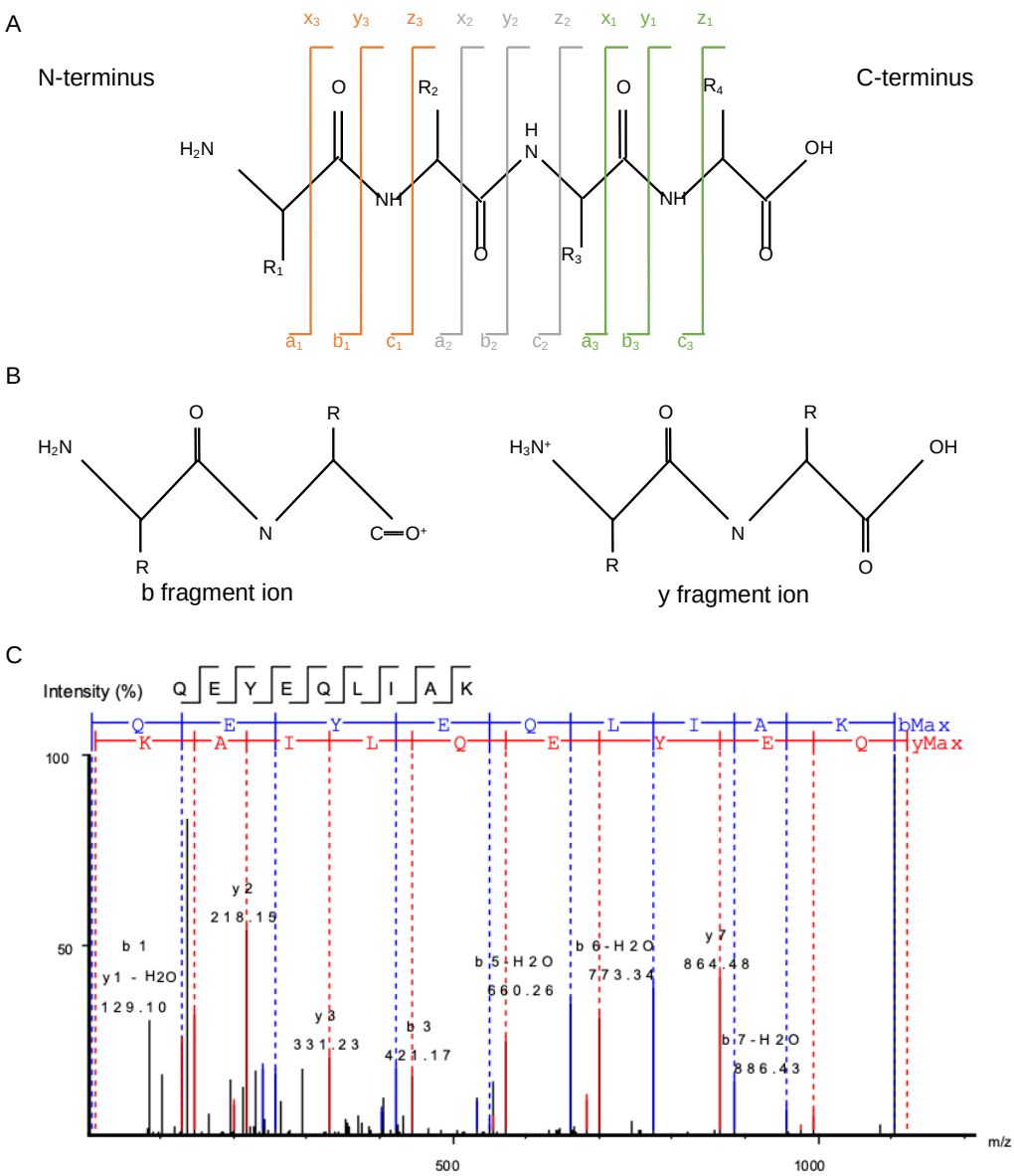

Figure 2-8 Example of fragmentation ions generated from a peptide. (A) An illustration of peptide backbone containing four amino acid residues (R_1, R_2 R_3, R_4) and the types of fragment ions generated in a MS/MS spectrum. Fragmentation ions are classified as a, b, or c ions if the charge is retained on the N-terminal side, a, b, or c. If the charge is retained on the C-terminal side of the peptide, the fragment ion is classified as either is x, y, or z. HCD fragmentation generates b and y ions. (B) The structures of a b and y fragmentation ions showing charge location. (C) An example MS/MS spectrum where the peptide amino acid sequence is determined based on the difference in mass between b or y ions.

spectrum most closely[87].

2.4.1.1 Constructing protein sequence databases for metaproteomics.

One of the most critical steps for a successful database search is the careful construction of a protein sequence database used to generate theoretical spectra for peptide spectrum matching. Ideally, the constructed database should contain any protein sequence that is expected to be present in the measured peptide mixture, including proteins from common mass spectrometry-based sample preparation contaminants[88]. The comprehensiveness and size of the database are two key metrics that influence the number of PSMs that can be identified using a particular database[89]. Database searching will only identify PSMs from the acquired MS/MS spectra if the proteins for those PSMs are included in the constructed database. Therefore, to maximize PSM identifications, it is essential to make sure the database is comprehensive as possible to ensure that all proteins that are present in the sample are represented in the database. While larger database sizes increase the number of identified PSMs, if the database contains protein sequences that are not expected to be present in the sample, this has the undesirable consequence of increasing the chance of high-scoring false positive identifications. In a similar fashion, a database that is not comprehensive and does not contain protein sequences that are present in the sample, can also lead to false identifications. If the database contains a sequence that generates a very similar theoretical spectrum with a similar precursor mass and shared b- and y-ions, this can result in a peptide spectrum match that would have a lower score if the real sequence was represented in the database[89]. In addition, the expanded search space associated with very large databases results in extended computational time for database search algorithms.

The database searching approach was initially developed to analyze proteomes of individual organisms. For single proteomes, the construction of a protein database is fairly straightforward, and in an ideal setting, the constructed database is derived from the organism's previously sequenced genome. Metaproteomics presents unique

challenges related to protein database construction that accurately captures the full protein content of organisms in a highly diverse microbial community. Protein database construction for metaproteomics should include protein information from the genomes of every organism present in the microbial community, as well as proteins from the host or dietary components if applicable and known. With the exception of defined artificial communities, such as the ones described in Chapter 5, fully comprehensive and representative databases cannot be constructed without DNA sequencing of the community.

There are four sources of sequence information that used to construct protein databases for metaproteomics: (1) matched metagenomes, (2) unmatched metagenomes, (3) unrestricted reference databases, and (4) restricted reference databases[89]. Matched metagenomes, where the same sample is used to generate both the metagenomics and metaproteomics datasets, provide the most accurate list of proteins that may be present in the sample. However, matched metagenome sequence information is the most expensive among the four sources in terms of the monetary cost of sequencing the samples, the time and computational resources, and the expertise to assemble the metagenomes. For unmatched metagenomes, sequence information is derived from metagenomic sequencing data from the same environmental system, but not the specific samples analyzed by metaproteomics. This includes metagenomics data included in gene catalogs for a particular system such the mouse or human gut[10,90,91]. Due to the high level of taxonomic variation in microbiome samples even within the same system[92], these databases may contain millions of protein sequences, so this source of sequencing information can potentially include large numbers of protein sequences that are not actually present in the samples. While this source of sequencing information has no monetary cost, it presents computational challenges related to the extremely large search space. Unrestricted reference databases, such as the NCBI RefSeq database, which contain all sequences that have been deposited in a sequence repository, cover a large sequencing search space. As this search space is not specific to the sample, or even to the system, suffers from a large number of false hits and therefore low PSM identifications after controlling for

a desired false discovery rate (FDR). In addition, public sequencing repositories many not have adequate representation for all members of the community[5,93]. The restricted reference database approach is based on prior knowledge about the taxonomic composition of the community, typically through 16S rRNA gene sequencing or other information, to build pseudo-metagenomes using publicly available reference genomes for taxonomically relevant organisms[94].

Finally, complex metaproteomics datasets often contain more mass spectra in a single study compared to studies of single organisms and require larger search spaces based on the large database size. Not all database search algorithms can handle the time or memory resources needed to process this data type. This computational limitation has prompted the development of metaproteomics-centric database search algorithms capable of processing datasets with massive search space and are compatible with operating on high-performance computing (HPC) cluster[95,96].

2.4.1.2 Evaluation of False Discovery Rates.

One downside of peptide-spectrum-matching is that it can sometimes lead to false positive identifications, especially with larger datasets that are common in metaproteomics[97]. Factors such as erroneous sequences in the database and low-quality spectra can contribute to false positive identifications. To avoid drawing incorrect biological conclusions from the data based on the inclusion of false positive identifications, statistical procedures are often implemented to limit the false discovery proportion (FDP). In essence, the FDP is the proportion of false positive identifications, or mismatches, among all PSMs identified in the dataset. It is impossible to precisely calculate the FDP due to the random nature of mismatches during peptide spectrum matching[97]. However, statistical procedures, such as false discovery rate (FDR) calculations can be used to estimate the FDP of a dataset.

To date, FDR is the most accepted global confidence measure to assess the error rate across a bottom-up proteomics dataset in order to estimate the PSM, peptide, and protein identification confidence [87]. The target-decoy method is the most common

approach used to determine the FDR in bottom-up proteomics[98] In this case, a decoy database is appended to the target database with the assumption that the number of false PSMs in a decoy search is equivalent to the number of false PSMs in a target search **(Figure 2-9A).** The decoy database is composed a reversed (or shuffled) version of the target database. Reversed decoy databases are commonly used over shuffled because the *in-silico* digest of a reversed database will have the same frequency and lengths of peptides which gives the same number of peptides being used in the decoy database as the target database. The reason this is a good model is because it is simple to implement, and it gives an equal chance of true and false PSMs being hit. Since matches to the decoy database (decoy hits) are false identifications, the FDR of the experiment can be calculated as FDR = (2*(# of decoy hits))/ (# of target hits)[98].

User-defined scoring thresholds can be set during the database searches based on the desired false discovery rate for the experiment **(Figure 2-9B).** Typically, thresholds are set at the PSM, peptide, and/or protein level. For example, all work in this dissertation implemented a peptide-level FDR threshold of 1%, meaning no more than 1% of the peptides in the entire dataset could be assumed to be false identifications. One consideration for this strategy is that it relies on the assumption that the size of the target database is representative of the true protein content in the sample. If the size of the target database is very large and contains a number of protein sequences that are not expected to be in the sample, the number of hits to the corresponding decoy database would increase compared to an appropriately sized database. To maintain the desired experimental FDR for an oversized database, a stricter scoring threshold would need to be set, and thus more real or true PSMs would be filtered out of the dataset[89]. Therefore, increasing the accuracy of the PSMs identification comes at the cost of sensitivity.

2.4.2 Peptide identification using *de novo* sequencing approaches.

While the database search strategy is the primary method to identify peptides and proteins in bottom-up proteomics workflows, the amino acid sequence of a peptide

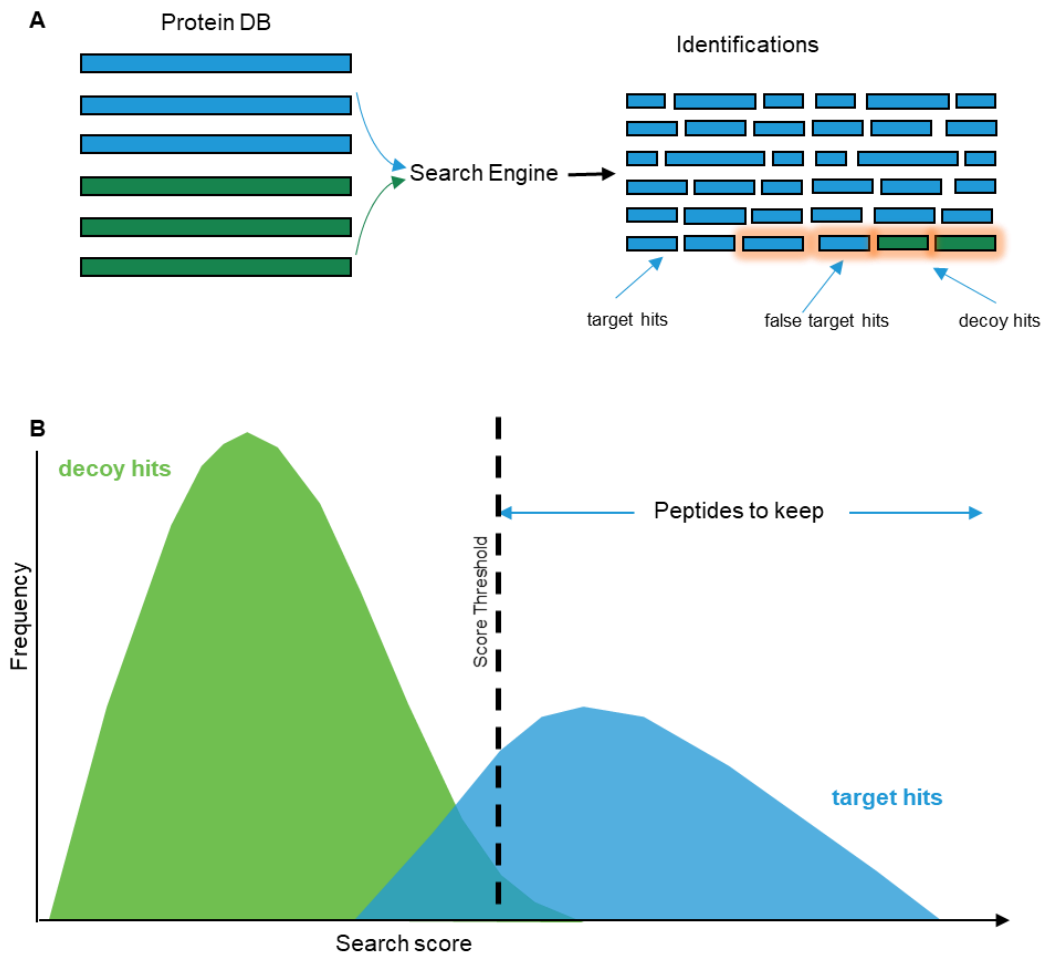

Figure 2-9 Experiment-wide false discovery rates (FDR) evaluation with a target-decoy strategy. (A) Database searching is conducted against a concatenated protein database containing target and decoy sequences. When the decoy database is constructed suitably, the false identifications should distribute between target hits and decoy hits. The resulting proportion of decoy hits compared to target hits can be used to estimate the experimental FDR. (B) Target and decoy scores are sorted using software scoring functions to separate true and false identifications and user-defined PSM score thresholds can then be applied until a desired FDR is achieved, where the FDR is the proportion of false positive above the threshold.

can also be determined from MS/MS fragmentation information by using *de novo* peptide sequencing algorithms. *De novo* peptide sequencing has gained popularity in recent years as a database-independent approach to peptide identification since it is not hindered by database-related limitations that remain a challenge for peptide-spectrum-matching approaches. Historically, *de novo* sequencing algorithms were considered inferior to database search algorithms due to factors related to MS instrumentation performance and computational resources. *De novo* sequencing from low-resolution MS/MS spectra generated results with poor accuracy, but as high-resolution MS instruments have been developed, the performance of *de novo* algorithms has improved. In addition, in the past, *de novo* sequencing algorithms were less efficient in terms of computational time and resources compared to database search algorithms due to the large unrestricted search space. As high-performance computing capabilities have improved, the current de novo algorithms are comparable in speed.

Among available *de novo* sequencing pipelines, the PEAKS platform has been developed as a high throughput method for identifying peptides using am automated *de novo* peptide sequencing based approach[99,100]. This platform incorporates algorithms for four distinct components for peptide identification: (1) *de novo* peptide sequencing, (2) database searching, (3) PTM analysis, and (4) peptide mutation analysis. To identify peptides, MS/MS data is first interrogated by automated *de novo* sequencing followed by databased searching. The combination of de novo sequencing with database searching has two primary benefits. First, peptides from proteins that are present in the sample can be identified even if the proteins are not included in the database. Second, the combination of the two complementary identification approaches results in improved accuracy and confidence of the identified peptides since a match between the *de novo* sequencing and the database search results is a good indicator that the database search results are correct. Two additional algorithms in the workflow are used to identify PTMs and peptide mutations. PEAKS PTM is an algorithm that enables searching an unlimited number of PTMs to improve the peptide identification rate of modified peptides[101]. SPIDER is a peptide mutation and

homology search algorithm that is incorporated into the PEAKS workflow that reconstructs the correct peptide sequence based on sequence tags identified during *de novo* sequencing and the homologous sequences in the database[102].

As the construction of an accurate protein database remains especially challenging for complex microbiomes, new *de novo* peptide sequencing pipelines customized for metaproteomic datasets, such as NovoBridge, have also been developed[103]. Due to the benefits of a database-independent approach, *de novo* sequencing strategies can be implemented into stand-alone or hybrid *de novo*-assisted database search workflows. One final consideration is that the *de novo* sequencing strategy as a stand-alone tool is limited to a peptide-centric view of the metaproteome, and a protein database is still required for protein inference and confident biological interpretation of the data[104].

2.4.3 Post-search processing of metaproteomics datasets.

Following the identification of peptides via database searching or *de novo* sequencing strategies, several procedures are implemented in bottom-up proteomics workflows to prepare the date for biological interpretation. First, the presence of specific proteins is inferred based on the identified peptides. Details on commonly used protein inference strategies are discussed in Chapter 4. The quantitative information for the PSMs or peptides used during inference procedures is used to quantify protein abundances. All work in this dissertation is based on label-free quantification (LFQ), which is considered a fairly robust method for protein identification and quantification with the laborious and expensive upstream labeling of proteins used in isotopic labeling strategies[105]. LFQ is a method for quantifying relative changes in abundance in two or more samples rather than measuring proteins that have been labeled with a stable isotope-containing compounds[106]. Although there are many different algorithms and workflows for LFQ in bottom-up proteomics, there is no gold standard for proteins quantification. However, all of these approaches use one of two predominant methods: (1) MS1-based methods, where the intensity of the

unfragmented precursor ion is used for quantification, and (2) MS2-based methods that count the number of MS/MS spectra identified for a specific protein as a quantitative metric[94]. For MS1-based approaches, where the peak area of the MS1precursor intensity is determined from an extracted ion chromatogram (XIC), these MS1 approaches are generally considered more sensitive to smaller abundance changes than MS2-based counting approaches[107]. Spectral counts are defined at the total number of MS/MS spectra identified for a protein and are heavily influenced by factors such as chromatographic separations, sample complexity, MS instrumentation choice, and dynamic exclusion parameters[108]. For example, spectral counting methods methods depend on MS/MS fragmentation, and thus relies on the TopN selection process during DDA analysis. Since the TopN selection process can be a stochastic process, especially with complex metaproteome samples, spectral counting is generally considered less accurate and robust than MS1 intensity-based methods for metaproteomics and should be considered semi-quantitative and semi-random[94]. The difference in sensitivity between the two strategies is in part due to the structure of the data captured by each method. MS1-based strategies are continuous data types, they provide more sensitivity than spectral counting, which is a discrete data type. As such, MS2-based spectral counting does not accurately capture low abundance analytes that may only be sampled once during measurement. MS1 intensity-based quantification methods will provide a non-digital quantification value to capture the variation between low abundance analytes, which is particularly useful for complex metaproteomes. One consideration for the selection of a quantification method for metaproteomics is the ease of implementation. Many software packages built for protein inference and quantification were designed for single proteomes, and still lack for the ability to handle complex metaproteomic datasets. Since metaproteomes generally have high proteome redundancies, the assignment of shared peptide abundance values for a particular protein is a challenging task. Procedures related to protein inference and quantification is discussed in detail in Chapter 4. In addition, since metaproteomic measurements often have very complex chromatograms where peptides are co-eluted, MS2-based counting methods are generally considered easier

to correctly assess peptide abundances in most generic bottom-up proteomics workflows72. Finally, the choice of quantification method used will have an impact on the accuracy of different statistical approaches and should be considered when selecting a downstream data processing workflow[108].

Following protein quantification, abundance values in LFQ datasets are often normalized across samples. This normalization process is aimed at accounting for the inherent biases of MS data, ranging from sample handling to instrumentation measurement differences, in order to make the samples more comparable[109]. There are numerous normalization approaches which should be selected based on the type of samples and measurements used to generate the data, but in general, all normalization methods have been adapted from DNA microarray techniques[109]. Missing values in quantitative data matrix often occur in due to the stochastic nature of sampling of DDA-based LC-MS/MS measurements, and up to 50% of the data matrix may be missing quantitative values[110]. As many statistical approaches require filled-in data matrices, this missing information prevents comprehensive and accurate assessment of the data. Thus, various procedures are used to impute missing values in the data matrix to facilitate downstream analyses.

Beyond evaluating abundance changes of individual protein sequences across datasets, functional and taxonomic annotation of the proteins is necessary to elucidate what biological processes are occurring in the community and which organisms are carrying out those functions. Numerous tools and databases are available for functional annotation of proteins, including eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups)[111], Kyoto Encyclopedia for Genes and Genomes (KEGG), and MetaCyc (metabolic Pathway Database)[112]. The functionally annotated proteins can then be used to determine what proteins in specific metabolic pathways are expressed and therefore, what biological activities are occurring at the time of sample collection.

2.5 Summary

Overall, the concepts and methodologies covered in this chapter provide the framework for conducting the various metaproteomic measurements necessary for characterizing microbiome functionality and interactions that are presented in the rest of this dissertation. In the following two chapters, several steps in the sample preparation and measurement workflow are evaluated and optimized to produce a robust workflow that provides high yields of peptides from high complexity metaproteome samples across a range of environmental microbiomes.

Chapter 3 - Technical considerations and optimization of sample preparation techniques for various complex environmental matrices.

Data presented in this chapter was generated for the following published journal articles and in-preparation manuscripts:

Van Den Bossche, T., Kunath, B. J., Schallert, K., Schäpe, S. S., **Peters, S.L**., Abraham, P. E., Armengaud, J., Arntzen, M. Ø., Bassignani, A., Benndorf, D., Fuchs, S., Giannone, R. J., Griffin, T. J., Hagen, L. H., Halder, R., Henry, C., Hettich, R. L., Heyer, R., Jagtap, P., Jehmlich, N., … Muth, T. (2021). Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. Nature Communications, 12(1), 7305–7305. https://doi.org/10.1038/s41467-021-27542-8

S.L.P contributions include metaproteomic measurements, data analysis, writing and editing of the original manuscript and response to reviewers.

Saunders, J., McIlvin, M., **Peters, S. L.**, Bertrand, E., Breier, J., Brisbin, M., Colston, S., Compton, J. R., Griffin, T., Hervey, J, Hettich, R., Jagtap, P., … Saito, M. Ocean Metaproteomic Laboratory Intercomparison from the North Atlantic Ocean using Data Dependent Acquisition. ISME. (Under preparation, Fall 2022)

S.L.P contributions include metaproteomic measurements, data analysis, and editing of the original manuscript.

3.1 Optimizing elements of sample preparation optimization for metaproteomes.

To enable the measurement campaigns for the studies detailed in chapters 5 through 7, a substantial amount of work for the dissertation focused on aspects of wet lab methodology advancements to increase the performance and depth of metaproteome measurements. In particular, work has been completed for developments and optimizations in sample preparation related to protein extraction, clean-up, quantification, and liquid chromatography peptide separation prior to MS measurement. These developments have increased reproducibility, throughput, and reduced measurement carry-over between samples for various environmental matrices, which have enabled high-quality data for in-depth characterization of microbial community function for the projects outlined in chapters 5 through 7. The methodological considerations and advancements explored in this chapter feature environmental matrices that differ from the fecal samples in the projects from chapters 5 through 7 but still represent complex media. Thus, the sample preparation factors presented in chapter 3 should be considered when embarking on any metaproteomic measurement campaign and can be broadly applied to all work presented in this dissertation.

The first portion of this chapter explores some of the challenges of extracting proteins from complex environmental matrices with limited microbial biomass. Second, the impact of interfering compounds on protein quantification during the sample preparation process is highlighted. Finally, the impact of different liquid chromatography-based peptide separation methods on mass spectrometry measurement depth is evaluated for samples with varying microbial complexity.

3.2 Extracting proteins from complex environmental matrices with limited microbial biomass.

As outlined in chapter 2, for proteomic analysis, the initial step of obtaining protein biomass during the preparation process is critical to accurately capture the full protein complement of the sample. This step is more complicated for metaproteomic analysis of microbial communities than single proteome samples due to the sample matrix. Environmental microbiome samples, such as feces and soil, are complex mixtures that can contain microbial cells, eukaryotic host cells, plant-derived fibrous materials, and other abiotic components, which means the up-front cellular lysis methods have significant impacts on downstream protein and peptide yields. Therefore, when choosing an appropriate strategy for cellular lysis and protein extraction, the composition of the sample in terms of these components must be considered. Commonly used methods for each step in the sample preparation workflow are relatively robust, with unbiased lysis of heterogeneous community microbial cells and extraction of proteins from complex matrices. However, the preparation approach needs to be customized for each environmental matrix. The key motivation for this project was to develop and optimize a robust method to extract proteinaceous biomass from complex soil matrices for proteomic interrogation by LC-MS/MS. Specifically, the methodological approach needed to be optimized for permafrost-derived soil samples that are limited in microbial biomass, high in clay content, and contain co-extracted humic substances and variable amounts of moisture.

3.2.1 Challenges with protein extraction from soil matrices.

Soils, in general, are one of the most challenging environmental matrices to sample for metaproteomics. Soils often feature spatial distribution and heterogeneity and temporal heterogeneity[113], which makes selecting representative and replicative samples difficult, especially for some of the questions addressed regarding fundamental soil biology. In addition, many soil types characteristically have very

high microbial diversity, with upwards of thousands of species of bacteria, archaea, and fungi identified in a single gram of soil[28]. In many cases, most microbes are on the edge of life, persisting in dormant or sporulated states, with only a few members dominating at a particular point in time. While there is a high diversity of microbes in the soil, the microbial biomass is often low. Therefore, the abundance of microbes and proteins that can be extracted from the soil is typically low compared to other environmental matrices.

The composition and physical characteristics of the soil matrix significantly impact protein extraction efficiency. The presence and amount of organic matter and clay minerals, and the pH of the soil have been previously shown to be the key factors influencing the recovery of proteins from different soil types as they all affect the adsorption of protein to soil particles and the ease of extraction[114]. Humic substances are heterogeneous organic matter that are products of microbial metabolism. Humic substances consist of a variety of aromatic and aliphatic compounds that form a complex array of structures with different molecular sizes that can be measured by the mass spectrometer along with any proteins if they are present in the sample. Humic substances are produced through biotic–abiotic transformations of postmortem plant, animal, and microbial debris[115]. Humic substances are divided into three fractions that can be separated by differential solubility, including humin, fulvic acids, and humic acids. The humin fraction is insoluble in both alkali or acid solvents. Fulvic acids are soluble in a wide range of pH conditions, and humic acids are soluble under alkali conditions but are insoluble in strong acids. Extractable humic acids and fulvic acids are the main humic interferences that impede protein characterization since the humin fraction is insoluble under a wide range of conditions. These humic substances confound protein characterization as they interfere with all steps in the sample preparation process, including protein extraction, clean-up, and identification by LC-MS/MS. First, many chemical methods used to enhance protein extraction, like detergents, also enhance humic and fulvic acid extraction. Proteins can become entrapped in the humic acid matrix during the sample preparation process, thereby protecting those proteins from proteolysis[116]. Humic acids also interfere with mass

spectrometry measurements. They can suppress the ionization of peptides and consist of readily ionizable low-molecular-weight compounds that are less than 2000 Da which can be easily detected by ESI mass spectrometry[56]. Finally, peptide samples containing a large amount of humic acids are difficult to load onto LC trapping columns and often leading to flow obstruction. Since many soil samples contain limited microbial biomass in some soil samples, it is often necessary to load the entire sample for a single MS to yield sufficient signal intensity in the mass spectra. Therefore, there is a high risk of column clogging and sample loss[56]. This means we have to consider humic substances while optimizing our protocol to minimize their impact on the measurements.

Clay minerals also have a significant impact on the extractability of proteins. While direct extraction of proteins from soil has been shown to be beneficial compared to indirect extraction[113,117], where the microbes are enriched/isolated from the soil matrix before protein extraction, direct extraction also raised additional challenges in terms of the soil mineralogy. One of the main problems with extracting microbial proteins directly from the soil is that biomolecules, such as proteins and nucleic acids, often strongly adhere or adsorb onto clay minerals. This process is reversible to a limited extent, but it obstructs protein extraction and purification and quantification, separation, and identification. Soil adsorption of proteins has implications for soil proteomics because the soil matrix stabilizes proteins and protects them against proteolysis.

Finally, additional characteristics, such as the hydration and the pH of the soil samples, need to be considered before choosing an extraction method. If not accounted for at the start of the extraction method, the hydration level of the soil will impact the weight of the starting material and influence the overall estimation of protein recovery from the initial sample mass. Depending on the soil type and variability of hydration levels between samples in the experiment, an additional upfront sample pre-processing step, such as lyophilization, can be added to remove water from the soil matrix with minimal impact on the microbial activity and protein structure[118,119]. This pre-processing step is beneficial in two ways. First, if the soil

matrix contains a high amount of clay, removing larger particles that do not contain much proteinaceous biomass, such as rocks and twigs, will be difficult when the soil matrix is fully saturated. Lyophilization will enable the sifting of the samples to remove large debris. Second, estimating the necessary amount of sample to be processed for sufficient protein extraction will be difficult using the wet weight of the soil. If there is variability in soil matrix hydration levels, pre-lyophilization of the samples will allow for dry weight measurements to estimate the necessary amount of material for protein extraction. The dry weight estimate will ensure the same amount of material is processed per sample regardless of the initial water saturation of the soil. Other studies have demonstrated that soil pH is an important factor in protein recovery, as pH influences the adsorption of protein to soil particles and the ease of extraction[114,120].

In soil metaproteomics, no universal extraction methodology has been established. Given the physicochemical diversity of soils, there is unlikely to be a single approach that works for all soil types. There are two primarily used approaches to extract proteins from a soil matrix. Indirect protein extraction is based on the enrichment of microbial cells from the environmental matrix prior to protein extraction using techniques such as density gradient centrifugation[113]. While this approach can reduce the number of proteins that bind to mineral particles or soil organic matter, it has been shown to introduce bias by selectively enriching actively growing cells[121]. In recent years, many soil protein extraction methods for metaproteomics rely on direct extraction of the proteins from the soil matrix for unbiased extraction of proteins from all organisms present in the community [28,117]. This direct extraction approach is often completed by making a soil slurry and vortexing with extractants such as SDS or NaOH prior to mechanical disruption of the cell membranes via sonication or bead beating. After cellular lysis, proteins are extracted and cleaned up via methods such as TCA precipitation or chloroform-methanol extraction before proteolytic digestion. There have been several methodological advancements outside of metaproteomics for the extraction of other soil components, such as DNA or chemical compounds, that may be applied to extract

proteins. For example, a pressurized liquid extraction (PLE) method using a hard cap espresso machine has been applied to extract polycyclic aromatic hydrocarbons (PAHs) from soil and sediment samples[122]. The introduction of heat, pressure, and upfront filtration to the extraction process may enhance protein extraction without the co-extraction of interfering compounds. The combination of lysis and extraction steps needs to be customized for each soil matrix type to ensure the best clean-up of co-extracted interfering compounds such as humic and fulvic acids. To date, most soil protein extraction methods still have some issues with unbiased protein recovery, and there is room for improvement in the methodologies, especially for challenging soil types.

The primary goal of this project was dedicated to enhancing protein extraction methods for soil samples that have (1) low microbial biomass, (2) variable clay content, and (3) the presence of co-extracted humic substances. Three components of the sample preparation process were evaluated. First, the adaptation of the protein aggregation capture (PAC) method for protein clean-up and digestion was assessed to include an additional acid crash step to remove fulvic acids. Second, the co-extraction of interfering substances from both soil types resulting from the chemical extractant used in the extraction was assessed. Third, vortex-based and pressure-assisted extraction methods were compared to determine the impact on protein recovery.

3.2.2 Results.

3.2.2.1 Adaptation of the protein aggregation capture method to remove co-extracted humic substances from soil samples.

Many soils, including the permafrost-affected soils in this project, have very low microbial biomass per gram of soil, so large amounts of soil are often processed to get sufficient peptide yields for LC-MS/MS measurements. Soil components such as humic substances and small clay particles interfere with detecting peptides that are co-extracted with proteins and should be removed to the best extent possible before

LC-MS/MS. The protein aggregation capture (PAC) method can be adapted for this workflow to work with substantial sample amounts (10-100 g of soil) to increase protein yields but not significantly increase the carry-over of non-proteinaceous material.

To successfully adapt the PAC protocol for soil samples, we first confirmed that co-extracted humic acids bind to the magnetic bead microparticles. Different amounts of a commercially available humic acid standard (0, 12.5, 25, 50, 100, 200, 400 µg) were added to 400 µg of *C. autoethanogenum* proteins and 400 µg of magnetic beads. Acetonitrile addition induced the immobilization of the protein/humic acid mixtures onto the magnetic beads. After the aggregation and wash steps, the size of the resulting bead pellet was visually inspected to determine if humic acids were binding to the magnetic bead microparticles. **Figure 3-1** shows that the size of the bead pellets increases with increasing amounts of humic acid standard. The bead pellet with 400 µg of protein and 400 µg of humic acid was ~3x larger than the bead pellet that only contained 400 µg, confirming that humic acids will aggregate on microparticles with the introduction of high organic solvent concentrations.

Fulvic acids and humic acids can be separated based on their differential solubility. Previous work using other protein clean-up methods has shown that incorporating an acidic solution during a key step in the digestion process can remove a significant fraction of these co-extracted humic substances[56]. Utilizing the knowledge that fulvic acids are soluble at a wide pH range while humic acids are insoluble at low pH, we set out to adapt the PAC protocol to remove fulvic acids from the sample by incorporating an acidified aqueous wash after the protein aggregation step. In theory, after the addition of an acidic solution, the proteins and humic acids, which are insoluble at low pH, would immobilize on the magnetic bead microparticles, and the fulvic acids would remain in the solution. The fulvic acids could then be easily removed before protein digestion. To confirm that incorporating this acidified aqueous wash does not lead to protein loss, proteins from *C. autoethanogenum* cell lysates were aggregated on magnetic beads and washed with either 1% formic acid solution or LC-MS/MS grade H_2O. After washes, proteins were detached from the magnetic beads

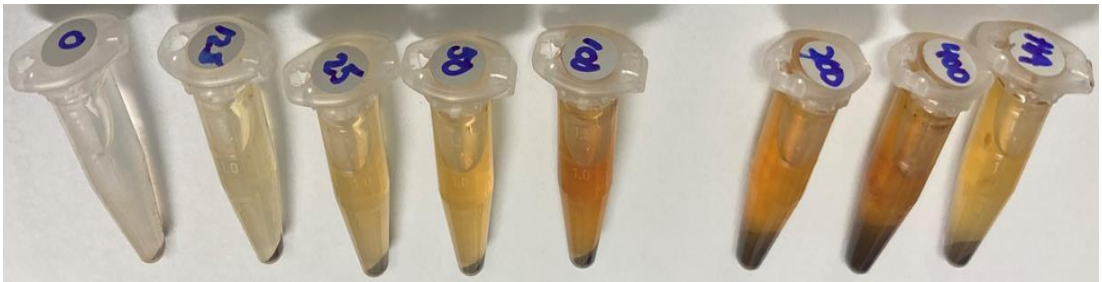

Figure 3-1 Aggregation of proteins and humic acids on magnetic bead microparticles. 400µg of C. autoethanogenum proteins immobilized on hydrophobic magnetic beads with increasing amounts of a humic acid standard (0-400µg). The tube on the far right (HA) is a control sample with only 400µg of humic acids (no proteins) immobilized on the beads.

and quantified with a UV-Vis spectrophotometer. No protein loss was observed with the addition of formic acid during the washing process, indicating that fulvic acids can be removed with minimal protein loss.

Reducing or removing humic acids is more challenging than removing fulvic acids at the protein level as they have similar characteristics to proteins, such as size and solubility in different pH ranges. Removal of humic acids at this stage in the sample preparation process also has the undesirable effect of protein removal. However, most humic acids have larger molecular weights than peptides and remain insoluble at low pH, so filtration with a 10kDa MWCO filter after the digestion process is stopped does remove many large molecular weight humic acids from the sample without impacting peptide recovery.

To assess the efficiency of the modified PAC protocol for removing humic substances from soil samples, 15g of topsoil samples were spiked with a mixture of unlysed bacterial cells from organisms listed in **Table 3-1** at a concentration of 1×10^8 bacterial cells per gram of soil. After vortexing and cellular lysis procedures, samples were processed by the unmodified PAC method or the modified PAC method including an acidified aqueous wash. We assessed the impact of humic substance removal from the samples using the charge states of the measured ions as a proxy for the amount of humic substances remaining in the samples at the time of LC-MS/MS measurement. Humic substances are primarily singly charged ions, while tryptic peptides are typically doubly or triply charged[56,123]. In addition, the LC-MS/MS charge distribution results from a fecal sample of a gnotobiotic mouse that was colonized with bacteria isolated from the human gut and fed a high-fat diet with limited fruits and vegetables was compared to the results from the soil samples as a type of control. Since fecal materials are expected to have lower concentrations of humic substances compared to some soil types, the mouse fecal pellet was included as a control sample to show the charge state distribution of precursor ions in a complex environmental sample with less expected humic substances present that was also processed with the standard PAC method.

Comparing the distribution of charge states of the measured ions in topsoil

Table 3-1 Organisms used for soil spike-in experiments

Organism	**Gram stain**
Bacteroides ovatus	Gram-negative
Clostridium autoethanogenum	Gram-positive
Collinsella aerofaciens	Gram-positive
Escherichia. coli K12	Gram-negative
Odoribacter splanchnicus	Gram-negative

between the two protocols shows that adding an acidified aqueous wash to the PAC protocol reduces the number of ions that couldn't be assigned a charge state and the number of singly charged ions **(Figure 3-2).** The percentage of singly charged precursor ions decreased by 9% with the addition of an acid wash, and the number of precursor ions that could not be assigned a charge was also reduced by 14%. At the same time, the percent of doubly or triply charged ions, which are the typical charge states of peptide ions, increased by 23% with the addition of the acidified aqueous wash. This increase in peptide-like precursor ion charge states is reflected in the peptide identifications from the database searches of the five organisms spiked into the sample. There were 130 peptides identified in the sample with an additional acidified aqueous wash compared to only two peptides identified in the soil sample prepared with the unmodified PAC method. These results show that adding an acidified aqueous wash didn't reduce peptide recovery. However, there is still an increased amount of precursor ions with unassigned or single charge states compared to other environmental samples containing fewer humic substances. Sixty-two percent of the precursor ions in the control sample were doubly or triply charged compared to 33% of ions in the acid-washed soil sample and 10% of ions in the unmodified PAC method soil sample. Due to the large percentage of unassigned and singly charged precursor ions still in the samples with the modified PAC protocol, our efforts turned to optimize the methodology to reduce the initial co-extraction of humic substances from the soil matrix in an attempt to increase the peptide identification rate.

3.2.2.2 Optimizing conditions for pressure-assisted protein extraction.

To increase protein extraction efficiency, we set out to adapt the espresso machine for a modified low-cost pressurized liquid extraction (PLE) method to extract proteins from soils. Typical PLE operate at elevated temperature (25–200°C) and pressure (500–3000 psi)[124]. Espresso machines operate at temperatures of 80°C and up to 200 psi, adding both a heat and pressure element to the extraction process compared to traditional vortex-based protein extraction methods. In addition, the additional

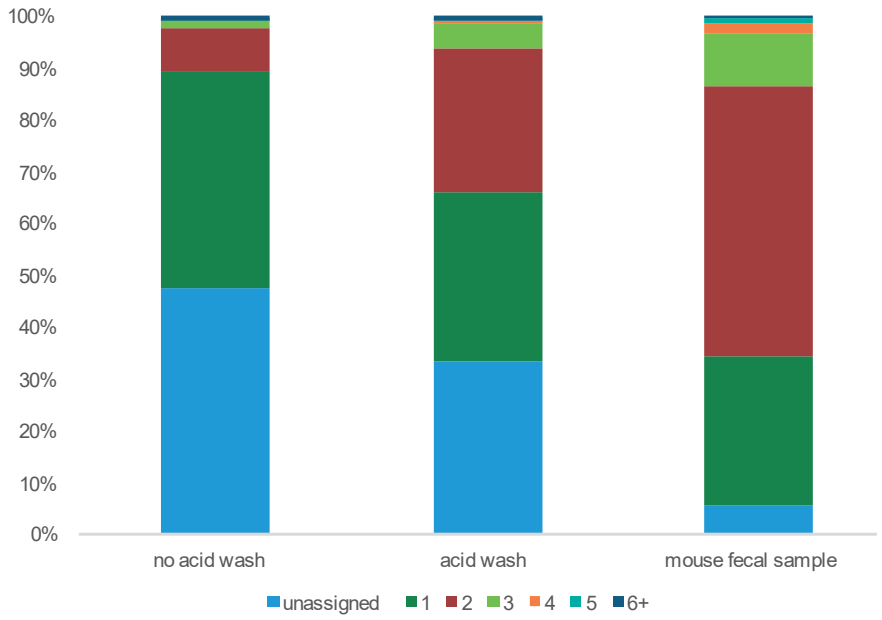

Figure 3-2 Charge state distribution of precursor ions detected by LC-MS/MS. The bar chart shows the relative proportion of each charge state for all precursor ions identified in each sample. If no charge state could be determined for a precursor ion, the charge state was marked as unassigned. All precursor ions with a charge state of +6 or higher were grouped into a single group (6+). Two topsoil samples were prepared with or without an incorporated acidified aqueous wash. The charge state distribution of a mouse fecal sample prepared with no incorporated acidified aqueous wash is shown as a control.

filtration provided by the espresso machine has the benefits of reducing the amount of soil mineral particles retained throughout the sample preparation process. To adapt the espresso machine to be compatible with protein extraction from a variety of soil types, we tested several parameters: additional filtration in the portafilter to reduce sample clogging, the ratio of sample to packing material to maintain appropriate pressure for extraction, and the wetness of the soil prior to extraction needed to prevent over-pressurization of the system.

First, we looked at implementing an additional layer of filtration in the portafilter during extraction since preliminary extractions with no filter additions led to soil particles clogging the portafilter holes during the extraction process. The clogging of the portafilter holes resulted in extraction failure due to over-pressurization of the system or inconsistent volumes extracted. Placing a Whatman 0.2µm filter paper at the bottom of the portafilter prior to soil addition did reduce the frequency and amount of portafilter holes clogging and therefore reduced extraction failures. Still, the extraction volumes were not consistent using this type of filtration. Lining the portafilter with a paper coffee filter prior to soil addition completely prevented portafilter clogging and yielded consistent extraction volumes of 20mL. Paper coffee filters generally have a large pore size of around 10-20µm to allow for liquid percolation[125] and do not remove most unlysed bacterial cells from the extracted sample. In terms of clay minerals in soil samples, while this approach does not remove primary particles, which are less than 2µm in size, it is sufficient to remove larger flocculi (10-20µm) and microflocs (50-500µm)[126]. Due to these filtration characteristics and the reproducibility of expresso machine extraction using this filtration method, we decided to implement coffee paper filters into the extraction workflow.

Second, we tested the implementation of packing material in the portafilter to maintain consistent pressure during the extraction. The back pressure acquired by the machine during extraction is determined by the particle sizes of the material packed into the portafilter[122]. Typically, PLE methods disperse the sample with an inert packing material, such as sand, prior to placement in the extraction vessel to ensure

better extraction solvent-matrix contact, prevent agglomeration, and reduce uneven extraction solvent flow[127,128]. Four sand:sample packing ratios were tested (0:1 sand:soil, 1:1 sand:soil, 2:1 sand:soil, 3:1 sand:soil). The extractions with no sand added or equivalent amounts of sand and soil did not provide even solvent flow, and soil samples with a high clay content tended to cake, which caused over-pressurization of the system. Both the 2:1 and 3:1 sand to soil ratios worked well to maintain consistent pressure during the extraction process. We decided to use the 2:1 ratio for future experiments as this provided the largest amount of soil in a single extraction attempt, which could be important for protein extraction with low microbial biomass soil samples.

Finally, we tested the compatibility of espresso machine-based extraction with commonly used extractants for soil samples. From early extraction trials, the wetness of the soil had an apparent impact on the pressure maintenance of the espresso machine during extraction. To determine if we could add the extraction buffer directly to the soil matrix after packing in the portafilter, we tested three different soil wetness levels (fully saturated with 20mL of extraction buffer, partially saturated with 4mL of extraction buffer, and dry soil samples with no extraction buffer added). Repeated extraction attempts with the different wetness levels showed that dry soil was critical for the robust operation of the espresso machine, as any amount of liquid added to soils with a high clay content caused over-pressurization of the system and uneven extraction solvent flow through the soil matrix. For the remaining experiments, samples were lyophilized after extraction buffer was added to ensure dryness before espresso machine extraction. Extractants for soil protein recovery include simple salts (sodium pyrophosphate), bases (sodium hydroxide (NaOH)), organic acids (sodium citrate), and surfactants (Sodium dodecyl sulfate (SDS)). Previous work has shown that the effectiveness of extractant type for protein recovery is highly dependent on the physicochemical properties of the soil matrix[114,129]. For this study, we looked at two of the most commonly used extractants, NaOH and SDS. Four extraction buffers were prepared for comparison: water only, 0.1M NaOH, 3% SDS, 0.1M NaOH + 3% SDS. Both extraction buffers containing SDS foamed during the extraction process

and caused over-pressurization of the espresso machine, and therefore deemed incompatible with this type of extraction. No observational differences were found in the operation of the espresso machine between the water extraction and NaOH extraction. As other studies have shown that NaOH aids in the extraction of proteins from soil samples, we decided to use this extraction buffer for future experiments. With all of these adaptations to the workflow to enable robust and reproducible operation of the espresso machine, we set out to determine differences in protein recovery based on extraction method (espresso machine vs. vortex).

3.2.2.3 Evaluation of mechanical and chemical extraction methods in topsoil.

To fully evaluate mechanical extraction methods and extraction buffers, an experiment was set up with known amounts of *E. coli* cells (10^8 cells per gram of soil) added to topsoil samples to compare traditional vortex-based vs. pressure-assisted espresso machine extraction using either water or 0.1M NaOH as an extractant **(Figure 3-3).** After extraction and sonication, samples were centrifuged, and the resulting supernatant was processed for proteolytic digestion via the modified PAC method.

After centrifugation to remove large soil debris, there were very visible differences between extraction approaches and extraction buffers **(Figure 3-4).** The vortex method was much darker in color and murkier compared to samples prepared with espresso machine extraction throughout the prep. Samples prepared with NaOH were also darker in color compared to the corresponding water extraction samples.

These color differences were due to non-proteinaceous material present after extraction, including humic substances and clay mineral particles. Darker supernatants had larger pellet sizes during PAC. These pellets were hard to resuspend before trypsin digestion, which could negatively impact the protein digestion efficiency if proteins are shielded in the pellet from trypsin during the digestion.

The results of the database searches against the theoretical *E. coli* proteome showed that the espresso machine pressure-assisted extraction with sodium hydroxide

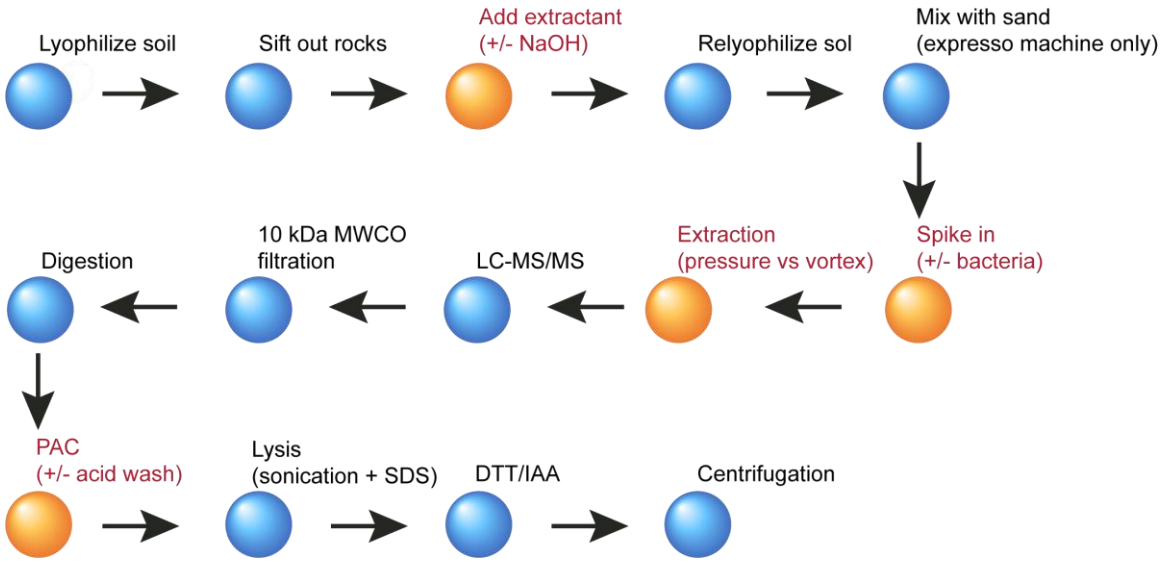

Figure 3-3 Schematic illustration of the sample preparation workflow used to assess protein extraction methods for the analysis of soil metaproteomes. The four steps marked in orange (red text) represent critical steps in the process where different variables were tested (extraction buffers, extraction method, acid wash) using both permafrost-affected soils and commercially available topsoil.

Figure 3-4 The extracted volume of topsoil samples prior to protein clean-up and digestion. Order (left to right): Espresso machine extraction with H2O, Espresso machine extraction with NaOH, Vortex extraction with H2O, Vortex extraction with NaOH. Vortex-based extractions resulted in darker supernatants after low-speed centrifugation to remove large soil particles than espresso machine-based pressure-assisted extractions. In addition, extractions with NaOH resulted in darker supernatants compared to H2O-based extractions.

pretreatment had the highest number of detected peptides in any sample. However, repeated extractions using this approach yielded inconsistent recoveries of peptides, indicating more optimization is needed for the method. Both of the vortex extraction samples (+/- NaOH) with bacterial spike-ins also had a large number of proteins recovered. However, based on the charge state distribution of the samples, there were many more singly charged ions than espresso machine pressure-assisted extraction samples, indicating a large amount of humic substances were co-extracted with proteins using this method. Furthermore, the vortex extraction sample prepared with NaOH had lower *E. coli* peptides detected compared to the vortex extraction prepared with water only. While NaOH enhances protein extraction from soils, it is also used to extract humic substances[130]. The presence of humic acids in the sample may explain the lower number of *E. coli* peptides detected in samples, as the humic substances may have suppressed the ionization of peptides. In addition, since the digested peptides from vortex-based extraction samples contained more non-proteinaceous material compared to the espresso samples, this decreased the longevity of the LC analytical column and the MS instrument cleanliness. Overall, the addition of NaOH to the extraction buffer appears to help the recovery of proteins from soil samples as long as efforts are taken to reduce the amount of interfering compounds, such as fulvic and humic acids. In addition, while preliminary extraction attempts showed pressure-assisted extractions had higher protein yields and less non-proteinaceous material compared to vortex-based extraction, these results were not consistently reproducible with repeated extraction attempts. Vortex-based extractions also recovered a high amount of proteins, but the presence of co-extracted non-proteinaceous material made sample processing challenging and decreased LC-MS/MS performance. As soil chemistry plays a significant role in determining the success of protein extraction, extractant choice and extraction method need to be customized for the particular soil type being analyzed.

3.2.2.4 Optimizing and evaluating protein extraction in permafrost-derived soils.

Preliminary measurements using a vortex-based extraction of 2g of soil from core samples collected from Bayelva, Svalbard, yielded only two detectable peptide-like features. These soil samples have high clay content, limited estimated microbial biomass (~10^4 cells per gram of soil), and humic substances in the soils also interfere with the protein extraction process and follow-up LC-MS/MS measurements. Therefore, the lack of detectable peptides without optimization was not unexpected. Based on the results of the methodology testing with topsoil samples, 15g of lyophilized and sifted permafrost affected soils were processed using both the vortex-based and pressure-assisted extraction methods with a NaOH-based extraction buffer and an acidified aqueous wash incorporated into the PAC procedure. **Figure 3-5A** shows that both modified extraction methods reduce the relative abundance of singly charged precursor ions from a sample even with 7.5x material processed compared to the original methodology. Doubly and triply charged ions, indicative of peptide-like ions, represented ~50% of precursor ions detected in each of the samples with optimized methodology, compared to 6% of the ion population in the original method. This is a much higher percentage compared to the fraction of doubly and triply charged ions found in topsoil samples from the previous experiments utilizing the same methodology, indicating these permafrost soils had fewer interfering compounds present in the measured peptide mixture. As no matching metagenomes were available for the permafrost samples used for methodology testing, *de novo* peptide sequencing was used to determine the amino acid sequences of all peptide-like features. In total, the original method yielded only two high-quality *de novo* sequence tags compared to 270 and 6900 *de novo* sequence tags identified in the pressure-assisted extraction and the optimized vortex-based extraction, respectively. The optimized vortex method had the most peptide-like features identified. However, the depth level is still limited compared to what would be expected for optimal metaproteome coverage in an environmental sample. Of the 6900 *de novo* sequence

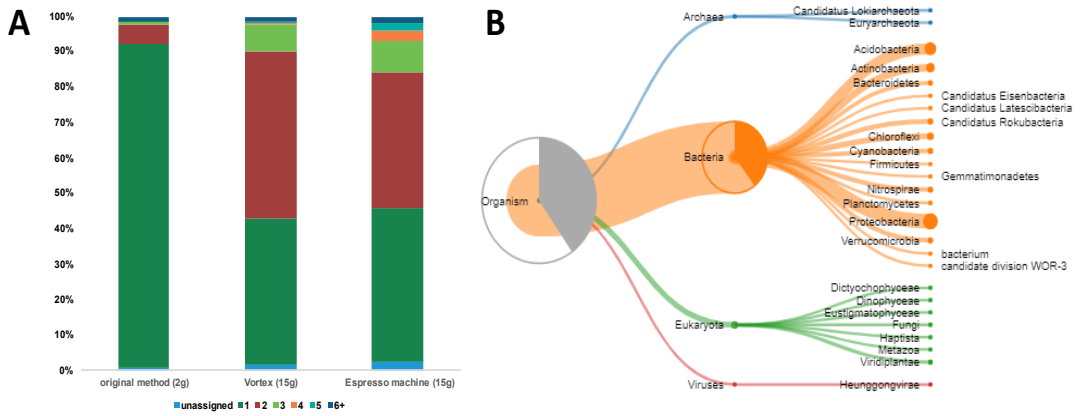

Figure 3-5 Evaluation of LC-MS/MS results for optimized protein extraction in permafrost soils. (A) The charge state distribution of precursor ions detected by LC-MS/MS for permafrost-affected soils using different extraction techniques. Three extraction methods were evaluated using soil collected in a single soil core at the same depth (30cm below the soil surface). Two grams of soil were prepared for the original vortex-based extraction. Fifteen grams of soil were prepared for the optimized vortex-based extraction and the espresso machine pressure-assisted extraction. (B) Unipept phylum-level taxonomy of high-quality de novo sequence tags with no matching database entries based on a lowest common ancestor (LCA) analysis. De novo sequence tags were considered high quality if they had an average local confidence score of greater than 80% and a minimum of six amino acids. Only de novo sequence tags with no matches to database entries for common mass spectrometry contaminants were analyzed.

tags, only 884 of these sequence tags could be matched to tryptic peptides with UniProt entries. Many of these peptides mapped to proteins related to core activities like translation, such as elongation factors and ribosomal proteins. **Figure 3-5B** shows a treemap of the phylum-level results from a biodiversity analysis where the lowest common ancestor of each UniProt entry for matching peptides was calculated. Most *de novo* sequence tags mapped to bacterial proteins for organisms commonly found in soil samples. Many of the microorganisms identified matched organisms identified in the metagenomic analysis of soil cores collected near the sampling site in this study[131]. At this shallow measurement depth, metaproteomics can provide insights into which organisms in a community are active but cannot provide many details about specific functions and processes being carried out by those organisms.

3.2.3 Discussion.

In total, this study highlights the challenges associated with extracting proteins from complex soil matrices. Factors such as low microbial biomass, co-extraction of interfering compounds, and soil mineralogy impact the protein extraction efficiency and resulting LC-MS/MS measurements. As shown in this study, the success of protein extraction and removal of interfering compounds can be very different between soil types when the same exact method is used for extraction, emphasizing that there is not a universal approach to extracting proteins from all soil types. In an ideal setting, each step in the methodology approach, including the type of extraction, chemical extractants, and protein clean-up method, should be assessed and optimized for the analyzed soil type. Furthermore, depending on the experimental questions being addressed in a study, alternative approaches may be necessary to achieve the metaproteome depth required to answer those questions. These alternative approaches include indirect protein extraction methods where microbial cells are enriched before extraction. Many indirect protein extraction methods introduce some bias since only a small fraction of the cells in the soil are extracted, and the extracted cells do not always accurately represent the total biodiversity in the soil sample. Care should be taken to

determine what measurement depth is necessary and to not come to false or incomplete interpretations of the data based on methodological bias introduced in sample preparation. Despite these limitations, metaproteomics directly measures the function of soil microbiota. It provides an excellent complement to other omics approaches where microbial activity cannot be directly measured.

3.2.4 Methods.

3.2.4.1 Soil selection.

For this experiment, two different soil types were selected for testing protein extraction methods to evaluate the robustness of the methodology across various soil chemistries. For the collection of permafrost-affected soils, an active layer core was collected from permafrost sites near the Bayelva River in the Leirhaugen glacier moraine in Ny Ålesund, Svalbard, Norway. The soil core consisted of the first 30cm of soil below surface level and was removed in core liners to retain the permafrost structure. The texture and physical characteristics of the soil varied with depth, and colored layers could be observed from different depths of the same borehole—indicating that different soil layers contained different geochemistry and *in situ* microbial metabolism. The core was sliced into 2-cm depth intervals using a sterile geological sampling hammer and chisel and frozen at -80C until processing. Commercially available topsoil was used in the experiments as a secondary soil type.

3.2.4.2 PAC adaption.

To assess whether humic acids bind to microparticles in addition to proteins, 1mg of commercial humic acid standard (Sigma Aldrich) was dissolved in 1mL of 1 mL of 100 mM Tris-HCl buffer (pH 8). Varying amounts of the dissolved humic acid standard (0, 12.5, 25, 50, 100, 200, and 400µg) were added to 400µg of proteins from *C. autoethanogenum* cell lysates. 300ug of magnetic beads (1 micron, SpeedBead

Magnetic Carboxylate; GE Healthcare UK) was added to each sample. Samples were then adjusted to 70% acetonitrile to induce aggregation of proteins and humic acids. Aggregated proteins and humic acids were washed on a magnetic rack with 1mL of 100% acetonitrile, followed by 1mL of 70% ethanol. The binding of humic acid standards to the magnetic beads during the aggregation process was evaluated by the size of the resulting bead/protein/humic acid pellet.

To assess the possibility that adding an acidified aqueous wash step following protein aggregation on the magnetic beads interferes with the protein digestion process and reduces protein recovery, 300µg of crude proteins from *C. autoethanogenum* cell lysates were aggregated on microparticle beads. Proteins were then washed according to the PAC protocol, with the optional wash of 1mL 1% formic acid/ LC-MS/MS grade H_2O. After washing, captured proteins were resuspended in 4% SDS/100 mM Tris-HCl buffer (pH 8) and boiled at 90°C for 10 minutes. Recovered proteins were quantified via NanoDrop Scopes A205.

3.2.4.3 Extractant compatibility.

The compatibility of various chemical extractants was evaluated for both soil and mechanical extraction types. For each extraction method tested, four different extraction buffers were used (LC-MS/MS grade H_2O, 0.1M NaOH, 3% SDS, 0.1M NaOH + 3% SDS). Compatibility was assessed by determining the resulting protein yields using each buffer, and the recovery of non-proteinaceous material in the digested peptide sample at the end of the extraction process.

3.2.4.4 Optimizing conditions for pressure-assisted extraction.

Several parameters were tested to develop a robust and reproducible method for pressure-assisted protein extraction using a Breville Barista Express® espresso machine. Two types of additional filters were tested to reduce portafilter clogging during the extraction process: a paper coffee filter and Whatman 0.2µm filter paper.

To evaluate the maintenance of the overall pressure of the system during the extraction process varying ratios of soil to packing material were tested (0:1 sand:topsoil, 1:1 sand:topsoil, 2:1 sand:topsoil, 3:1 sand:topsoil). The impact of soil wetness on the pressure maintenance of the system and the resulting extraction volume was tested. For each 15g of lyophilized permafrost soil or topsoil, three levels of soil wetness were evaluated: fully dry (no extraction buffer added), partially saturated (4mL of extraction buffer added), and fully saturated (20mL extraction buffer added).

3.2.4.5 Comparison of mechanical extraction methods.

To compare protein extraction efficiency between traditional vortex methods and the optimized pressure-assisted espresso machine extraction method, a series of experiments were set up following the general workflow outlined in **Figure 3-1**. Although some of the procedures were not necessary for the vortex method, such as re-lyophilization of the samples after the addition of NaOH, they were included for all samples as they were essential for successful pressure-assisted espresso machine extractions. Standardizing these steps enabled uniform comparison between the two methods. The methods outlined in the workflow were tested with both topsoil and permafrost-affected soil samples.

3.2.4.6 Spiking of Soil with Known Bacterial Isolates.

To track microbial proteins in the absence of a sample matched metagenome, selected soil samples were prepared for methodology testing with the addition of 3 Gram-negative and 2 Gram-positive bacterial isolates listed in **Table 3-1.** Bacterial cells from one organism or mixtures of organisms with equivalent cellular biomass were added to soil samples either at the concentration estimated to be necessary for successful protein extraction (1×10^8 cells/gram of soil) or at the concentration of cells estimated to be present in the permafrost affected soils (1×10^4 cells/gram of soil).

3.2.4.7 Protein extraction and digestion.

To ensure complete lysis of cells from the environmental matrix, samples were centrifuged and subjected to sonication and chemical lysis vis SDS after vortexing or pressure-assisted espresso machine extraction. Protein extraction, clean-up, and digestion were conducted with modifications to the protein aggregation capture (PAC) method[59] to account for the large volumes and co-extracted humic substances present in these samples. Most importantly, the addition of a 1% formic acid acidified aqueous was incorporated after initial aggregation on the magnetic beads to remove soluble fulvic acids from the sample. Trypsin was selected as the proteolytic enzyme for protein digestion. The resulting peptides were adjusted to 0.5% formic acid and filtered on a 10kDa MWCO filter to remove any under-digested proteins and remaining humic acids.

3.2.4.8 Peptide measurement by LC-MS/MS and data analysis.

3ug of proteolytic peptides were analyzed via automated 1D-RP-LC−MS/MS with a 180-minute linear gradient as previously described in chapter 2. The resulting MS/MS spectra were interrogated by database searching in Proteome Discoverer 2.5 (SequestHT/Percolator) and through *de novo*–assisted database searching[99] in PEAKS Studio 10.6 (Bioinformatics Solutions). Custom-built proteome databases were derived from the combination of each bacterial isolate used in the spike-in experiments **(Table 3.1)** and against common LC-MS/MS contaminant proteins. No soil-specific protein databases were available for database searching. Identified peptides and *de novo* sequence tags were required to have six amino acids minimum length of 6 amino acids. Sequence tags were also required to have a minimum average local confidence score of >80% to only retain high-quality *de novo* sequencing tags. Database searching in Proteome Discoverer was used to assess the charge state distribution of identified consensus features. *De novo*–assisted database searching in PEAKS aided in identifying non-endogenous peptides resulting from the bacterial additions and

endogenous *de novo* peptide sequence tags resulting from *in situ* soil organisms not represented in the protein databases. *De novo* sequence tags that didn't match any protein database entries were subjected to a lowest common ancestor (LCA) analysis in Unipept[132] to assess the potential taxonomic diversity of organisms present in the soil samples.

3.3 Protein clean-up and quantification in the presence of co-extracted interfering molecules.

Data generated and analyzed for the Ocean Metaproteomic Laboratory Intercomparison as presented in Saunders et al., 2022 (anticipated). ISME.

The quantification of proteins and peptides throughout the sample preparation process is an important consideration during the sample preparation process for metaproteomics. While this is typically considered a more trivial aspect of the sample preparation process for low-complexity samples such as individual proteins or bacterial isolates, it plays a more prominent role when working with environmental samples containing co-extracted molecules that may interfere with the protein and peptide quantification process. For example, feces often contain bilirubin, soils contain humic substances, and plant samples can contain chlorophyll. Different assays, such as the Bradford assay, BCA assay, fluorometer, and UV-Vis spectroscopy, all have sensitivities that may interfere with specific compounds. The appropriateness of each assay for a particular sample type should be considered in the sample preparation process. Protein/peptide quantitation should also be conducted at an appropriate time in the sample preparation process, depending on the sample type. At the protein level, quantification of proteins helps determine the proper amount of proteolytic enzyme to be added during digestion procedures to prevent under or over-digestion of proteins. As some of the co-extracted molecules can be removed or reduced in the sample during the clean-up process, for some samples, quantifying the amount of protein extracted from the samples after clean-up procedures can be crucial

for accurate estimation of protein concentrations. At the peptide level, the amount of peptides loaded onto the LC columns directly influences measurement depth and the number of peptides detected during an LC-MS/MS measurement. Under- or overloading the columns can decrease peptide separation, identification rates, and the reliability of quantification[71]. For metaproteome samples, this reduced identification rate can hamper biological insights into the protein expression and function of low-abundance members of a community. While it is good practice to adjust the amount of peptides loaded based on particular sample types and the LC set-up utilized, when and how proteins and peptides are being quantified during the sample preparation process influences the reproducibility of measurements for complex environmental samples.

For example, in a recent intercomparison effort to assess the reproducibility of metaproteome measurements of environmental ocean samples work was conducted in collaboration with eight (meta)proteomics laboratories worldwide that have different sample preparation processes and instrumentation. Each lab was given a portion of a filter used to collect samples from the euphotic zone of the North Atlantic Ocean with instructions to process and measure 2ug of the prepared peptide mixture by LC-MS/MS. Only the measurements and analysis collected for this dissertation are presented here. Figures **3-6A and 3-6B** show the variation in quantification for a single sample at both the protein and peptide level, depending on the quantification method used. At each level, the estimated amount recovered according to quantification using NanoDrop Scopes A205, which are sensitive to interferences in the sample that absorb around 205nm in the UV-Vis spectrum (like peptides), was more than 2.6x greater than the estimated amount recovered by BCA assay. This highlights the notion that the amount of peptides loaded should not be standardized for reproducibility unless the quantification method is also standardized. Figures **3-6C and 3-6D** show how variable amounts of peptides loaded onto an LC column impact the peptide identification in a single sample containing hundreds of organisms and drastically impact AUC quantification, especially among lower abundance members.

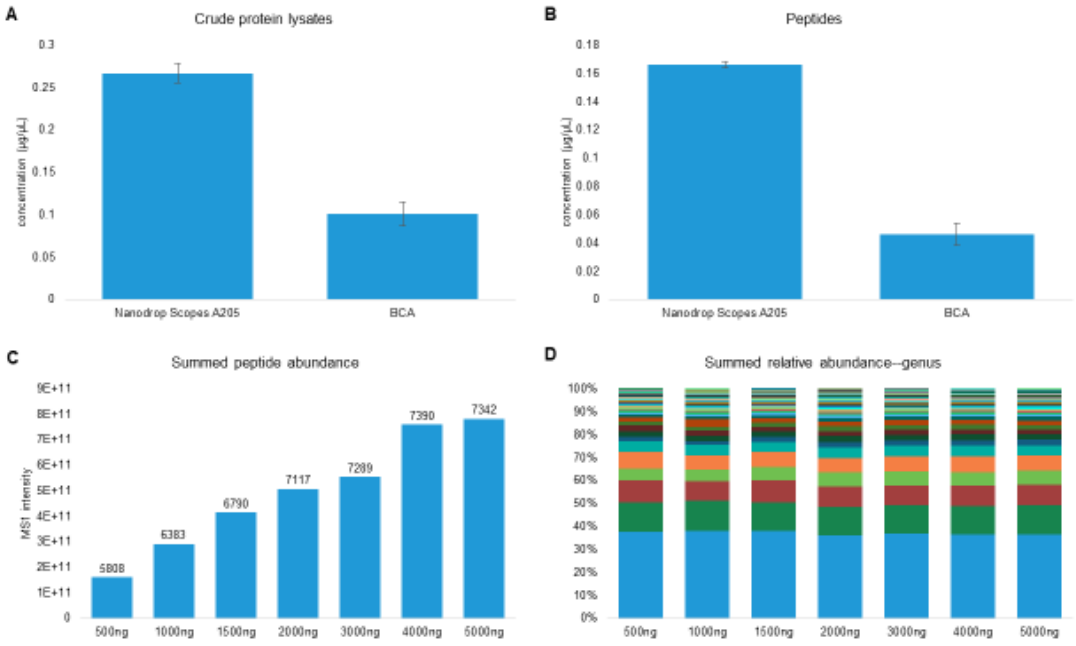

Figure 3-6 Impact of protein quantification method on LC-MS/MS peptide identification and quantification. (A) Protein concentration of an ocean filter sample quantified by corrected absorbance (Scopes) at 205 nm (NanoDrop OneC) or by the bicinchoninic acid (BCA) protein assay after cellular lysis. (B) Peptide concentration of an ocean filter sample quantified by corrected absorbance (Scopes) at 205 nm (NanoDrop OneC) or by the bicinchoninic acid (BCA) protein assay after digestion and MWCO filtration. (C) Summed MS1 intensities of an ocean filter peptides loaded onto an RP-LC column at various amounts (500ng-5000ng). Peptides were loaded onto the column based on the peptide concentrations determined by the BCA protein assay. The number of detected peptide analytes detected for each load amount is listed above each column. (D) Genus-level organismal relative abundances based on summed AUC protein abundance for each load amount based on the peptide concentrations determined by the BCA protein assay.

3.4 Separation and fractionation of complex peptide mixtures before MS.

Data generated and analyzed for the Critical Assessment of Metaproteome Investigation (CAMPI) as presented in Van Den Bossche et al., 2021. Nature Communications.

As mentioned in chapter 2, while 1D-RP-LC is robust and reproducible for separating peptide mixtures, the peak capacity of the PR-LC column may be exceeded as the sample complexity increases and may not be sufficient to separate metaproteomic samples that contain thousands to tens of thousands of peptide analytes in a single sample. For highly complex samples, additional peptide separation strategies, either with pre-fractionation of the peptide mixtures with SDS-PAGE or multidimensional liquid chromatography, are often necessary to improve the taxonomic and functional characterization of the microbial community being characterized[133]. The limitation of using additional separation strategies to increase identification rates and quantification accuracy is that it comes at the cost of increased instrument measurement time, potential sample loss, and increased resource cost. Due to the varying complexity of metaproteomic samples, the presence of interfering compounds, and the wide dynamic range of protein abundances within the sample, it is necessary to optimize liquid chromatography-based peptide separation without compromising throughput. This optimization requires the evaluation of optimal peptide loads, gradient lengths, and fractionation strategy for the specific samples based on the complexity of the microbial community to maximize data acquisition times without sacrificing the overall instrument run times.

To highlight how the functional taxonomic insights derived from MS/MS measurements are impacted by the LC separation strategy applied, different metaproteomic samples of varying microbial complexity were measured by several common LC separation techniques. This work was conducted as part of an international multi-lab comparison of metaproteomic workflows using two samples: a simplified, laboratory-assembled human intestinal model and a human fecal sample. The low complexity microbial community is a simplified mock community simulating

the gut microbiome (SIHUMIx)[134] composed of eight organisms initially isolated from the human gut. Since the separation time on a 1D-RP-LC directly impacts the number of peptides and proteins detected, this mock community was measured via 1D-RP-LC-MS/MS with a linear gradient of 120 minutes or 240 minutes. At the protein level, 2,674 and 3,636 proteins were identified for the 120- minute and 240-minute linear-gradient measurements, respectively **(Figure 3-7A).** Doubling the retention time of the LC separation led to a 36% increase in protein identifications. 61% of all proteins identified were found in both samples. 7% of proteins were only found in the 120-minute gradient sample, and 32% of identified proteins were only found in the 240-minute gradient measurement. **Figure 3-7B** shows the protein abundance distribution of organisms in the community. At the community level, this increase in protein identification enables more accurate quantification where the relative protein abundance of each organism more closely matched the abundance distribution of the stabilized community assessed in the original study of this community[134].

For some lower abundance community members, such as *Clostridium ramosum*, the increase in protein identifications allowed in-depth characterization of metabolic activity that would have otherwise been missed using a shorter gradient. Among identified proteins that are unique to *C. ramosum*, identifications increased from 156 protein groups to 492 protein groups when the linear gradient was doubled **(Figure 3-8A).** At a functional level, these identification gains lead to functional evidence of this member in the community at a pathway level. **Figure 3-8B** shows the KEGG pathways of *Clostridium ramosum* proteins identified in each of the two samples. As noted in the figure, several pathways were only identified with deeper measurements. In addition, **Figure 3-8C** shows GO terms related to biological processes found in proteins exclusive to the 240-minute gradient sample. Many of the proteins that were unique to the extended gradient run were lower abundance proteins that matched to GO terms related to stress response pathways such as a response to heat, methylglyoxal biosynthetic process, protein refolding, and oxidation-reduction processes. At a high level, detecting low abundance stress-related proteins in lower abundance members may give more insight into the overall dynamics of the

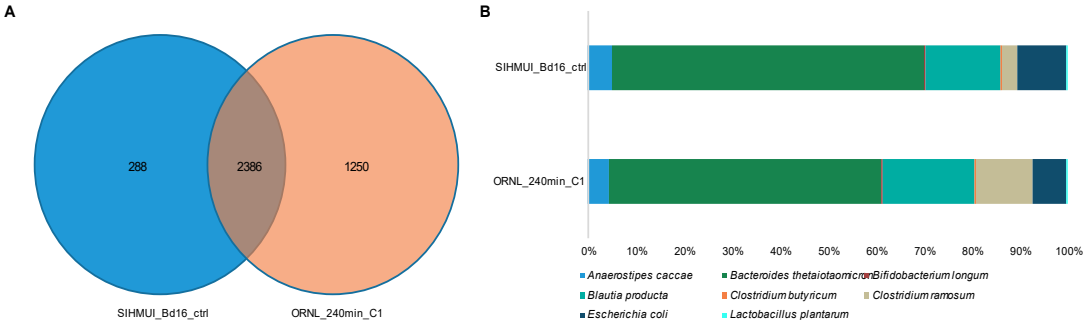

Figure 3-7 Comparison of 1D-RP-LC measurements of SIHUMIx mock community. (A) Venn diagram showing the overlap of identified proteins in the 120-minute linear-gradient measurement (SIHMUI_Bd16_ctrl) and the 240-minute linear-gradient measurement (ORNL_240min_C1). (B) Organismal relative abundances based on summed AUC protein abundance.

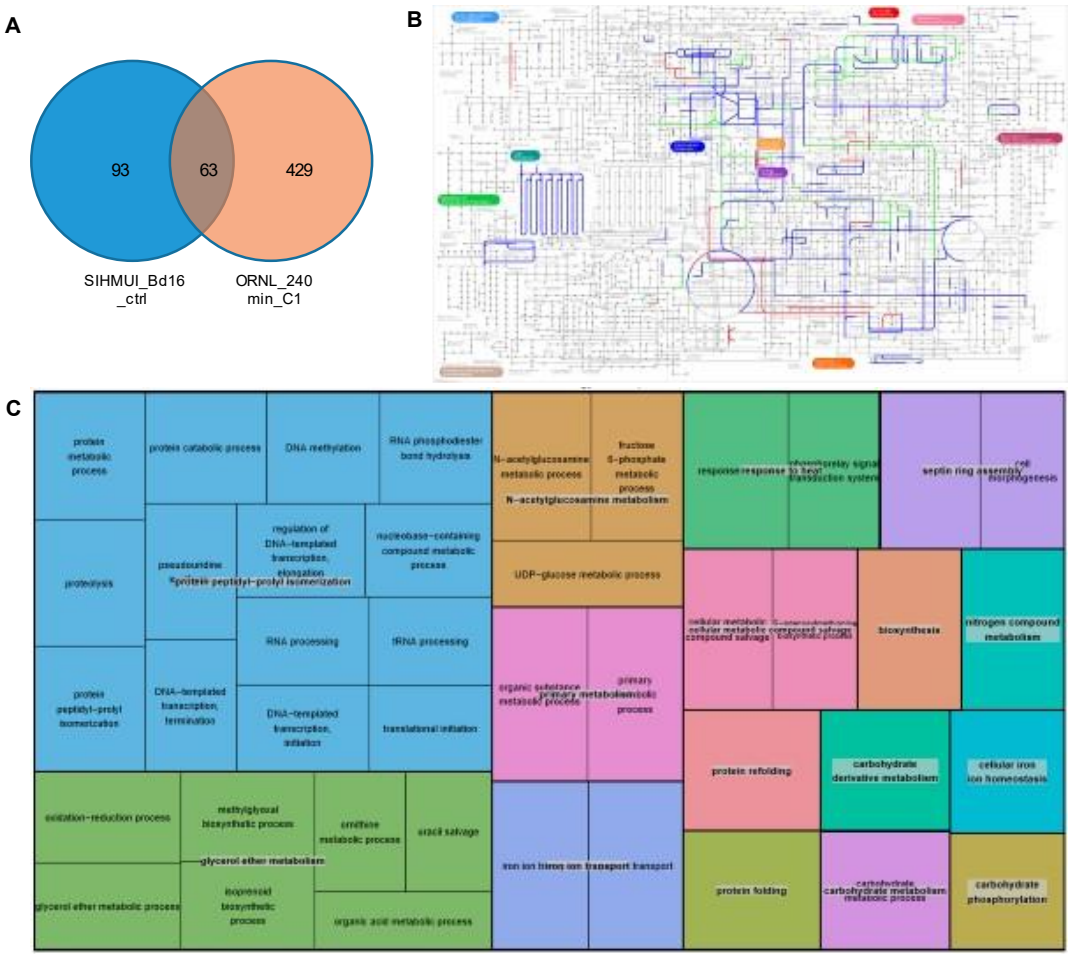

Figure 3-8 Functional profiles of identified Clostridium ramosum proteins. (A) Venn diagram of C. ramosum proteins identified in each LC gradient (B) KEGG pathways of Clostridium ramosum proteins. Pathways noted in blue were exclusively identified in the 240-minute gradient run. Pathways noted in blue were identified solely in the 120-minute gradient run. Pathways noted in green were identified in both LC measurements. (C) GO terms related to biological processes found in proteins exclusive to 240-minute gradient run. The colored blocks are clusters of semantically similar GO terms (195 terms identified in the sample). The term listed in each box is representative of the cluster based on its semantic 'uniqueness' compared to the other terms in the cluster. The size of the box refers to the frequency the representative GO term appears in the UniProt-GOA annotation database. Some of the boxes have the same color if they can be further collapsed into super-clusters of loosely related terms. The super-cluster is also given a representative GO term based on uniqueness.

community than the detection of high abundance proteins in a dominant member.

To compare the incorporation of orthogonal separation before RP-LC, three LC set-ups were used to measure a fecal sample from a healthy adult donor. This sample has an unknown composition of microbial cells, host cells, plant-derived fibrous materials, and other abiotic components, all of which increase the complexity of the sample. Two of the conducted measurements incorporated SDS-PAGE separation of the peptide mixture prior to LC-MS/MS measurement. The first SDS-PAGE sample contained one gel band measured over a 120-minute RP-LC gradient. The second SDS-PAGE sample contained four gel band fractions, each measured for 60 minutes by RP-LC for a combined separation time of 240 minutes. In addition to SDS-PAGE fractionation, the same fecal sample was processed and measured by a 4-step MudPIT strategy, with peptide eluted off the SCX resin in four pulses of increasing salt concentration followed by 210 minutes RP-LC gradient.

Figures 3-9A-C show the database search results at the PSM, peptide, and protein levels. At the PSM level, both SDS-PAGE gel band samples outperformed the MudPIT sample, with more than 5,500 more PSMs detected in each of these samples compared to the MudPIT sample. However, many of these additional PSMs mapped to peptides with multiple PSMs, so the SDS-PAGE samples did not outperform the MudPIT samples at the peptide and protein levels. In addition, while all three samples identified around 210 human proteins, the MudPIT sample had around 500 more microbial proteins identified than either of the SDS-PAGE gel band samples. In terms of protein quantification, these additional microbial protein identifications meant that 90% of the summed protein intensity was microbial in origin, compared to 81-82% in the SDS-PAGE samples (**Figure 3-9D**). Comparing the two SDS-PAGE gel band samples to each other, while the protein abundance ratios between human and microbial proteins were similar, incorporating fractionation of the gel bands enabled the identification of 180 additional microbial proteins. Preparing SDS-PAGE gel bands for LC-MS/MS analysis is relatively labor-intensive. It requires a lot more handling of the samples prior to LC-MS/MS measurement compared to the MudPIT separation approach. Therefore, the chance of sample loss and contamination

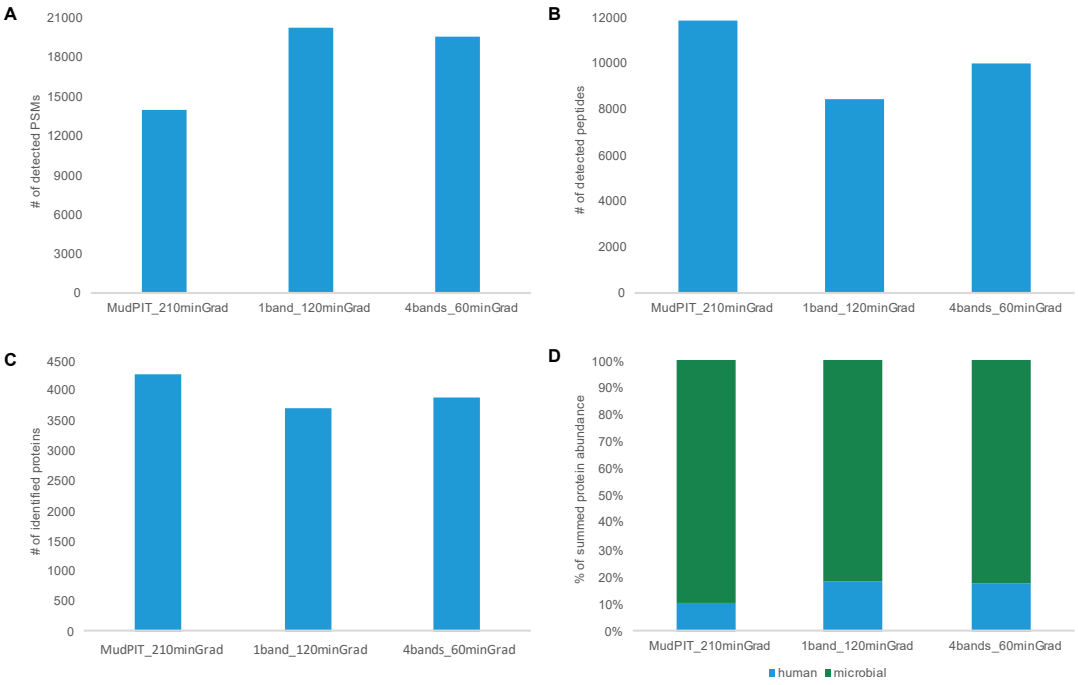

Figure 3-9 Identified PSMs, peptides, and proteins in a human fecal sample using different LC separation approaches. (A). Detected PSMs per sample (B). Detected peptide analytes per sample (C). Identified proteins per sample, clustered at 100% amino acid sequence identity. (D). Relative abundance of microbial and human proteins per sample, based on summed peak apex MS1 intensities.

of samples with common mass spectrometry contaminants such as human keratins increases with the SDS-PAGE approach, which may explain why fewer microbial proteins were identified in these samples in the presence of highly abundant host proteins. In addition, the increased measurement time by both the MudPIT approach and the gel band fractionation approach gave an additional level of separation that led to a higher number of identifications. Without this extra level of separation, several lower abundance peptides were not measured or quantified, including many low abundance microbial peptides. In total, while additional LC separation approaches add to the overall measurement time, for complex microbiome samples with a wide dynamic range of protein abundances, these approaches may be necessary to detect low abundance proteins of interest and low abundance organisms in the community.

In total, these measurements were conducted as part of a community-driven, multi-laboratory comparison in metaproteomics. This laboratory intercomparison across several different labs with very different workflows showed that for metaproteomic samples, variability of peptide level identifications is predominantly due to sample processing, with bioinformatic pipelines making a smaller impact on the identification rates. This finding was observed regardless of the complexity of the community. In general, sample preparation strategies with more extensive sample fractionation and faster instruments correlated with deeper analytical depth. Thus, separation and fractionation levels should be considered when measuring complex microbiome samples. Despite the variation in peptide-level identification rates, this multi-laboratory comparison found that much of the variation between workflows disappeared when viewing the data at the protein and functional profile level.

Chapter 4 - Challenges of metaproteomic datasets and informatics approaches.

Data presented in this chapter was generated for the following published journal articles:

Van Den Bossche, T., Kunath, B. J., Schallert, K., Schäpe, S. S., **Peters, S.L**., Abraham, P. E., Armengaud, J., Arntzen, M. Ø., Bassignani, A., Benndorf, D., Fuchs, S., Giannone, R. J., Griffin, T. J., Hagen, L. H., Halder, R., Henry, C., Hettich, R. L., Heyer, R., Jagtap, P., Jehmlich, N., … Muth, T. (2021). Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. Nature Communications, 12(1), 7305–7305. https://doi.org/10.1038/s41467-021-27542-8

S.L.P contributions include metaproteomic measurements, data analysis, writing and editing of the original manuscript and response to reviewers.

Patnode, M. L., Beller, Z. W., Han, N. D., Cheng, J., **Peters, S. L**., Terrapon, N., Henrissat, B., Le Gall, S., Saulnier, L., Hayashi, D. K., Meynier, A., Vinoy, S., Giannone, R. J., Hettich, R. L., & Gordon, J. I. (2019). Interspecies Competition Impacts Targeted Manipulation of Human Gut Bacteria by Fiber-Derived Glycans. *Cell*, 179(1), 59–73.e13. https://doi.org/10.1016/j.cell.2019.08.011

S.L.P contributions include metaproteomic measurements, data analysis, editing of the original manuscript, revisions, and response to reviewers.

4.1 Introduction to adapting proteomic approaches for metaproteomic datasets.

Many computational challenges associated with processing datasets for a single proteome become even more challenging for metaproteomes as sample complexity increases. Several database search features designed to address these issues in single proteomes need to be carefully evaluated and adapted for metaproteomic samples to ensure accurate interpretation of both the functional and taxonomic results. Over the course of this dissertation, several back-end bioinformatic approaches have been evaluated for compatibility with metaproteomic datasets and implemented to improve data analysis of the collected measurements for projects in chapters 5-7. Database search features that were investigated include aspects related to peptide identification, quantification, and protein inference. The primary data analysis platform under investigation was Proteome Discoverer, which is a Thermo Scientific data analysis platform used to analyze spectral data from a wide variety of mass spectrometers. This platform is compatible with metaproteome datasets, including datasets from fractionated LC-MS/MS measurements, which are common in metaproteomics. It has several features designed to enable increased proteome coverage and increased measurement depth independent of instrumentation advancements.

4.2 Evaluating the impact of database search strategies on peptide identification rates.

Data generated and analyzed for the Critical Assessment of Metaproteome Investigation (CAMPI) as presented in Van Den Bossche et al., 2021. Nature Communications.

The choice of database search strategy implemented can influence the outcome of the number of identified peptides and proteins in a metaproteome. Bioinformatic challenges related to database searching are exacerbated in metaproteome samples compared to single proteomes due to the extremely large search space associated with

metaproteomes. The bioinformatic pipeline selected must be able to operate with very large protein databases containing up to millions of proteins and very large data files that are often generated for metaproteomic measurements, including fractionated data if additional LC separation was necessary for sample measurement. In addition, protein inference becomes particularly challenging with metaproteomes due to the number of homologous proteins that may be present in a sample[135,136], thus the protein inference implementation by each search strategy needs to be able to resolve protein redundancies. To assess the impact of database search algorithms on the identification rates, the three simplified defined community samples used for measurement evaluations in Chapter 3 were compared using three distinct search platforms: Crux/moFF, MyriMatch/IDPicker (MM/IDP) and MS Amanda 2.0/Elutator).

The Crux software toolkit is an open-source search workflow developed to analyze a variety of types of shotgun proteomics data[136], and different search algorithms and PSM validation tools can be incorporated into workflows. For our analysis, we used the Tide search algorithm, which implements SEQUEST-style searching[137], for peptide identifications, and Percolator for PSM scoring and protein inference[138]. The identified peptides were then quantified using modest Feature Finder (moFF), which is an open-source tool for quantification of label-free proteomics data[139]. The MyriMatch database search is another open-source database search algorithm[140] that can be paired with the PSM validation and protein assembly tool IDPicker[141]. The MS Amanda 2.0 search algorithm is paired with the Elutator tool for PSM scoring and validation[142]. This is one of the search pipelines implemented in the Proteome Discoverer platform.

To enable a reliable and accurate comparison of the three database search pipelines, we fixed the search parameters such as minimum lengths of peptides and the number of tryptic miscleavages allowed, as well as parameters related to the instrumentation (mass tolerances of precursor and fragmentation ions, etc.). **Figure 4-1** shows the variation in the identifications between the three search pipelines at the PSM, peptide and protein levels. In general, while there is more variability between search platforms at the PSM and peptide analyte levels, all three platforms were fairly

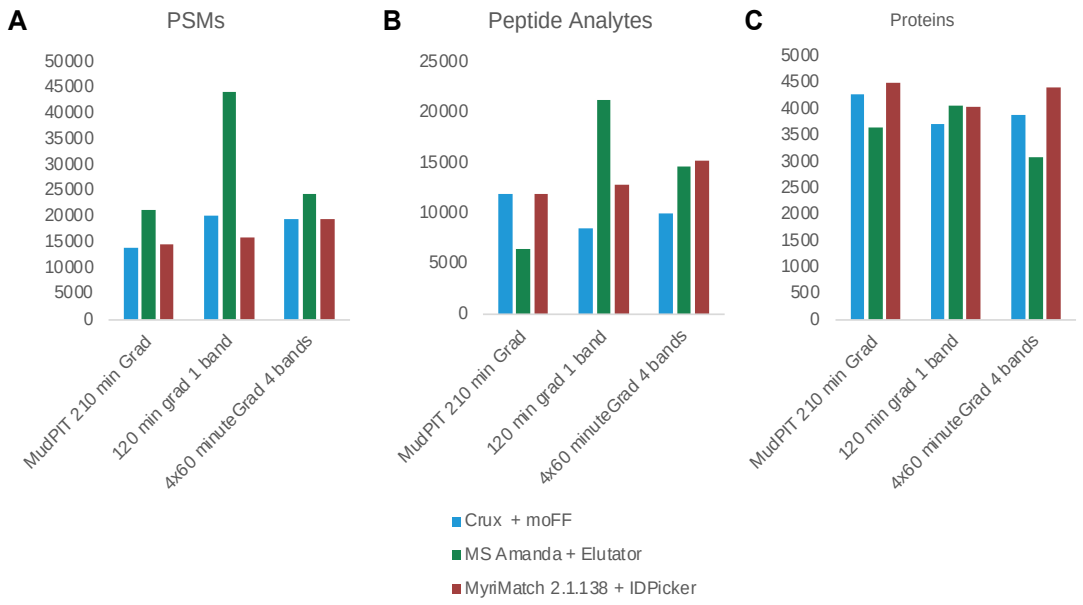

Figure 4-1 PSMs, peptides analytes, and proteins identified from the SIHUMIx mixture by three different search algorithm and PSM validation tools. Each search pipeline was compared to three LC-MS/MS measurements of the defined microbial community (a MudPIT measurement with four salt cuts and a 210-minute RP-LC gradient, an SDS-PAGE gel band measured over a 120-minute RP-LC gradient, and a fractionated SDS-PAGE gel band sample with four gel band fractions each eluted over a 60-minute RP-LC.

consistent at the protein level. Interestingly, at the protein level, there was the least variation between search pipelines for the unfractionated SDS-PAGE sample. The larger variation between search pipelines for both fractionated samples may be due in part to parameters implemented into some of the search platforms related to factors such as retention time realignment, which would be more likely to impact fractionated samples compared to unfractionated samples with shorter gradients.

To elucidate differences between search algorithms, the unfractionated sample was more closely examined to look at the protein identification overlap and detected protein abundance. For this sample, twenty-five percent of identified proteins for the unfractionated SDS-PAGE gel band sample were shared between the three search pipelines **(Figure 4-2A).** Among the identified proteins, MM/IDP and the MS Amanda/Elutator pipelines both identified around 300 more proteins than the Crux/moFF pipeline. In addition, an overlapping histogram of the relative protein abundances of identified proteins in the sample by each search algorithm (**Figure 4-2B**) shows that the dynamic range of proteins identified with MS Amanda/Elutator pipeline in Proteome Discoverer is the largest of the three pipelines, with six orders of magnitude difference between the lowest abundance and highest abundance proteins. The Crux/moFF search workflow has smallest dynamic range but identifies more proteins in the lower abundance region compared to the other two workflows. In summary, the choice of database search approach has an impact on the peptides and proteins identified in a sample. As complex microbiome samples have dynamic ranges at both the organism and protein level, the fact that the MS Amanda/Elutator pipeline in Proteome Discoverer can detect a dynamic range of six orders of magnitude makes it particularly amenable to this sample type. However, factors beyond the choice of search algorithm that are implemented in database search pipelines, such as the ability to identify peptides from chimeric spectra and how quantification is applied, should also be considered when choosing a database search strategy for a particular sample and LC-MS/MS measurement type.

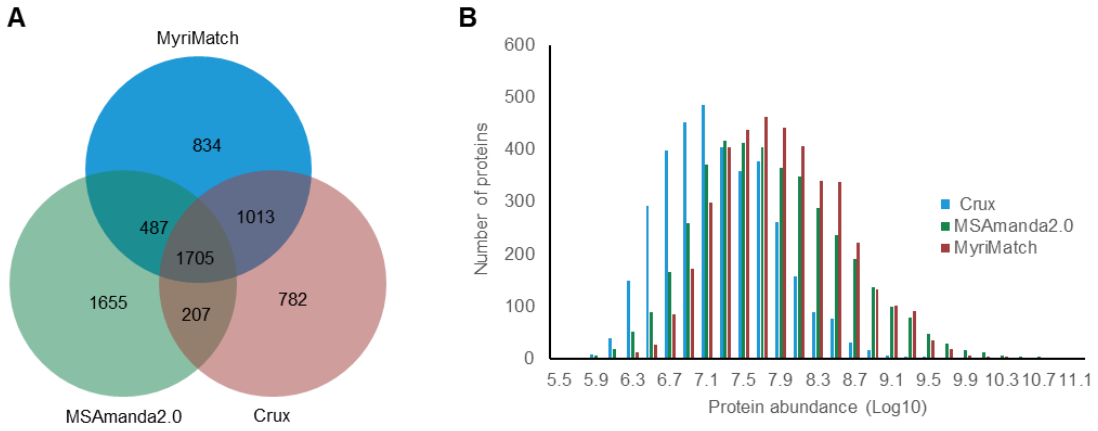

Figure 4-2 120-minute gradient unfractionated SDS_PAGE gel band identified proteins. (A) Venn diagram recording the number of identified proteins by pipeline. (B) The relative amount of protein abundances with each search algorithm.

4.3 Identification of peptides from chimeric spectra.

Data generated, analyzed, and adapted from Patnode et al., 2019. Cell.

In typical DDA-based bottom-up shotgun proteomics experiments, precursor ions are selected for MS/MS fragmentation within a narrow isolation window of 1-3 m/z. Due to the complexity of peptides present in a mixture, especially in metaproteomic samples, there are often many peptide ions with similar m/z that co-elute at any given retention time. These peptides are co-isolated within the same m/z window and co-fragmented **(Figure 4-3)**. The resulting MS/MS spectra of the co-fragmented peptides are known as chimeric spectra[143]. Advancements in instrument precision, chromatography separation schemes, and narrower isolation windows can reduce the number of chimeric spectra present in a measurement of a complex peptide mixture. However, analyses have shown that even within a peptide mixture of a single proteome, up to 50% of MS/MS spectra collected are chimeric[63,143,144]. Chimeric spectra reduce peptide identification rates and present a computational challenge as most database search algorithms operate under the assumption that there is one peptide match per MS/MS spectrum[144].

To address the low peptide identification rates caused by chimeric spectra, tools have been developed to identify and validate multiple peptide matches for an MS/MS spectrum (mPSM)[145]. Among these tools is the CharmeRT workflow[142], which first identifies chimeric spectra using a second search approach with the MS Amanda database search algorithm[146] and validates the detected PSMs using retention time prediction with the tool Elutator. MS Amanda differs from other search algorithms as it was developed for searching high-resolution tandem mass spectrometry data, leveraging both high mass accuracy and consideration of fragment ion intensities[147]. The second search approach strategy involves removing all peaks of the highest-scoring PSMs after the initial search is completed before re-searching the data for the detection of additional PSMs. To validate detected mPSMs, Elutator was built upon the principles of Percolator[148]. Elutator has been optimized for MS

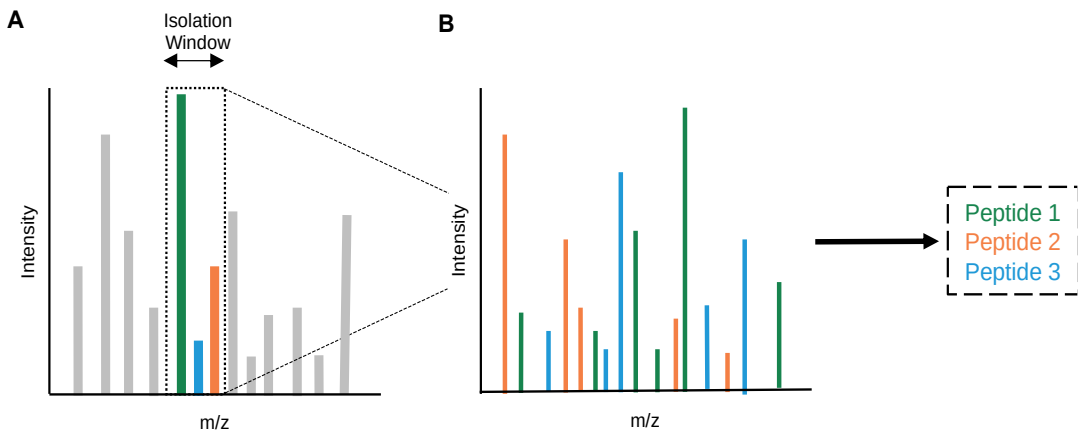

Figure 4-3 Co-isolation and co-fragmentation of precursor ions in LC-MS/MS bottom-up proteomics experiments. Chimeric spectra are generated by the co-isolation of precursor ions with similar m/z that are selected for MS/MS fragmentation. (A) MS1 spectrum of co-eluting peptide ions, including three precursor ions within the same narrow m/z isolation window selected for MS/MS fragmentation. (B) MS/MS spectrum of fragment ions from the three co-eluted and co-isolated precursor ions. These fragmentation ions belong to three distinct peptides.

Amanda search results using several features such as the difference between the measured and predicted retention time of an eluted peptide and recalibration of masses for all precursor ions and fragment ions.

To assess the impact of chimeric spectra on the peptide identification rate in metaproteomic samples, the CharmeRT workflow was compared to workflows that only identify one peptide per MS/MS spectrum using MS data acquisition file collected for gnotobiotic mouse feces. Details about these samples are further described in detail in Chapter 5.4. In brief, germ-free mice were colonized with a defined community of up to fourteen bacteria. For this analysis, samples containing the full community were compared to drop-out samples, where one of the fourteen members was omitted from the community. For a single fecal sample collected from a mouse fed a pea fiber supplemented diet, six database search workflows were compared (**Table 4-1**).

The MM/IDP workflow has previously been used for peptide identification in fecal metaproteomes[49,149,150]. The AP workflow utilizes the MS Amanda search engine, which is capable of detecting multiple PSMs per MS/MS spectrum and Percolator, a commonly used PSM validation tool. cRT0, cRT1, cRT2, and cRT3 are different executions of the CharmeRT workflow, using Elutator for PSM validation. Each of the CharmeRT workflows differs by the number of additional searches implemented.

Each of the six workflows performed similarly in terms of the distribution of peptide sequence lengths for peptides containing more than 17 amino acids. However, there were differences in identification rates between search pipelines for smaller peptides (<17 amino acids), with the greatest differences between workflow for identification peptide of peptides with fewer than 10 amino acids. Among the CharmeRT workflows, additional searches to identify peptides resulting from chimeric spectra were particularly beneficial for peptides with lengths of 10-15 amino acids. (**Figure 4-4**)

Among the identified peptides, the largest gains in identification rates were based on the implementation of the MS Amanda 2.0 search engine and PSM validation

Table 4-1 Comparison of database workflows

Workflow	Database Search Algorithm	PSM Validation Tool	# of Additional Searches
MM/IDP	MyriMatch	IDPicker	0
AP	MS Amanda 2.0	Percolator	0
cRT0	MS Amanda 2.0	Elutator	0
cRT1	MS Amanda 2.0	Elutator	1
cRT2	MS Amanda 2.0	Elutator	2
cRT3	MS Amanda 2.0	Elutator	3

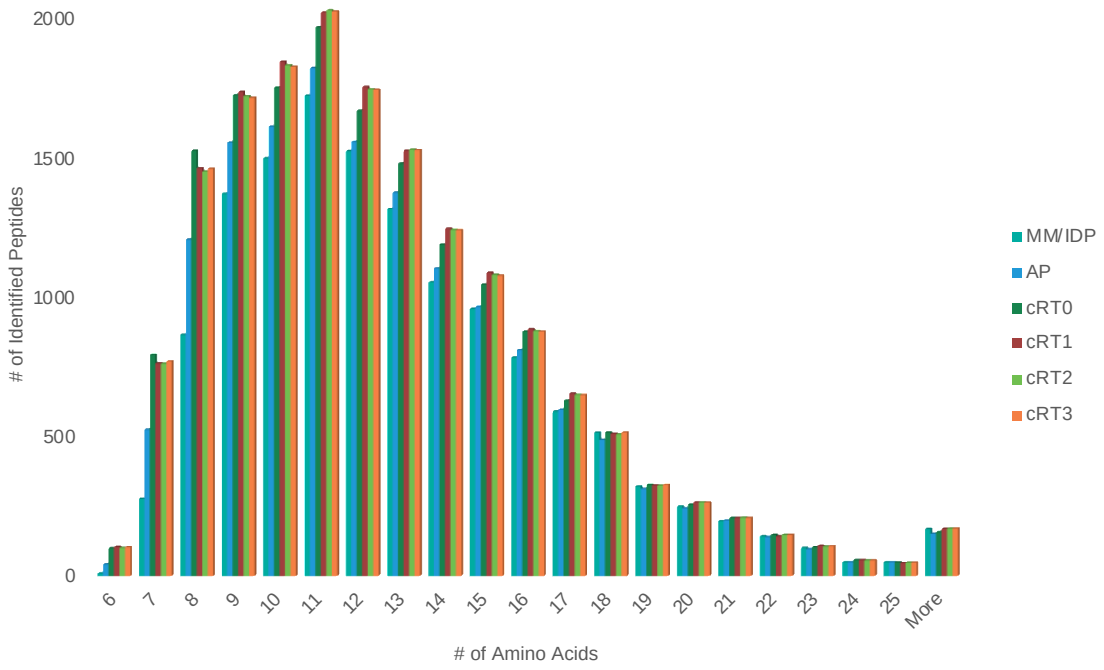

Figure 4-4 Distribution of peptide sequence lengths (number of amino acids) of peptides identified in each database search workflow.

using Elutator to leverage precursor ion mass recalibration and retention time prediction. Adding one additional search in the CharmeRT workflow (cRT1) increased peptide identifications to 16,879 from 16,538 peptides identified in the initial search (cRT0).

Transitioning from a traditional database search pipeline (MM/IDP) to a pipeline tailored for the identification of chimeric spectra from high-resolution MS/MS data (cRT1) increased peptide identifications by 23% **(Figure 4-5A).** At the protein level, the observed identification trends between search workflows were similar to the performance trends at the peptide level **(Figure 4-5B).** At the protein level, the number of quantified proteins increased by 42% with the cRT1 workflow compared to the MM/IDP workflow. Incorporating a second search strategy to identify additional peptides from chimeric spectra resulted in the identification of nearly 300 additional proteins with quantitative information in the fecal sample.

In total, 5,355 distinct proteins were found in this sample using either the MM/IDP workflow or the cRT1 workflow. Only 55% of proteins were shared between the two search pipelines. 1,918 proteins were only found using the cRT1 workflow compared to 477 proteins exclusively found in the MM/IDP workflow. Since additional peptides identified from chimeric spectra are typically low abundance analytes, the identification of these peptides improves metaproteome measurement depth and measurement reproducibility. **Figure 4-6A** shows the number of quantified proteins identified in the fecal sample using either the MM/IDP or cRT1 workflow. The identification of more low abundance peptides that uniquely map one protein in the metaproteome by searching chimeric spectra resulted in roughly double the quantifiable proteins for some low abundance organisms in the community, such as *Odoribacter sphanchnicus* and *Ruminococcaceae* species. In addition, for some bacteria with a high level of proteome redundancy, such as *Bacteroides finegoldii* and *Bacteroides massiliensis*, the increased measurement depth using second search strategies enabled the identification of peptides that uniquely map to proteins from specific organisms. Figure **4-6B** shows that increasing the number of unique peptide identifications with a second search approach results in microbial proteins

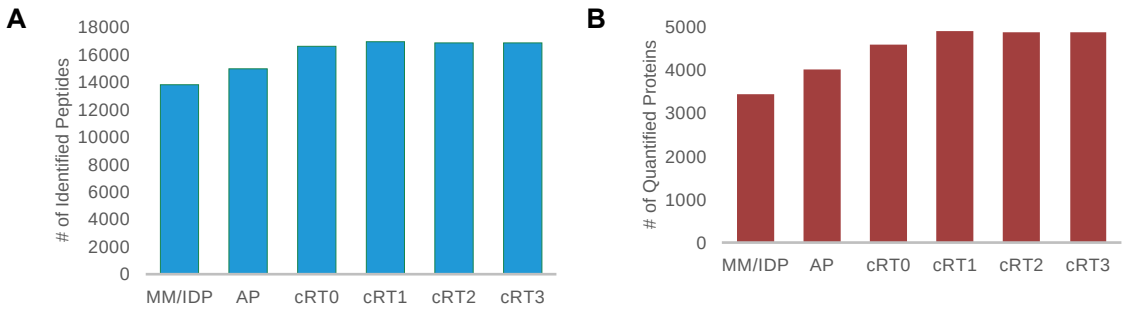

Figure 4-5 Comparison of database search strategy identifications of a gnotobiotic mouse fecal sample. (A) Peptides identified using each database search workflow. (B) Proteins quantified from unique peptides identified in each database search workflow.

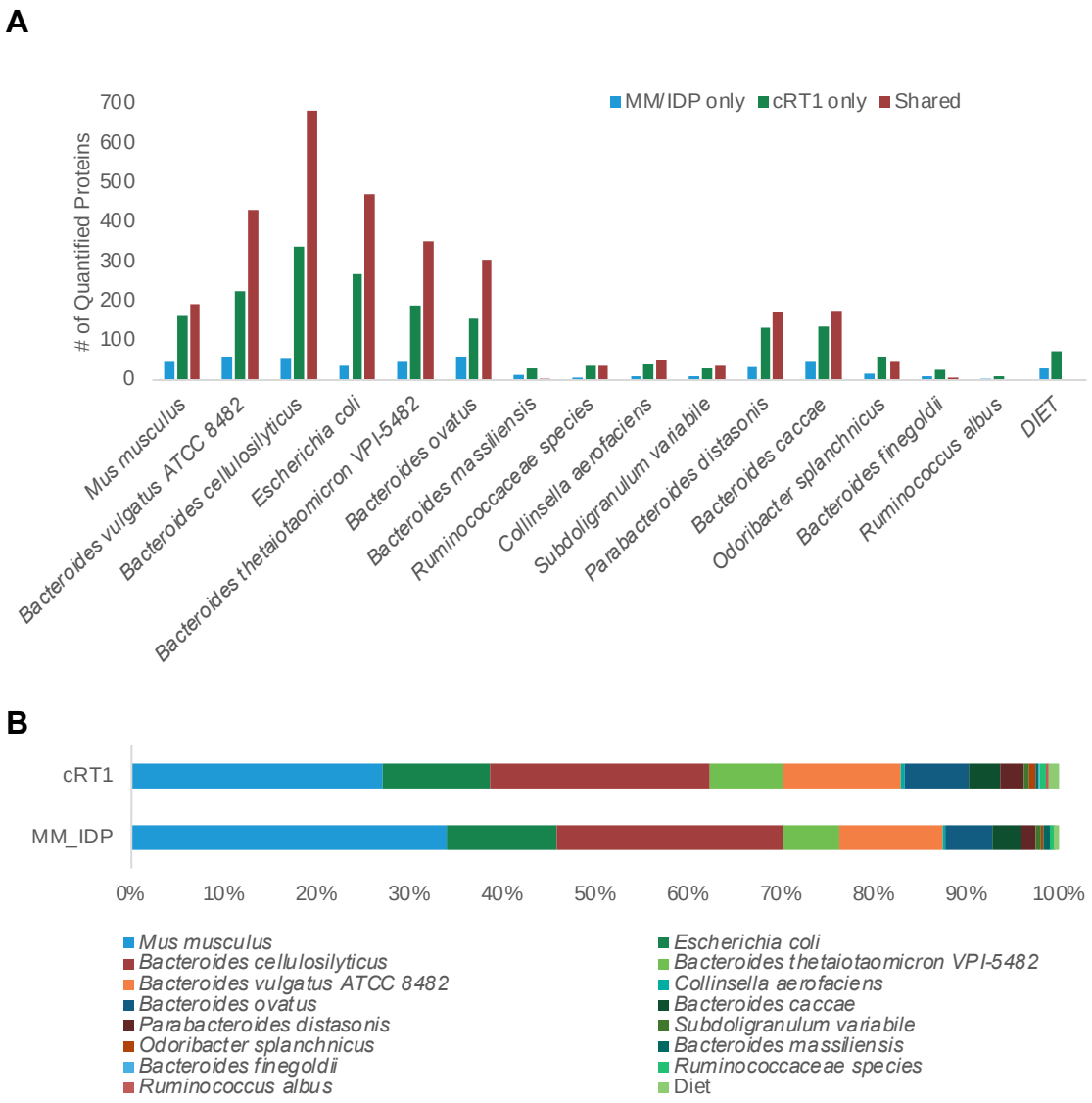

Figure 4-6 Community-level comparison of database search strategy identifications of a gnotobiotic mouse fecal sample. (A) The number of quantifiable proteins for each organism in the defined community found exclusively in the MM/IDP or cRT1 workflows or found using both workflows. (B) Relative organismal abundances based on protein abundances derived from uniquely mapping peptides.

contributing to more of the overall sample abundance. When there are more uniquely mapping peptides detected in a metaproteome sample, the microbial origin of proteins is more likely to be resolved during the protein inference process, resulting in a protein data matrix that more closely resembles the true proportions of organismal abundance in a community.

One caveat with implementing second search strategies is that peptide quantification can be inaccurate in some database search platforms, such as Proteome Discoverer, based on how quantitative MS1 intensities are matched to peptides identified from MS/MS spectra. Additional testing of the second search approach was conducted using defined metaproteome mixtures with exclusion of some proteomes from the mixture in some samples. The drop-out samples (where *B. cellulosilyticus* was omitted from community) showed that while peptide identifications of mPSMs from chimeric spectra based on the MS/MS fragmentation events were accurate, the quantitative MS1 peak intensities (consensus features) of the precursor ions associated with chimeric spectra were only accurate for the most abundant precursor ion that was co-fragmented. Specifically, we found that the Proteome Discoverer platform assigned the consensus feature derived for the most abundant precursor ion to every precursor ion identified in a given chimeric spectrum. Using the defined microbial mixtures, we found that the misassignment occurs irrespective of the complexity of the sample but seems to be exacerbated with complex metaproteomes with high redundancy. We found in the drop-out community validation experiments those missing members appeared to be prevalent members of the community when a second search strategy (cRT1, cRT2, cRT3) was implemented for quantitative values. Therefore, with the current strategy used by the Proteome Discover platform to match consensus features to MS/MS identifications, second search approaches to detect additional peptides derived from chimeric spectra should only be used for peptide identification and not peptide quantification for metaproteome samples.

4.4 Peptide quantitation and protein inference.

In addition to increasing the number of peptides identified in metaproteome samples, we assessed the accuracy of emerging tools used for peptide quantification and protein inference. Many aspects of peptide and protein inference are evaluated with data collected from measurements of single proteomes. Validation of platforms that perform peptide quantitation using single proteome datasets may not be appropriate to address challenges associated with complex metaproteomes with high protein redundancy. Assessment of critical database search decisions for protein inference and quantitation in metaproteomics using "ground-truth" defined microbial communities is necessary when implementing new database search strategies for metaproteomics. To evaluate peptide quantification and protein inference aspects for metaproteomics, we used a gnotobiotic mouse model composed of the host and fourteen colonized microorganisms including several *Bacteroides* species. Measurements were made for the evaluation of peptide quantitation and protein inference with some samples composed of all members in the community and some samples composed of the same organisms, but with *Bacteroides cellulosilyticus WH2* removed from the community.

4.4.1 Match-between-run (MBR) for peptide quantification.

Data generated, analyzed, and adapted from Patnode et al., 2019. Cell.

Peptide quantification for bottom-up DDA shotgun proteomics experiments is typically based on label-free analyses, in which the peptide's m/z and chromatographic information are leveraged. Stochasticity is an important issue for quantitative multi-run label-free data-dependent acquisition[151] and especially for metaproteomic datasets. Due to the stochastic nature of sampling, there are often lots of missing values in the quantitative data matrix for metaproteomics. Traditionally, statistical imputation is implemented to fill in the missing values in the quantitative data matrix in order to

perform downstream statistical analyses of the data. An alternative to statistical imputation that has gained popularity over the past several years is to perform an identification transfer, known as match-between-runs (MBR) between all samples measured in the experiment. MBR is implemented to reduce missing values in the data matrix by leveraging m/z and chromatographic information to quantify peptides without MS/MS information by matching unidentified peaks in samples that have similar m/z and charge states within a narrow RT window as high-quality PSMs in other samples. **Figure 4-7** shows an example of XICs from a precursor ion found in two biological replicates of gnotobiotic mouse fecal samples. The selected ion had MS/MS sequencing events in two out of three replicates, and the corresponding MS1 peak area was found in the third replicate in the replicate set using MBR in Proteome Discoverer. These extracted ion chromatograms (XICs) demonstrate that MBR is matching real peaks in the samples where MS/MS events were not triggered to fill in the quantitative data matrix.

MBR implementation across a dataset of samples that included gnotobiotic mice colonized with the full 14-member community or with the drop-out community led to a more populated data matrix (**Figure 4-8**). Across the entire dataset, MBR implementation increased the quantitative values in the data matrix by 23%, equating to 134,379 "found" data points in the data matrix. In addition, when looking at the relative abundance of each organism in the community among biological replicates, the wobble observed in organismal abundances within replicate sets is decreased using MBR. The average RSD values of organismal abundance within replicate sets decreased by 2.7%. In total, MBR is an effective computational approach to address stochasticity issues that arise in bottom-up shotgun metaproteomics experiments and is a useful alternative to statistical imputation to fill in missing values in a data matrix.

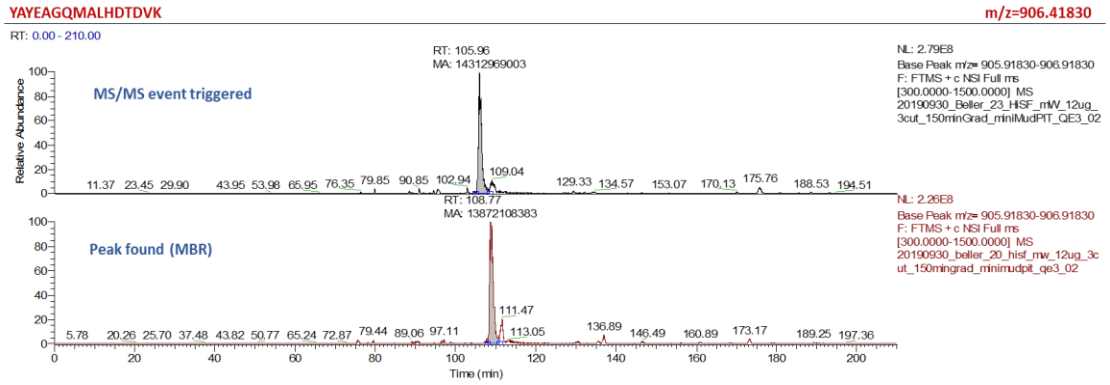

Figure 4-7 MS1 XICs of precursor ion m/z=906.41830. This is the precursor ion of peptide YAYEAGQMALHDTDVK. XICs of this precursor ion that was selected for MS/MS fragmentation (top) and found by MBR (bottom).

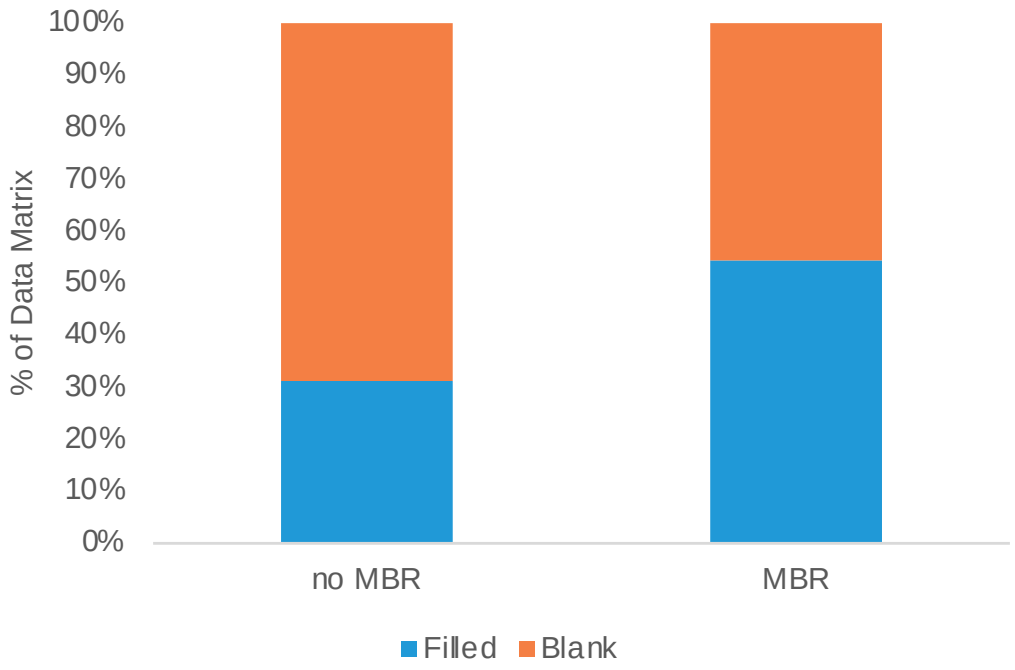

Figure 4-8 Percent of data matrix filled using MBR implementation. MBR applied across all ten samples examined, including five biological replicates for the full community samples and five biological replicates for the *B. cellulosilyticus* drop-out samples. MBR implementation increased the percent of quantitative values in the data matrix.

4.4.2 Evaluating false transfer rates associated with MBR.

Data generated, analyzed, and adapted from Patnode et al., 2019. Cell.

A previous study evaluating the authenticity of identification transfers with MBR using a two-proteome dataset composted of human cell lysates (HCT116 cells) and yeast cell lysates (*Saccharomyces cerevisiae*) showed that false transfers from the MBR algorithm implemented in the MaxQuant proteomics platform do occur at a measurable rate [152]. In this study, 44% of proteins from yeast cells in the two-proteome mix were incorrectly transferred to measurements in the human-only cells. However, the authors found that most of the false transfers were spurious and resulted from MS/MS identifications that were only found once across the entire dataset. Implementation of additional quantitation filters in this platform prevented the spurious transfers, which prevented incorrect quantitation of peptides in the dataset.

Initial testing of the false transfer rate from MBR implementation in the Proteome Discoverer platform with this same two-proteome model showed that false transfers also occur during the MBR process in Proteome Discoverer similar to the false-transfer issues observed in the MaxQuant platform. This demonstrates that the false transfer issue associated with MBR is platform agnostic. In Proteome Discoverer, these false transfers can be controlled by addressing ambiguous extracted ion chromatogram (XIC) features (consensus features) in the dataset. One way to address ambiguous consensus features is to implement a rule that a certain number of MS/MS fragmentation events must occur in the dataset in order for quantitative values to be transferred to samples that have MS1 precursor ion evidence but no MS/MS fragmentation evidence of the peptide. The specific number of MS/MS fragmentation events should be assessed for each dataset based on the complexity of the samples, the size of the overall dataset, and the number of replicates for each condition tested.

The gnotobiotic mouse model dataset was used to further evaluate MBR false transfer rates in complex metaproteomes. To assess the level of false transfers in the dataset, we looked at quantitative values assigned for *B. cellulosilyticus* peptides, since

this was the organism omitted from the community in some samples **(Figure 4-9).** MBR implementation alone increases the number of quantitative values for *B. cellulosilyticus* within replicate sets by ~20%. However, in the samples *where B. cellulosilyticus* was absent, the gains achieved by MBR were due to false transfers. Requiring at least one MS/MS fragmentation event within the replicate set for MBR values to be included in the dataset reduces the number of false transfers while still increasing the number of high-quality quantitative values in the full-community data matrix. This MS/MS fragmentation event requirement reduced the identification transfers of quantitative values for *B. cellulosilyticus* peptides from 11,185 to 45 erroneous quantitative values. The remaining quantitative values were low abundance and can be further filtered out of the data matrix with either stricter MBR thresholds or during downstream protein inference processes. Control of MBR implementation during searching reduced false transfers, while still decreasing missing values by more than five percent across the entire dataset and up to ten percent within replicate sets compared to no MBR implementation. MBR with additional filtering of matched XIC features based on MS/MS fragmentation events should be implemented to ensure stochastic hits causing erroneous transfers are minimized.

4.4.3 Benchmarking the accuracy of common protein inference strategies for metaproteomic datasets.

Data generated, analyzed, and adapted from Patnode et al., 2019. Cell.

One of the major challenges with metaproteomics is the correct assignment of peptides into proteins due to the redundancy of the proteomes. When a peptide maps to several proteins across organisms, relying on the quantitative values of ambiguous identifications can lead to incorrect interpretation of functional and taxonomic results. For example, in a concept known as protein “hitchhiking”, proteins included in the database used for database searching that are not actually present in the measured sample can “hitchhike” and be inferred as present in the samples if they share peptides

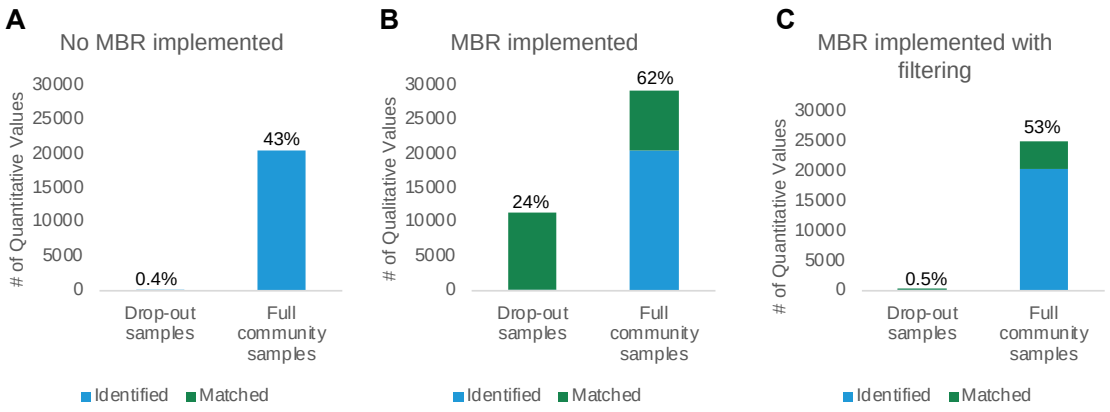

Figure 4-9 ***B. cellulosilyticus*** **peptides utilizing different implementations of MBR.** Percentage values show the proportion of the respective data matrix that is filled with quantitative values. MBR implemented with additional post-search filters (minimum MS/MS fragmentation events per replicate set requirement) led to the highest increase in quantitative values with minimal false transfers.

with other proteins that are present in the sample **(Figure 4-10).** This phenomenon is common in metaproteomes, as there are often a number of closely related organisms in a community that share a number of homologous proteins[153]. Peptides are classified into two types, shared or unique peptides. Unique peptides only map to one protein in the protein database, while shared peptides can map to multiple proteins in the database. Many protein inference methods rely on the principle of parsimony in order to explain the detected peptides with a minimal set of proteins. Proteome Discoverer can implement different styles of protein inference for shared peptides which cannot be unambiguously assigned to one protein in the database. The first method of protein inference only uses unique peptides that only map to one protein in the database for protein inference. All shared peptides, which can compose more than half of the identified peptides in a metaproteome with high protein redundancy, are excluded.

The second method, based on rules of parsimony, distinguishes peptides as unique, razor, or shared. In this case, a razor peptide is defined as a shared peptide that has been assigned to a protein with the largest number of total peptides identified and is not assigned to any other matching proteins in the database. Shared protein are any remaining peptides that cannot be assigned as a razor peptide. In these cases, the proteins these peptides map to have the same number of total peptide identifications. The final protein inference method available in Proteome Discoverer is to use all identified peptides for protein inference. In the case of peptides that map to multiple proteins in the database, they are assigned to all matching proteins.

To assess how protein quantities change based on the peptides used for protein inference, the same gnotobiotic mouse model colonized with 14 microorganisms, with samples composed of all members and samples lacking *B. cellulosilyticus* used to assess peptide identification from chimeric spectra and match-between-runs implementation, was used as a ground-truth dataset to validate the accuracy of protein inference strategies in the Proteome Discoverer platform. Within this community, a significant number of peptides are shared between community members, especially among the seven *Bacteroides* species in this defined community. **Figure 4-11** shows the pairwise similarities of tryptic peptidomes based on minimum similarity clustering

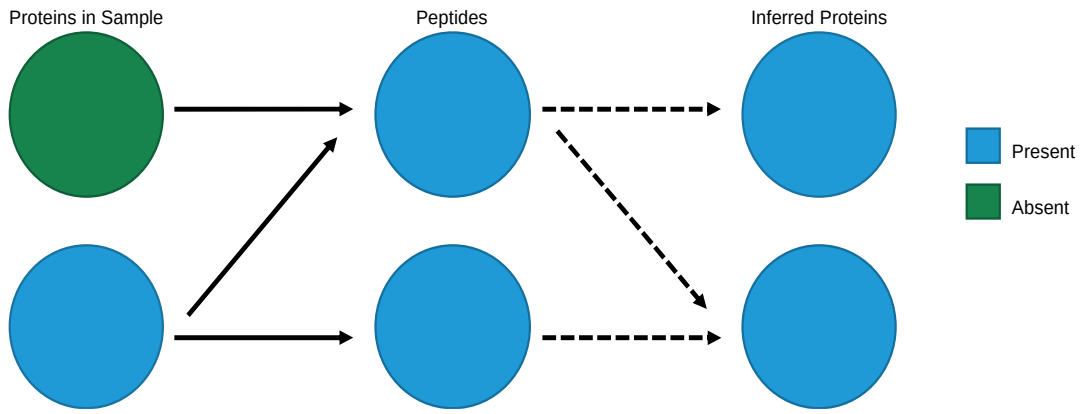

Figure 4-10 Hitchhiking proteins. During the protein inference process, proteins that are included in the protein database but are not present in the sample can be inferred as present if they share peptides with another protein that is included in the database and present in the sample.

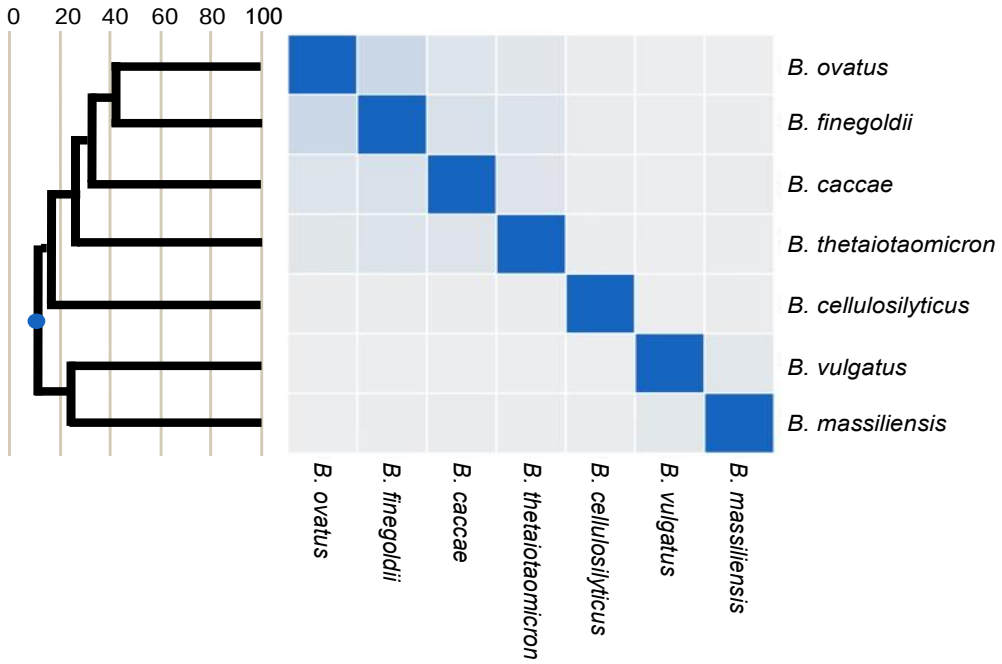

Figure 4-11 Theoretical tryptic peptidome of *Bacteroides* species in a gnotobiotic mouse model community. The theoretical tryptic peptidome is the complete set of all tryptic peptides encoded in an organism's predicted proteome. Pairwise similarities between tryptic peptidomes were calculated using the minimum similarity measure where the size of the intersection is divided by the minimum size of the two unions. On the left, a phylogenetic tree is shown based on the unweighted pair group method with arithmetic mean (UPGMA) clustering of the pairwise similarities. The x-axis on the phylogenetic tree represents the percent similarity for each pairwise similarity calculated.

of all seven *Bacteroides* in the gnotobiotic mouse dataset. Many of the theoretical tryptic peptidomes of the seven *Bacteroides* species share peptides with proteins from other *Bacteroides* species and the other seven organisms in the community, and based on the inference method used, these shared peptides may or may not be used for quantitation. For example, 40% of *B. finegoldii* predicted peptides are shared with *B. ovatus*. This overlap means that up to 34,093 shared peptides between these two organisms would be discarded during protein inference if only unique peptides are used for protein inference. However, if these shared peptides are used for protein inference, they could be inaccurately assigned to the wrong organism if only one organism is expressing the source organism of the shared peptides. Therefore, the assignment of the shared peptides significantly impacts the quantitative accuracy of these highly related organisms in microbial communities.

To assess the accuracy of the various protein inference methods implemented in Proteome Discoverer, protein inference was tested using only unique peptides, unique plus razor peptides, or all peptides regardless of whether they are unique or shared peptides. Protein inference relies on the correct assignment of peptides to proteins and makes a significant impact on protein quantitation in communities with members containing highly redundant proteomes. Protein quantitation based on ambiguous peptides can skew peptide assignment results. Choices made at the peptide level have a major impact on the interpretation of both protein-level and organism-level results in functionally redundant communities.

Figure 4-12 shows the relative distribution of abundances for the seven *Bacteroides* members in the community using each protein inference method. Even though *B. cellulosilyticus* was not present in the sample, because of its proteome redundancy with other members in the community, using razor or shared peptides for protein inference resulted in the apparent presence of this organism in the sample. Therefore, the inclusion of razor peptides and shared peptides for label-free quantitation should be avoided for high-confidence functional and taxonomic interpretation of metaproteomes.

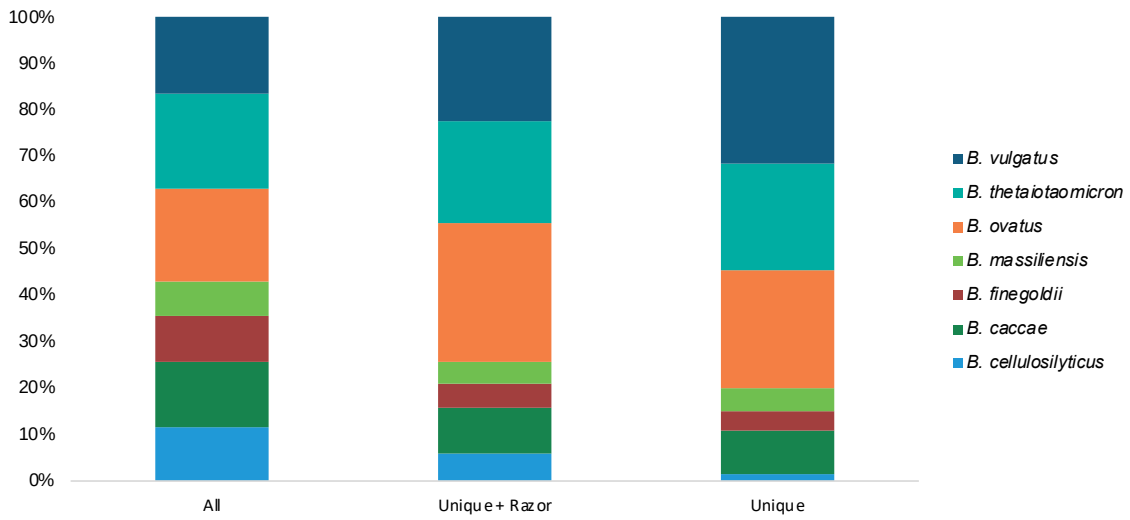

Figure 4-12 Relative summed organismal protein abundance of *Bacteroides* members in the samples lacking *B. cellulosilyticus* based on three quantitation methods. Protein quantitation based all peptides, unique and razor peptides, or unique peptides. When using razor and shared peptides, *B. cellulosilyticus* appeared to make up more than 11% of *Bacteroides* abundance, even though this member was not present in the sample.

4.5 Conclusions.

In total, decisions made during the database search process can have large impacts on the interpretation of the functional composition of the measured sample. Parameters related to peptide identification, quantification, and protein inference should be considered for every metaproteomics campaign based on the type of LC-MS/MS data collected and the functional redundancy of organism present in the sample. Careful evaluation of these parameters has resulted in database search strategy enhancements which have simultaneously improved metaproteome measurement depth and measurement reproducibility. These enhancements have provided a robust foundation to help address the biological questions within the projects covered in Chapters 5-7.

Chapter 5 - Interrogation of bacterial cooperation and competition through diet manipulation of human gut communities in gnotobiotic animals.

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

Wolf, A. R., Wesener, D. A., Cheng, J., Houston-Ludlam, A. N., Beller, Z. W., Hibberd, M. C., Giannone, R. J., **Peters, S. L.,** Hettich, R. L., Leyn, S. A., Rodionov, D. A., Osterman, A. L., & Gordon, J. I. (2019). Bioremediation of a Common Product of Food Processing by a Human Gut Bacterium. *Cell Host & Microbe*, 26(4), 463–477.e8. https://doi.org/10.1016/j.chom.2019.09.001

S.L.P contributions include metaproteomic measurements, data analysis, figure generation, editing of the original manuscript, revisions, and response to reviewers.

Wolf, A. R., Wesener, D. A., Cheng, J., Houston-Ludlam, A. N., Beller, Z. W., Hibberd, M. C., Giannone, R. J., **Peters, S. L.,** Hettich, R. L., Leyn, S. A., Rodionov, D. A., Osterman, A. L., & Gordon, J. I. (2019). Bioremediation of a Common Product of Food Processing by a Human Gut Bacterium. *Cell Host & Microbe*, 26(4), 463–477.e8. https://doi.org/10.1016/j.chom.2019.09.001

S.L.P contributions include metaproteomic measurements, data analysis, editing of the original manuscript, revisions, and response to reviewers.

Wesener, D. A., Beller, Z. W., **Peters, S. L.,** Rajabi, A., Dimartino, G., Giannone, R. J., Hettich, R. L., & Gordon, J. I. (2021). Microbiota functional activity biosensors for characterizing nutrient metabolism in vivo. *eLife*, 10(na). https://doi.org/10.7554/eLife.64478

S.L.P contributions include metaproteomic measurements, data analysis, writing and editing of the original manuscript, revisions, and response to reviewers.

5.1 Introduction.

Massive changes in food preferences in recent years have implications on our biology and the disease risks associated with those dietary patterns. One method to study the biological implications of food choice is to use diet oscillation studies to define the role of gut microbiota on host nutritional status. The core focus of the projects presented in this chapter is the interrelationships between the food we eat in Westernized diets and their interactions with our gut microbiota. Understanding of these interactions should give insights into host physiology and metabolism and may emphasize how food is a driver or modulator of health. These projects are based on gnotobiotic models of the human gut microbiota and a combination of experimental tools to look at how different dietary components added to a representative Western diet configure the microbial community structure and function. The main goal of these projects is to identify dietary components that shape gut microbial community composition and function in ways that improve host wellness. In this chapter, the proteomics and metaproteomics work is presented from three projects that were conducted in collaboration with Jeffrey Gordon's laboratory at Washington University in St. Louis. The scope of the first project is to understand how a member of the human gut microbiota (*Collinsella sp*.) responds to a common chemical modification of food (fructoselysine) introduced by food processing. Proteomics work was conducted to determine the extent of this chemical modification in the processed dietary whey fed to the mice during the study. The second project was aimed at identifying different dietary fibers and their bioactive components that increase the fitness of particular *Bacteroides* species in the context of a representative Western diet using humanized gnotobiotic mice. Longitudinal metaproteomic measurements of

fecal samples were conducted for gnotobiotic mice colonized with a defined microbial community including various *Bacteroides* members and fed a representative Western diet supplemented with food-grade fiber preparations from the byproducts of food manufacturing. The third project focuses on defining the responses of community members to two structurally distinct arabinan polysaccharides, one isolated from the endosperm of pea and the other from sugar beet. We used metaproteomics to compare the degradation of these two glycans by several *Bacteroides* spp. in the bacterial consortium when fed a glycan-supplemented representative Western diet.

5.2 Controlled investigation of microbial community dynamics in gnotobiotic animals.

One approach for mechanistic studies of the gut microbiome is the use of germ-free or gnotobiotic models. Genetically identical germ-free and gnotobiotic animals are tractable models in gut microbiome research in which precise experimental analysis can be conducted in a highly controlled environment. Studies with germ-free models can implicate whether or not the presence of microbiota in the gut has an impact on host phenotype. When the phenotype of germ-free mice is different than conventionally raised counterparts, this suggests that the microbiota are involved in the development of the phenotype. A number of studies have implicated gut microbiota in the development of numerous host phenotypes[154]. Mono-associated models, where germ-free mice are colonized with a single bacterium, can aid in identifying of bacterial species that are responsible for the production of specific products that are putative drivers of disease, as well as other host-microbe interactions. A notable example of host interactions that were elucidated through studies of mono-associated mice is the induction of host T regulatory cells by polysaccharide A that is produced by *Bacteroides fragilis*[155]. Defined and complex community models may validate the causal role of bacterial agents in a context that is more translatable to human health[156]. Typically, gnotobiotic mice are germ-free mice that have been colonized with custom-designed communities with a defined microbial membership

to answer a specific research question. These models circumvent some of the physiologic abnormalities that can be observed in germ-free or mono-associated mice[156]. In addition, germ-free mice can be humanized by colonizing with microbiota from the human gut[157,158]. In total, these models provide a complex *in vivo* system to study microbiome interactions in a highly controlled and defined environment.

5.3 Characterizing the prevalence of protein post-translational modifications that form in the manufacturing of processed foods.

Text and figures adapted from Wolf et al., 2021. Cell Host & Microbe.

5.3.1 Project scope.

Previous work has shown that fructoselysine forms during the processing of milk to whey with heat treatment through a process known as a Maillard reaction[159]. Some Maillard reaction products (MRPs) in processed foods have been associated with several diseases such as cancer and diabetes[160,161]. In this study, gnotobiotic mice colonized with 54 phylogenetically diverse human gut bacterial strains, including several Collinsella species and were fed defined sugar-rich diets with whey as the sole protein source. While *Collinsella* spp. are normal inhabitants of the human gut, increases in the abundances of some *Collinsella* species have been implicated in Type 2 diabetes disease progression[162]. In this model system, fructoselysine moieties in the whey protein isolate serve as a representative MRP which were used to illustrate how chemical modifications of processed food can impact gut microbiota and may lead to harmful impacts on host health.

To confirm that hexose-modified lysine residues exist in the whey protein isolate preparation used for diet supplementation in the gnotobiotic mouse experiments, an analysis of post-translational modifications of whey proteins was conducted. Tryptic peptides of powdered bovine whey protein isolates were subjected to multidimensional LC-MS/MS analysis.

5.3.2 Results and discussion.

In this sample, over 150 proteins were identified. The five most abundant proteins identified in the sample comprised 62% of the total peptide abundance in the sample **(Figure 5-1A).** These five proteins also showed the largest frequency of hexose modifications at any residue among all proteins in the sample, with 87% of the total hexose modifications identified in the sample belonging to these five proteins. To quantify the PTMs across the five most abundant proteins, we looked at three different views of the data: (1) The number of peptide sequences identified per protein either unmodified or modified with a particular PTM. (2) The number of peptide-spectral matches (PSM) per protein either unmodified or modified with a particular PTM. A PSM is a more quantitative measure relative the identified peptides as each peptide can have multiple PSMs measurements (i.e.MS/MS fragmentation events). This is usually correlated with the abundance of each particular peptide. (3) Area-under-the-curve (AUC) peptide abundance per modification per protein. AUC abundance is most accurate of the three views to quantify proteins and peptides and uses the chromatographic area of each peptide peak to establish peptide and therefore protein abundance.

The PTM assessment of these five proteins determined the relative frequency of the two glycosylation modifications of interest, fructosylation (one hexose with a mass shift of +162.0528 Da) or lactosylation (two hexose with a mass shift of +324.1056 Da), which occurred on lysine and arginine residues. In terms of the number of modified PSMs compared to the unmodified counterparts, the majority of PSMs were unmodified in the five proteins, with 95% of the PSMs per protein lacking a Hex(1) or Hex (2). The majority of modifications occurred on lysine residues **(Figure 5-1B and 5-1D).** At the peptide identification level, 80% of peptides were

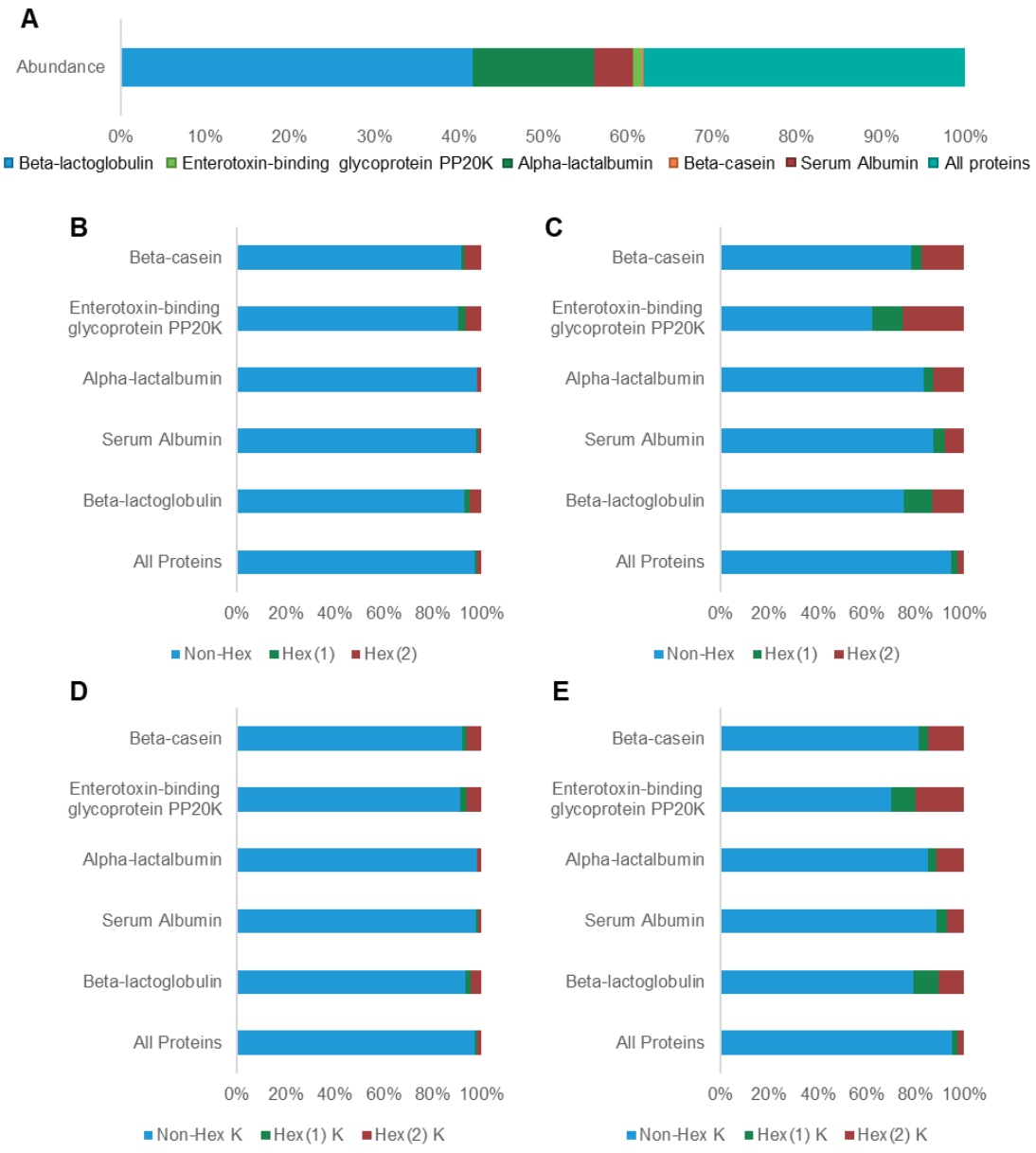

Figure 5-1 Mass spectrometry analysis of the whey protein isolate. Tryptic peptides generated from proteins in the whey protein isolate preparation. Protein abundance distribution of the sample, show the abundances of the five most abundant proteins in the sample and the combined abundance of all other proteins (A). The percentage of peptide spectral matches (PSMs) or peptides modified by zero, one or two hexoses at any residue are quantified in panels (B) and (C), respectively. Peptides modified by one or two hexoses at lysine residues were also quantified by PSMs (D) or peptides (E).

unmodified at either lysine or arginine residues **(Figure 5-1C)** and 83% of lysine residues were unmodified **(Figure 5-1E).** In general, hexose modification of proteins represented a small portion of the whey proteome compared to unmodified proteins. However, the hexose-modified proteins encompassed a large proportion of the identified PTM space per protein.

Inspection of both the number of identified peptides and the AUC abundances of peptides for each of the most abundant proteins, there was a generally low modification frequency relative to unmodified peptides. However, the hexose modifications appeared to encompass a large proportion of the total identified PTM space per protein. For example, looking at the most abundant protein, β-lactoglobulin, peptides with PTMs that were not induced by the sample preparation method represented more than half of the identified peptides. Of those peptides with PTMs, peptides with Hex(1) or Hex(2) modifications on lysine residues represented 38% of the identified peptides **(Figure 5-2A).** Looking at protein AUC abundances, only 1% of the total protein abundance came from modified peptides. However, of that small percentage, the majority of the signal was from hexose-modified peptides **(Figure 5-2B).**

In summary, proteomics confirms that the whey protein isolate being fed to the mice during the study is being chemically modified to form fructoselysine during the manufacturing process. In total, this study shows that among different *Collinsella* spp. in the defined community, both *C. intestinalis* and *C. aerofaciens* are able to convert fructoselysine to glucose 6-phosphate and lysine, and that glucose 6-phosphate is shuttled through glycolysis and exported in part as acetate and formate. In an environment containing glucose and fructoselysine versus glucose alone, *C. intestinalis* induces expression of fructoselysine utilization genes so it could use it as a carbon source, while *C. aerofaciens* does not. Additionally, *C. intestinalis* grows at a faster rate *in vitro* on fructoselysine than on glucose. Overall, this study shows a nice example of how closely related microbes are able to handle chemical modifications on food that are introduced during the food manufacturing process differently, and how this can lead to changes in microbiome structure and function.

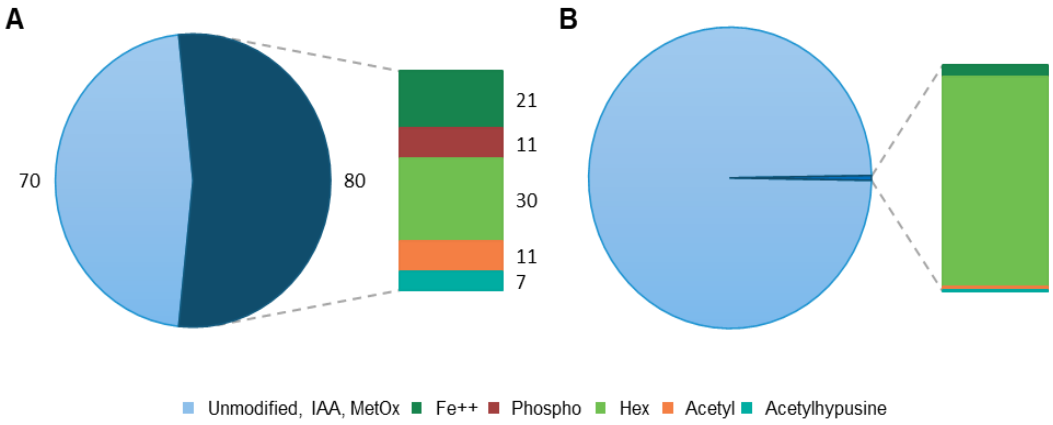

Figure 5-2 β-lactoglobulin modified peptides. The number of PTMS identified in the targeted PTM search of the glycosylation modifications of interest and the most abundant modifications identified in the open search. As oxidation of methionine residues (MetOx) and carbamidomethylation of cysteine residues (IAA) are generally artifacts of the sample preparation, these two modifications were counted with the unmodified peptides. The number of peptide sequences identified for β-lactoglobulin that were either unmodified or modified with a particular PTM (A). Area-under-the-curve (AUC) peptide abundance per modification for β-lactoglobulin (B).

5.4.3 Methods.

Powdered bovine whey protein isolate (Fonterra Co., New Zealand) was resolubilized in denaturation buffer (2% sodium deoxycholate, 100-mM ammonium bicarbonate, 10 mM dithiothreitol, pH 8.0) and incubated at 80°C for 10 minutes. The sample was then adjusted to 30 mM iodoacetamide and incubated at room temperature for 15 minutes in the dark. Five hundred micrograms of the denatured and reduced whey protein isolate were then transferred to a 10 kDa MWCO spin filter (Vivaspin 500, Sartorius) for in situ clean-up and digestion with sequencing-grade trypsin (1:50 (w/w); Pierce) using previously described conditions[163]. Tryptic peptides were collected by centrifugation (12,000 x g at 4°C). Un(der)digested proteins that remained atop the filter were re-solubilized in denaturation buffer and digested again with chymotrypsin (1:50 (w/w); Pierce) to obtain a complementary set of peptides for better protein sequence coverage and to recover larger peptides that may have blocked lysine and arginine residues due to the hexose post-translational modifications[164]. Chymotryptic peptides were collected by centrifugation (12,000 x g at 4°C). Tryptic and chymotryptic peptides were treated with formic acid (to 1%) to precipitate residual sodium deoxycholate, and the precipitate was removed from the peptide solutions with water-saturated ethyl acetate. Peptide solutions were concentrated by SpeedVac and quantified by BCA assay (Pierce).

Whey-derived tryptic peptides were sequenced by LC-MS/MS using a Vanquish ultra-high performance liquid chromatography (UHPLC) system plumbed directly in-line with a Q Exactive Plus mass spectrometer (Thermo Scientific) outfitted with a triphasic MudPIT back column (RP-SCX-RP) coupled to a nanospray emitter packed with 30 cm of 5 mm Kinetex C18 RP resin (Phenomenex). For each sample, 3 mg of peptides were loaded, desalted, separated and analyzed by one salt cut of ammonium acetate (500 mM), followed by a 100-minute organic gradient and column re-equilibration[163]. Eluting peptides were measured and sequenced by data-dependent acquisition on the Q Exactive MS. Acquired MS/MS spectra of tryptic peptides were first searched against the Bos taurus proteome (UniProt ID: UP000009136) using

MyriMatch/IDPicker[165] to identify the major components of the bovine whey protein isolate. A sub-data- base of the top 400 proteins was created and used for our subsequent PTM analysis. Reversed decoy sequences were appended to the sub-database to assess the false discovery rate in a generic open search with a wide precursor mass tolerance to identify peptides with all possible modifications[166]. The MSFragger open search was configured with the following parameters to identify common mass shifts on peptides: precursor mass tolerance of 500 Da, fragment mass tolerance of 10 ppm, and a maximum number of three variable modifications. In addition, an open search was performed on *Clostridium thermocellum* with the same MSFragger settings to confirm the specificity of fructosylation and lactosylation of lysine residues in the whey protein isolate sample relative to an organism that does not contain proteins with such modifications.

Following the open search, a targeted PTM search of MS/MS spectra of tryptic and chymotryptic peptides was performed using MS Amanda 2.0, Percolator, and ptmRS in the Proteome Discoverer 2.2 software package (Thermo Scientific), focusing on the two glycosylation modifications of interest as well as several abundant modifications identified in the MSFragger open search. The specific modifications targeted in the closed search included a static modification of carbamidomethylation (C+57.0214 Da) and the following dynamic modifications; oxidation of methionine (M+15.9949 Da), acetylhypusine on lysine (K+113.0841 Da), phosphorylation of serine, threonine or tyrosine (STY+79.9663 Da), acetylation of lysine (K+42.0106 Da), Fe++ cation replacement of proton on aspartate or glutamate, fructosylation of lysine or arginine (KR+162.05280), and lactosylation of lysine or arginine (KR+324.10560). Post-processing of the data was completed with Percolator using peptide-spectrum match filtering to <1% peptide-level FDR. Precursor ion intensity (area under the curve) was used to derive relative abundance values of peptides. PTM site localization was assessed with ptmRS.

Subsequent analysis focused on the five proteins in the sample with the most hexose modifications, which include the most abundant proteins identified: beta-lactoglobulin, alpha-lactoglobulin, albumin, enterotoxin-binding glycoprotein PP20K,

and beta-casein. For the hexose-modified proteins of interest, the number of peptide-spectral matches (PSMs) and peptides with an identified Hex(1) or Hex(2) modification were analyzed for all lysine and arginine residues. In cases where there was an ambiguous modification [Hex(1) and a Hex(2) on the peptide], the peptide was counted as being both Hex(1)-modified and Hex(2)- modified.

5.4 Integration of metaproteomics with other techniques to assess community degradation of dietary fibers.

Text and figures adapted from Patnode et al., 2019. Cell.

5.4.1 Project scope.

In a second study using defined human gut-derived bacterial communities, the scope of the project was to identify different dietary fibers and their bioactive components that increase the fitness of particular *Bacteroides* species in the context of a representative Western diet. Gnotobiotic mice were colonized in a defined 14-species community of human gut-derived bacterial strains that are robust colonizers of the gut in a high-fat/low-fiber diet and that capture the majority of bacterial taxa present in a human fecal microbiota sample[167]. Included in this microbial consortium are *Bacteroides* species that were previously correlated with protection from the increased obese (Ob) phenotypes that developed in co-housed Ob-Ob control mice[168]. These *Bacteroides* species were originally correlated with protection against host obesity phenotypes in a study where germ-free mice were colonized with fecal microbiota from twin pairs that were stably discordant for obesity. When mice colonized with the lean (Ln) or Ob phenotype microbiota were co-housed, the Ob recipient mice did not develop obesity or metabolic abnormalities. This outcome occurred because of the invasion of Ln-donor species (*B. thetaiotaomicron*, *B. vulgatus*, *B. caccae*, *B. ovatus* and *B. cellulosilyticus*) into the Ob mice under specific diet contexts. Specifically, the invasion of these *Bacteroides* species only occurred with a diet is low in saturated fats and high fruits and vegetables (LoSF/HiFV).

Invasion did not occur in a diet that is high in saturated fats and low in fruits and vegetable with (HiSF/LoFV).

Through the combination of forward genetic screens and quantitative LC-MS/MS of bacterial gene expression, we conducted a series of diet oscillation experiments paired with direct community manipulation to develop an understanding of the effects of different bioactive food components on the gut microbiota as well as community modulation as a result of fiber administration. A base representative Western diet (HiSF/LoFV) was constructed based on the result of the consumption patterns from U.S. National Health and nutrition survey (NHANES) database[168]. This base diet was supplemented with thirty-two food-grade fiber preparations from the byproducts of food manufacturing and screening experiments were conducted with 16S rRNA sequencing to analyze the relative abundance of each community member at the end of each diet treatment. This resulted in 21 fibers that increased some species by 1% for every 1% increase in fiber administration, and supplementation with four of the fibers significantly increased the abundance of at least one of the targeted *Bacteroides* species. The four lead fiber preparations (pea fiber, citrus pectin, orange peel, tomato peel) were selected for metaproteomics analyses.

5.4.2 Results and discussion.

In an initial metaproteomics screening experiment where mice were fed the baseline HiSF/LoFV diet and supplemented with the four lead fibers prior to LC-MS/MS measurements, there were distinct shifts in the functional relative abundance of some community members based on summed protein expression. Since pea fiber and citrus pectin had the most pronounced effects on distinct sets of taxa, these two fiber preparations were selected for detailed functional studies of bioactive polysaccharide utilization. Structural analysis of the two lead fibers using permethylation and GC-MS identified arabinan as the most abundant polysaccharide in pea fiber, after accounting for starch, which is typically degraded and absorbed by

the host, and cellulose, which is not metabolized by *Bacteroides* spp. The citrus pectin contained predominantly homogalacturonan.

To assess whether or not any species use these specific polysaccharides as nutrient sources, we set up a longitudinal strain-omission experiment for metaproteomics analysis **(Figure 5-3A).** Metaproteomic measurements were coupled with multi-taxon insertion sequencing (INSeq) of the transposon (Tn) libraries of four *Bacteroides* species (*B. thetaiotaomicron, B. vulgatus, B. ovatus* and *B. cellulosilyticus)* to characterize the effects of monotonous feeding of selected fiber preparations on the community's expressed proteome and on the fitness of Tn mutants. The goal of integrating results from the two techniques is that by identifying polysaccharide processing genes whose expression is increased and that function as key fitness determinants, we could infer which components of the fiber preparations were bioactive. **Figures 5-3B-I** show the combined results of the metaproteomics and INSeq measurements for fecal material collected from gnotobiotic mice fed the HiSF/LoFV diet supplemented with either pea fiber or citrus pectin. Several genes were identified in the INSeq analysis that had significant effects on the fitness of the organism depending on the supplementation of different fiber preparation, and many of these corresponded with proteins that showed a significant change in abundance upon fiber supplementation compared to the control diet (colored dots on the figures). Significant genes were classified as responsive to fiber supplementation if they showed a significant positive fold-change in protein abundance and negative fitness effect when mutated. These genes appear in the bottom right quadrants of each orthogonal protein-fitness plots.

Interestingly, among differentially enriched proteins, 85 proteins that were significantly altered by pea fiber and 134 proteins that were significantly affected by citrus pectin administration were encoded by designated polysaccharide utilization loci (PULs). Most *Bacteroides* species contain multiple polysaccharide utilization loci (PULs) in their genomes, which provide a fitness advantage by endowing a species with the ability to sense, import, and process complex glycans. PULs encode genes such as carbohydrate-responsive transcription factors, SusC/SusD-like transporters,

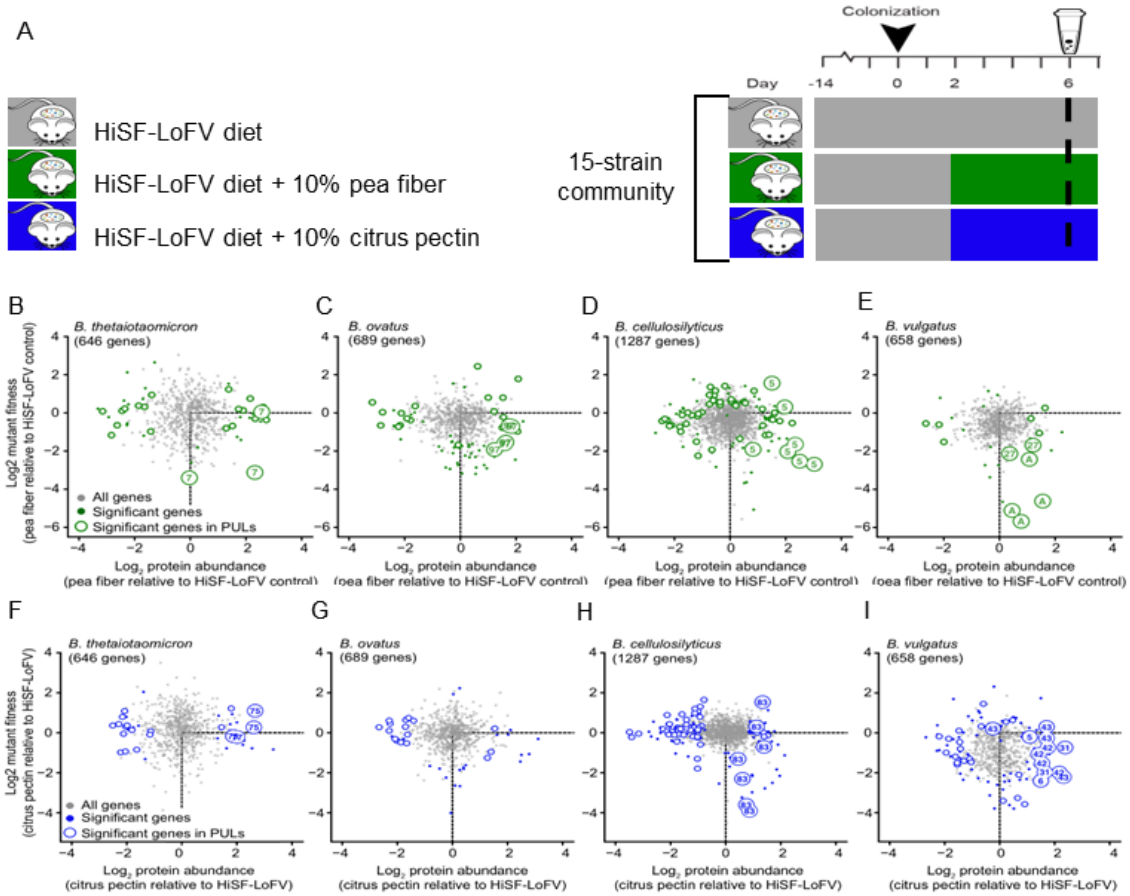

Figure 5-3 Proteomic and INSeq analyses of fecal samples collected on day 6. Monotonous feeding of the HiSF-LoFV control diet supplemented with or without pea fiber or citrus pectin to mice colonized with the whole community (A). Genes from *B. thetaiotaomicron* (B), *B. ovatus* (C), *B. cellulosilyticus* (D), and *B. vulgatus* (E) represented in both the protein dataset and INSeq mutant pool from the pea fiber supplemented diet vs. the control diet. Genes from *B. thetaiotaomicron* (F), *B. ovatus* (G), *B. cellulosilyticus* (H), and *B. vulgatus* (I) represented in both the protein dataset and INSeq mutant pool from the citrus pectin supplemented diet vs. the control diet. All genes found in both the proteomics and INSeq dataset are plotted as gray dots. Colored dots highlight genes that are significantly affected by the supplemented fiber as judged by levels of their protein products or their contribution to fitness; open circles mark the subset of these genes that are encoded by PULs. Genes that are present in predicted arabinan, arabinose, rhamnogalacturonan I, or homogalacturonan-processing PULs are labeled with a PUL number.

and CAZymes**.** Previous work has shown that regulated expression of PULs allows bacteria to acquire nutrients within the highly competitive gut environment[169,170].

Analysis of the fiber responsive genes encoded in the PULs showed that each species in the community displayed different carbohydrate prioritizing strategies for the carbohydrates present in pea fiber (arabinan and rhamnogalacturonan I (RGI)) and citrus pectin (galacturonan and starch). For example, we observed overlapping reliance of arabinan degradation pathways for three organism (*B. thetaiotaomicron*, *B. vulgatus*, and *B. cellulosilyticus*) which indicates that these species might directly compete with each other for arabinan. Competition for nutrients may explain the observed dominance of *B. cellulosilyticus* in relative community abundance compared to other *Bacteroides* species with similar PULs encoded in their genomes.

To assess whether or not *B. cellulosilyticus* directly competes for pea fiber arabinan, we set up another longitudinal strain-omission experiment **(Figure 5-4A).** For this experiment we tested for interactions between *B. cellulosilyticus* and other species by comparing fecal samples from mice colonized with the full community or a derivative community lacking *B. cellulosilyticus*. The abundance of *B. thetaiotaomicron* didn't increase when *B. cellulosilyticus* was omitted from the community, suggested these two organisms don't compete for arabinan **(Figure 5-4B).** This observation was supported by expression of proteins in *B. thetaiotaomicron* PUL7 that increased with fiber supplementation did not increase further when *B. cellulosilyticus* in the absence of *B. cellulosilyticus* **(Figure 5-4C).** In contrast, *B. vulgatus* the only member in the community that was significantly increased in abundance upon pea fiber administration when *B. cellulosilyticus* was omitted from the community **(Figure 5-4D).** Metaproteomic analysis showed that the abundances of *B. vulgatus* arabinose operon and PUL27 encoded proteins were persistently increased with pea fiber administration regardless of *B. cellulosilyticus* presence **(Figure 5-4E).** In total, these results show a negative interaction between *B. vulgatus* and *B. cellulosilyticus* and suggest that the suppression of *B. vulgatus* in the community in the presence of *B. cellulosilyticus* because of persistent competition between these organisms for arabinan in pea fiber. Competition for arabinan between

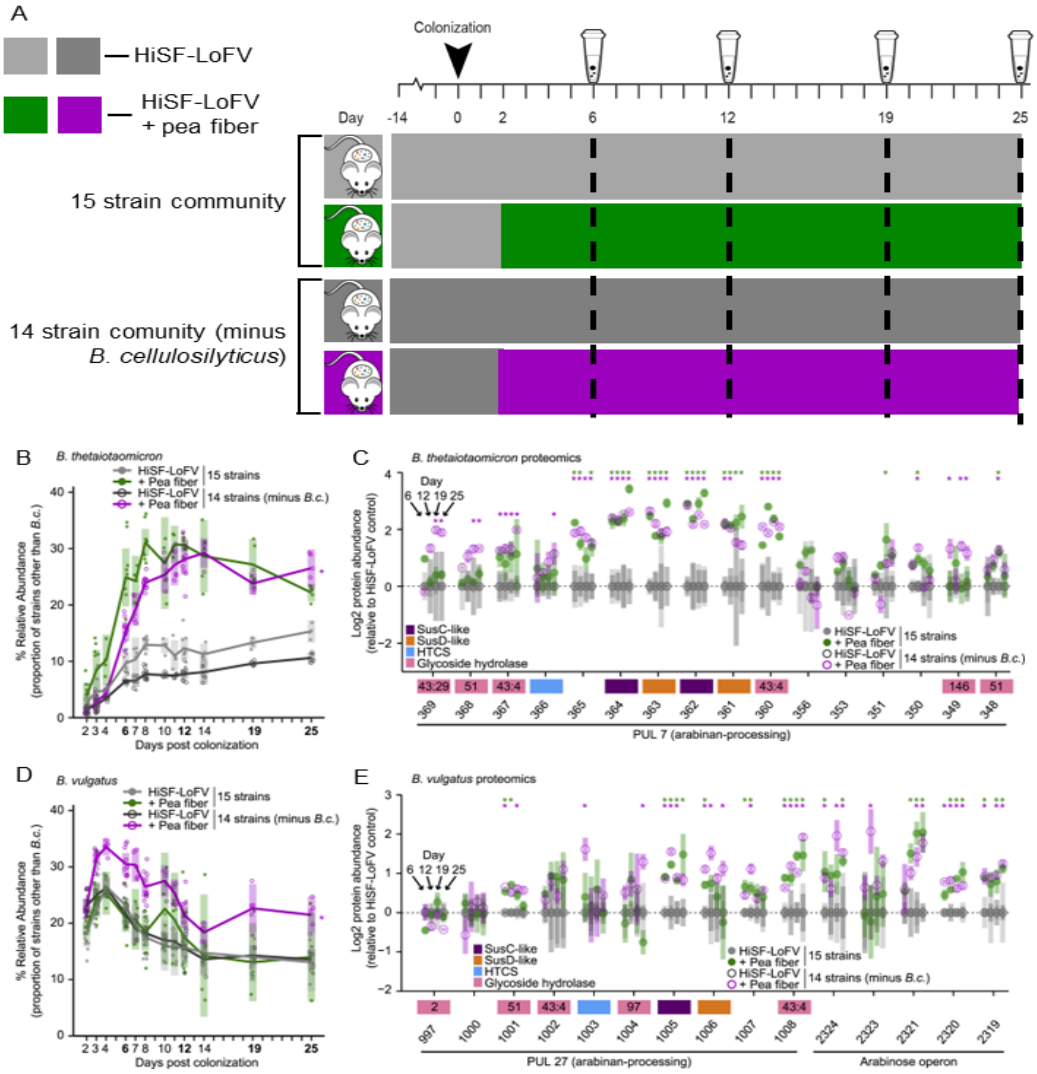

Figure 5-4 Deliberate Manipulation of Community Composition Demonstrates Interspecies Competition for Pea Fiber Arabinan. (A) Experimental design for feeding of the HiSF-LoFV control diet supplemented with pea fiber to mice colonized with the full or derivitive community. (B and D) Relative abundance of *B. thetaiotaomicron* (B) and *B. vulgatus* (D) in fecal samples. Key: circles, individual mice; lines, mean values; shading, 95% CI (n = 4–10 mice per group). *p < 0.05; (diet-by-community interaction; ANOVA). (C and E) Proteomic results from feces sampled on experimental days 6, 12, 19, and 25. Genes in *B. thetaiotaomicron* (C) and *B. vulgatus* (D) PULs of interest are shown along the x axis. Mean values ± SD (vertical shading) for each day are indicated in each of the four columns presented for each protein (n = 5 animals/treatment group). *p < 0.05, fold change > log2(1.2) (HiSF-LoFV + pea fiber versus HiSF-LoFV diet; limma).

B. cellulosilyticus and *thetaiotaomicron* does not suppress B *thetaiotaomicron* abundance in the community when *B. cellulosilyticus* is present.

Based on the observation that *Bacteroides* species in our community encode PULs in their genomes related to arabinan degradation proteins from these PULs increase in abundance upon pea fiber supplementation we assessed whether absence of *B. cellulosilyticus* compromised the efficiency of community arabinan utilization. To do this, an *in vivo* degradation assay was set up where mice colonized with the full community, or the derivative community were fed pea fiber or arabinoxylan coated beads. Wheat arabinoxylan, composed of 38% arabinose and 62% xylose, was used as one of the experimental controls as it has been demonstrated to support *in vitro* grown of *B. cellulosilyticus*[171] but not *B. vulgatus*[172]. Analysis of the beads recovered from the cecal and colonic contents of the mice showed that the presence of *B. cellulosilyticus* did not impact the level of community pea fiber degradation. In contrast, the capacity of the community to process arabinoxylan decreased in the absence of *B. cellulosilyticus.* This result was confounding as *B. ovatus* is significantly expanded in the community when *B. cellulosilyticus* is omitted. *B. ovatus* which encodes PULs involved in the degradation of arabinoxylan, so it is interesting that the increased abundance of this organism did not rescue arabinoxylan in the derivative community. This suggests *B. cellulosilyticus* and *B. ovatus* may exhibit a different nutrient harvesting relationship compared to the direct competition for arabinan observed between *B. cellulosilyticus* and *B. vulgatus.*

To test this potential interspecies relationship, another experiment was set up for metaproteomic analysis of PUL protein expression in the organisms **(Figure 5-5A).** As arabinoxylan is present in the HiSF-LoFV diet, no additional fiber supplementation was necessary. Fecal samples collected from mice colonized with the full community were compared to two derivative communities; the first community lacked *B. cellulosilyticus* (a potential competitor for arabinoxylan) and the second community lacked *B. vulgatus* (an organism unable to process arabinoxylan).

B. ovatus showed persistently increased abundance when *B. cellulosilyticus* was absent, but not difference in abundance between the full community and the

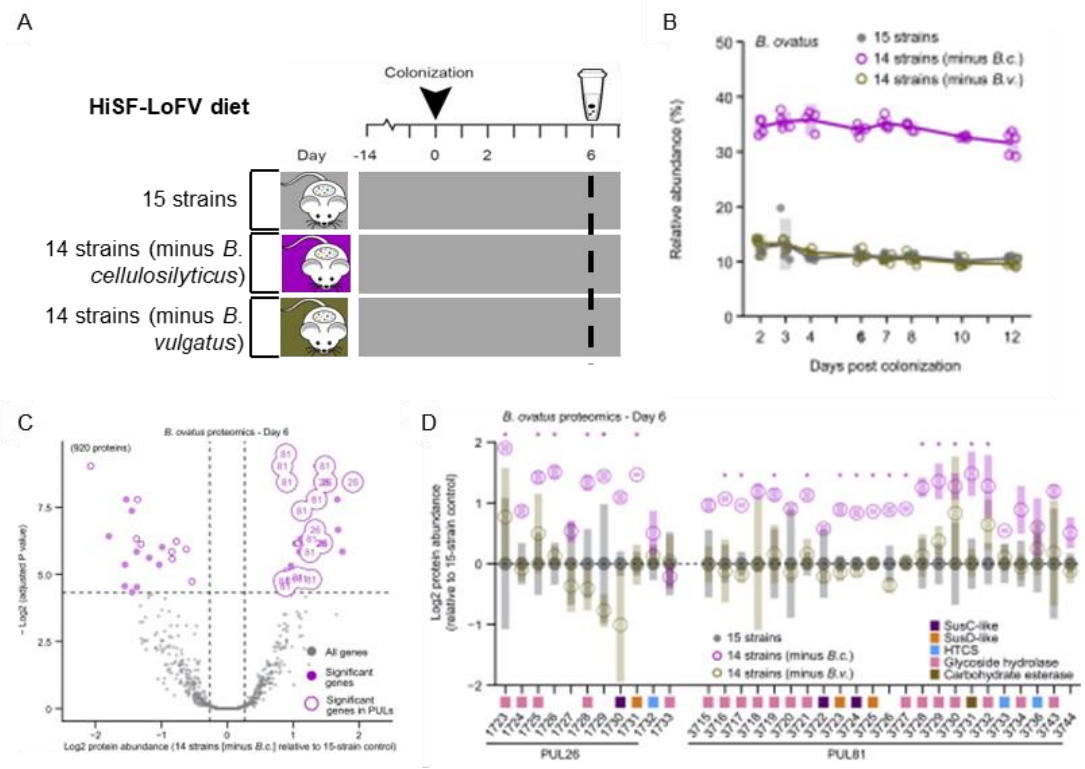

Figure 5-5 Detecting Acclimation to the Presence of a Potential Competitor Using Proteomics. (A) Experimental design monotonous feeding of the HiSF-LoFV control diet to mice colonized with the whole community or derivative communities. (B) Relative abundance of *B. ovatus* in fecal samples from gnotobiotic mice harboring derivatives of the full community. Mice received the control HiSF-LoFV diet in the presence (gray closed circles) or absence of *B. cellulosilyticus* or *B. vulgatus* (open circles; magenta and brown respectively). Key: circles, individual mice; lines, mean values; shading, 95% confidence interval. (C) Mean protein abundance in fecal samples from day 6 in the minus *B. cellulosilyticus* group relative to the *plus B. cellulosilyticus* group; proteins whose levels are significantly different between the groups and encoded by genes in PULs are highlighted with open circles, while those encoded by genes in arabinoxylan processing PULs are labeled with their PUL number. (D) Protein abundance for *B. ovatus* genes in arabinoxylan PULs shown along the x axis. Mean values ± SD (vertical shading) are indicated (n = 5 animals/treatment group). Genes are color-coded according to functional annotation. HTCS stands for hybrid two-component system. Key for circles: gray, 15-member community; magenta or brown, mice harboring communities without *B. cellulosilyticus* or *B. vulgatus*, respectively. *$p < 0.05$, fold change > log2(1.2) (plus *B. cellulosilyticus* versus minus *B. cellulosilyticus*.; limma).

community lacking *B. vulgatus* **(Figure 5-5B).** In contrast to our previous observations that *B. vulgatus* PUL protein expression is not changed in response to the presence of a direct competitor, B. *ovat*us exhibited metabolic flexibility when *B. cellulosilyticus* was absent versus present. Proteins encoded by two arabinoxylan-processing PULs (PUL26 and PUL81) predominated among those whose abundances were increased based on the presence of this potential competitor **(Figure 5-5C).** Looking at the metaproteome datasets with the full and *B. cellulosilyticus* omitted communities, the metabolic flexibility displayed by *B. ovatus* was apparent regardless of diet (HiSF-LoFV, pea fiber, or citrus pectin), which is consistent with the presence of arabinoxylan in the base diet. In addition, metaproteomics results demonstrated that omission of *B. vulgatus* did not induce changes in either the levels of proteins encoded by *B. ovatus* PULs, but omission of *B. cellulosilyticus* induced persistent increases in the expression of those proteins **(Figure 5-5D).**

Complementary to the metaproteomic results regarding proteins from *B. ovatus* PULs PUL26 and PUL81, the INSeq results showed that genes in these two arabinoxylan PULs were the most affected by omission of *B. cellulosilyticus*. In total, these results indicate that *B. ovatus* exhibits a marked decrease in its reliance on arabinoxylan in the full community context and increases its arabinoxylan processing capabilities in the absence of a potential competitor.

To test whether or not metabolic flexibility allows *B. ovatus* to acclimate to the presence of *B. cellulosilyticus* by shifting its nutrient harvesting strategies to mitigate competition between the two species for arabinoxylan, we performed an experiment with gnotobiotic mice colonized with the full community or defined communities omitting *B. cellulosilyticus*, *B. ovatus*, or both species from the full bacterial consortium (Figure 5.6A). Based on measurements collected 6 days after gavage, *B. ovatus* increased in abundance in the absence of *B. cellulosilyticus* **(Figure 5-6B),** but *B. cellulosilyticus* did not significantly change in abundance in the absence of *B. ovatus* **(Figure 5-6C)**. The analysis of proteins expressed by arabinoxylan-processing PULs showed an increase in abundance of 16 proteins from *B. ovatus* PUL26 and PUL81 **(Figure 5-6C)**. Only one protein encoded by one of the two

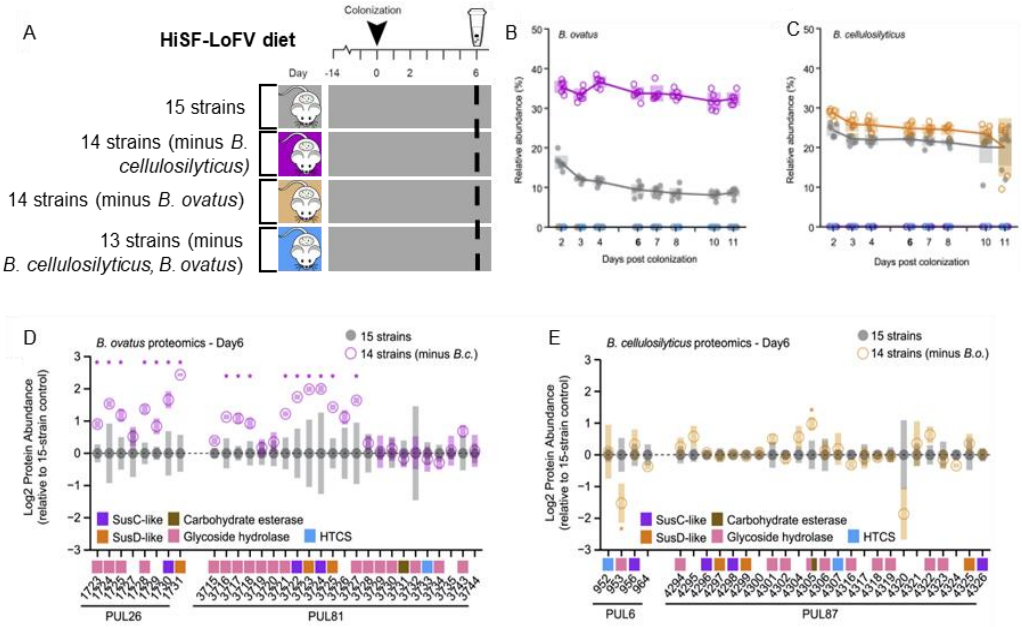

Figure 5-6 Alleviation of Competition between Arabinoxylan-Consuming *Bacteroides*. (A) Experimental design monotonous feeding of the HiSF-LoFV control diet to mice colonized with the whole community or derivative communities. (B and C) Relative abundance of *B. ovatus* (B), and *B. cellulosilyticus* (C) in fecal samples from gnotobiotic mice harboring derivatives of the full community. Animals fed the control HiSF-LoFV diet in the presence (closed circles) or absence of *B. cellulosilyticus* or *B. ovatus* or both species (open circles; magenta, orange, or cyan respectively). Key: circles, individual mice; lines, mean values; shading, 95% confidence interval. (D and E) Mean protein abundances in fecal samples obtained on experimental day 6. Proteins in *B. ovatus* (D) and *B. cellulosilyticus* (E) arabinoxylan-processing PULs are shown along the x axis relative to full community condition. Mean values ± standard deviation (vertical shading) are indicated (n = 5–7 animals/treatment group). Proteins are color-coded according to functional annotation (see key). HTCS stands for hybrid two-component system. Key for circles is the same as in (B-C). *p < 0.05 (full community versus derivative community]; limma.

arabinoxylan-processing PULs in *B. cellulosilyticus*, PUL6 and PUL87, increased in abundance when *B. ovatus* was omitted from the community. The results from this experiment combined with the INSeq observation in the previous experiment that arabinoxylan-processing genes are only important for the fitness of *B. ovatus* when *B. cellulosilyticus* is not present indicate that the metabolic flexibility of *B. ovatus* mitigates competition with *B. cellulosilyticus*, even though these two species both have the capacity to process the same nutrient resource.

In summary, this paper showed that plant polysaccharides (PPS) that can produce distinct shifts in the relative abundance of targeted *Bacteroides* species with diet supplementation were identified. Based on these identified bioactive fiber components, we were able to directly characterize how *Bacteroides* with distinct, as well as overlapping, nutrient harvesting capacities respond to specific plant polysaccharides, including arabinan and arabinoxylan. In addition, many of the techniques used to study this microbiome provide similar results, such as organismal abundances as determined by COPRO-Seq and metaproteomics, providing orthogonal validation of the observations gained with each approach. However, each of these techniques provided unique, but complementary information, such the functional importance of PUL genes as measured by metaproteomics and INSeq, which enabled key biological insights that may not have been elucidated without integration of these approaches. Finally, this study provided an *in vivo* demonstration of using artificial food particles as biosensors of community-wide glycan metabolism. In total the direct community manipulation paired with quantitative multi-omics measurements provided insights into specific microbe-microbe interactions in response to the administration of plant polysaccharides. These findings open the door to developing microbiota-directed foods that selectively increase the abundance of beneficial microbes and provide metabolic benefits to the host.

5.4.3 Methods.

Lysates were prepared from fecal samples by bead beating in SDS buffer (4% SDS, 100 mM Tris-HCl, 10 mM dithiothreitol, pH 8.0) using 0.15 mm diameter zirconium oxide beads, followed by centrifugation at 21,000 x g for 10 minutes. Pre-cleared protein lysates were further denatured by incubation at 85°C for 10 minutes, and adjusted to 30 mM iodoacetamide to alkylate reduced cysteines. After incubation in the dark for 20 minutes at room temperature, protein was isolated by chloroform-methanol extraction. Protein pellets were then washed with methanol, air-dried, and re-solubilized in 4% sodium deoxycholate (SDC) in 100 mM ammonium bicarbonate (ABC) buffer, pH 8.0. Protein concentrations were measured using the BCA (bicinchoninic acid) assay (Pierce). Protein samples (250 mg) were then transferred to a 10 kDa MWCO spin filter (Vivaspin 500, Sartorius), concentrated, rinsed with ABC buffer, and digested in situ with sequencing-grade trypsin[163]. The tryptic peptide solution was then passed through the spin-filter membrane, adjusted to 1% formic acid to precipitate the remaining SDC, and the precipitate removed from the peptide solution with water-saturated ethyl acetate. Peptide samples were concentrated using a SpeedVac, measured by BCA assay and analyzed by automated 2D-LC-MS/MS using a Vanquish UHPLC with autosampler plumbed directly in-line with a Q Exactive Plus mass spectrometer (Thermo Scientific) outfitted with a 100 mm ID triphasic back column [RP-SCX-RP; reversed-phase (5 mm Kinetex C18) and strong-cation exchange (5 mm Luna SCX) chromatographic resins; Phenomenex] coupled to an in-house pulled, 75 mm ID nanospray emitter packed with 30 cm Kinetex C18 resin. For each sample, 12 mg of peptides were autoloaded, desalted, separated and analyzed across four successive salt cuts of ammonium acetate (35, 50, 100 and 500 mM), each followed by a 105-minute organic gradient. Eluting peptides were measured and sequenced by data-dependent acquisition on the Q Exactive Plus[163]. MS/MS spectra were searched with MyriMatch v.2.2[140] against a proteome database derived from the genomes of the strains in the defined model community concatenated with major dietary protein sequences, common protein contaminants, and reversed

entries to estimate false-discovery rates (FDR). Since the relative abundance of *B. thetaiotaomicron* 7330 was low on day 6 [0.11% ± 0.22% (mean ± SD) for all groups], we chose to analyze all peptides that mapped to the *B. thetaiotaomicron* VPI-5482 proteome, regardless of whether they also mapped to *B. thetaiotaomicron* 7330. Peptide spectrum matches (PSM) were required to be fully tryptic with any number of missed cleavages, and contain a static modification of 57.0214 Da on cysteine and a dynamic modification of 15.9949 Da on methionine. PSMs were filtered using IDPicker v.3.0[141] with an experiment-wide FDR < 1%at the peptide-level. Peptide intensities were assessed by chromatographic area-under-the-curve (label-free quantification option in IDPicker). To remove cases of extreme sequence redundancy, the community meta-proteome was clustered at 100% sequence identity post-database search (UCLUST)[173] and peptide intensities were summed to their respective protein groups (seeds) to estimate overall protein abundance. Proteins were included in the analysis only if they were detected in more than three biological replicates in at least one experimental group. After considering only peptides that uniquely mapped to a single seed protein, the summed abundances proteins advanced to quantitative analysis yielded 59% from community members, 36% from mouse, and 2% from diet. Missing values were imputed to simulate the limit of detection of the mass spectrometer, using mean minus 2.2 x standard deviation with a width of 0.3 x standard deviation. Four additional imputed distributions produced results that were in general agreement with this approach in terms of fold-abundance change induced by fiber treatment and statistical significance.

5.5 Microbiota Functional Activity Biosensors (MFABs).

Text and figures adapted from Wesener et al., 2021. eLife.

5.5.1 Project scope.

The primary focus of this project was to develop a novel method for measuring biochemical activities expressed by the gut microbiome in an *in vivo* setting in order

to quantitatively characterize function. Specifically, the study was focused on the development of a recoverable, chemical probe based on microscopic silica paramagnetic beads, known as 'Microbiota Functional Activity Biosensors' (MFABs). MFABs can be designed to contain different covalently bound polysaccharide preparation and distinct fluorophores based on the research question of interest. While the previous study presented in Chapter 5.4 provided an *in vivo* demonstration using glycan coated beads to assess community-wide glycan degradation, the current approach for covalent attachment of glycans to microscopic paramagnetic glass beads with different covalently bound fluorophores eliminates the proteinaceous component of the previous bead design that relies on streptavidin-coated beads and biotin-conjugated polysaccharides and provides more attachment sites on the bead surface. For *in vivo* gut microbiome research, MFABs can be pooled and gavaged into gnotobiotic mice colonized with a defined consortium bacterial species and the beads can be recovered from the cecum or feces. Recovered beads can be analyzed using gas chromatography mass spectrometry to quantify the amount of remaining bead bound polysaccharides to assess bacterial community metabolism in different human diet contexts. In this study, the utility of MFABs was illustrated in three different experimental settings. Metaproteomics was used a complementary approach in the first application to assess which community members were expressing proteins to actively utilize the polysaccharides that were degraded on the gavaged beads.

Using the same defined community from the study described in Chapter 5.4, a series of experiments was set up using the gnotobiotic mice to further characterize community degradation of arabinan, the primary polysaccharide component found in pea fiber. We used MFABs, DNA sequencing approaches, and metaproteomics to compare community degradation of arabinan from a raw or unfractionated pea fiber preparation to two structurally distinct arabinan polysaccharides, one isolated from the endosperm of pea (PFABN) and the other from sugar beet (SBABN). Gnotobiotic mice were colonized with the defined bacterial consortium and fed HiSF/LoFV control diet with or without glycan supplementation.

5.5.2 Results and discussion.

COPRO-Seq was used to assess colonization of each community member in response to different diets, and results showed that five members of the community exhibited statistically significant differences in abundance in response to diets with supplementation of different glycan preparations. To complement these results, we measured matching fecal samples collected at six days post gavage (dpg) by LC-MS/MS-based metaproteomics in two independent experiments with a total of eleven mice per treatment arm **(Figure 5-7A).** Looking at the summed proteins abundances of each organism in the community, metaproteomics confirmed the five members identified by COPRO-Seq displayed distinct responses to different glycan preparations. Specifically, *B. thetaiotaomicron* increased in abundance for all three supplemented diets compared to the control diet. *B. ovatus* responded to exposure to both intact pea fiber and PFABN, but did not increase with SBABN supplementation. *B. vulgatus* increased in protein abundance only with SBABN supplementation. Both. *B. cellulosilyticus WH2* and *Ruminococcaceae sp.* increased in abundance when exposed to intact pea fiber supplementation, but not to either of the isolated arabinan preparations.

To explain underlying mechanisms of this diet-driven shift in community composition, we looked at protein expression within PULs possessed by the four *Bacteroides* species in the community that responded to arabinan supplementation. Using gene set enrichment analysis, there were 14, 12, 11, and 8 PULs identified that were considered responsive to at least one of the diet supplements in *B thetaiotaomicron*, *B. ovatus*, *B. cellulosilyticus WH2*, and *B. vulgatus*, respectively **(Figure 5-7B).** The results of the *in vivo* metaproteomics analysis comparing the differential expression of PULs was also consistent with a follow-up *in vitro* MFAB-based PFABN degradation experiment which demonstrated the ability of MFABs to induce the expression PULs involved in PFABN utilization.

In total, this study demonstrates an approach to identify bioactive components

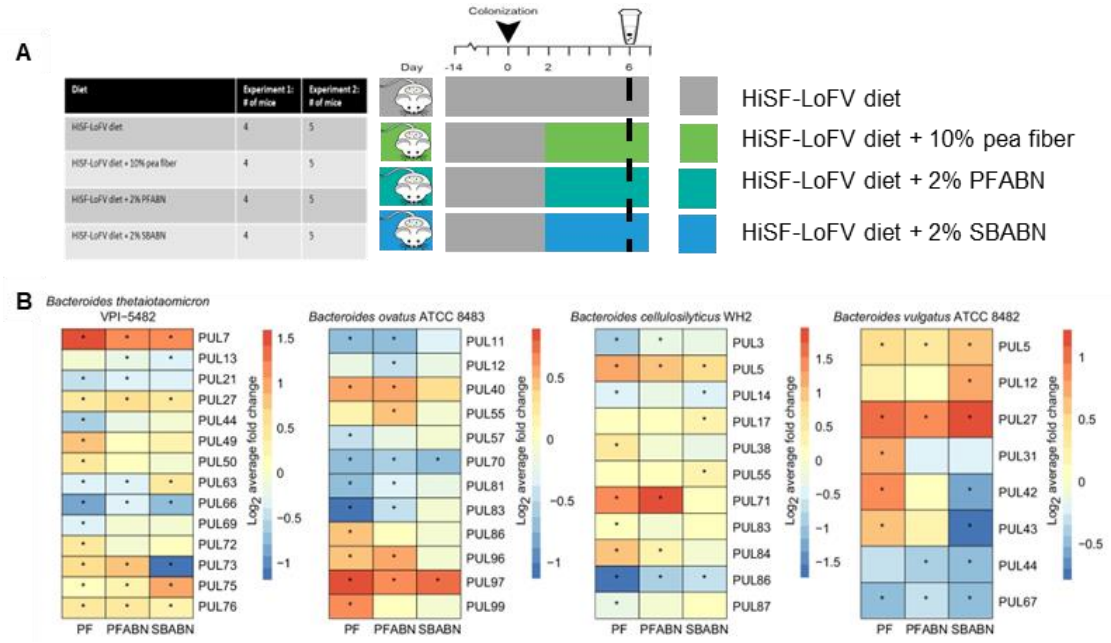

Figure 5-7 The effects of supplementing the HiSF-LoFV control diet with unfractionated pea fiber, PFABN, or SBABN on PUL gene expression. (A) Monotonous feeding experiments of the HiSF-LoFV control diet supplemented with or glycan supplementation for metaproteomic analyses. (B) Heat maps for *Bacteroides* species of interest, showing the average log2 fold change in abundance of proteins found within PULs identified as supplement-responsive using GSEA. *p<0.05 (unpaired one-sample Z-test, FDR corrected) compared to PUL protein abundance when mice were fed the base HiSF-LoFV diet.

of dietary fibers and define how they affect and are utilized by human gut microbiota. Specifically, this approach illustrates how to isolate bioactive arabinan-enriched fractions from plant fibers (pea fiber and sugar beet), define their structures, and characterize how gut microbiota respond to arabinans from unfractionated pea fiber versus isolated fractions of two structurally distinct arabinans. For the characterization, we analyzed feces collected from gnotobiotic mice colonized with a defined human gut model community, including several saccharolytic *Bacteroides* species, using forward genetic screens and metaproteomic analyses to demonstrate the response sensitivity of distinct organisms to structural differences in arabinans. This study also presents a generalizable method to design collections of artificial food particles (MFABs), where microscopic paramagnetic glass beads have identifying fluorescent tags and different glycans covalently bound to the surface of the beads. The unique design of these beads enables: (1) retrieval of the beads following oral gavage in *in vivo* experiments based on their magnetism, (2) bead sorting based on the surface-bound fluorophores, and (3) the quantification of bound polysaccharide degradation through the comparison of the amount of glycan on input beads to the amount remaining on the beads after recovery of beads from feces. In addition, as the bead design allows multiple types of ligands to be simultaneously attached to the bead surface, this study showed that pairing different combinations of glycans together enhanced the degradation of each glycan compared to supplementation with one glycan alone. In total, pairing bead-based MFAB glycan degradation experiments to quantify <u>what</u> bioactive dietary components are being utilized with forward genetic and metaproteomic analyses to define <u>how</u> each member of the community is utilizing those bioactive components provides a holistic approach to studying bacterial metabolism *in vivo*. Overall, the results presented in the study indicate that MFABs can be used as a complement for multi-omic approaches for characterizing biochemical activities of the human gut microbiome and in the future, MFABs may be used in the development of microbiome-directed diagnostics and therapeutics.

5.5.3 Methods.

The protocol for mass spectrometry-based metaproteomic analysis of the fecal samples has been described in detail in the study presented in Chapter 5.4 (Patnode et al., 2019, *Cell*).

Chapter 6 - Probing the diversity of interkingdom interactions in the preterm gut.

Text and figures were adapted from the following published journal articles and in-preparation manuscripts:

West, P.T., **Peters, S. L.,** Olm, M. R., Yu, F. B., Gause, H., Lou, Y. C., Firek, B. A., Baker, R., Johnson, A. D., Morowitz, M. J., Hettich, R. L., & Banfield, J. F. (2021). Genetic and behavioral adaptation of *Candida parapsilosis* to the microbiome of hospitalized infants revealed by in situ genomics, transcriptomics, and proteomics. Microbiome, 9(1), 142–142. https://doi.org/10.1186/s40168-021-01085-y

S.L.P contributions include metaproteomic measurements, data analysis, figure generation, writing and editing of the original manuscript and response to reviewers.

Peters, S.L., Morowitz, M.J., and Hettich, R.L., (2022). Understanding how the developing human preterm infant gut microbiome responds to antibiotic administration and host immune system-induced metal bactericidal control. *Frontiers in Microbiology* (in submission)

S.L.P contributions include study conception, data analysis, writing and editing of the original manuscript.

6.1 Introduction to the preterm gut environment.

The human microbiome is composed of bacteria, archaea, microbial eukaryotes, and viruses that inhabit all environment-facing parts of our bodies. The

intestinal microbiota's diversity and function play significant roles in host health and disease. The microbiome is linked closely linked to inflammation and metabolism—with diet, environment, age, and geography all having an impact on the diversity of the microbiome. The hospitalized preterm infant microbiome can be used as a tractable model to look at microbial colonization and development within the host compared to the more fully established adult gut microbiome**.** While an adult gut is colonized by hundreds to thousands of species[174], the infant gut progresses from sterility at birth to colonization levels similar to an adult over the first year of life[175]. Hospitalized preterm infants make a particularly amenable model to study colonization as samples are readily accessible for collection in a controlled hospital environment. The study of establishment dynamics is of broad interest, as the gut microbiome develops during the early stages of life and plays an essential role in host health through the production of metabolic resources and stimulation and training of the immune system.

The two projects in this chapter probe interkingdom interactions in the hospitalized infant gut environment to elucidate how these interactions impact persistence and functionality of the gut microbiota in early life. First, a time course metaproteomics study on fecal samples from hospitalized premature infants diagnosed with *Candida parapsilosis* blood or urine infections revealed distinct functional partitioning among microbiota in the presence of low abundance microbial eukaryotes. Second, the second study is an exploration of temporal relationships between microbial activities and the host immune system during the normal and abnormal establishment of the gut microbiome.

6.2 Multi-omics characterization of the establishment of the preterm infant gut microbiome in the context of eukaryotic membership.

Text and figures adapted from West et al., 2021. Microbiome.

6.2.1 Introduction.

Studies of the gut microbiome tend to focus on bacteria without consideration of microbial eukaryotes, despite their functional importance to the community[176]. Microbial eukaryotes are commonly found within the human gut microbiome but can have significant impacts on host health. Studies have linked the presence of fungal pathogens within the gut microbiome to substantial effects on morbidity and mortality among newborns[177]. Despite this correlation with host health, very little is known about eukaryotic establishment within the preterm gut microbiome and their interactions with bacterial microbiota and the host[178]. This study investigates the metagenomic and metaproteome signatures of preterm infant gut microbiomes from two infants who presented a clinical diagnosis of blood or urine fungal infection while hospitalized in a neonatal intensive care unit to examine and identify the presence of eukaryotes and to examine how community members interact, partition function, and establish stability over time.

6.2.2.1 Metrics and trends of the metaproteomic measurements.

Over 7,000 protein groups were quantified per infant. Around 25,000 peptides were identified per sample, with 21,000 unique peptides. There was considerable proteome variation across samples at the taxonomic and protein group levels with distinct differences between individuals (**Figure 6-1A**) and across time points within an individual (**Figure 6-1B**).

Annotation databases such as KEGG can be used to classify protein groups into orthologous groups based on their shared biochemical function[179]. Utilization of

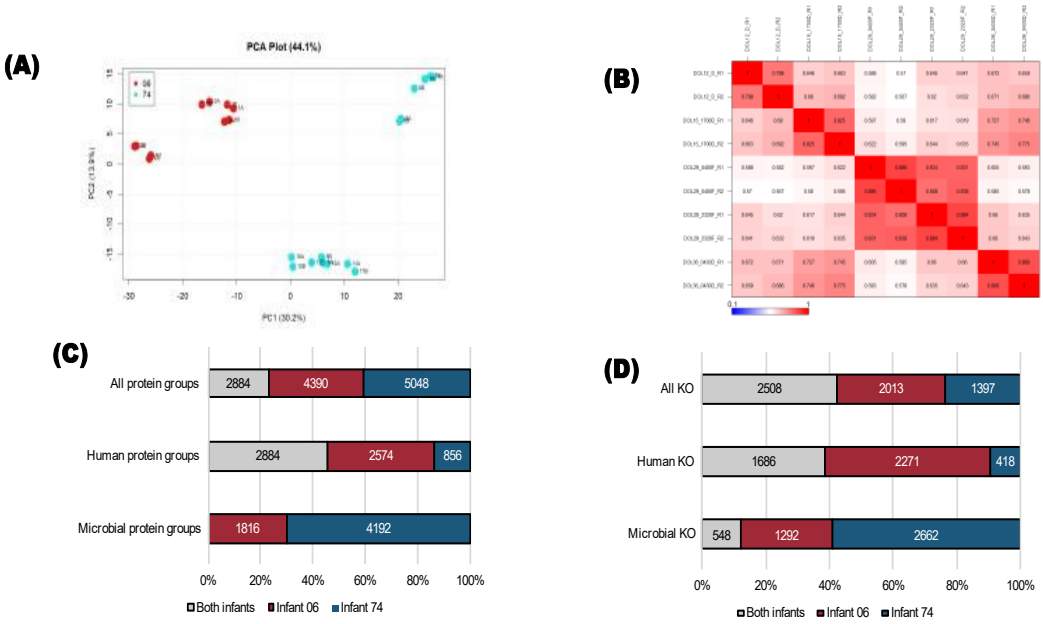

Figure 6-1 Principal component analysis of all metaproteomic measurements. (A). Distinct metaproteome composition between individuals and high reproducibility between technical replicates were observed. There was also a distinct separation between samples collected at early and late time points in both infants. Pearson correlation plot of infant 06 measurements (B). Within an individual, there is a higher correlation between close time points but an overall separation of the metaproteome across time. Variation of protein groups (C) and KEGG orthologs (KO) (D). The percent of the total quantified protein groups or KO terms found in both infants or found in only one infant was assessed at each functional level. The relative proportions of assignments are shown for all protein groups/KOs or protein groups/KOs belonging to the host or the microbiota.

these annotations can be a useful way to assess the similarity of functional activities between microbiomes with limited taxonomic overlap. To assess the difference in the metaproteomic measurements between individuals at different functional levels, the relative proportion of the total protein group of KO term identifications was compared at three levels: (1) all identifications, (2) human identification, and (3) microbial identifications. There was a high level of overlap of host protein groups, but no overlap of microbial protein groups **(Figure 6-1C)**. This finding was expected, as the host proteome is conserved between samples, while each individual may be colonized by very distinct microbiota. Restricting the dataset to the functional level by assigning KO terms to protein groups enabled the comparison of microbial functions between infants (**Figure 6-1D**).

6.2.2.2 Longitudinal metaproteomic characterization simultaneously reveals the presence and functions of bacteria and eukaryotes in the gut microbiomes of preterm infants.

Genomic and protein evidence of microbial eukaryotes was detected in one of the individuals. The functional taxonomy of bacterial community members was very distinct between the two infants, Firmicutes comprising the source of the majority of quantified proteins in the infant co-colonized with microeukaryotes and quantified proteins from Proteobacteria dominating the gut of the infant with no evidence of fungal colonization (**Figure 6-2A and 6-2B**). There were clear indications of robust establishment across all measured time points as well as active function and stability of these microbes within the gut microbiome. Infant 06 shows early establishment of microeukaryotes, with eukaryotic proteins representing up to 15% of the microbial abundance at any time point and overall community stability across time (**Figure 6-2C)**. The normal pH range of the intestinal tract is between 4 and 7. In vitro studies have shown that *C. parapsilosis* has optimal enzymatic activity at pH 4.5[180]. Since Firmicutes are more tolerant of acidic conditions than Proteobacteria, environment conditions produced by the bacteria may impact the favorability of *C. parapsilosis* co-

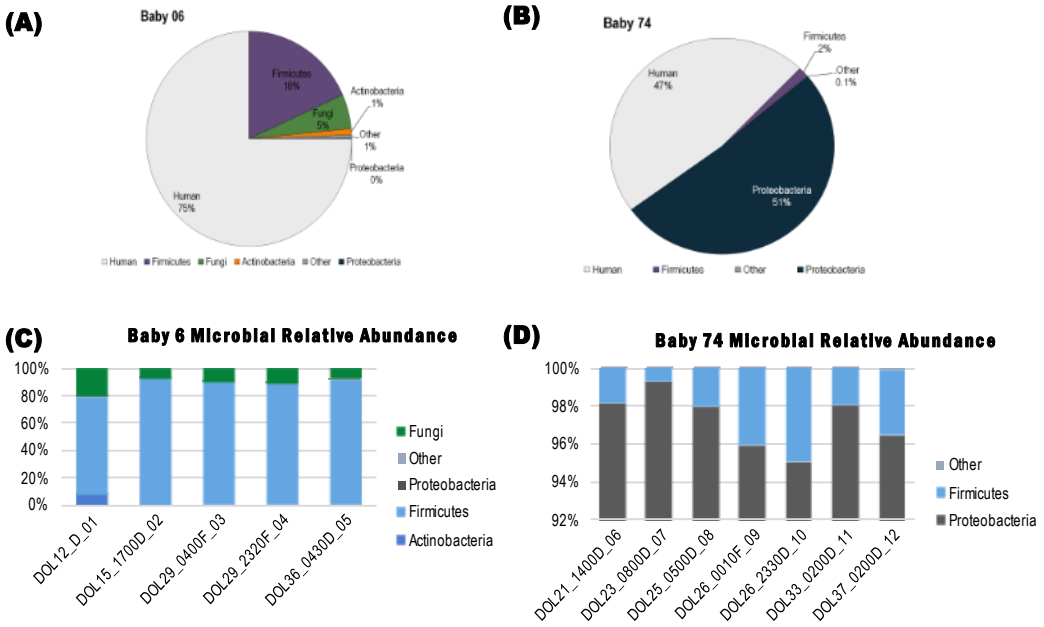

Figure 6-2 Taxonomic distribution of quantified proteins across all measurements for infant 06 (A) and infant 74 (B). 7274 proteins were quantified for infant 06 and 7932 proteins were quantified in infant 74 using the CharmeRT workflow. The taxonomic profiles of the gut metaproteomes of the two infants were very different and microbial eukaryotes were only detected in one individual. Relative abundance of microbial proteins across time for infant 06 (C) and infant 74 (D).

colonization. Infant 74 showed much more variation in community membership between the two dominant phyla across time **(Figure 6-2D).**

A gut-specific framework (GOMixer) was applied to infer species-associated gut metabolic modules (GMMs). A module is defined as a set of tightly related enzymatic functions which represent a specific cellular process with defined metabolic inputs and outputs. Across both infants, metabolic processes including central metabolic processes, protection against oxidative stress, the production of butyrate, and acetogenesis were overrepresented in the infant six community compared to the infant 74 community (**Figure 6-3A**). Triplots were generated to visualize the relative community metabolic investment of user-defined metabolic modules. **Figure 6-3B** is based on the relative abundance associated with the metabolism of the three predominant SCFAs (butyrate, propionate, and acetate). Short-chain fatty acids are major energy sources and immune propionate, and acetate). Short-chain fatty acids are major energy sources and immune modulators for the host. Distinct priorities in SCFA production are observed between infants. Infant 06 microbes showed clear investment in butyrate metabolism while members of the infant 74 microbes showed a preference for propionate metabolism. There was also longitudinal separation in SCFA production preference with a shift towards more generalized SCFA production at later days of life in both infants.

Among the functional activities of individual microbiota in each community, there was ample evidence of eukaryotic activity related to several processes, including central metabolism, alcohol metabolism, and organic acid metabolism (**Table 6-1**). Expressional evidence of these metabolic activities indicates stability of *C. parapsilosis* within the microbiome of infant 06. In addition, several bacterial functional activities were observed within each infant, including the production of lipopolysaccharide, Microbial-derived lipopolysaccharide is thought to be a primary driver of host immune response, and evidence of microbial LPS biosynthesis was observed in infant 74 across most time points (**Figure 6-4).**

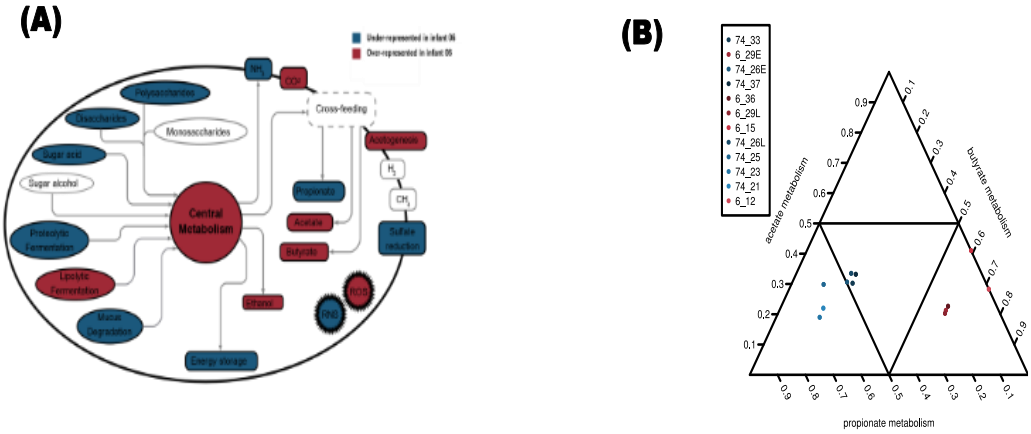

Figure 6-3 Tri-plot representation of short-chain fatty acid production. (A). Overview of the main metabolic processes between infants. (B). The sum of metabolic module differences was used to determine a global representation difference between infants. The color scale reflects the representation of the metabolic processes. Modules that were over-represented in infant 06 compared to infant 74 in all phyla are highlighted in red. Under-represented modules in infant 06 are listed in blue.

Table 6-1 Fungal metabolic modules identified in GoMIXER analysis.

Hierarchy Level 1	Hierarchy Level 2	Name
alcohol metabolism	ethanol metabolism	ethanol production (CO2 pathway)
		ethanol production (formate pathway)
amino acid degradation	nonpolar, aliphatic amino acid degradation	isoleucine degradation
		valine degradation
carbohydrate degradation	sugar alcohol degradation	xylitol degradation
central metabolism	energy metabolism	bifidobacterium shunt
		Glycolysis (pay-off phase)
		Glycolysis (preparatory phase)
		pentose phosphate pathway (non-oxidative branch)
		pentose phosphate pathway (oxidative branch)
		pyruvate dehydrogenase complex
		TCA cycle
		TCA cycle (Mycobacterium pathway)
lipid degradation	glycerol degradation	glycerol degradation (dihydroxyacetone pathway)
	glyoxylate bypass	glyoxylate bypass
organic acid metabolism	acetate metabolism	acetyl-CoA to acetate
	butyrate metabolism	acetyl-CoA to crotonyl-CoA
protection against oxidative stress	peroxidase	peroxidase
	superoxide dismutase	superoxide dismutase

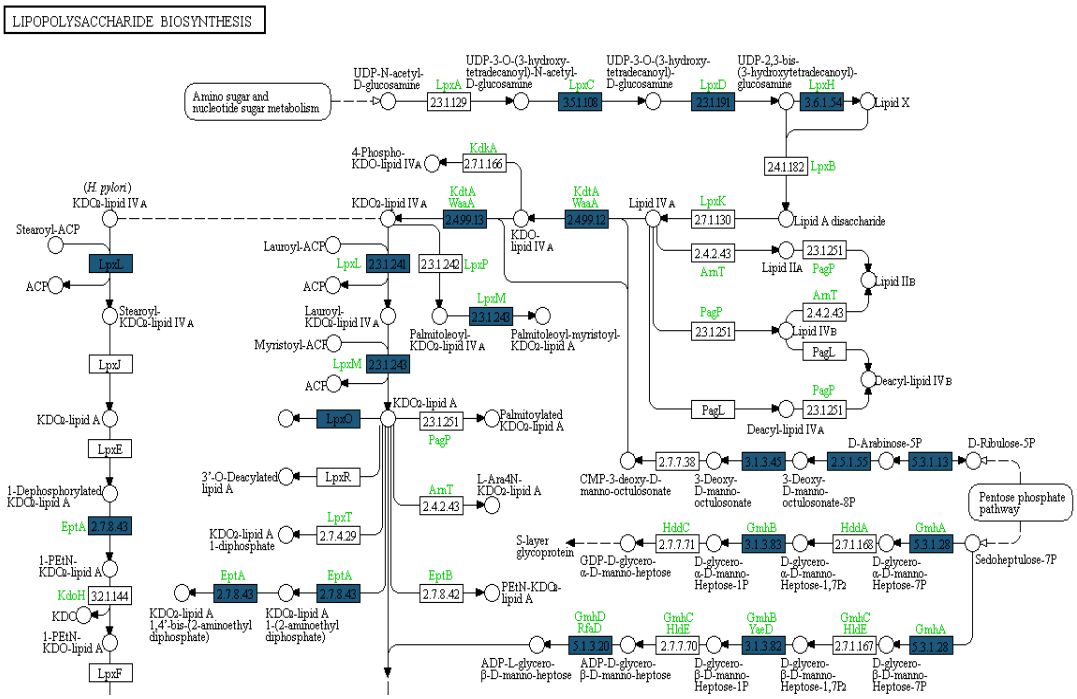

Figure 6-4 KO terms related to lipopolysaccharide biosynthesis in infant 74. Protein evidence of microbial production of lipopolysaccharides (LPS) is found at multiple time points (blue).

In response to microbial proteins, several host proteins related to the initiation of inflammation and the innate immune response were in high abundance in both infants. This indicates a robust host immune response to microbial organisms in both individuals. A transition from antimicrobial peptide production at early time points to the Toll-like receptor cascade was observed in infant 74. Adaptive immune response pathways varied across time in infant 06. These infant-specific immune responses mirror the observed infant-specific microbial metabolic processes.

6.2.2.3 Proteomic analysis of one infant with evidence of Candida parapsilosis colonization.

Further analysis was conducted for the individual with genomic and proteomic evidence of microeukaryotic colonization of the gut environment. Infant 06 was diagnosed with a *Candida parapsilosis* blood infection during hospitalization, and there were several proteins from this organism. detected in the fecal samples collected from this infant. Across the entire proteome dataset at all time points collected for this infant, 7063 host and microbial protein groups were quantifiable, with an average of 4872 protein groups at each time point. Among these quantifiable proteins, 5312 human protein groups and 1751 bacterial/microbial proteins were detected in the dataset. **Figure 6-5** shows the relative abundance of each organism in each sample. The limited amount of fecal material available for metaproteomic sample processing precluded the depletion of human cells before cellular lysis and protein extraction. Due to the inability to enrich the microbiome fraction before measurement, human proteins constituted more than 90% of the peptide abundance in each sample.

While this reduces the coverage depth of the microbial membership, it does allow simultaneous examination of both host and microbiome metabolic activities. Among the microbial proteins, there were 349 *C. parapsilosis* proteins quantified across all samples measured in the dataset, with a minimum of 126 *C. parapsilosis* proteins per sample. The relative protein abundance of this organism was highest in samples collected on day of life 29. While 349 proteins only represent

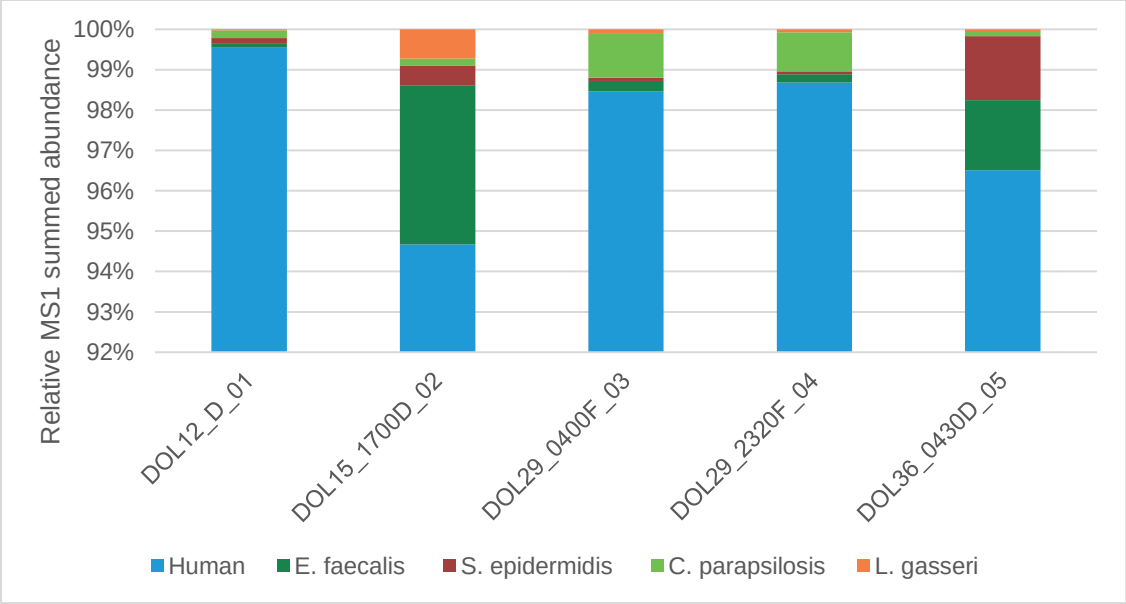

Figure 6-5 Organismal relative abundance based on summed protein abundances. In addition to highly abundant host proteins, protein evidence *of C. parapsilosis* and bacterial members was detected in all sampling time points. All members were established by day of life 12 and persisted across time.

around 6% of the predicted *C. parapsilosis* proteome, there was still ample evidence across time points to observe this organism's establishment and stability across time and decipher some of the general metabolic activities of the organism in the gut environment.

Among human proteins detected, there was ample evidence of neutrophil degranulation among the human proteins detected, with 324 out of 480 human proteins involved in the neutrophil degranulation pathway quantified across the entire dataset, with an average of 294 proteins detected at each sampling timepoint as annotated in the Reactome database. Based on an over-representation analysis, proteins related to neutrophil degranulation were found to be statistically enriched, with proteins detected in all ten of the reactions associated with this pathway. Taken together, functional information detected in the dataset indicates an active host immune response during the course of sample collection.

Among microbial proteins, **Figure 6-6** shows the metabolic activities of the microbiota across all sampling time points. Pathways containing reactions with protein evidence from more than one organism are denoted as a shared activity. Pathways that include reactions with associated protein evidence from an individual organism are marked as an organism-specific pathway. As noted in **Figure 6-6**, there is detectable evidence of *C. parapsilosis* core metabolic activities such as glycolysis, the TCA cycle, and organic acid metabolism. Repeated detection of similar abundances of these proteins across the 24-day timespan of collected samples indicates the establishment and stability of *C. parapsilosis* in the gut environment. Superoxide dismutase and heat shock proteins were among the most abundant *C. parapsilosis* proteins across all samples, suggesting an active response to oxidative stress. The presence of these proteins indicates *C. parapsilosis* was actively responding to, and adapting to, environmental stressors. In addition to *C. parapsilosis* metabolic activities, it is clear based on the protein evidence from participating metabolic reactions that the other microbes were also active within the community.

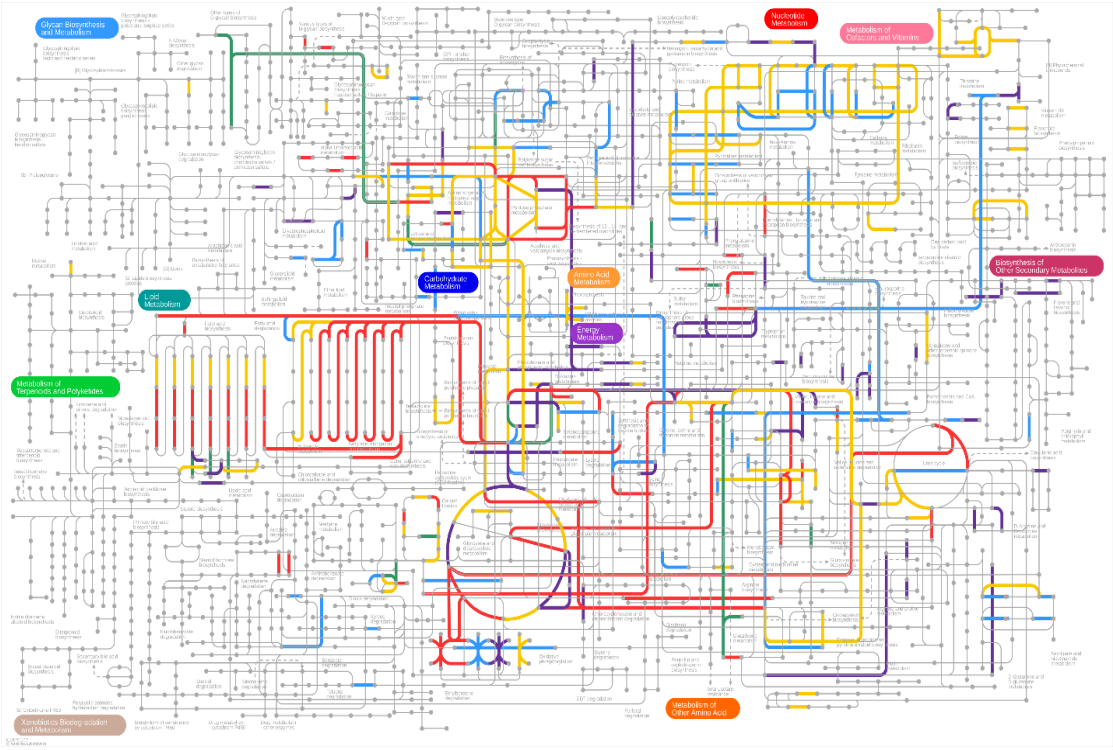

Figure 6-6 Global KEGG map of metabolism. Pathways are colored based on the organisms with protein evidence of participation in those pathways—shared metabolism (yellow), *C. parapsilosis* (red), *E. faecalis* (blue), *L. gasseri* (green), *S. epidermidis* (purple). Metabolic activities show distinct evidence of taxa-specific functional partitioning between *C. parapsilosis* and other microbiota.

In summary, segregation of functional activities was observed within the metaproteome of infant 06 between *C. parapsilosis* and bacterial members of the community. Taxa-specific metabolic partitioning within the metaproteome gives insights into cooperation and potential competition between community members. This metaproteome information provides insights into the kingdom-level functional partitioning of the preterm gut microbiome—specifically the functional importance and interactions of microbial eukaryotes within the intestinal microbial community during early life.

In addition to the metaproteomics work presented here, this study utilized metagenomics and metatranscriptomics to compare five unique *C. parapsilosis* genomes assembled from preterm infant microbiomes and environmental samples. Genome structures, population diversity, and *in situ* activity relative to *C. parapsilosis* reference strains grown in isolation were assessed. All five of the assembled genomes contained hotspots of single nucleotide variants (SNVs) that were shared by *C. parapsilosis* strains from multiple hospitals. Four of the newly reconstructed *C. parapsilosis* genomes have multiple copies of the RTA3 gene, which is implicated in antifungal resistance, potentially indicating *C. parapsilosis* adaptation to hospital antifungal use. Metatranscriptomic analyses showed that *C. parapsilosis* has a highly distinct profile *in situ* vs. in pure culture. Specifically, genes related to biofilm formation were relatively less expressed *in situ* and genes involved oxygen utilization differentially expressed between the two settings, indicating growth adaptations in response to the oxygen availability in the environment. Overall, these multi-omics results demonstrate that *in situ* study of *C. parapsilosis* and other microeukaryotes is necessary for a more holistic understanding of their biology in a community context.

6.2.3 Methods.

Longitudinal fecal samples collected from two preterm infants that had clinical isolation of *Candida parapsilosis* in blood or urine samples were analyzed using metagenomic sequencing and metaproteomic measurements. Untargeted DNA

shotgun sequencing was performed in technical replicate for matching samples for all time points analyzed by MS. Metagenomic sequencing reads were assembled and binned to generate genomic representations of the bacteria and eukaryotes present. For the generation of the proteomic datasets, lysates were prepared from ~ 50 mg of fecal material by bead beating in SDS buffer (4% SDS, 100 mM Tris-HCl, pH 8.0) using 0.15–mm diameter zirconium oxide beads. Cell debris was cleared by centrifugation (21,000×g for 10 min). Pre-cleared protein lysates were adjusted to 25 mM dithiothreitol and incubated at 85 °C for 10 min to further denature proteins and reduce disulfide bonds. Cysteine residues were alkylated with 75 mM iodoacetamide, followed by a 20-min incubation at room temperature in the dark. After incubation, proteins were isolated by chloroform-methanol extraction. Protein pellets were washed with methanol, air-dried, and resolubilized in 4% sodium deoxycholate (SDC) in 100 mM ammonium bicarbonate (ABC) buffer, pH 8.0. Protein samples were quantified by BCA assay (Pierce) and transferred to a 10-kDa MWCO spin filter (Vivaspin 500; Sartorius) before centrifugation at 12,000×g to collect denatured and reduced proteins atop the filter membrane. The concentrated proteins were washed with 100 mM ABC (2× the initial sample volume) followed by centrifugation. Proteins were resuspended in a 1× volume of ABC before proteolytic digestion. Protein samples were digested in situ using sequencing-grade trypsin (G-Biosciences) at a 1:75 (wt/wt) ratio and incubated at 37 °C overnight. Samples were diluted with a 1× volume of 100 mM ABC, supplied with another 1:75 (wt/wt) aliquot of trypsin, and incubated at 37 °C for an additional 3 h. Tryptic peptides were then spin-filtered through the MWCO membrane and acidified to 1% formic acid to precipitate the residual SDC. The SDC precipitate was removed from the peptide solution with water-saturated ethyl acetate extraction. Samples were concentrated via SpeedVac (Thermo Fisher), and peptides were quantified by BCA assay (Pierce) before LC-MS/MS analysis.

Twelve micrograms of each peptide sample were analyzed by automated 2D LC-MS/MS using a Vanquish UHPLC with an autosampler plumbed directly in line with a Q Exactive Plus mass spectrometer (Thermo Scientific). A 100-µm inner

diameter (ID) triphasic back column [RP-SCX-RP; reversed-phase (5 μm Kinetex C18) and strong-cation exchange (5 μm Luna SCX) chromatographic resins; Phenomenex] was coupled to an in-house pulled, 75 μm ID nanospray emitter packed with 30 cm Kinetex C18 resin. Peptides were autoloaded, desalted, separated, and analyzed across four successive salt cuts of ammonium acetate (35, 50, 100, and 500 mM), each followed by a 105-min organic gradient. Mass spectra were acquired in a data-dependent mode with the following parameters: a mass range of 400 to 1500 m/z; MS and MS/MS resolution of 35 K and 17.5 K, respectively; isolation window = 2.2 m/z with a 0.5-m/z isolation offset; unassigned charges and charge states of + 1, + 5, + 6, + 7, and + 8 were excluded; dynamic exclusion was enabled with a mass exclusion window of 10 ppm and an exclusion duration of 45 s.

MS/MS spectra were searched against custom-built databases composed of the concatenated sequenced metagenome-derived predicted proteomes from all time-points, the human reference proteome from UniProt, common protein contaminants, and reversed-decoy sequences using Proteome Discover 2.2 (Thermo Scientific), employing the CharmeRT workflow, with second search parameters turned off[142]. Peptide spectrum matches (PSMs) were required to be fully tryptic with two miscleavages, a static modification of 57.0214 Da on cysteine (carbamidomethylated) residues, and a dynamic modification of 15.9949 Da on methionine (oxidized) residues. False-discovery rates (FDRs), as assessed by matches to decoy sequences, were initially controlled at 1% at the peptide level. To alleviate the ambiguity associated with shared peptides, proteins were clustered into protein groups by 100% identity for microbial proteins and 90% amino acid sequence identity for human proteins using USEARCH[173]. FDR-controlled peptides were then quantified according to the chromatographic area under the curve (AUC) and mapped to their respective proteins. Peptide intensities were summed to estimate protein-level abundance based on peptides that uniquely mapped to one protein group. Protein abundance distributions were then normalized across samples using InfernoRDN[181], and missing values were imputed to simulate the mass spectrometer's limit of detection using Perseus[182]. KO terms were generated using GhostKOALA[183]. GOMixer was used to

evaluate taxonomic specific metabolic modules[184]. A metabolic module was inferred if more than 1/3 of the enzymatic steps in the module were covered within a single phylum. The phylum abundance of each module was calculated in GOMixer by using the summed area under the curve (AUC) abundance of each phylum's KO terms that mapped to the module. Pathview was used to visualize detailed pathway maps[185], and iPath3 was used for global maps[186]. The Reactome database (version 68) was used for annotation of human immune system proteins[187].

6.3 Understanding how the developing human preterm infant gut microbiome responds to antibiotic administration and host immune system-induced metal bactericidal control.

Text and figures adapted from Peters et al., 2022. Frontiers in Microbiology (in submission).

6.3.1 Introduction.

Early-life microbial colonization of mucosal surfaces in the gut during infancy is critical for balanced priming and education of the host immune system. Recent studies are defining the period in early development, a so-called "window of opportunity," where disruption of critical host-commensal interactions have lasting and sometimes irreversible impacts on the training of specific immune subset[188]. Studies using rodent models have shown that a brief, postnatal germfree period caused permanent changes in levels of systemic regulatory T cells, natural killer cells, and cytokine production[189] and that microbial exposure during early life has persistent effects on natural killer T cell function, which was shown to be crucial in the development of intestinal inflammatory disorders such as colitis[190]. Other intestinal inflammatory diseases, such as necrotizing enterocolitis, can result from exaggerated immune responses[191,192]. Necrotizing enterocolitis (NEC) is a fatal disease of neonatal preterm infants in 30-50% of cases[193]. It is associated with intestinal inflammation driven by the microbiota, making it an ideal model for studying early-

life host-microbe interactions in a dysbiotic gut[194,195]. Most interactions between microbiota and the host immune system occur on the mucosal surface of the large intestines. This surface is lined with intestinal epithelial cells (IECs), which create a physical barrier between the microbiota in the intestinal lumen and the surrounding tissue. IECs secrete antimicrobial peptides such as chemokines, cytokines, and defensins that function to prevent bacterial from breaching the gut wall[196.] Bacterial toxins and pathogen-associated molecular patterns (PAMPs), including bacterial flagellin, unmethylated CpG oligodeoxynucleotides, lipopolysaccharide (LPS), lipoteichoic acid, are recognized by host toll-like receptors, which can activate pro-inflammatory responses. There is a delicate balance related to these host-microbe interactions in order to maintain homeostasis in the environment. However, the underpinnings of this balance are largely unknown, especially related to the temporal aspects of host-microbe interactions during the development of dysbiosis and inflammation. It is still unclear whether microbial community and functional shifts precede inflammation or are a consequence of it.

In addition to the still poorly understood interplay between the host immune system and microbiota, other external factors, such as antibiotic administration impact the developing gut environment. Preterm infants commonly receive antibiotics early and frequently in life due to their vulnerability to infection[197,198]. Previous metagenomics studies have shown enrichment in antibiotic resistance genes in preterm infants that receive early-life antibiotics[197,199]. This perturbation of optimal microbiota colonization likely has long-term impacts on microbial establishment patterns and, ultimately, host health outcomes[200,201]. In addition, intrapartum antibiotics are often given to the mother as a preventative measure during the delivery process[202,203]. Previous work has demonstrated that the administration of antibiotics before birth profoundly impacts the composition of the gut microbiome and ultimately patient health status[204–206]. While host health outcomes based on gut microbial perturbations caused by both maternal and infant antibiotic administration have been widely demonstrated[207], the exact mechanism by which it plays on the priming of the developing immune system is still unclear.

Several studies have shown that antibiotic resistance is linked to tolerance of high environmental concentrations of metals such as copper[208,209]. The phenomenon of co-resistance to antibiotics and excess metals has been observed in several organisms commonly found in the human gut[210–214], including one study that reported that multi-drug-resistant Enterobacterales carried metal resistance genes up to seven times more frequently compared to antibiotic-sensitive strains[215]. Disrupting the availability of metals in the environment is an evolutionarily ancient strategy to limit bacterial growth, and vertebrates have developed immune-related mechanisms, in processes referred to as nutritional immunity, that restrict the availability of some metals to microbiota in the environment while at the same time flooding infection sites with antimicrobial levels of other metals[216,217]. Concurrently, many organisms have developed mechanisms to maintain intracellular metal homeostasis, including resistance to excess environmental concentrations, as the balance of metals in the cell is necessary to prevent mismetallation and cell damage[218,219]. In vivo characterization of metal homeostasis in the gut environment is challenging due to the technical limitations of most traditional microbiome research methods. Nevertheless, this is an understudied area that needs to be considered for its unrecognized role in shaping the early life gut.

To this end, this study was designed to examine the gut microbiota response to human host-imposed immune control during establishment of the developing gut microbiome over the first few months of life, specifically utilizing a high-resolution LC-MS/MS approach that enables simultaneous investigation of host and microbial functions that impact this colonization process. This was based on a genome-assembled metaproteomics longitudinal study of fecal microbiome samples collected from hospitalized premature infants, including a subset of infants who developed NEC during the study. The developing infant gut is an ideal model system because infant microbiomes are far less complex than adult microbiomes. The early-life period is critical in the maturation of both the immune system and microbial communities. Earlier investigations on a subset of the samples used in the present analysis predominantly focused on microbial metabolic functions and community

composition[150,220]. These investigations led to a few key findings that expand understanding of the establishment dynamics in the early-life preterm gut. First, the initial investigation focused on four of the infants used in the present analysis demonstrated that the premature infant microbiome is highly variable, with metaproteome measurements showing drastic differences in functional composition between infants and across time. Second, following these findings, an expanded analysis was conducted for paired metagenomic and metaproteomic measurements for samples collected from fifteen infants, which revealed that genetically similar organisms colonizing different infants have very distinct proteomes with unique metabolic profiles, supporting the idea that microbiome function largely depends on microbial responses to the physiological conditions present at a given point in time in the infant gut and is not solely dependent upon which organisms are present. Third, microbiome diversity was lower in samples collected during or within five days of antibiotic treatment. Using replication rates and overall abundances of organisms within these infants demonstrated that some organisms in the preterm infant gut persisted and continued replicating in the presence of administered antibiotics, indicating some level of resistance to those antibiotics. Finally, in several instances, there were shifts in both taxonomy and microbial function preceding NEC diagnosis; however, no specific species or metaproteome type could explain all infants who developed this condition during the study. However, these works largely ignored the contribution of the host immune system, which plays a crucial role in shaping the early-life gut environment[189,221–224], although the exact mechanisms of this process are still poorly understood. A clear understanding of host-microbe interactions during the colonization process depends upon the intrinsic capabilities of metaproteomics measurements to capture host and microbiota functional information simultaneously and link measured entities back to source organisms. In the present analysis, we set out to elucidate the host immune system's specific roles on microbial establishment dynamics and to study how microbes modulate their function during the colonization process in response to and as a stimulant of the developing host immune system.

6.3.2 Results.

Metaproteomics measurements reveal simultaneous information about host immune processes and microbial functions in the fecal samples. Samples were collected during the first three months of life from hospitalized premature infants, including six infants that went on to develop necrotizing enterocolitis (NEC) during the study. In total, 91 stool samples from 17 infants were measured and analyzed **(Figure 6-7)**. Due to the nature of sample collection in this clinical study, there were unavoidable differences between infants related to the number of samples collected and the times that samples were collected among infants in the study.

Across the entire dataset, 595,324 peptide sequences were measured across all 91 samples. This equated to, on average, 2000 human and 5000 microbial protein groups quantified per sample, respectively (**Figure 6-8A**). Although a newer distinct database searching strategy was utilized in the present study, the average number of peptides and protein groups identified per sample was similar to results obtained with the alternative approaches used in previous investigations. In each sample, human and microbial protein group identifications had similar relative proportions, with microbial proteins composing ~ 70% of the identified protein groups. Both the number of samples collected per infant and the timespan between timepoints impacted protein identifications. Individuals with several collected samples spanning an extensive time range had more unique protein identifications than infants with only a few samples collected over a very short period. However, when looking at a higher level (pathways and protein classes), the functional categories were fairly uniform across samples from all infants. Among human protein groups, immune protein groups make up around 10-15% of relative abundance when looking at protein groups that are uniquely mapped to one category (**Figure 6-8B**), with ~45% of quantified proteins unique to each specific immune pathway.

Functional β-diversity of host and microbial proteins reveals distinct partitioning which is quite variable and is highly infant-specific. To narrow the investigation to focus on biological drivers of microbiome function related to host

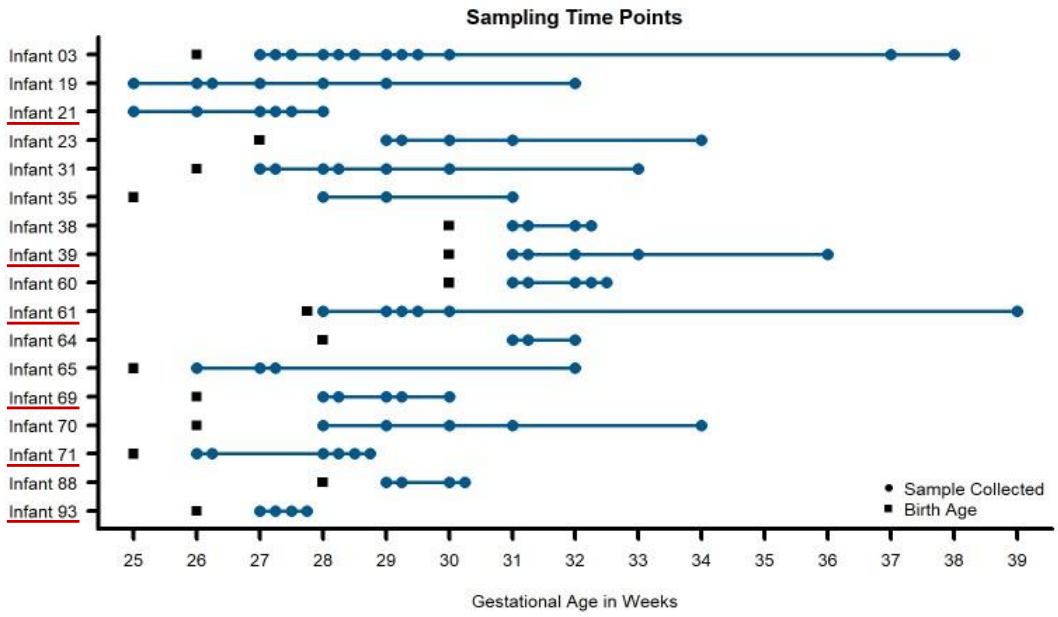

Figure 6-7 Fecal samples collected for metaproteomic measurements for all infants in the study. Using necrotizing enterocolitis (NEC) as a representative dysbiotic condition, 91 fecal samples (blue circles) were taken from seventeen preterm infants over the first 90 days of life. Multiple time points were measured per infant to assess intraindividual variability over time. Six infants developed NEC during the early stages of life (underlined in red on the y-axis). Birth age (in gestational weeks) is noted for each infant with black squares.

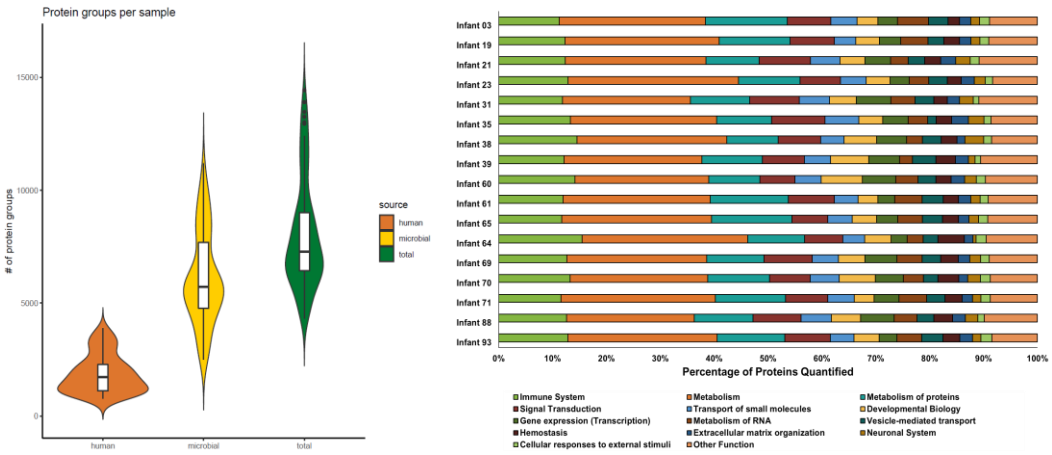

Figure 6-8 Distribution of protein groups identified in each sample for each protein source (A) and the percentage of human proteins quantified in each Reactome annotation category (B). In panel A, protein group identifications were based on uniquely mapping peptides. Microbial and human proteins were clustered at 100% and 85% sequence similarity, respectively. Each violin plot illustrates the kernel probability density for all samples collected (the width represents the proportion of the data) and the horizontal bars on the box plot depict the median and interquartile range of the distribution. In panel B, the Reactome database was used to annotate human proteins based on proteins that uniquely map to one category. The bar plots show the proportion of proteins in each Reactome category quantified per infant. Among the quantified human proteins, proteins related to immune functions (lime green bars) represent around 10-15% of human protein identifications.

immune mechanisms, we used non-metric multidimensional scaling of Jaccard distances to assess the functional β-diversity of samples based on the presence quantified in each sample. The functional dissimilarities between samples were explored on four different levels: (1) all quantified human protein and microbial KEGG orthology (KO) terms, (2) only microbial KO terms, (3) all human proteins, and (4) only immune-related human proteins. In this study, NEC was used as a representative dysbiotic condition to compare normal and abnormal establishment of the gut environment. Due to the heterogeneity in colonization patterns among infants who did or did not go on to develop this condition, it is apparent that other health factors can also have critical impacts on both host and microbial protein expression during this phase of life. To explore the impact of these other factors, ordination analysis was paired with PERMANOVA statistical testing to see if any health metadata categories were significant drivers of separation between samples. Among the metadata factors that played a role in the functional β-diversity of the samples, the ordination analysis revealed functional partitioning among samples based on several metadata categories including infant number, delivery mode, feeding type, and the development of infection, sepsis, or NEC. Other factors such as the age of the infants based on gestational age or day of life, birth age, and birth weight did not significantly impact the functional dissimilarity of the samples at any of the four protein levels considered. The largest factor contributing to the variance between samples considering all four functional levels was the infant each sample was collected from, with 79% of the variance across all proteins measured in the dataset explained by this factor.

Due to the heterogeneity in both infant health characteristics (delivery mode, gestational age, disease status, antibiotic administration, time of sample collection, etc.), many of the observed trends were largely infant-specific. Based on this level of global analysis, it was clear that the approach of evaluating all samples from the infants in the cohort together could lead to a confounded interpretation of the data, as host-microbial interactions are highly dependent on the environmental context at the time of sample collection. In addition, grouping samples without accounting for sampling

time washed out some observations of highly dynamic, time-dependent processes. The ability to interrogate these temporal dynamics was a key strength of this sample set, which led us to the conclusion that an alternative approach needed to be taken to analyze the data further. In light of this conclusion, we set out to investigate some of the unanswered questions that arose in previous analyses of this dataset related to antibiotic resistance and shifts in community abundance that were not associated with times of disease onset.

6.3.2.1 Antibiotic resistance mechanisms help selected microbes overcome susceptibility to antibiotics and persist in the environment.

Based on previous findings[150,220], it was apparent that there were antibiotic resistance-related proteins in *E. faecalis* within some of the infants in this cohort and the replication rates of selected microbes during periods of antibiotic administration indicated antibiotic resistance might be highly prevalent in organisms dominating the early-life gut environment. However, protein expression of the specific antibiotic resistance mechanisms used by any microbes in this dataset was not interrogated. Here, we set out to determine the extent of antibiotic resistance in the gut microbiome for all infants in this cohort and to characterize the diversity of antibiotic resistance mechanisms being utilized by the microbiota. All quantified microbial proteins were searched against the comprehensive antibiotic resistance database (CARD). In total, 241 protein sequences were identified across all 17 infants that mapped to 93 antibiotic resistance ontology (ARO) terms. These proteins represented six different resistance mechanisms. This protein overlap illustrates the functional redundancy of gut microbiota—with several organisms simultaneously expressing the same antibiotic resistance protein. It also supports the previous inference that the infant's gut is a reservoir for antibiotic resistance genes[201,225,226].

On average, 17 AROs were found in each timepoint across infants. In general, while the overall abundances of antibiotic resistance proteins changed over time, the relative contribution of individual resistance mechanisms utilized by microbiota

remained unchanged across time in a specific individual unless perturbation of the system by administration of antibiotics. In addition, the overall ARO composition was highly variable among infants. Among the entire infant cohort **(Figure 6-9)** only four infants (3, 35, 65, and 71) had evident variations in the resistance mechanisms over time without antibiotic administration. In these infants, over time there was a change in the types of resistance mechanisms being expressed that did not correspond with periods of administered antibiotics. In addition, in several infants, there was a change in the expression of AROs (based on summed abundance) that was not proportional to the change in the overall abundance of organisms expressing those proteins at those times. Overall, the antibiotic resistance patterns are somewhat chaotic over time, and observations regarding antibiotic resistance are best made at the level of the infant.

Upon examining the relative summed protein abundance of each organism in the community across the longitudinal series for each infant, some changes in community composition were observed, as would be expected based on antibiotic treatment of the infants over the first several weeks of life. In samples collected after the antibiotics were administered, there was a dramatic shift in the relative community composition based on this perturbation. While the susceptibility to antibiotics and associated resistance mechanisms expressed by microbes played an essential role in establishment dynamics, several dramatic changes in organismal abundance were not explained by the expression, or lack thereof, of antibiotic resistance-related proteins in individual microbes. This indicates there must be other drivers of community establishment dynamics, and we hypothesized that microbial responses to host immune mechanisms might explain some shifts in relative organismal abundance across time in each infant.

Looking at samples across time within each individual yielded unique patterns of community dynamics in response to antibiotic administration and host processes when patient medical information was considered. Leveraging patient metadata regarding disease onset and the timing and types of antibiotics administered to each infant, functional shifts of individual microbiota could be differentiated between microbial responses to known and external perturbations on the gut environment

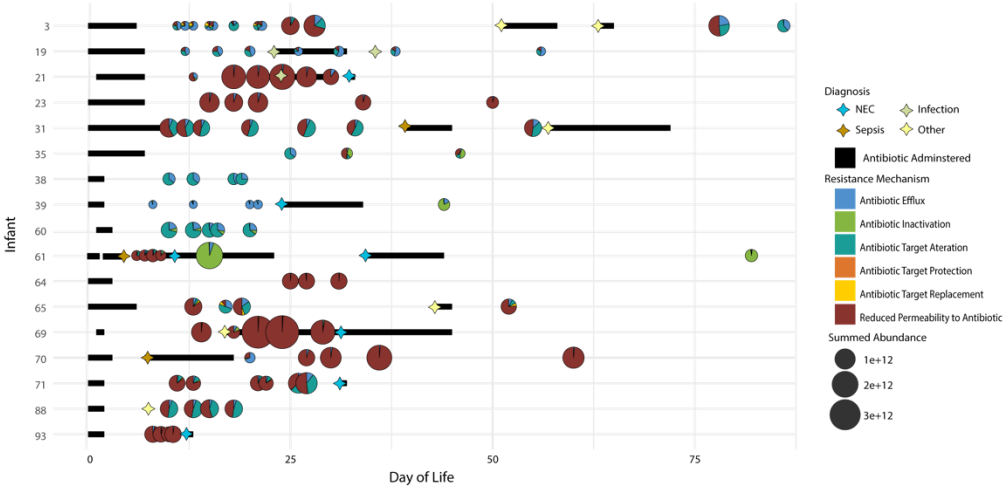

Figure 6-9 Distribution of antibiotic resistance orthologs (AROs) found in each sample for all 17 infants in the study. Samples are plotted based on time of collection (DOL) on the x-axis and by the infant on the y-axis. The size of the bubble indicated the summed ARO abundance in the sample. The pie charts indicate the contribution of each CARD database antibiotic resistance mechanism identified in the sample to the overall summed ARO abundance. Also plotted are markers indicating dates of disease diagnosis and bars to indicate periods of time when antibiotics were administered.

analysis is highly individual-specific, it provides insights into community establishment caused by antibiotic administration and unknown and internal perturbations caused by interactions with host processes and other community members. While this type of dynamics that would otherwise be missed if samples from infants were generalized to broad analysis classes without accounting for the heterogeneity of all longitudinal and environmental factors which have a non-negligent impact on the colonization of the gut environment. To highlight this type of health-status-based longitudinal analysis, a few example infants will be highlighted and discussed that have a wide range of sampling time points leading up to disease diagnosis and subsequent antibiotic treatment and samples collected during and after antibiotic administration. Although only two infants are described in detail for the following analysis, the protein expression profiles are not unique to these example infants but are found in multiple infants across the cohort.

Metal homeostasis mechanisms help microbes overcome host-imposed copper toxicity and iron restriction Out of the 17 infants in the cohort, infant 61 was the only infant diagnosed with sepsis and recurrent NEC. Several samples were collected from subsequent days prior to the initial NEC diagnosis as well as samples collected during and after antibiotic administration. This sampling allowed us to examine the observed changes in microbial functional composition preceding antibiotic administration. In infant 61, after antibiotic treatment is initiated on day nine, there is a dramatic shift in the relative community composition based on this perturbation with both a reduction of some community members and a subsequent rebound in the relative abundance of those members observed after antibiotic treatment is completed **(Figure 6-10)**. However, not all changes in organismal abundance can be explained by susceptibility to antibiotics and associated resistance mechanisms being expressed. Among the three dominant organisms in this infant (*Enterococcus faecalis, Klebsiella pnuemoniae*, and *Enterobacter cloacae*), all three are typically susceptible to the types of antibiotics delivered during the course of the study. *E. faecalis* was the dominant organism leading up to NEC diagnosis on day nine and decreased in relative abundance following subsequent initiation of antibiotic treatment. After six days of continuous

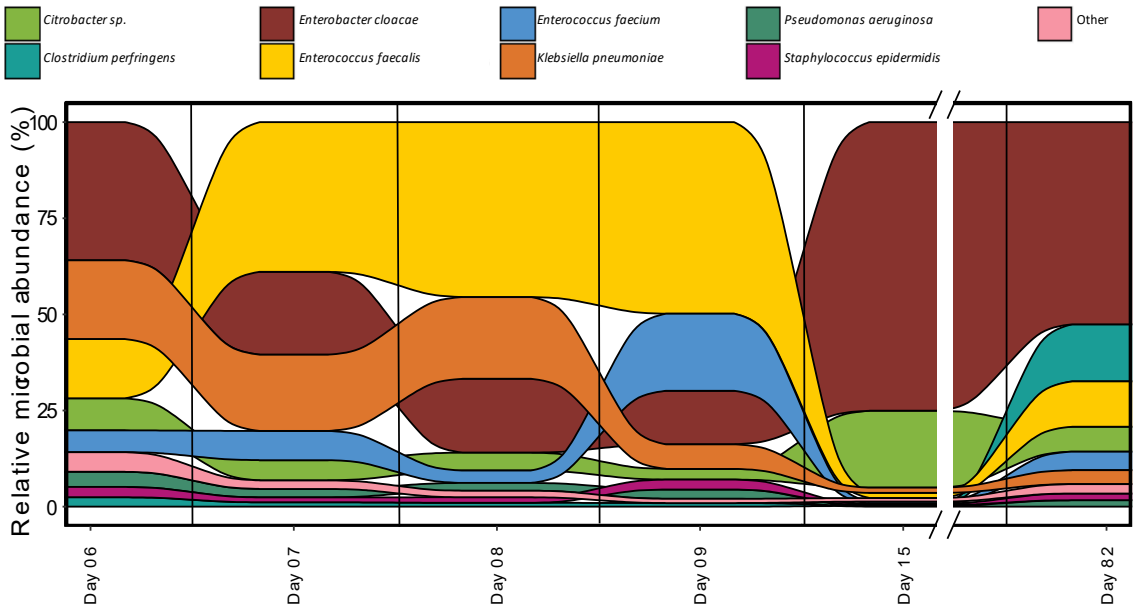

Figure 6-10 Relative abundance of organisms at each time point for samples collected for infant 61.

antibiotics, it was drastically reduced in relative abundance, corresponding to its susceptibility to the antibiotics administered and the lack of resistance mechanisms. *E. cloacae* increased in relative abundance after the administration of antibiotics on day nine and persisted as the dominant organism in the community until the measurement on day 82. *K. pnuemoniae* was the dominant organism with protein expression relating to antibiotic resistance mechanisms**.** However, this organism's relative abundance was already decreasing by day nine, before the start of antibiotic treatment, which indicates there are additional drivers of community establishment. We hypothesized that host immune responses might be contributing to the community shift.

Among host proteins related to immune response, there was increased host expression of proteins related to copper trafficking for antimicrobial activity starting at day eight. Expression of host proteins related to copper trafficking, including PDZD11 and binding partner PLEKHA5, was observed in five out of the six timepoints collected for infant 61. In addition, there was also evidence of host trafficking of iron observed in this infant, with all timepoints in this infant displaying the highest average abundance of the siderophore binding protein, lipocalin-2 (LCN2), observed across the entire infant cohort. LCN2 is known for its ability to bind bacterial siderophores, microbial proteins that assist in iron harvesting, such as enterobactin and salmochelin, to limit the bacterial iron acquisition and subsequently hamper the growth of microbes[227].

Based on the indications that some infants were mounting an active immune response utilizing copper and iron trafficking-related antimicrobial activities, we focused on related metal trafficking activities in the microbiota. In each infant based on the increased host expression of proteins related to copper toxicity as an immune response, including in infant 61, there was a corresponding microbial expression of expressed proteins associated with copper toxicity resistance or tolerance, including CueR, CueO, and CopA in some bacteria. These proteins are specifically involved in the efflux of Cu(I) from the cell and the transformation of Cu(I) into Cu(II). This

resistance mechanism explains how some organisms within these infants might be persisting despite pressure from both host defenses and antibiotic administration.

Comparing the functional activities of the individual microbes in this infant, we found that *E. cloacae* expressed many proteins related to copper efflux. These proteins were either not observed, in reduced abundance, or with fewer observed proteins in the other microbes in this infant. This suggests copper efflux and resistance were not as prominent in these organisms, even if there was an active response to excess copper. In addition, some of the proteins involved in the copper response identified in the two other dominant organisms do not have definitive physiological roles, unlike the proteins observed in *E. cloacae*, where there is a definitive path of efflux and detoxification of copper. Among the other dominant organisms in infant 61, *E. faecalis* was also found to express copper homeostasis-related proteins (CopA, CopB, and CutC), but at lower levels than *E. cloacae*. In addition to proteins involved in mitigating intracellular copper toxicity, only *E. cloacae* showed expression of multiple proteins related to the biosynthesis of the siderophore enterobactin among the organisms in this infant. As only *E. cloacae* was the only organism with observed expressional evidence of producing and importing enterobactin in this infant due to the elevated demand for iron-based on high levels of Cu(I) ions produced through host defense mechanisms, the other species were compromised by their inability to offset Cu(I) levels intracellularly with iron and the lack of different mechanisms to export Cu(I).

Overall, the longitudinal measurements of microbial functional dynamics in this infant indicate that following the increase in the apparent host-imposed copper toxicity on day eight, the relative abundance of the microbiota is driven by the ability to respond to this host immune response. In summary, the observed organismal abundance trends in this infant during periods of antibiotic administration and host immune-mediated metal toxicity show that microbial resistance and homeostasis mechanisms to both of these perturbations are critical for persistence and maintaining establishment in the developing gut. For example, lacking either antibiotic resistance mechanisms (*E. faecalis*) or copper efflux proteins (*K. pnuemoniae*) relate to a

decrease in relative abundance, while expression of proteins related to both of these mechanisms (*E. cloacae*) relates to increased relative abundance in the community after environmental stress.

6.3.2.2 Host sequestration of manganese and zinc does not impact microbes with enhanced import capabilities.

Infant 39 had multiple time points sampled across the first several weeks of life, with samples collected before and after NEC diagnosis and subsequent antibiotic administration. There was a shift in community structure at the time of NEC diagnosis on day 24 and the subsequent treatment, which we presume was predominantly caused by the administration of antibiotics **(Figure 6-11).** For example, *C. freundii* expressed the majority of the antibiotic resistance-related proteins in this infant and dramatically increased in abundance after administration of these antibiotics. This supports the idea that organisms with antibiotic resistance mechanisms may persist and expand while the organisms lacking these mechanisms become dramatically reduced in abundance after antibiotic treatment. While the expression of antibiotic resistance genes by organisms during periods of antibiotic administration plays a role in shaping overall community composition, there are also changes in organismal abundance that are not correlated to microbial response to antibiotic treatment.

There was also an unexplained shift in community functional composition between samples on DOL 13 and 20. During this time period, there was a corresponding host expression of host immune proteins related to manganese and zinc trafficking. Two host proteins, S100A8 and S100A9, can be involved in metal sequestration by antimicrobial peptides and were found in high abundance across all 91 samples in the dataset. These two proteins form the heterodimer calprotectin, which is often used as a fecal biomarker of inflammation, and sequesters both zinc (Zn) and manganese (Mn) when it is released by neutrophils at infection sites [228]. In infant 39, calprotectin showed more than a three-fold increase during this period before returning to early-life levels at DOL 44. The most dramatically shifting

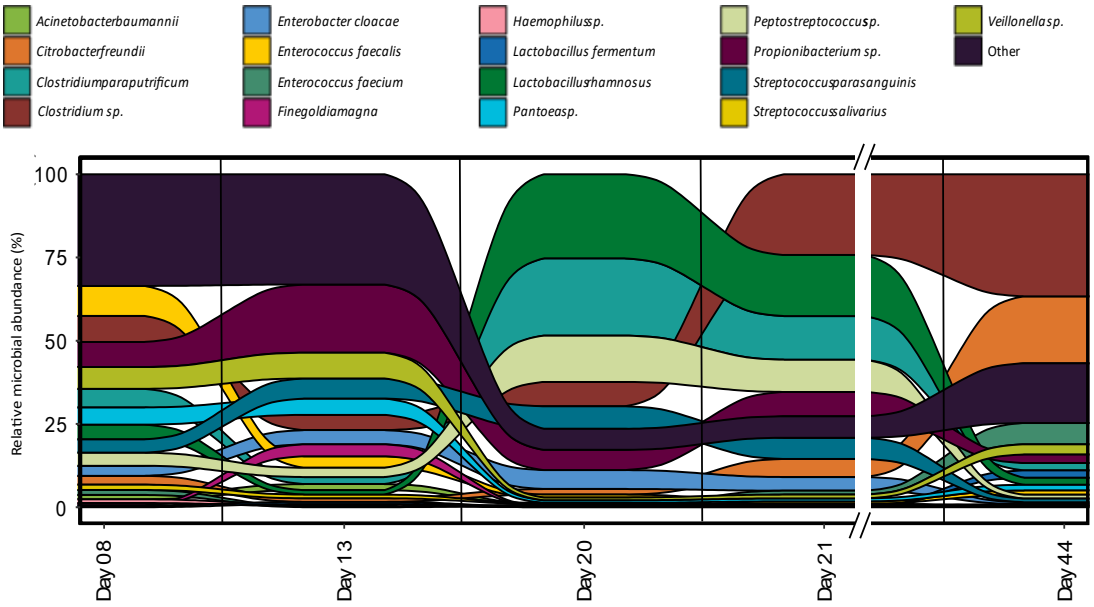

Figure 6-11 Relative abundance of organisms at each time point for samples collected for infant 39.

organism between days 13 and 20 was *Lactobacillus rhamnosus*. We hypothesized that the dramatic increase and later decrease of calprotectin might be influencing the relative abundance of *L. rhamnosus* over the first several weeks of life. As calprotectin inhibits bacterial growth through chelation of manganese and zinc, we looked to associated mechanisms, such as membrane transporters, in *L. rhamnosus* related to manganese and zinc import that helped this organism import these metals into the cell to overcome the host-imposed metal limitation. This organism was expressing mntA, mntB, mntC, and mntH. MntABC has been demonstrated to import manganese during low Mn supply, and mntH specifically competes with calprotectin for luminal manganese[229]. In addition to manganese-specific transport proteins, several proteins were either specific to zinc transport or related to the non-specific transport of zinc and manganese, which indicates the microbes in this gut environment were actively coping with zinc and manganese limitations in the surrounding environment. *L. rhamnosus* expressed non-specific zinc/manganese transport system substrate-binding proteins and a manganese-dependent inorganic pyrophosphatase (ppaC), which indicates the microbe was actively importing manganese and utilizing it within the cell. In summary, based on the presence of these transporters and the expression of several manganese-dependent proteins in microbes who increased in abundance during periods of high host expression of calprotectin, we can infer host sequestration of this metal does not impact the growth of *L. rhamnosus* in this infant. This enables *L. rhamnosus*, and other microbes with these enhanced manganese and zinc import capabilities, to outcompete other microbes lacking these mechanisms (*C. freundii*) in manganese and zinc-limited environments.

Maternal intrapartum antibiotic administration drives grouping of samples based on host and microbial proteins Due to the scope of the study, there was limited focus on the impact of intrapartum antibiotic administration on the development of the gut microbiome and impact on infant health outcomes at the time of sample collection. Considering this, comprehensive records of maternal antibiotics administered prior to delivery were not collected for all mothers for infants in this cohort. Despite this limitation, some interesting results emerged when evaluating the

limited metadata for this environmental factor. Conclusions about the findings presented for this infant cohort might be different with more complete metadata collection and therefore should not be generalized to all preterm infants. However, the findings of this analysis are still presented in the current study for two reasons. First, the analysis methods utilized in this study can be applied to other studies where more comprehensive medical information is collected at the time of sample collection to further interrogate how external environmental factors, such intrapartum antibiotic administration, help shape the development of the early-life gut environment. Second, we attempted to reduce the impact of missing metadata on the findings, by removing infants from the analysis with missing maternal metadata, so the findings might still be the same even with better recording of maternal medical information.

Among the metadata factors that played a role in the functional β-diversity of the samples, the ordination analysis revealed functional partitioning among infants based on the timing of maternal antibiotic administration **(Figure 6-12 A-C)**. Within this cohort, infants were selected for further analysis based on maternal medical information detailing when prophylactic intrapartum antibiotics (IAP) administration was initiated. Four out of the 17 infants were excluded based on a lack of information related to IAP timing. Another infant was excluded from further analysis due to maternal administration of antibiotics in addition to ampicillin and gentamycin for the treatment of a group B streptococcus infection (GBS) and chlamydia. Of the twelve infants where sufficient metadata was provided, the nMDS ordinations showed 9% of the variance between the samples was explained by whether or not maternal antibiotic administration was initiated more than 24 prior to delivery. Of the 62 samples included in the down-selected cohort, the sampling distributions based on maternal antibiotic timing were fairly even between the two groups based on the day of life timescale compared to the gestational weeks timescale **(Figure 6-12D)**. As many of the infants whose mothers received intrapartum antibiotics less than 24 hours prior to delivery were younger in gestational age than infants in the other group, further analysis was conducted considering the day of life timescale due to the even sampling distribution. Both the human immune proteins and the microbial KOs had significant

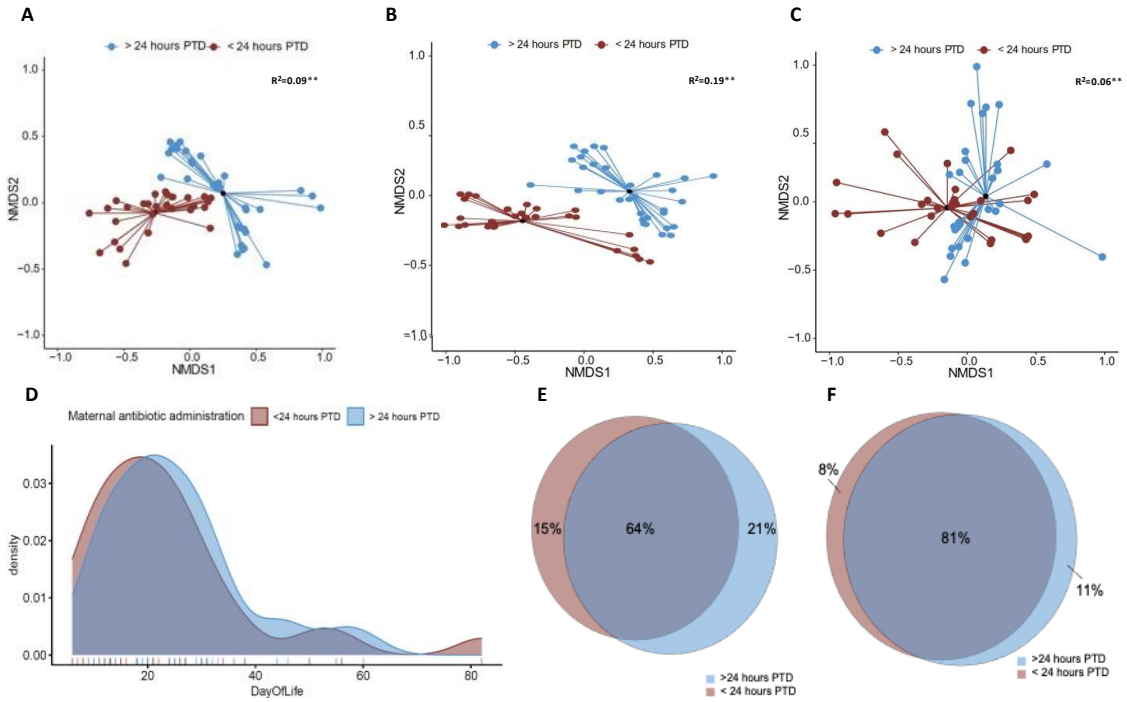

Figure 6-12 Functional β-diversity by proteins source for the 12 infants included in the analysis based on maternal antibiotic administration information. Non-metric multidimensional scaling (NMDS) of Jaccard distances for the collective functionality (based on presence or absence of proteins or KEGG ortholog groups [KOs]) of each sample for (A) Human immune proteins and microbial KOs, (B) human immune proteins, (C) microbial KOs. ** indicates a p-value < 0.05 by Adonis test. (D) Density plot of DOL sampling distribution for the 62 samples included in the analysis. Venn diagrams of (E) human immune proteins or (F) microbial KOs were quantified in groups of samples based on whether or not maternal antibiotic administration was initiated more than 24 prior to delivery.

results. However, 19% of the variance in detected immune proteins was explained by the maternal antibiotic delivery timing (< or > 24-hour PTD). In contrast, only 6% of the variance in microbial KOs across samples was explained by this metadata category. As both host and microbial protein expression were significantly different between samples in each group, further analysis was conducted to determine if specific pathways or protein classes significantly differed across time between infants based on the timing of maternal antibiotics.

Both antibiotic timing groups (< or > 24-hour PTD) had considerable overlap with many immune proteins or microbial KO terms shared between samples in both groups **(Figure 6-12E-F)**. However, many of the proteins or KO terms that were only found in samples of a particular group were largely infant-specific and not necessarily representative of each group as a whole. Other approaches beyond looking at the presence or absence of specific proteins were incorporated into the analysis to understand better which specific pathways or functions were different between groups across all infants within each group. To test whether the trajectories of host immune and microbial pathways were different between groups of infants in the maternal antibiotic administration sample subset, spline-based statistical analysis was used to determine if the protein pathways follow different trajectories between the two groups of infants over time than would be expected by random chance. For the host immune pathways, six pathways had significantly different trajectories between infants based on the timing of maternal antibiotics prior to delivery. These include interferon signaling, MHC class II antigen presentation, cytosolic sensors of pathogen-associated DNA, neutrophil degranulation, interferon signaling, class I

MHC mediated antigen processing & presentation, and ROS and RNS production in phagocytes. In addition to host proteins, longitudinal trajectories of several microbial pathways were significantly different between the two infant groups, including the expression of antibiotic resistance genes and some ABC transporters.

As a complementary approach to the trajectory significance testing between the maternal antibiotic administration subset, an indicator value analysis was

performed using the IndVal package in R to determine which specific immune proteins are driving the separation of groups the NMDS ordination/PERMANOVA analysis. Of the 1140 immune-related proteins identified across the 62 samples in the maternal antibiotic administration subset, 529 of these proteins were considered significant indicators based on an adjusted p-value of <0.05. 28 of these proteins were indicators for the > 24 hours PTD group and 501 proteins were indicators of the <24 hours PTD group. To further refine the results, an additional filter of an IndVal>0.5 was included to increase stringency as many of the significant indicators were low abundance proteins or were found in one or a few samples. Using these thresholds, 235 indicator proteins were found to be significant indicator proteins. Hierarchical clustering of these 235 proteins resulted in three main groups. Within each maternal antibiotic administration group, there was no pattern related to the age of the infant for the samples. The first cluster of 149 proteins is dominated by proteins that are in low abundance in the <24 hours PTD group and are mostly absent in the >24 hours PTD group. The top over-represented pathways in this cluster included neutrophil degranulation, MHC class II antigen presentation, and signaling by interleukins. In addition to immune pathways, cellular responses to chemical stress pathways were overrepresented, including ROS sensing by NFE2L2. Cluster 2 contains 64 proteins that were of medium abundance in the <24 hours PTD group and lower abundance in the > 24 hours PTD group. Enriched pathways include neutrophil degranulation, binding and uptake of scavenger receptors (which is not an immune pathway, but shares many proteins with immune pathways), FC epsilon receptor signaling, and signaling by the B cell receptor. The third cluster contained 22 proteins that were present in all 62 samples and high abundance in the <24 hours PTD group and lower abundance proteins in the >24 hours PTD group. In this cluster, within the < 24-hour PTD group, proteins were highest at early in life, with stable or decreasing abundances over time. Protein abundances were relatively stable across time in the >24 hours PTD group. Of the 22 proteins, neutrophil degranulation, the initial triggering and regulation of the complement cascade, and antimicrobial peptides (specifically metal

sequestration by antimicrobial peptides) were enriched based on these significant indicator value proteins.

In addition, this analysis based on microbial proteins yielded 634 indicator proteins to differentiate between groups of infants based on maternal antibiotic administration. 255 proteins were found to be indicators of the microbial metaproteome of an infant with administration <24 hours PTD. However, most of these were part of core metabolic processes and were mostly low abundance proteins. They most likely met indicator value significance thresholds due to presence or absence in a particular group as they were near the limit of detection, and not because they are indicative of a particular group.

6.3.3 Discussion.

Across the entire dataset, there were thousands of microbial and human peptides measured per sample, which equated to an average of 3000 human proteins and 9000 microbial proteins identified per infant. This proteomic coverage achieved allowed in-depth elucidation of both host and microbiota function in the early-life preterm gut. Among the host proteins related to immune function, which represented 10-15% of human proteins quantified in the dataset, 40-50% of these quantified proteins were unique to specific immune pathways, which led to the confident interpretation of the host functional data related to the research questions addressed in this study. In the previous investigations of some samples in this cohort, it was apparent that the microbial metaproteome is highly variable over the first several weeks of life and between individuals. The functional β-diversity analysis in the present investigation with the inclusion of additional infants reaffirmed this finding. The functional β-diversity analysis showed that this intra- and inter-individual functional variability is not restricted to the microbiota. Host protein expression related to immune function was variable and dependent on the environmental context each infant faced during this early-life period. In spite of the inter-individual variability, the longitudinal measurements collected in this study enabled an in-depth

investigation of intra-individual variability of microbial and host immune function simultaneously. Intra-individual variability corresponded with microbial responses to both external perturbations of the gut environment, through the administration of antibiotics, and also to host immune functions.

The present study demonstrated that the timing of maternal antibiotic administration before delivery could significantly impact an infant's levels of expressed proteins related to immune processes, such as neutrophil degranulation and toll-like receptor cascades over the first few months of life. While previous research has established that maternal administration, and the timing of those antibiotics, can have profound effects on the colonization of the infant gut[206,230], very few studies have also investigated the impact of this administration on the development of specific immune processes. Further research is needed to fully understand the impact maternal antibiotic timing has on both the developing host immune system and colonizing microbiota.

Microbial mechanisms, such as antibiotic resistance, play a pivotal role in shaping the early-life preterm gut. The present study indicates that, in addition to antibiotic resistance mechanisms, metal homeostasis mechanisms correlate with community establishment dynamics. In fact, metal homeostasis may play a more prominent role in early-life community establishment than previously recognized. Metals are involved in many reactions and are essential to life. They are essential cofactors required for many reactions for both the host and microbes in the intestinal environment, and lack of these metals leads to organismal damage and death[231]. However, excess levels of metals also have negative impacts on cellular processes, such as the mismetallation of proteins, and high metal concentrations can be toxic due to their ability to disrupt normal metabolic functions and cause spontaneous redox cycling[216]. Organisms that have developed mechanisms that lead to resistance to host defenses involved in the trafficking of metals, based on either toxicity or starvation, have a better chance of survival and establishment in the gut. These early-life community dynamics play critical roles in developing the host immune system and

ultimately host health through the interactions between microbiota and the host immune proteins.

Among metal trafficking activities, the interplay between host immune responses and gut microbiota for the utilization of copper and iron appears to play a vital role in community assembly for some infants in this cohort. While copper is essential for biological processes, excess intracellular copper is toxic because of its potential to mismetallate proteins since it has a higher affinity for noncognate ligands due to its location in the Irving-Williams series, which allows it to disrupt the binding of other transition metals like iron, manganese, and zinc[232]. Previous studies of bacterial isolates have demonstrated that copper toxicity targets iron-sulfur-containing proteins via iron displacement from solvent-exposed iron-sulfur clusters[233–235]. Organisms have developed several strategies based on the trafficking of both copper and iron to deal with host-imposed copper toxicity. Based on the observed protein expression infant 61, **Figure 6-13A** is a simplified representation of interactions between the host immune response and *E. cloacae* in this infant. Following phagocytosis, inflammatory agents, such as lipopolysaccharide stimulate copper uptake in macrophages by inducing the expression of copper importers at the plasma membrane. Cytoplasmic Cu(I) is then delivered to copper pumps which undergoes trafficking to the phagolysosome compartment where copper ions are loaded. The accumulation of Cu(I) within the phagolysosome contributes to bacterial killing through membrane damage, displacement of iron by copper in iron-sulfur (Fe-S) clusters, and formation of reactive oxygen species[236,237]. In the bacterial cytoplasm, sensing of cytoplasmic Cu(I) by CueR induces the expression of copper transporter CopA, which effluxes Cu(I) into the periplasm. It is then transformed into the less toxic form of Cu(II) by the multicopper oxidase CueO or exported out of the cell. Both CopA and CopB are copper P-type ATPases; however, the exact physiological function of CopB has not been determined[238]. CopA has a low Cu(I) affinity and high turnover rate and is involved in exporting excess Cu(I) into the periplasm. CopB has a high Cu(I) affinity and low turnover rate and is suggested to transport Cu(I) into the periplasm for subsequent insertion into cuproenzymes)[239]. To complement strategies

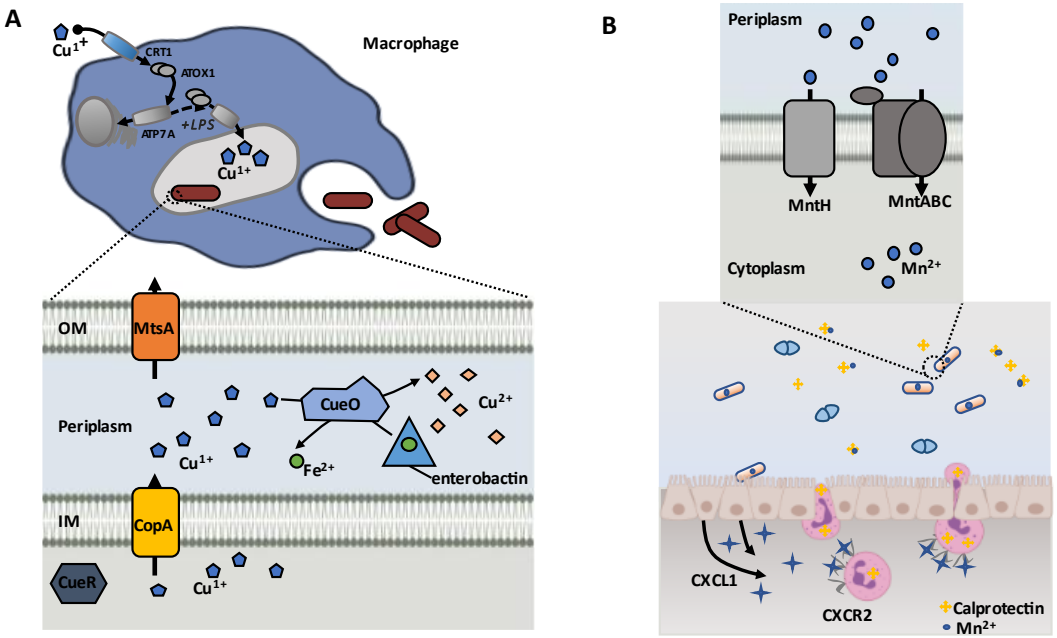

Figure 6-13 Metal homeostasis mechanisms of microbiota and corresponding host immune responses. (A) Graphical representation of copper/iron-related activities between the host immune system and *E. cloacae* in infant 61 based on the pathways and proteins detected in samples from this infant. Additional proteins related to these processes were also identified. (B) Graphical representation of manganese-related activities between the host immune system and *L. rhamnosus* in infant 39 based on the pathways and proteins detected in samples from this infant. Additional proteins related to these processes, as well as microbial manganese/zinc non-specific transporters were also identified.

to export excess Cu(I) out of the cell or transform it to a usable form, previous work has demonstrated that some microbes have developed tactics to oxidize siderophores as a mechanism to deal with high levels of intracellular copper[240]. In the presence of copper, the multicopper oxidase CueO oxidizes enterobactin, which circumvents enterobactin-mediated reduction of Cu(II) to Cu(I), and the resulting oxidation product, 2-carboxymuconate, sequesters copper. The oxidation process also releases chelated iron from the siderophores, maintaining the proper balance of copper and iron concentrations in the periplasm, which helps prevent mismetallation and damage to the cell. Studies have shown that *E. coli* lacking or defective in enterobactin biosynthesis and ferric-enterobactin uptake are more sensitive to copper toxicity[226,232].

While there are many similarities in the usage of copper and iron by bacteria, the host has developed very different mechanisms to utilize these transition metals to restrict bacterial growth through either toxicity or limitation. During the bacterial infection process, Cu(I) is accumulated in cytoplasmic vesicles of phagocytic cells that partially fuse with the phagolysosome, flooding invading microbes with toxic levels of copper ions[237]. The accumulation of Cu(I) in the phagolysosome may depend upon the trafficking of ATP7A to the membranes of these vesicles[241]. ATP7A is a Cu-ATPase of silencing ATP7A expression in mouse macrophages attenuated bacterial killing, which suggests a role for ATP7A-dependent copper transport in the bactericidal activity of macrophages[242]. In addition, another study has shown that expression of ATP7A was downregulated in mouse colon tissue following antibiotic treatment, further supporting the idea that the presence of microbiota is essential in initiating this host immune response[243]. PDZD11 and binding partner PLEKHA5 interact with ATP7A to influence its activity in cellular copper trafficking[244,245]. In total, six out of 17 infants expressed proteins related to copper trafficking for antimicrobial activities, including ATP7A, PDZD11, CTR1, and PLEKHA5. In this study, the increase in the expression of host proteins related to ATP7A corresponding to shifts in microbial dietary changes or antibiotic treatment suggest that initiation of this host immune response was correlated to shifts in the abundance of these organisms. The host utilizes several mechanisms for maintaining iron homeostasis to

limit pathogenic growth at mucosal interfaces through the restriction of metal-chelating proteins such as lactoferrin and lipocalin-2 (LCN2). As lactoferrin is a typical mass spectrometry contaminant (found in exocrine secretions such as sweat), this protein was removed from the analysis as the source of this protein could not be confirmed in the measurement. One proposed mechanism to enhance the host's antimicrobial activities of copper toxicity is the increased expression of the inflammation-associated protein LCN2 to control the competition for iron during infection by binding and sequestering various siderophores produced by enteric bacteria[246]. In this study, LCN2 was present in 90 samples, and the expression of LCN2 by almost all infants in the study indicates an active attempt by the host to limit iron harvesting and ultimately bacterial growth by the microbes producing siderophores. Overall, the observation that a microbe's ability to circumvent the host's attempt to limit some microbial growth through copper-mediated toxicity and limitation of iron harvesting contributed to its persistence and establishment in the environment.

Several studies investigating the activity of calprotectin during bacterial infections have found that it can be detected in the feces of patients with other types of intestinal inflammation, such as colon cancer and inflammatory bowel disorders[228,247,248]. Breastmilk contains high levels of calprotectin, and this heterodimer has been suggested to be involved in host immune regulation[201]. In this cohort, we observed changes in fecal calprotectin levels that were not associated with changes in consumption of formula and breastmilk, supporting the idea that changes in calprotectin levels were due to an inflammatory response and not due to diet. During the periods of altered host expression of calprotectin, there was a corresponding pattern of increased manganese import in selected organisms in several infants based on the expression of proteins related to the import of this metal. Figure 6-13B illustrates how the host can impose manganese and zinc starvation, and one mechanism microbes might use to overcome this, as observed in infant 39 during the period of increased host expression of calprotectin. During the infection process, enterocytes are stimulated to express the cytokines, such as IL-17 and IL-22, which

bind to receptors on neutrophils, thereby promoting transmigration of the neutrophils to the site of infection where they release calprotectin into the intestinal lumen where it binds to zinc and manganese, making these nutrients unavailable to the microbes[249].

Some microbes have overcome this strategy by utilizing additional membrane transporters such as the ones shown in the figure that enhance their ability to import manganese into the cell. As both manganese and zinc are cofactors for many enzymes involved in critical processes required for bacterial growth, including central metabolism, mechanisms for bacteria to compete with the host for these metals are essential for the survival of the organisms[216]. Both manganese and zinc are also cofactors for several proteins directly involved in microbial response to other innate immune responses and to antibiotic intervention for the treatment of infection. Specifically, enteric pathogens require manganese as a cofactor for superoxide dismutase to protect the cell from host-imposed oxidative stress[248]. Interestingly, zinc is a necessary cofactor for metallo-beta-lactamases, which are capable of inactivating beta-lactam antibiotics[250]. This may explain why some of the organisms in this study that expressed both the relevant antibiotic resistance-related proteins and manganese import proteins increased in relative abundance during periods of antibiotic administration in contrast to organisms with similar antibiotic resistance capabilities that lacked manganese importers, which decreased in abundance. In general, as both manganese and zinc are attractive targets for the host to limit as part of an immune response, microbial homeostasis mechanisms for these metals may play a more significant role in community assembly in the preterm infant gut than previously recognized.

The observations relating to antibiotic administration and metal homeostasis cannot be generalized to a larger cohort as other environmental factors (delivery mode, feeding type, hospital environment, etc.) can have non-negligent influences on the composition and functionality of gut microbiota and host. To fully explore the observations in the present study, more longitudinal studies with the recruitment of larger cohorts and a more balanced experimental design in terms of the timing of sample collection will enable more types of longitudinal analyses capable of handling

uneven sampling distributions. Tracking the longitudinal trajectories of proteins early in life can clarify colonization processes and community dynamics compared to single timepoint measurements. In general, longitudinal measurements can improve our understanding of disease progression to better predict host health outcomes. Future studies can leverage longitudinal measurements of real clinical samples where known perturbations to the system (i.e., antibiotic administration) are incorporated into the sampling time series to study the impact of these perturbations on community establishment dynamics.

6.3.4 Methods.

6.3.4.1 Sample selection, preparation, and measurement.

Sample collection, processing, and measurements by LC-MS/MS were already described[150,220]. Briefly, using necrotizing enterocolitis (NEC) as a representative dysbiotic condition, 91 metaproteomic measurements were taken from 17 preterm infant fecal samples over the first 90 days of life. Six infants developed NEC during the study. 0.3 g of raw fecal stool was processed using an indirect enrichment strategy[251], and 50ug of digested peptides were analyzed in technical duplicate via two-dimensional nanospray liquid chromatography-tandem mass spectrometry (LC-MS/MS) on an LTQ-Orbitrap Elite mass spectrometer (Thermo Scientific).

6.3.4.2 Database searching and construction of protein datasets.

Matched metagenome-derived protein databases for each sample were compiled into infant-specific databases containing any predicted proteins for that individual. In addition, the databases contained protein sequences from the human reference proteome (UniProtKB/TrEMBL, UP000005640), common mass spectrometry contaminants, and reverse sequences of all entries in the database to

control the false discovery rate (FDR). Previously collected MS/MS spectra were searched using Proteome Discover 2.3 (Thermo Scientific), employing MSAmanda 2.0 and Elutator[142,146] with the second search feature in MSAmanda2.0 turned off. Peptide spectrum matches (PSMs) were required to be fully tryptic with up to two miscleavages, a static carbamidomethylation modification of 57.0214 Da for cysteine residues, and a dynamic oxidation modification of 15.9949 Da for methionine residues. FDR was assessed by matches to reverse decoy sequences and controlled at 1% on the peptide level. To alleviate the ambiguity associated with shared peptides, proteins were clustered into protein groups by 100% identity for microbial proteins and 85% amino acid sequence identity for human proteins using UCLUST[173]. FDR-controlled peptides were then quantified according to the chromatographic area under the curve (AUC). Peptide AUC values obtained through match-between-runs (MBR) were included for each sample if there was one MS2 event per sample group (infant). Peptide intensities were summed to estimate protein-level abundance based on peptides that uniquely mapped to one protein group. Technical replicates were merged at the protein level after protein roll-up. One of the files for infant 65 (DOL 17) was not searchable, so only the other replicate was included in the analysis. Protein abundances from all 91 samples were log-transformed, normalized at the sample level by LOESS, and standardized across the entire dataset by median absolute deviation (MAD) and median centering. Missing values were imputed to simulate the mass spectrometer's limit of detection with a downshift of 2.4 and a width of 0.3.

6.3.4.3 Functional annotation of proteins.

Annotation of immune-related proteins was performed using the Reactome Database (version 77). Proteins that mapped to several immune pathways were counted for each pathway. Further analysis was performed at the pathway level with the requirement that at least one protein in the pathway is unique. For annotation of microbial proteins, taxonomy assignments were performed as previously described[220]. KO terms were identified using the eggNOG database emapper function (version

emapper-1.0.3-35-g63c274b)[111] . All quantified microbial protein groups were searched against the Comprehensive Antibiotic Resistance Database (CARD)[252] to identify antibiotic resistance-related proteins. Protein sequence hits to the database that mapped to more than one antibiotic resistance ortholog (ARO) were counted for each matching ARO. Both perfect and strict hits were included in the analysis.

6.3.4.4 Data analysis.

The samples' dissimilarities were assessed using Jaccard distances' NMDS ordinations using the metaMDS() function in the vegan R package. Before the Jaccard distance calculations, the microbial protein groups were summed to the KO term level. Permutational multivariate analysis of variance (PERMANOVA), applying 999 permutations, was used to assess statistical significance in beta diversity between categorical variables of interest—related to patient health metadata, including delivery mode, antibiotic usage, health/disease status, etc. The betadisper package in R was used as a complementary assessment to the PERMANOVA testing to test for homogeneity of multivariate dispersion to ensure the within-group dispersion was not significant using 999 permutations. A category was considered to be significant in the analysis if there was a significant PERMANOVA result (p-value <0.05) and an insignificant betadisper result (p-value>0.05) after Benjamini-Hochberg correction. The IndVal package in R was used to evaluate what specific proteins were driving the separation between samples based on metadata factors determined as significant in the functional β-diversity analysis. For the relative organismal abundance alluvial plots, the “other” category all proteins with no genus or species level taxonomy classification or if the total organismal abundance was less than 1% of total abundance for any timepoint in the infant.

Chapter 7 - Validation that human microbiome phages use alternative genetic coding.

Text and figures were adapted from the following manuscript currently under review:

Peters, S.L., Borges, A.B., Giannone, R.J, Morowitz, M.J., Banfield, J.F. and Hettich, R.L., (2022). Validation that human microbiome phages use alternative genetic coding. Nature Communications (under revision).

S.L.P contributions include experimental design, metaproteomic measurements, data analysis, figure generation, writing and editing of the original manuscript, revisions and response to reviewers.

7.1 Introduction.

7.1.1 The role of bacteriophages in the environment.

Bacteriophages, or phages for short, are viruses that infect bacteria and are considered distinct from cellular life due to their inability to carry out biological processes necessary for reproduction. Phages are the most abundant biological entity on earth with phage to bacteria ratios ranging from 10:1 in soil and aquatic environments to 1:1 in the mammalian gut[253,254]. Bacteriophages modulate the composition of microbial communities through the selective predation of bacteria, alteration of host metabolism, and redistribution of cellular lysis products in the environment during the infection process[255]. Despite their recognized importance as components of ecosystem dynamics, phages remain one of the most poorly understood members of microbiomes[256,257] due to the limitations of the methodologies used to study them. Most of our knowledge of environmental viruses comes from marine environments. These environments are fundamentally different ecosystems from the mammalian gut which has a number of factors that influence phage-host population

dynamics including the complex anatomy of the gut, the actions of the local immune system, the influx of new phages and hosts from the external environment, as well as the chemical composition and amounts of dietary inputs for host metabolism[258]. Methodological advancements are needed to shed light on the gut-associated phageome and its importance to mammalian gut homeostasis. Furthermore, fundamental questions remain regarding how phages interact with and redirect, the translation systems of their host bacteria.

7.1.2 Uncultivated gut bacteriophages with alternative genetic code.

There is evidence from metagenomic studies that some phages appear to use the bacterial ribosome to translate their proteins using both the standard and an alternative code. In these phages, proteins that require a stop codon to be read as an amino acid are thought to be only translated late in the infection cycle after code switch machinery has been deployed[258]. Some phages are predicted to reassign the normal stop codon TAG to be translated as glutamine (Q), and others reassign the TGA stop codon to tryptophan (W). This phenomenon appears to be common in human and animal microbiomes[255,258–262] and is particularly prevalent in phages that infect Firmicutes and Bacteroidetes[258].

If not recognized, stop codon reassignment can limit our ability to identify phages in metagenome sequences and restrict our understanding of phage gene inventories. Specifically, incorrect code usage leads to low predicted coding densities, truncated gene products, and genes predicted in incorrect reading frames. Alternative code choices can be tested to determine if they restore full-length open reading frames, and the amino acid to which a stop codon is reassigned can be predicted based on amino acid alignments with homologous proteins from related phages. Direct proteomic confirmation of gene predictions of stop codon reassignment has been primarily restricted to bacteria[263,264], but predictions for this type of alternative coding event in phages have never been experimentally validated. Nonetheless, validation of alternative genetic coding is crucial for the accurate understanding of the translation

of gene products, and the utilization of experimental evidence provided by proteomics for genome reannotation based on atypical codon usage has been reported in several studies[265–267]. For example, proteomics-based systematic characterization of the N-termini of proteins validated more than 60 non-canonical translation initiation codons in the *Deinococcus deserti*[268]. Bioinformatic studies have inferred the use of genetic code 15 in some bacteriophages[258,259,262] and the protist *Iotanema spirale*[269]. However, there are no reports validating the expression of this alternate code, and it has been actively rejected in the summary of genetic codes recognized by NCBI (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi).

Our previous metagenomic study[270] identified two unrelated human microbiome samples that each contained an abundant crAss-like phage. The adult sample contained a 191 kilobase crAss-like phage genome with the potential to circularize, and the infant sample had a 94 kilobase crAss-like phage genome, which was curated to completion (**Figure 7-1**). These samples were prioritized for metaproteomic measurements to address two key questions: 1) can proteins of phages be detected in the presence of highly abundant bacterial, human, and dietary proteins, and 2) can phage proteins be detected that confirm the expression of alternative genetic code 15?

7.2 Results and Discussion.

7.2.1 Selective enrichment of virus-like particles from human feces.

Phages contribute a relatively small proportion of proteinaceous biomass in fecal samples, making detecting their proteins by shotgun proteomics particularly challenging. In fact, initial measurements of the fecal samples detected no phage proteins using a standard workflow for the preparation of fecal samples for metaproteomics. To detect phage proteins in feces, optimization of the sample preparation methodology was essential to identify phage proteins among highly abundant bacterial, human host, and dietary proteins present in the sample. The

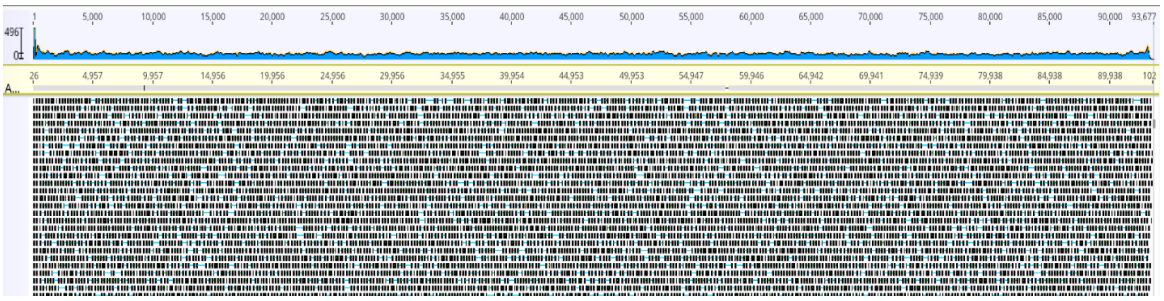

Figure 7-1 Genome curation and variation. Overview of the final curated genome showing complete and relatively even coverage by paired 150 bp Illumina reads (mean insert size: 391 bp) when reads are mapped, allowing ≤ 2% single nucleotide polymorphisms.

dynamic range of proteins present in feces has prompted metaproteomics methodological advancements to selectively enrich or deplete some proteins prior to LC-MS/MS measurements based on their source organism[251]. However, several enrichment strategies require a large amount of starting fecal material (300mg) which may not be possible during the sample collection or available if some of the material is reserved for paired experiments. Previous work has shown alternatively coding phages have genome sizes, and presumably corresponding physical sizes, that range from very small to very large[258]. Alternative coding is predicted to be prevalent during late infection for the expression of structural proteins[261]. Our goal was to optimize a methodology that enabled the enrichment of virus-like particles (VLPs) and their proteins to enable deeper measurements of relatively abundant phage structural proteins regardless of the phage's physical size. There are already several existing protocols for the recovery of phages from biospecimens[115,271,272]. However, these protocols were designed for metagenomic analyses of the enriched VLPs and also do not account for the variable physical size of alternative coding phages that might be present in the gut[257]. In addition, several techniques are primarily useful for filtering small sample volumes as they rely on impact filtration onto small pore-size filters which are often prone to clogging issues[273].

Initial work on phage enrichment techniques for metaproteomics was conducted on human feces using a non-endogenous dsDNA cyanophage (*Cylindrospermopsis raciborskii* Virus (CrV)) and a non-endogenous bacterium (*Clostridium autoethanogenum*) spike-in methodological controls. CrV was selected as a spike-in phage as it has a genome size of 104 kb, which is similar to the two target phages in our study identified by metagenomics that are predicted to use alternative genetic coding. As previous studies have shown genome size is proportional to the physical sizes of virion[274,275], CrV should be captured in the same sample fraction as the target phages during the phage enrichment process. The addition of the non-endogenous phage particles and bacterial cells to various fecal samples, collected from infants and adults, allowed us to track where each entity ended up during the enrichment process. The variety of fecal specimens allowed us to test the robustness

of the method, in terms of filter clogging, across a range of fecal specimens with varying degrees of microbiome complexity and dietary components. Several parameters were evaluated during methodological testing, including pre-homogenization of the fecal samples prior to enrichment, the addition of NaCl and other components to the enrichment buffer, centrifugation speeds, filter pore sizes, incubation times, and filtration temperatures. **Figure 7-2** is the generic workflow that was the most robust across all fecal samples.

Notably, the addition of NaCl to the enrichment buffer minimized the loss of the spiked bacteriophage during the homogenization pre-processing step. In addition, filtration at a warmer temperature (37°C) reduced filter clogging, especially when processing infant samples. After homogenization and low speed-centrifugation at 3,000 x g, the majority of the spiked C. auto cells remained in the pellet while the phage particles were still suspended in the supernatant. After the depletion of bacterial cells, the remaining human and phage proteins were further processed using a combination of filters with different pore sizes. Two filters were included in the enrichment process. The initial 0.8µm filter removes a large amount of the non-proteinaceous debris and reduced downstream filter clogging. Intermediate filter sizes (0.45µm and 0.22µm) were also evaluated, but these filters were prone to clogging and retained the virions in the fecal debris. The 300kDa pore size filter was used to capture the majority of VLP proteins from the flow-through of the 0.8µm filtration step. Using this combination of centrifugation and filtration-based enrichment, MS/MS was able to detect phage peptides in the presence of highly abundant host peptides when the phage was spiked into feces at a starting amount of less than 0.5% of proteinaceous biomass. **Figure 7-3** shows the extracted ion chromatograms of a CrV peptide ion. This ion was among the most abundant CrV peptides found in the samples. While it was found in both the 300kDa filter top and flow-through samples, it was 2.6x more abundant on the filter top compared to the flow-through. Overall, this enrichment strategy enables the successful detection of low abundance phage proteins in the presence of highly abundant proteins from the human host and bacteria.

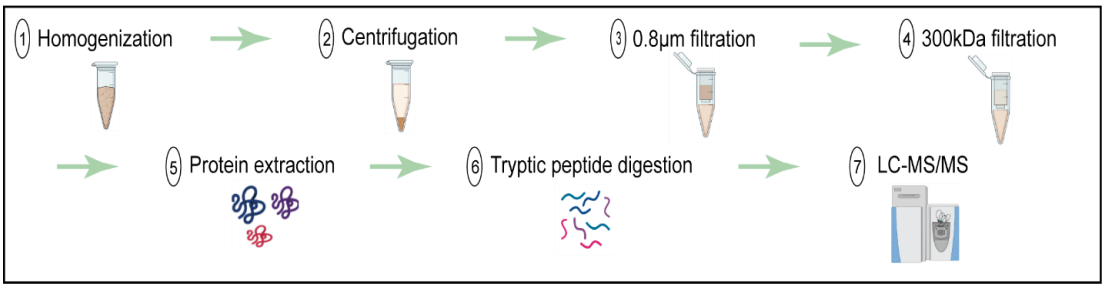

Figure 7-2 Workflow for phage enrichment from human feces.

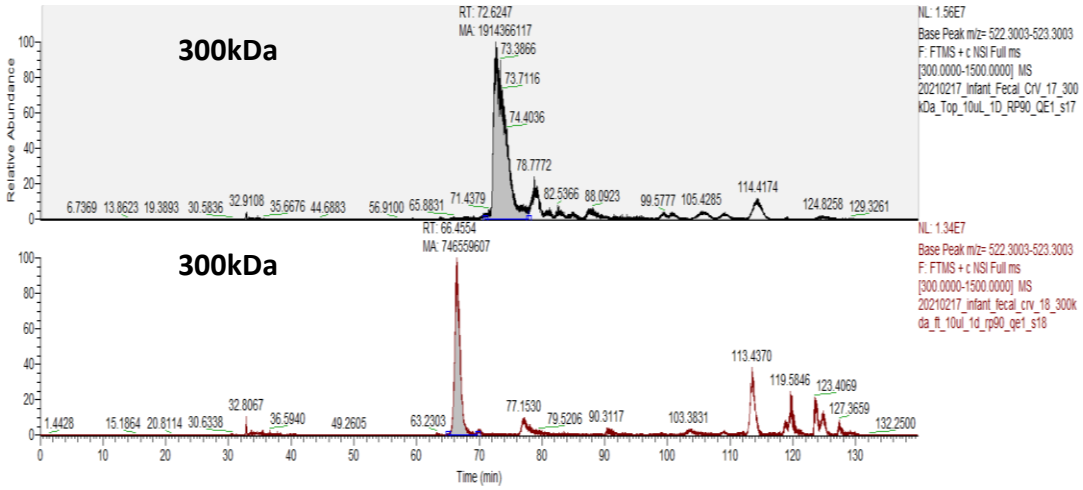

Figure 7-3 Extracted-ion chromatograms of m/z 522.8003 ion. This is the ion for peptide sequence TQVVILED, which is a peptide from CrV uncharacterized protein gp028. This ion was in 2.6x increased abundance in the 300kDa filter top fraction compared to the 300kDa flow through for infant fecal samples with a spike-in of CrV.

7.2.2 Confirmation of genetic code 15 usage by gut bacteriophage.

After a successful phage enrichment strategy was developed, paired metagenomic and metaproteomic measurements were conducted on fecal samples containing abundant crAss-like phages from one infant and one adult. Metagenomic data indicated these phages are predicted to use genetic code 15, based on the increased coding density observed with translations using genetic code 15 relative to genetic code 11. To ensure accurate peptide identifications from the metaproteomes, assembled metagenomic data from the same samples were used to generate databases that included phage proteins that were predicted using either the standard genetic code 11 (TAG→stop) or alternative genetic code 15 (TAG→Q), as well as all other bacterial proteins in the sample, the human reference proteome, and proteins commonly found as contaminants.

The LC-MS/MS data was searched against the comprehensive sample matched databases that included phage proteins predicted using either code 11 or code 15. Identified peptides were evaluated codon by codon to determine whether translation using standard or alternative genetic code was appropriate. To complement the database search strategy, *de novo* peptide sequencing, which derives peptide sequence information directly from the MS/MS spectra, was incorporated into the traditional database search workflow to provide a database-independent confirmation of phage translation that is agnostic to the translation code used for gene predictions.

Database searching of the phage-enriched fraction of the samples yielded 167 phage-specific peptides in total, with peptide-level false discovery rates at <1%. These peptides mapped to 13 and 14 phage proteins in the infant and adult samples, respectively. In addition, numerous peptides and proteins from bacteria and humans were identified. Many of the phage peptides identified by database searching were further supported by *de novo* sequencing tags. Roughly half of the identified phage peptides in each sample mapped only to proteins predicted using genetic code 15. **Figure 7-4** shows the genome maps of the target phages in each sample, with the locations of predicted and detected proteins using either code 11 or code 15 translation.

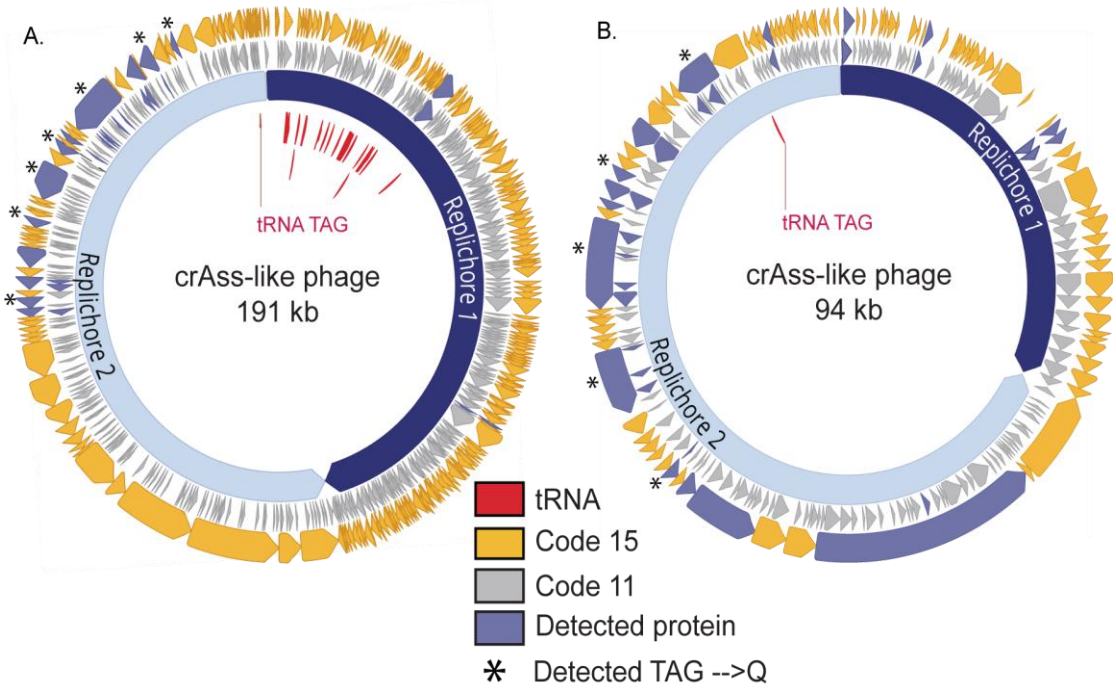

Figure 7-4 Proteomic detection of alternatively coded proteins from two phage genomes. L2_026_000M1_scaffold_35 (A) and L3_063_250G2_scaffold_974_ curated (B) are alternatively coded crAss-like phages. Prediction of genes in code 11 (inner grey ring) leads to gene fragmentation and low coding density, while gene prediction in code 15 (outer yellow genes) restores open reading frames. Genes with detected peptide evidence are colored purple. Some detected peptides contain glutamines encoded by reassigned TAG codons, and these genes with these validated recoding events are marked with stars. Suppressor tRNAs (red labels) are predicted to suppress translation termination at recoded TAG stop codons. Individual replichores were identified based on GC skew patterns indicative of bidirectional replication.

Some of the proteins identified with code 15 predictions were annotated as structural proteins, including capsid, portal, and tail-associated proteins, while the remaining proteins were unannotated. The detection of mostly late infection structural proteins was expected based on the enrichment for viral-like particles employed for sample preparation. Additional LC-MS/MS measurements were conducted on the unenriched fraction of the fecal samples that primarily contained unlysed bacterial cells and host proteins to determine if any additional phage peptides could be detected for early infection proteins that would be present in the host bacterium at the time of sample processing. Measurements detected phage peptides in the infant sample, including peptides for three additional proteins that were not identified in the original phage enriched samples. These three proteins included two hypothetical proteins and one ribosomal protein found in a region of the genome predicted to use genetic code 11. Across all identified phage proteins in these samples, 67% of genetic code 15 proteins could be confidently annotated, while only 34% of standard genetic code 11 proteins could be confidently annotated. As incorrect code prediction leads to genes predicted in incorrect reading frames and truncated gene products, this discrepancy in annotations levels is alarming. It does emphasize the need for correct code usage during gene predictions in order to accurately catalog phage gene inventories, as very few insights on biological function can be elucidated from incorrectly, and poorly, annotated genomes.

Figure 7-5 shows the protein sequence coverage map from the alternatively coded phage tail fiber protein (L3_063_250G2_scaffold_974_curated_39.code15) identified in the infant fecal sample. The region of the phage genome corresponding to this single protein contained six truncated proteins when predicted using the standard code. However, when using code 15, the full-length alternatively coded protein contained 23 peptides identified through database matching, of which, 11 were exclusively identified using code 15. Four peptides, highlighted in red boxes, directly confirm that the TAG stop codon is reassigned to glutamine. The identification of several *de novo* sequencing tags provides additional evidence of the existence and expression of recoded stop codons in this alternatively coded protein.

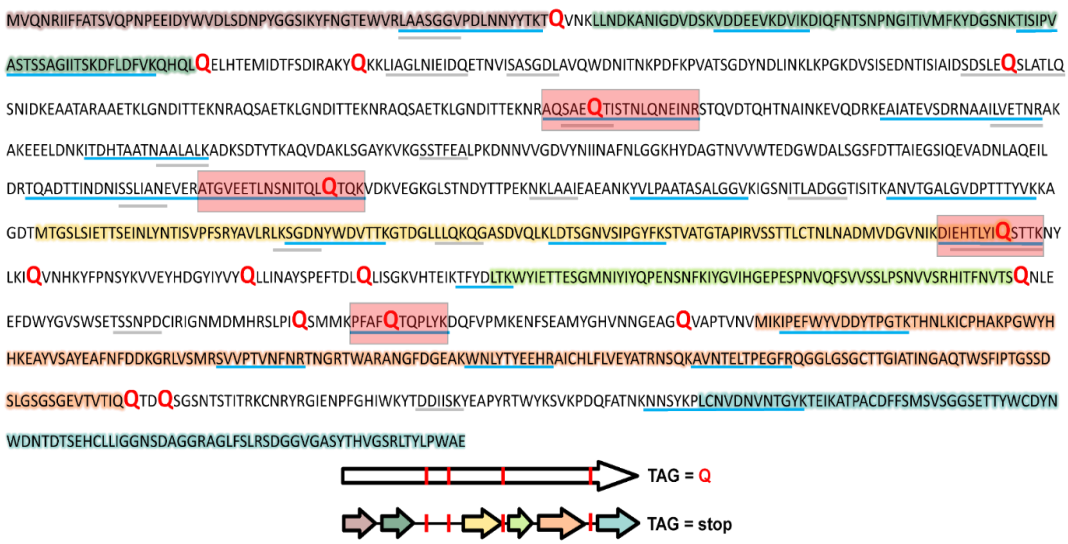

Figure 7-5 Protein sequence coverage map of alternative code phage tail-related protein. Highlighted sections of the code 15 predicted protein sequence (top) show the corresponding proteins that would have been predicted using standard code 11 (predicted open reading frames), also depicted in the graphical representation (bottom). Blue lines illustrate regions covered by tryptic peptides identified through LC-MS/MS database matching, whereas gray lines represent regions of the predicted protein sequence with matching de novo sequence tag coverage. Red text in the sequence indicates the location of glutamine residues from reassigned stop codons. Red boxes on the sequence coverage map and red bars on the graphical representation indicate the recoded glutamine residues with peptides identified through database searching.

Numerous identified peptides in both the infant and adult fecal samples further substantiate phage reassignment of the TAG stop codon to glutamine. **Figures 7-6** and **Figure 7-7** shows two examples of high-quality MS/MS spectra for alternatively coded phage peptides. In both instances, the glutamine residue from the recoded stop codon was positioned in the middle of a tryptic peptide. In the figure, only the direct y-type fragment ion series was chosen for annotation due to their preferential generation in higher-energy C-trap dissociation (HCD) fragmentation during MS/MS measurement[276].

Figure 7-6A shows a peptide containing a methionine from a predicted start codon using standard code residing in the middle of the peptide sequence in addition to a glutamine from a recoded stop codon. As several amino acids depicted here map to codons upstream of the standard code methionine start codon, this tryptic peptide would not exist if the phage was using standard code translation. The peptide in **Figure 7-6B** contains three glutamine residues; one canonical glutamine and two glutamines from recoded stop codons. One of the recoded glutamines was predicted as a stop codon at the end of a protein predicted through standard code translation. With a nearly complete fragmentation ion series, the detected tryptic peptide shows several amino acids flanking this recoded stop codon, covering an amino acid sequence that would not exist in a standard code open reading frame. In addition, a *de novo* sequencing tag matching nearly the entire length of the database match had high local confidence scores for every amino acid residue, including the recoded glutamines, providing additional support that this peptide, and others like it, do in fact exist (**Figure 7-7A**). Finally, as genetic code 15 only utilizes ATG as a start codon for translation initiation, confirmation of expression of this start codon was necessary to validate this genetic code. **Figure 7-7B** shows an example of direct peptide sequencing of a peptide containing a methionine from the ATG start codon for a genetic code 15 predicted protein, confirming translation initiation at this site in the genome. To confirm that alternative start codons were not being utilized, additional databases were generated to determine if translation was being initiated upstream of the predicted ATG start codon. Searches with databases that extended the protein-coding sequences

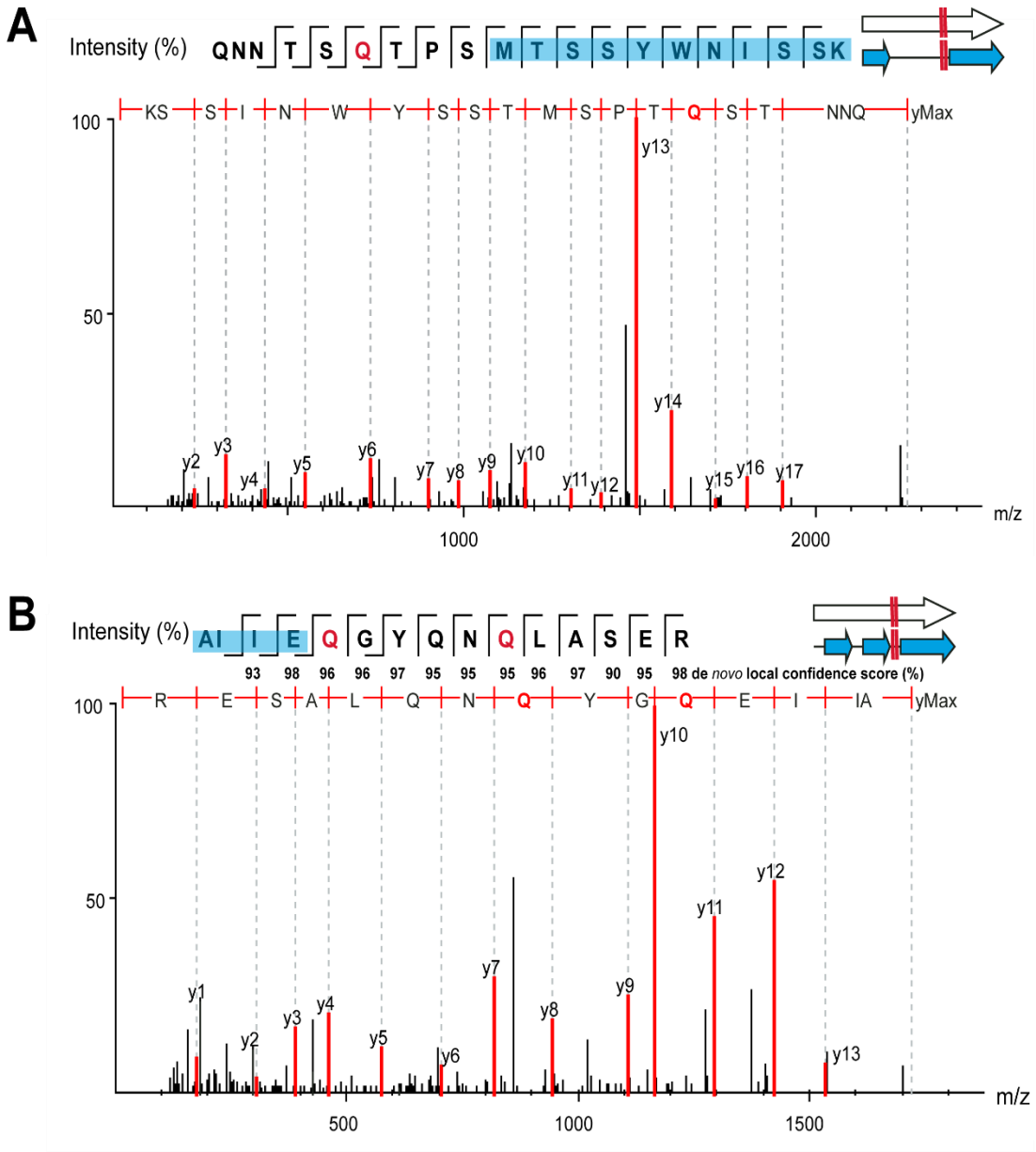

Figure 7-6 Example MS/MS spectra of alternative coding tryptic peptides. In all panels, red "Q" represents a stop codon that has been reassigned to glutamine using code 15 translation. Blue "Q" represents canonical glutamine residues. Note—the annotated y-ion series are read from the c-terminus to the n-terminus. (top left) Residues highlighted in green on the amino acid sequence fragmentation ladder depicting b-and y-ion fragmentation series represent portions of the peptide also predicted through code 11 (standard code). (A) Read-through of a standard genetic code 11 start codon (L3_063_250G2_scaffold_974_curated_32). (B) Read-through of a standard genetic code 11 stop codon (L2_026_000M1_scaffold_35_232).

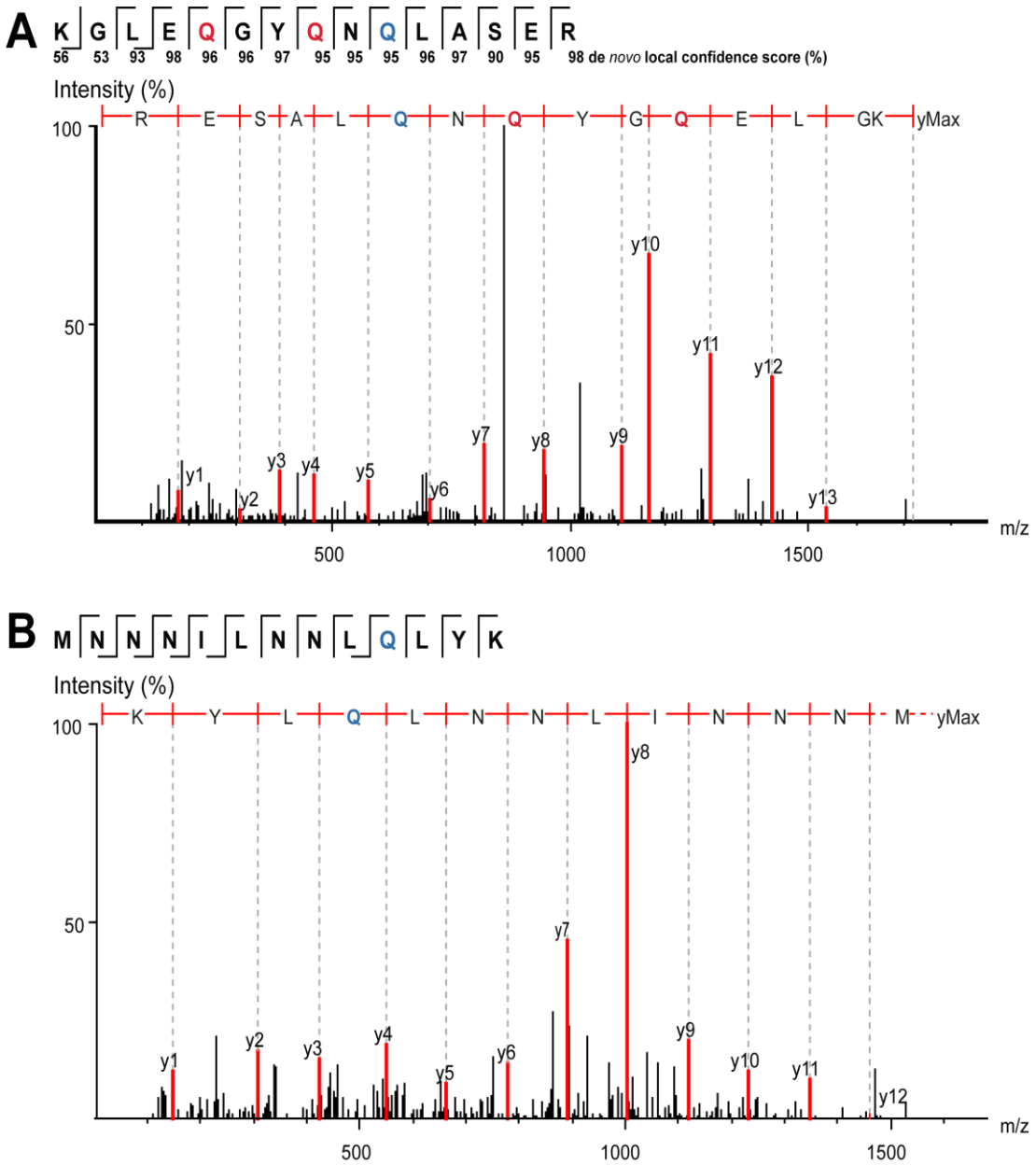

Figure 7-7 MS/MS spectrum of de novo sequence tag of an alternative coded tryptic peptide (A). The de novo sequence tag matches nearly the entire amino acid sequence of the database search identified peptide in Fig. 3A. The canonical glutamine and both glutamine residues from recoded stop codons, as well as flanking residues, have very high de novo local confidence scores. MS/MS spectrum of a tryptic peptide covering a start codon of an alternatively coded protein (B). The tryptic peptide contains the first thirteen amino acids of protein L3_063_250G2_scaffold_974_curated_25.code15. This demonstrates the methionine from the predicted ATG start codon of genetic code 15 is being translated.

several amino acids upstream of the predicted start codons yielded no peptide evidence that translation was occurring upstream of the predicted code 15 open reading frame.

These examples provide experimental validation that standard genetic code 11 is not being utilized by the phage in the translation of this region of the genome, and instead genetic code 15 is being used. In total, there is copious expressional evidence of genetic code 15, including direct evidence of stop codon readthrough and peptides existing outside of genetic code 11 predicted open reading frames, in regions of the genome with increased coding density using genetic code 15 predictions compared to code 11 predictions. There is no peptide evidence of TAG stop codon recoding in genome regions predicted to use standard genetic code 11 based on a similar coding density for each of the genetic codes. This peptide evidence supports the assignments of genetic codes based on relative coding densities for these regions of the genome.

7.3 Conclusions.

It has been suggested that alternatively coded structural proteins may be a strategy employed by phages to prevent premature expression of structural and lytic phage genes during the replication process[260]. In this study, the combination of metagenomics and metaproteomics confirmed that when it occurs within genes, the TAG stop codon is translated as glutamine. Direct metaproteomic confirmation of alternative codes has rarely been performed, but it is easy to imagine extending this approach to validate other types of alternative coding, such as the use of alternative start codons or incorporation of non-standard amino acids such as selenocysteine and pyrrolysine. A more complete understanding of how phages modulate the genetic code, likely to ensure appropriate translation of their proteins, may have applications in synthetic biology (e.g., where non-standard reading of codons can be used to create non-biological polymers[277] and in phage engineering.

7.4 Methods.

7.4.1 Sample selection.

In a previous study, human adult and infant stool samples were collected and sequenced with short-read shotgun sequencing[270]. The adult and infant samples prioritized for proteomics here were chosen because they both had alternatively coded crAss-like phages present at high abundance. Phage L2_026_000M1_scaffold_35 is the most abundant genome in the adult sample at 659x sequencing coverage, and the next highest coverage genome is a *Bacteroides vulgatus* genome at 307x coverage. Phage L3_063_250G2_scaffold_974 is the most abundant genome in the infant sample at 4752x sequencing coverage, and the next highest coverage genome is a *Bacteroides vulgatus* genome at 242x coverage.

7.4.2 Genome predictions and phage genome curation.

Coding sequences were predicted by Prodigal[278] using standard genetic code 11 and alternative genetic code 15. Code assignments were determined based on the relative coding density of contigs predicted using genetic code 11 or code 15. Coding density was calculated by summing the length of all genes in each contig and dividing that length by the total contig length. Contigs of total length 5-100 kb were assigned to use alternative genetic code 15 if there was an increase of greater than 10% coding density for genetic code 15 predictions relative to genetic code 11 predictions. Contigs >= 100 kb in length were required to have a coding density increase of greater than 5% with genetic code 15 predictions relative to genetic code 11 predictions to be assigned to be using genetic code 15. HMMER[279] (hmmsearch) was used to annotate protein sequences with the PFAM, pVOG, VOG, and TIGRFAM HMM libraries. In some cases, BLAST searches against the NCBI database, and remote homology searches using the HHPred[280] and Phyre2[281] web portals were used to augment initial annotations. tRNAs were predicted using tRNAscan-s.e. V.2.0 in general mode[282].

Replichores were identified by calculating GC skew (G-C/G+C) and cumulative GC skew using the iRep package (gc_skew.py)[283]. Genome curation was performed using previously described methods[284]. Curation and genome figure generation were completed using Geneious Prime® 2021.0.3 (https://www.geneious.com/). Additional databases were generated to determine if the translation was occurring upstream of the predicted genetic code 15 start codons, with the inclusion of several codons upstream of the ATG start codon. When possible, a maximum of thirty amino acids upstream of the predicted start codon were appended to the protein sequence. In most cases, there were in-frame stop codons (TAA, or TGA) upstream of the start codon, so the translations included any amino acids between the upstream stop codons and the predicted start codon, which in many cases was less than 30 amino acids.

7.4.3 Sample preparation for LC-MS/MS.

100 mg stool sample was resuspended in 1200uL 100mM Tris-HCl, pH 8.0 and homogenized with 0.9-2.0mm stainless steel beads (NextAdvance, part #SSB14B). Homogenized samples were incubated for 60 minutes before centrifugation at 3,000xg for 30 minutes. After centrifugation, the pre-cleared supernatant was filtered on a 300kDa MWCO PES filter (Pall, Omega Membrane 300K, part # OD300C34). The filtered eluate and the residual proteinaceous biomass remaining on top of the filter (resuspended in 100mM Tris-HCl, pH 8.0) were collected for downstream processing. Samples were adjusted to 4% (wt:wt) sodium dodecyl sulfate (SDS)/5mM dithiothreitol and incubated at 95°C for 10 min. Samples were alkylated with 15mM iodoacetamide (IAA) for twenty minutes at room temperature in the dark. The crude protein sample volume was processed by the protein aggregation capture (PAC) method[59]. Briefly, 300ug of magnetic beads (1 micron, SpeedBead Magnetic Carboxylate; GE Healthcare UK) was added to each sample. Samples were then adjusted to 70% acetonitrile to induce protein aggregation. Aggregated proteins were washed on a magnetic rack with 1mL of 100% acetonitrile, followed by 1mL of 70% ethanol. The washed proteins were then resuspended in 4%

SDS/ 100mM Tris-HCl, pH 8.0, and boiled off the magnetic beads at 95°C for 10 minutes. Protein amounts were quantified by corrected absorbance (Scopes) at 205 nm (NanoDrop OneC; Thermo Fisher). The cleaned proteins were re-aggregated back on the magnetic beads and washed with 1mL of 100% acetonitrile and 1mL of 70% ethanol to remove detergent from samples. The aggregated protein pellet was resuspended in 100mM Tris-HCl, pH 8.0, and digested with 1:75 (wt:wt) proteomics-grade trypsin (Pierce) overnight at 37°C and again for four hours the following day. Tryptic peptides were filtered on a 10kDa MWCO filter plate (AcroPrep Advance, Omega 10K MWCO) at 12,000xg and adjusted to 0.5% formic acid before quantification by NanoDrop OneC.

7.4.4 LC-MS-MS.

Digested peptides were analyzed by automated 1D LC-MS/MS analysis using a Vanquish ultra-HPLC (UHPLC) system plumbed directly in line with a Q Exactive Plus mass spectrometer (Thermo Scientific). A trapping column (100 µm inner diameter; packed with 5 µm Kinetex C18 reverse-phase resin (Phenomenex) packed to 10 cm) was coupled to an in-house-pulled nanospray emitter (75 µm inner diameter; 1.7 µm Kinetex C18 reverse-phase resin (Phenomenex) packed to 15 cm). For each sample, 10uL of peptides were loaded, desalted, and separated by uHPLC under the following conditions: sample injection followed by 100% solvent A (95% H2O, 5% acetonitrile, 0.1% formic acid) from 0-30 minutes to load and desalt, a linear gradient from 0% to 30% solvent B (70% acetonitrile, 30% water, 0.1% formic acid) from 30-220 minutes for separation, and 100% solvent A from 220-240 minutes for column re-equilibration. Eluting peptides were measured and sequenced with a Q Exactive Plus MS under the following settings: data-dependent acquisition; mass range 300 to 1,500 m/z; MS and MS/MS resolution 70K and 15K, respectively; MS/MS loop count 20; isolation window 1.8 m/z; charge exclusion unassigned, 1, 6 to 8.

7.4.5 Proteomics data analysis.

MS/MS spectra were interrogated by *de novo*–assisted database searching[99] in PEAKS Studio 10.6 (Bioinformatics Solutions) against a custom-built proteome database derived from the combination of the sequenced metagenome-derived predicted proteomes for all contig except the target phage contig, the phage proteome predicted in standard code (code 11), and the phage proteome predicted in the alternative code (code 15), the human reference proteome from UniProt (UP000005640), common LC-MS/MS protein contaminants, and reversed-decoy sequences of all proteins in the database. Secondary databases were searched against that included several amino acids upstream of the genetic code 15 start codon to confirm if translation was occurring upstream of the predicted start. In all database searches, he parent and fragment ion mass error tolerances were set to ± 10 ppm and ±0.02 Da, respectively. Peptide spectrum matches (PSMs) were required to be tryptic with semi-specific digestion needed and a maximum of three missed cleavages. Accepted modifications included a fixed modification of carbamidomethylation (+57.02) of cysteine residues and a variable modification of methionine oxidation (+15.99), with a maximum of three variable modifications. A false discovery rate of 1% was applied to accept the peptide and protein sequences, and a minimum of one unique peptide was required to identify a protein. *De novo* only parameters were left at default settings with average local confidence (ALC) scores of >50% and *de novo* sequence tags displayed if at least six amino acids were shared with the database sequence. The resulting database-identified peptides and corresponding *de novo* sequence tags were manually validated to generate the final list of phage peptide sequences present in the sample. A database hit passing the 1% FDR threshold was required for a peptide sequence to be considered as detected. *De novo* sequence tags were regarded as complementary evidence for the code 15 database-identified peptides to confirm the expression of this code only if the residue local confidence scores of the amino acids in the *de novo* sequence tag matching the database sequence were greater than 90% confidence for each residue. In the cases where a TAG stop codon

was reassigned to glutamine, the glutamine and several flanking amino acids in the *de novo* sequence tag had to pass the >90% residue local confidence score threshold.

Chapter 8 - Overview and perspectives on MS-based metaproteomics.

8.1 Conclusions.

Microorganisms across diverse environments are major contributors to fundamental processes such as biogeochemical cycling in ocean and soil environments and drivers of health and disease in the human gut. Microbiomes often exhibit extraordinary complexity due to the presence of diverse taxa in widely varying abundances at any given point in time. The human gut microbiome is one of the most studied microbial ecosystems due to its integral role in health and disease. Microbiome research using meta-omics technologies can provide a vast window into the composition, structure, and function of these communities. While genomic approaches characterize the diversity and potential metabolisms, transcriptomic and proteomic methods providing insights into expression and function of that potential. Of these technologies, metaproteomics is increasingly being applied to study a range of environments for insights into the metabolic activities of genomic potential of community members. Metaproteomics provides a distinct perspective from other omics technologies since it is a direct measurement of a final functional output which can be linked back to the genomic content of the source organism. Measurement of these proteins can be used to examine ecosystem functional diversity, determine the community response to and utilization of nutrient inputs and environmental stressors, and be used to estimate potential enzymatic activity. On the basis of the central dogma of biology for the formation of proteins, and the nature of metaproteomic data interpretation, this technology is inherently linked to other omics methods and often provides complementary information. Likewise, it also is complementary to the downstream omics approach of metabolomics as it provides insights into the metabolic outputs to better explain the observed phenotype of the biological system. Therefore, integration of metaproteomics data with data collected using other omics approaches

provides crucial biological insights into microbiomes. The advent and application of instrumentation components such as nano-spray multidimensional liquid chromatography and high-resolution mass spectrometry has improved the measurement of proteins, enabling measurement depths necessary for interrogation of complex microbiomes with wide dynamic ranges.

The primary focus of this dissertation was to advance metaproteomics research methodologies to interrogate interactions between microbiota across the kingdoms of life in the human gut environment. While the primary ecosystem for this dissertation was the human-associated gut microbiome, technological advancements achieved over the course of this dissertation can be applied to microbiomes across several environments. For example, Chapter 3 presented sample preparation and measurement advancements for diverse ecosystems ranging from euphotic zone ocean microbiomes to warming Artic permafrost soils. Chapter 4 described several informatic considerations that were optimized for metaproteomic measurements to provide high-quality, deep measurements for the microbiome research studies presented in later chapters. Chapters 5-7 highlighted several projects that interrogate microbial interactions across the domains of life using tractable gut models and unmanipulated human microbiomes. These include interactions at the host-microbe interface, interactions between closely related bacterial species with common functional potential, and closer examination of non-bacterial community members including bacteriophages and microeukaryotes. Importantly, each of the presented studies demonstrated an integrated multi-omic experimental design. Metaproteomics not only provided complementary information to the other omics techniques used in the analyses, but also was used validate findings from other approaches and to benchmark the capabilities of novel methods for studying microbial resource utilization in the gut environment.

All projects from Chapter 5 utilize tractable gnotobiotic models to study microbial interactions and functionality in the gut environment. All three of the studies call attention to the importance of diet as key regulators of bacterial metabolism. Previous research has shown that diet-microbiome interactions have the

potential to impact host health in either a negative or positive way depending on the composition of the gut microbiome and dietary context.

The first study, which highlights how chemical modifications of dietary components introduced during food processing impact the metabolism of the gut microbiome. While this study focused on the interactions between different *Collinsella* species and one particular chemical contaminant, with the prevalence of food processing, there are likely many interactions between other food modifications and commensal bacteria. As research into these relationships continues, the relevance to human physiology must be considered, as the effects of food modifications due to processing on the microbiome and human physiology is still not well known. Future studies are needed to investigate the role of processed foods in driving inflammatory or metabolic diseases through alterations of the gut microbiome.

The second and third studies focused on community degradation of plant polysaccharides, paving a path to the development of microbiota-directed foods. The second study used a used a multi-omic approach to identify bioactive nutrients in fiber and their utilization by the gut microbiota. The findings of this study shows that interspecies competition and cooperation control the outcome of diet-based microbiota manipulation. This target-driven approach of selecting specific bacterial targets for modulation by diet to better understand how the gut microbiome functions, provides the foundation for developing therapeutic diets. In fact, lead fiber preparations identified this study that selectively increased key *Bacteroides* species which are under-represented obesity-associated microbiomes, were later tested in two pilot studies focusing on controlled diets in humans[285]. In these pilot studies, obese or overweight participants consumed a similar diet to the ones consumed in the gnotobiotic mouse study presented in this dissertation, with snack proteotypic that contained up to four of the identified lead fibers. The outcomes of the pilot studies showed that fiber-driven changes in the gut microbiome were correlated to changes in participants' plasma proteome. While additional studies are needed to fully assess the translatability of gnotobiotic model studies to human research, it is easy to imagine a

future where disease treatment may include tailored diets aimed at altering bacterial composition and function.

The third study presented in this chapter primarily focused on the design and application of bead-based biosensors (MFABs) to determine *in vivo* how microbial communities process different food components such as dietary glycans. Future studies are needed to determine the safety and utility of using MFABs for human research. However, if these studies find that MFABs are safe for human consumption, they could be used to identify how different substrates are processed by the microbiome. For example, MFABs could be used in personalized medicine to diagnose how well a person's specific gut microbiota are able to utilize different food. They also may be used in research as a component of in vivo assays to develop supplements, such as prebiotics and synbiotics, that promote the growth of selected microbes for host health.

Chapters 6 and 7 presented research that used unmanipulated human microbiomes to explore interkingdom interactions. Both of the studies presented in Chapter 6 exclusively examined preterm infant gut microbiomes to assess community-wide functionality. The first study explored microbial establishment and functional partitioning in the context of microeukaryotic membership. Very few integrated metagenomics/metaproteomics studies have evaluated eukaryotic membership in microbial communities due to challenges associated with assembling eukaryotic genomes from metagenomes. One of the primary benefits of metaproteomics is that this approach is agnostic to the origin of the detected peptides—as long as there is corresponding protein information represented in the search database, peptides derived from any organism can be identified, regardless of whether the source organism is prokaryotic or eukaryotic. In this study, we used an integrated omics approach utilizing a workflow to assemble eukaryotic metagenome-assembled genomes (MAGs)[286] to look at functional inter-kingdom interactions between the fungus *Candida parapsilosis* and the bacterium *Enterococcus faecalis.* One of the primary challenges with studying microbiome interactions of *Candida* species that are medically relevant in a laboratory setting is that most model infection systems, such

as mice, are not natural hosts to *Candida* species. Therefore, any observations that come out of these studies may not translate to humans. In addition, pure culture experiments, which are an accessible way to study *Candida* species, lack the community context, which can greatly influence the behavior and functionality of an organism. Indeed, in this study, *C. parapsilosis* exhibited distinct differences in *in situ* expression patterns compared to similar *C. parapsilosis* strains grown in pure culture. Therefore, it is imperative that more microbiome studies are designed and conducted to reflect the actual diversity of the community, including eukaryotic membership, for a holistic view of community function.

The second study in Chapter 6 exploited longitudinal metaproteomics datasets of fecal samples to study the interplay between the host immune system and the early-life bacterial colonizers of the preterm infant gut. One of the unique features of metaproteomics is the simultaneous measurement of temporally connected host and microbial functional activities. In this study, measurement and host and microbial proteins enabled the incorporation of information of host immune responses to better explain colonization dynamics in the variable early-life gut environment. In addition, as the developing infant gut is an extremely dynamic environment, longitudinal sampling provided context to microbial establishment and persistence patterns. Overall, this genome-resolved metaproteomics study provided a novel approach to study the temporal expansion and resilience of microbial colonization and provides one of the first detailed studies to elucidate the microbiome's potential immunomodulatory roles relevant to necrotizing enterocolitis and other dysbiotic conditions in preterm infants. Moving forward, studies that incorporate both microbial and host information in longitudinal measurements should be critically important for the enhanced predictions of microbial community development patterns and ultimately, host phenotypes.

Finally, Chapter 7 provides in depth investigation of the proteome of an uncultivated member of the human gut. In this study, fecal samples from two individuals, one infant and one adult, that each contained a crAss-like phage predicted to use genetic codes different than their bacterial hosts in some genome regions. First,

this study a novel methodology for the enrichment of viral proteins from feces, regardless of the physical size of the bacteriophage. This methodology opens the door for more *in situ* metaproteomic studies of this underrepresented and understudied member of microbiomes. Most importantly, this study provides the first experimental validation of alternate coding for gene expression in any bacteriophage and the first confirmation of alternate code 15 (i.e., the translation through a TAG stop codon to generate glutamine) in any biological system. Previous reports of alternate coding in phage were based on genomic predictions alone, and in terms of genetic code 15, NCBI will not list it as a recognized genetic code without experimental validation. Since metagenome predictions do not constitute actual functional validation, this highlights the need to complement genomic findings with measurement techniques, such as metaproteomics, which can demonstrate the existence of a translated product. In terms the roles phages play in ecosystems, accurate gene predictions to construct complete gene inventories is critical to accurately understand phage biology as very few insights on biological function can be elucidated from incorrectly, and poorly, annotated genomes. Since alternative coding is predicted to be highly prevalent among gut phages, it is of vital importance that future investigations characterize the different types of genetic codes used by phages and to look into how and when phages switch from normal to alternate coding to determine what role this plays in their infection cycles. In addition to augmenting our understanding of the roles of phages as drivers of ecosystem change, expanding our knowledge of genetic code diversity has potential applications in synthetic biology for the design of new genetic codes and in phage therapy.

8.2 Remaining challenges and future outlook of metaproteomics.

As metaproteomics is a key technology for connecting genomic potential to metabolic information, and for understanding how microbiomes function in both a spatial and temporal context, this field will continue to grow and gain popularity in the microbial ecology community. In fact, from the initial mention of metaproteomics,

this omics approach has become increasingly popular in the microbiome research community, with increasing publication rates every year, from one initial publication in 2004 to 939 publications mentioning metaproteomics as of mid-2022 (PubMed; "metaproteomic∗" as search query; July 2022). In light of this growing interest, due to the remaining hurdles with this technology, there are three areas where research focus will be centered: (1) advancement of sample preparation, analytical measurements, and informatics approaches, (2) community education of this technology, and (3) new applications of metaproteomics for microbiome research.

The metaproteomic research field still faces many challenges related to the intricate complexity and large dynamic range of microbiota found in microbial communities. Among these challenges includes improved, unbiased protein extraction from complex environmental matrices with biotic and abiotic interferences. As LC-MS/MS-based metaproteomics is implemented for the study of more environmental systems, novel methodologies for extraction of proteins from a range of challenging marine and soil environments need to be developed since current methods are insufficient to for protein extraction from these environmental matrices that contain a highly diverse microbial community.

In terms of outstanding informatics limitations for metaproteomics research, there is still an underperformance of computational platforms when it comes to protein identification from datasets with extremely large search spaces. Not all database search algorithms are capable of handling large protein databases, the high number of spectra collected from large metaproteomics campaigns, or spectra collected from fractionated samples. In addition, many emerging features for enhanced protein identification in proteomics, such as retention time prediction algorithms or quantification methods for peptides identified from chimeric spectra, are still not equipped to handle extremely complex spectral data. In addition, as successful metaproteomic analysis is reliant on confident sequence information, there is still an outstanding need for high-quality sample-specific protein databases generated from upstream omics methods that capture all members of the microbial community including microeukaryotes and viruses. This will be facilitated by advancements in

metagenome sequencing, assembly, and annotations. Innovations in data interpretation, including data analysis components such as protein inference, quantification, and functional annotation, are still needed to accurately unravel microbial networks and understand interaction of microbiota with each other and their environment. Due to these informatic limitations, a significant portion of the collected data is not analyzed, and future efforts need to be focused on recovering currently unutilized spectra.

The microbiome research community will need better education on the technical details and capabilities that can be afforded by metaproteomics to produce sound scientific research. Along with education, over the coming years we will see broad benchmarking efforts within the metaproteomics community to standardize methodologies and accelerate experimental and informatic advancements that are tailored to complex metaproteomics samples. In fact, a portion of the work presented in Chapters 3 and 4 of this dissertation was conducted a part of two ongoing global collaborative efforts among several early adopters of this technology to shape this burgeoning field to steer it in a direction where it can be widely used in the broader microbial ecology community.

Over the next decade, major analytical and bioinformatic developments are likely to facilitate unprecedented insights into microbiome function and dynamics. Focused LC-MS/MS-based methods for studying protein expression that are fairly established for the study of single proteomes, but only have limited demonstrations in metaproteomics, such as targeted and DIA methods, will become more mainstream as instrumentation and informatic platforms improve. In addition, as measurement depth increases for metaproteomics, research focuses may turn to comprehensive characterization of all proteoforms expressed in a community through interrogation of post-translational modifications, alternative gene products produced through polypeptide cleavages, proteins of unknown function, and amino acid sequence variation. Finally, as the cost and implementation of upstream omics approaches improve over the coming years, this will spur new research questions that can be addressed when looking at uncultivated organisms in their native environment. For

example, with the combination of long-read sequencing and proximity ligation sequencing methods, high-quality phage genomes can be assembled from metagenomes to improve protein database quality for metaproteomic measurements and directly link phage-host pairs in environmental systems. This will enable higher-confidence interrogation of *in situ* host-phage interactions for organisms that cannot be cultivated in the laboratory. Other approaches, such as stable isotope probing (SIP) of proteins will enable the direct linkage of specific organisms in a microbial community to a specific function, by following the incorporation and transfer of isotopically labeled substrates. Overall, as metaproteomics reaches maturity, the full potential of the use of this technology will be recognized and applied for probing the functionality of microbial communities.

References

1. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A* **113**, 5970–5975 (2016).
2. Nguyen, J., Lara-Gutiérrez, J. & Stocker, R. Environmental fluctuations and their effects on microbial communities, populations and individuals. *FEMS Microbiology Reviews* **45**, 1–16 (2021).
3. Prosser, J. I. *et al.* The role of ecological theory. *Nature Reviews Microbiology* **5**, 384–392 (2007).
4. Allison, S. D. & Martiny, J. B. H. Resistance, resilience, and redundancy in microbial communities. *In the Light of Evolution* **2**, 149–166 (2009).
5. Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* **3**, (2018).
6. Liu, S. *et al.* Opportunities and challenges of using metagenomic data to bring uncultured microbes into cultivation. *Microbiome* **10**, 1–14 (2022).
7. Louca, S. *et al.* Function and functional redundancy in microbial systems. *Nature Ecology and Evolution* **2**, 936–943 (2018).
8. Marchesi, J. R. & Ravel, J. The vocabulary of microbiome research: a proposal. *Microbiome* **3**, 1–3 (2015).
9. Berg, G. *et al.* Correction to: Microbiome definition re-visited: old concepts and new challenges. *Microbiome* **8**, 1–22 (2020).
10. Peterson, J. *et al.* The NIH Human Microbiome Project. *Genome Research* **19**, 2317–2323 (2009).
11. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
12. Vogel, T. M. *et al.* TerraGenome : a consortium for the sequencing of a soil metagenome. *Nature Reviews Microbiology* **7**, 2009 (2009).

13. Knight, R. *et al.* Best practices for analysing microbiomes. *Nature Reviews Microbiology* **16**, 410–422 (2018).
14. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* **12**, 1–12 (2014).
15. Bilen, M. *et al.* The contribution of culturomics to the repertoire of isolated human bacterial and archaeal species. *Microbiome* **6**, 1–11 (2018).
16. Lagier, J. C. *et al.* Microbial culturomics: Paradigm shift in the human gut microbiome study. *Clinical Microbiology and Infection* **18**, 1185–1193 (2012).
17. Milo, R. What is the total number of protein molecules per cell volume? A call to rethink some published values. *BioEssays* **35**, 1050–1055 (2013).
18. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics* **13**, 227–232 (2012).
19. Gillet, L. C., Leitner, A. & Aebersold, R. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annual Review of Analytical Chemistry* **9**, 449–472 (2016).
20. Zuñiga, C., Zaramela, L. & Zengler, K. Elucidation of complexity and prediction of interactions in microbial communities. *Microbial Biotechnology* **10**, 1500–1522 (2017).
21. Wilmes, P. & Bond, P. L. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environmental Microbiology* **6**, 911–920 (2004).
22. Ram, R. J. *et al.* Microbiology: Community proteomics of a natural microbial biofilm. *Science (1979)* **308**, 1915–1920 (2005).
23. Herbst, F. A. *et al.* Enhancing metaproteomics-The value of models and defined environmental microbial systems. *Proteomics* **16**, 783–798 (2016).
24. Zhang, X. *et al.* Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nature Communications* **9**, 1–14 (2018).

25. Moya, A. & Ferrer, M. Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance. *Trends in Microbiology* vol. 24 Preprint at https://doi.org/10.1016/j.tim.2016.02.002 (2016).

26. Verberkmoes, N. C. *et al.* Shotgun metaproteomics of the human distal gut microbiota. *ISME Journal* **3**, 179–189 (2009).

27. Saito, M. A. *et al.* Needles in the blue sea: Sub-species specificity in targeted protein biomarker analyses within the vast oceanic microbial metaproteome. *Proteomics* **15**, 3521–3531 (2015).

28. Starke, R., Jehmlich, N. & Bastida, F. Using proteins to study how microbes contribute to soil ecosystem services: The current state and future perspectives of soil metaproteomics. *Journal of Proteomics* **198**, 50–58 (2019).

29. Benndorf, D. *et al.* Improving protein extraction and separation methods for investigating the metaproteome of anaerobic benzene communities within sediments. *Biodegradation* **20**, 737–750 (2009).

30. Herold, M. *et al.* Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nature Communications* (2020) doi:10.1038/s41467-020-19006-2.

31. Wenzel, L. *et al.* SDS-PAGE fractionation to increase metaproteomic insight into the taxonomic and functional composition of microbial communities for biogas plant samples. *Engineering in Life Sciences* **18**, 498–509 (2018).

32. Shrestha, H. K. *et al.* Metaproteomics reveals insights into microbial structure, interactions, and dynamic regulation in defined communities as they respond to environmental disturbance. *BMC Microbiology* **21**, 1–17 (2021).

33. Hugo Roume, Anna Heintz-Buschart, Emilie E L Muller, Patrick May, Venkata P Satagopam, Cédric C Laczny, Shaman Narayanasamy, Laura A Lebrun, Michael R Hoopmann, James M Schupp, John D Gillece, Nathan D Hicks, David M Engelthaler, Thomas Sauter, Paul S Kei, R. L. M. & P. W. Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *npj Biofilms and Microbiomes* **1**, 1–11 (2015).

34. Muller, E. E. L. *et al.* Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nature Communications* **5**, (2014).

35. Martínez Arbas, S. *et al.* Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics. *Nature Microbiology* **6**, 123–135 (2021).

36. Laskay, Ü. A., Lobas, A. A., Srzentić, K., Gorshkov, M. v. & Tsybin, Y. O. Proteome digestion specificity analysis for rational design of extended bottom-up and middle-down proteomics experiments. *Journal of Proteome Research* **12**, 5558–5569 (2013).

37. Schaffer, L. v. *et al.* Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics* **19**, 1–15 (2019).

38. Lermyte, F., Tsybin, Y. O., O'Connor, P. B. & Loo, J. A. Top or Middle? Up or Down? Toward a Standard Lexicon for Protein Top-Down and Allied Mass Spectrometry Approaches. *J Am Soc Mass Spectrom* **30**, 1149–1157 (2019).

39. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).

40. Silvestre, D. di, Brambilla, F., Agnetti, G. & Mauri, P. *Bottom-Up Proteomics*. *Manual of Cardiovascular Proteomics* (2016). doi:10.1007/978-3-319-31828-8.

41. Aakko, J. *et al.* Data-Independent Acquisition Mass Spectrometry in Metaproteomics of Gut Microbiota - Implementation and Computational Analysis. *Journal of Proteome Research* **19**, 432–436 (2020).

42. Tsou, C. C. *et al.* DIA-Umpire: Comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods* **12**, 258–264 (2015).

43. Kaur, G. *et al.* Extending the Depth of Human Plasma Proteome Coverage Using Simple Fractionation Techniques. *Journal of Proteome Research* **20**, 1261–1279 (2021).

44. Villalobos Solis, M. I., Chirania, P. & Hettich, R. L. In silico evaluation of a targeted metaproteomics strategy for broad screening of cellulolytic enzyme

capacities in anaerobic microbiome bioreactors. *Biotechnology for Biofuels and Bioproducts* **15**, 1–15 (2022).

45. Wang, D. Z., Kong, L. F., Li, Y. Y. & Xie, Z. X. Environmental microbial community proteomics: Status, challenges and perspectives. *International Journal of Molecular Sciences* **17**, 1–20 (2016).
46. Field, L. M., Fagerberg, W. R., Gatto, K. K. & Anne Böttger, S. A comparison of protein extraction methods optimizing high protein yields from marine algae and cyanobacteria. *Journal of Applied Phycology* **29**, 1271–1278 (2017).
47. Zhou, J. Y. *et al.* Simple sodium dodecyl sulfate-assisted sample preparation method for LC-MS-based proteomics applications. *Analytical Chemistry* **84**, 2862–2867 (2012).
48. Tanca, A., Biosa, G., Pagnozzi, D., Addis, M. F. & Uzzau, S. Comparison of detergent-based sample preparation workflows for LTQ-Orbitrap analysis of the Escherichia coli proteome. *Proteomics* **13**, 2597–2607 (2013).
49. Blakeley-Ruiz, J. A. *et al.* Metaproteomics reveals persistent and phylum-redundant metabolic functional stability in adult human gut microbiomes of Crohn's remission patients despite temporal variations in microbial taxa, genomes, and proteomes. *Microbiome* **7**, 1–15 (2019).
50. Nickels, J. D. *et al.* Impact of Fatty-Acid Labeling of Bacillus subtilis Membranes on the Cellular Lipidome and Proteome. *Frontiers in Microbiology* **11**, 1–13 (2020).
51. Salvachúa, D. *et al.* Outer membrane vesicles catabolize lignin-derived aromatic compounds in Pseudomonas putida KT2440. *Proc Natl Acad Sci U S A* **117**, 9302–9310 (2020).
52. Tanca, A. *et al.* A straightforward and efficient analytical pipeline for metaproteome characterization. *Microbiome* **2**, 1–16 (2014).
53. Zhang, X. *et al.* Assessing the impact of protein extraction methods for human gut metaproteomics. *Journal of Proteomics* **180**, 120–127 (2018).

54. Cañas, B., Piñeiro, C., Calvo, E., López-Ferrer, D. & Gallardo, J. M. Trends in sample preparation for classical and second generation proteomics. *Journal of Chromatography A* **1153**, 235–258 (2007).

55. Annesley, T. M. Ion suppression in mass spectrometry. *Clinical Chemistry* **49**, 1041–1044 (2003).

56. 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* **16**, 2537–2546 (2017).

57. FOLCH, J., LEES, M. & SLOANE STANLEY, G. H. A simple method for the isolation and purification of total lipides from animal tissues. *J Biol Chem* **226**, 497–509 (1957).

58. Dagley, L. F., Infusini, G., Larsen, R. H., Sandow, J. J. & Webb, A. I. Universal Solid-Phase Protein Preparation (USP3) for Bottom-up and Top-down Proteomics. *Journal of Proteome Research* **18**, 2915–2924 (2019).

59. Batth, T. S. *et al.* Protein aggregation capture on microparticles enables multipurpose proteomics sample preparation. *Molecular and Cellular Proteomics* **18**, 1027–1035 (2019).

60. Miller, R. M., Ibrahim, K. & Smith, L. M. ProteaseGuru: A Tool for Protease Selection in Bottom-Up Proteomics. *Journal of Proteome Research* **20**, 1936–1942 (2021).

61. Michalski, A., Neuhauser, N., Cox, J. & Mann, M. A systematic investigation into the nature of tryptic HCD spectra. *Journal of Proteome Research* **11**, 5479–5491 (2012).

62. Shao, C., Zhang, Y. & Sun, W. Statistical characterization of HCD fragmentation patterns of tryptic peptides on an LTQ Orbitrap Velos mass spectrometer. *Journal of Proteomics* **109**, 26–37 (2014).

63. Villalobos Solis, M. I., Giannone, R. J., Hettich, R. L. & Abraham, P. E. 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 (2019).

64. Furey, A., Moriarty, M., Bane, V., Kinsella, B. & Lehane, M. Ion suppression; A critical review on causes, evaluation, prevention and applications. *Talanta* **115**, 104–122 (2013).
65. Xie, F., Smith, R. D. & Shen, Y. Advanced proteomic liquid chromatography. *Journal of Chromatography A* **1261**, 78–90 (2012).
66. Zhang, X. *et al.* Multi-dimensional liquid chromatography in proteomics-A review. *Analytica Chimica Acta* **664**, 101–113 (2010).
67. Sanders, K. L. & Edwards, J. L. Analytical Methods and recent applications in omics investigations. *Analytical Methods* **12**, 4404–4417 (2020).
68. Motoyama, A. & Yates, J. R. Multidimensional LC separations in shotgun proteomics. *Analytical Chemistry* **80**, 7187–7193 (2008).
69. Yang, F., Shen, Y., Camp, D. G. & Smith, R. D. High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis. *Expert Review of Proteomics* **9**, 129–134 (2012).
70. Kislinger, T., Gramolini, A. O., MacLennan, D. H. & Emili, A. Multidimensional protein identification technology (MudPIT): Technical overview of a profiling method optimized for the comprehensive proteomic investigation of normal and diseased heart tissue. *J Am Soc Mass Spectrom* **16**, 1207–1220 (2005).
71. Maia, T. M. *et al.* Simple Peptide Quantification Approach for MS-Based Proteomics Quality Control. *ACS Omega* **5**, 6754–6762 (2020).
72. Hinzke, T., Kouris, A., Hughes, R. A., Strous, M. & Kleiner, M. More is not always better: Evaluation of 1D and 2D-LC-MS/MS methods for metaproteomics. *Frontiers in Microbiology* **10**, 1–13 (2019).
73. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science (1979)* **246**, 64–71 (1989).

74. Karas, M. & Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10 000 Daltons. *Analytical Chemistry* **60**, 2299–2301 (1988).

75. Awad, H., Khamis, M. M. & El-Aneed, A. Mass spectrometry, review of the basics: Ionization. *Applied Spectroscopy Reviews* **50**, 158–175 (2015).

76. Kebarle, P. & Verkcerk, U. H. Electrospray: From Ions in solution to Ions in the gas phase, what we know now. *Mass Spectrometry Reviews* **28**, 898–917 (2009).

77. Banerjee, S. & Mazumdar, S. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *International Journal of Analytical Chemistry* **2012**, 1–40 (2012).

78. Kebarle, P. & Tang, L. From Ions in Solution To Ions in the Gas Phase. *Analytical Chemistry* **65**, 972A-986A (1993).

79. King, R., Bonfiglio, R., Fernandez-Metzler, C., Miller-Stein, C. & Olah, T. Mechanistic investigation of ionization suppression in electrospray ionization. *J Am Soc Mass Spectrom* **11**, 942–950 (2000).

80. Gosetti, F., Mazzucco, E., Zampieri, D. & Gennaro, M. C. Signal suppression/enhancement in high-performance liquid chromatography tandem mass spectrometry. *Journal of Chromatography A* **1217**, 3929–3937 (2010).

81. Wilm, M. & Mann, M. Analytical properties of the nanoelectrospray ion source. *Analytical Chemistry* **68**, 1–8 (1996).

82. Niessen, W. M. A. & Falck, D. Introduction to Mass Spectrometry, a Tutorial. *Analyzing Biomolecular Interactions by Mass Spectrometry* 1–54 (2013) doi:10.1002/9783527673391.ch1.

83. Han, X., Aslanian, A. & Yates, J. R. Mass spectrometry for proteomics. *Current Opinion in Chemical Biology* **12**, 483–490 (2008).

84. Zubarev, R. A. & Makarov, A. Orbitrap mass spectrometry. *Analytical Chemistry* **85**, 5288–5296 (2013).

85. Michalski, A. *et al.* Mass spectrometry-based proteomics using Q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer. *Molecular and Cellular Proteomics* **10**, M111.011015 (2011).

86. Wöhlbrand, L., Trautwein, K. & Rabus, R. Proteomic tools for environmental microbiology - A roadmap from sample preparation to protein identification and quantification. *Proteomics* **13**, 2700–2730 (2013).

87. Couté, Y., Bruley, C. & Burger, T. Beyond Target-Decoy Competition: Stable Validation of Peptide and Protein Identifications in Mass Spectrometry-Based Discovery Proteomics. *Analytical Chemistry* **92**, 14898–14906 (2020).

88. Mellacheruvu, D. *et al.* The CRAPome: A contaminant repository for affinity purification-mass spectrometry data. *Nature Methods* **10**, 730–736 (2013).

89. Blakeley-Ruiz, J. A. & Kleiner, M. Considerations for constructing a protein sequence database for metaproteomics. *Computational and Structural Biotechnology Journal* **20**, 937–952 (2022).

90. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology* **32**, 834–841 (2014).

91. Xiao, L. *et al.* A catalog of the mouse gut metagenome. *Nature Biotechnology* **33**, 1103–1108 (2015).

92. Bäckhed, F. *et al.* Defining a healthy human gut microbiome: Current concepts, future directions, and clinical applications. *Cell Host and Microbe* **12**, 611–622 (2012).

93. Tanca, A. *et al.* The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome* **4**, 1–13 (2016).

94. Capelo, J. L. *Emerging sample treatments in proteomics*. (2019).

95. Muth, T. *et al.* The MetaProteomeAnalyzer: A powerful open-source software suite for metaproteomics data analysis and interpretation. *Journal of Proteome Research* **14**, 1557–1565 (2015).

96. Zhang, X. *et al.* MetaPro-IQ: A universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **4**, 1–12 (2016).

97. Cargile, B. J., Bundy, J. L. & Stephenson, J. L. Potential for false positive identifications from large databases through tandem mass spectrometry. *Journal of Proteome Research* **3**, 1082–1085 (2004).

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

99. Tran, N. H. *et al.* Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nature Methods* **16**, 63–66 (2019).

100. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc Natl Acad Sci U S A* **114**, 8247–8252 (2017).

101. Han, X., He, L., Xin, L., Shan, B. & Ma, B. PeaksPTM: Mass spectrometry-based identification of peptides with unspecified modifications. *Journal of Proteome Research* **10**, 2930–2936 (2011).

102. Han, Y., Ma, B. & Zhang, K. SPIDER: Software for protein identification from sequence tags with de novo sequencing error. *Journal of Bioinformatics and Computational Biology* **3**, 697–716 (2005).

103. Kleikamp, H. B. C. *et al.* Database-independent de novo metaproteomics of complex microbial communities. *Cell Systems* **12**, 375-383.e5 (2021).

104. Zhang, J. *et al.* PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular and Cellular Proteomics* **11**, M111.010587 (2012).

105. Domon, B. & Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nature Biotechnology* **28**, 710–721 (2010).

106. Millán-Oropeza, A., Blein-Nicolas, M., Monnet, V., Zivy, M. & Henry, C. Comparison of Different Label-Free Techniques for the Semi-Absolute Quantification of Protein Abundance. *Proteomes* **10**, (2022).

107. Blein-Nicolas, M. & Zivy, M. Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics. *Biochimica et Biophysica Acta - Proteins and Proteomics* **1864**, 883–895 (2016).

108. Branson, O. E. & Freitas, M. A. A multi-model statistical approach for proteomic spectral count quantitation. *Journal of Proteomics* **144**, 23–32 (2016).

109. Välikangas, T., Suomi, T. & Elo, L. L. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Briefings in Bioinformatics* **19**, 1–11 (2018).

110. Webb-Robertson, B. J. M. *et al.* Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of Proteome Research* **14**, 1993–2001 (2015).

111. Huerta-cepas, J. *et al.* eggNOG 5 . 0 : a hierarchical , functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. **47**, 309–314 (2019).

112. Krieger, C. J. *et al.* MetaCyc: A multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research* **32**, 438–442 (2004).

113. Keiblinger, K. M. *et al.* Soil metaproteomics - Comparative evaluation of protein extraction protocols. *Soil Biology and Biochemistry* **54**, 14–24 (2012).

114. Greenfield, L. M., Hill, P. W., Paterson, E., Baggs, E. M. & Jones, D. L. Methodological bias associated with soluble protein recovery from soil. *Scientific Reports* **8**, 3–8 (2018).

115. Kulikova, N. A. Interactions between Humic Substances and Microorganisms and. *Molecules* **26**, 1–32 (2021).

116. Zang, X., van Heemst, J. D. H., Dria, K. J. & Hatcher, P. G. Encapsulation of protein in humic acid from a histosol as an explanation for the occurrence of organic nitrogen in soil and sediment. *Organic Geochemistry* **31**, 679–695 (2000).

117. Benndorf, D., Balcke, G. U., Harms, H. & von Bergen, M. Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. *ISME Journal* **1**, 224–234 (2007).

118. Deschamps, C. *et al.* Comparative methods for fecal sample storage to preserve gut microbial structure and function in an in vitro model of the human colon. *Applied Microbiology and Biotechnology* **104**, 10233–10247 (2020).

119. Kasper, J. C. & Friess, W. The freezing step in lyophilization: Physico-chemical fundamentals, freezing methods and consequences on process performance and quality attributes of biopharmaceuticals. *European Journal of Pharmaceutics and Biopharmaceutics* **78**, 248–263 (2011).

120. Masciandaro, G. *et al.* Comparison of extraction methods for recovery of extracellular β-glucosidase in two different forest soils. *Soil Biology and Biochemistry* **40**, 2156–2161 (2008).

121. Taylor, E. B. & Williams, M. A. Microbial protein in soil: Influence of extraction method and C amendment on extraction and recovery. *Microbial Ecology* **59**, 390–399 (2010).

122. Armenta, S., de La Guardia, M. & Esteve-Turrillas, F. A. Hard Cap Espresso Machines in Analytical Chemistry: What Else? *Analytical Chemistry* **88**, 6570–6576 (2016).

123. Tabb, D. L. *et al.* Determination of Peptide and Protein Ion Charge States by Fourier Transformation of Isotope-Resolved Mass Spectra. *J Am Soc Mass Spectrom* **17**, 903–915 (2006).

124. Usenko, S., Subedi, B., Aguilar, L. & Robinson, E. *High-throughput analysis of PPCPs, PCDD/Fs, and PCBs in biological matrices using GC-MS/MS. Comprehensive Analytical Chemistry* vol. 61 (Elsevier B.V., 2013).

125. Ottenhall, A., Henschen, J., Illergård, J. & Ek, M. Cellulose-based water purification using paper filters modified with polyelectrolyte multilayers to remove bacteria from water through electrostatic interactions. *Environmental Science: Water Research and Technology* **4**, 2070–2079 (2018).

126. Tan, X. *et al.* Evaluation of the particle sizes of four clay minerals. *Applied Clay Science* **135**, 313–324 (2017).

127. Shams, K. A. *et al.* Review Article Green technology : Economically and environmentally innovative methods for extraction of medicinal & aromatic plants (MAP) in Egypt. **7**, 1050–1074 (2015).

128. Zhang, K. & Wong, J. W. *Solvent-based extraction techniques for the determination of pesticides in food*. *Comprehensive Sampling and Sample Preparation* vol. 4 (Elsevier, 2011).

129. Bastida, F., Jehmlich, N., Torres, I. F. & García, C. The extracellular metaproteome of soils under semiarid climate: A methodological comparison of extraction buffers. *Science of the Total Environment* **619–620**, 707–711 (2018).

130. Asing, J. *et al.* Optimization of extraction method and characterization of humic acid derived from coals and composts. *J. Trop. Agric. and Fd. Sc* **37**, 211–223 (2009).

131. Sipes, K. *et al.* Permafrost Active Layer Microbes From Ny Ålesund, Svalbard (79°N) Show Autotrophic and Heterotrophic Metabolisms With Diverse Carbon-Degrading Enzymes. *Frontiers in Microbiology* **12**, 1–14 (2022).

132. Gurdeep Singh, R. *et al.* Unipept 4.0: Functional Analysis of Metaproteome Data. *Journal of Proteome Research* **18**, 606–615 (2019).

133. Wenzel, L., Heyer, R., Schallert, K., Löser, L. & Reichl, U. SDS-PAGE fractionation to increase metaproteomic insight into the taxonomic and functional composition of microbial communities for biogas plant samples. 498–509 (2018) doi:10.1002/elsc.201800062.

134. Schäpe, S. S. *et al.* The simplified human intestinal microbiota (Sihumix) shows high structural and functional resistance against changing transit times in in vitro bioreactors. *Microorganisms* **7**, 1–19 (2019).

135. Heyer, R., Kohrs, F., Reichl, U. & Benndorf, D. Metaproteomics of complex microbial communities in biogas plants. *Microbial Biotechnology* **8**, 749–763 (2015).

136. McIlwain, S. *et al.* Crux: Rapid open source protein tandem mass spectrometry analysis. *Journal of Proteome Research* **13**, 4488–4491 (2014).

137. Diament, B. J. & Noble, W. S. Faster SEQUEST searching for peptide identification from tandem mass spectra. *Journal of Proteome Research* **10**, 3871–3879 (2011).

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

139. Argentini, A. *et al.* MoFF: A robust and automated approach to extract peptide ion intensities. *Nature Methods* **13**, 964–966 (2016).

140. Tabb, D. L., Fernando, C. G. & Chambers, M. C. MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of Proteome Research* **6**, 654–661 (2007).

141. Ma, Z. Q. *et al.* IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *Journal of Proteome Research* **8**, 3872–3881 (2009).

142. Dorfer, V., Maltsev, S., Winkler, S. & Mechtler, K. CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. *Journal of Proteome Research* **17**, 2581–2589 (2018).

143. Houel, S. *et al.* Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *Journal of Proteome Research* **9**, 4152–4160 (2010).

144. 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* **10**, 1785–1793 (2011).

145. Wang, J., Bourne, P. E. & Bandeira, N. MixGF: Spectral probabilities for mixture spectra from more than one peptide. *Molecular and Cellular Proteomics* **13**, 3688–3697 (2014).

146. Dorfer, V., Strobl, M., Winkler, S. & Mechtler, K. MS Amanda 2.0: Advancements in the standalone implementation. *Rapid Communications in Mass Spectrometry* **35**, 1–7 (2021).

147. Dorfer, V. *et al.* MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *Journal of Proteome Research* **13**, 3679–3684 (2014).

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

149. Patnode, M. L. *et al.* Interspecies Competition Impacts Targeted Manipulation of Human Gut Bacteria by Fiber- Derived Glycans Article Interspecies Competition Impacts Targeted Manipulation of Human Gut Bacteria by Fiber-Derived Glycans. *Cell* **179**, 59-73.e13 (2019).

150. Xiong, W., Brown, C. T., Morowitz, M. J., Banfield, J. F. & Hettich, R. L. Genome-resolved metaproteomic characterization of preterm infant gut microbiota development reveals species-specific metabolic shifts and variabilities during early life. *Microbiome* **5**, 72 (2017).

151. Webb-Robertson, B. J. M. *et al.* Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of Proteome Research* **14**, 1993–2001 (2015).

152. Lim, M. Y., Paulo, J. A. & Gygi, S. P. Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. *Journal of Proteome Research* **18**, 4020–4026 (2019).

153. Schiebenhoefer, H. *et al.* A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and Prophane. *Nature Protocols* **15**, 3212–3239 (2020).

154. Luczynski, P. *et al.* Growing up in a bubble: Using germ-free animals to assess the influence of the gut microbiota on brain and behavior. *International Journal of Neuropsychopharmacology* **19**, 1–17 (2016).

155. Mazmanian, S. K., Cui, H. L., Tzianabos, A. O. & Kasper, D. L. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* **122**, 107–118 (2005).

156. Franklin, C. L. & Ericsson, A. C. Microbiota and reproducibility of rodent models. *Lab Animal* **46**, 114–122 (2017).

157. Al-Asmakh, M. & Zadjali, F. Use of germ-free animal models in microbiota-related research. *Journal of Microbiology and Biotechnology* **25**, 1583–1588 (2015).

158. Goodman, A. L. *et al.* Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc Natl Acad Sci U S A* **108**, 6252–6257 (2011).

159. Renzone, G., Arena, S. & Scaloni, A. Proteomic characterization of intermediate and advanced glycation end-products in commercial milk samples. *Journal of Proteomics* **117**, (2015).

160. Luévano-Contreras, C., Gómez-Ojeda, A., Macías-Cervantes, M. H. & Garay-Sevilla, M. E. Dietary Advanced Glycation End Products and Cardiometabolic Risk. *Current Diabetes Reports* vol. 17 Preprint at https://doi.org/10.1007/s11892-017-0891-2 (2017).

161. Kumar, J., Das, S. & Teoh, S. L. Dietary Acrylamide and the Risks of Developing Cancer: Facts to Ponder. *Frontiers in Nutrition* vol. 5 Preprint at https://doi.org/10.3389/fnut.2018.00014 (2018).

162. Candela, M. *et al.* Modulation of gut microbiota dysbioses in type 2 diabetic patients by macrobiotic Ma-Pi 2 diet. *British Journal of Nutrition* **116**, (2016).

163. Clarkson, S. M. *et al.* Construction and optimization of a heterologous pathway for protocatechuate catabolism in escherichia coli enables bioconversion of model aromatic compounds. *Applied and Environmental Microbiology* **83**, (2017).

164. Giannone, R. J., Wurch, L. L., Podar, M. & Hettich, R. L. Rescuing Those Left Behind: Recovering and Characterizing Underdigested Membrane and Hydrophobic Proteins To Enhance Proteome Measurement Depth. *Analytical Chemistry* **87**, 7720–7728 (2015).

165. Holman, J. D., Ma, Z. Q. & Tabb, D. L. Identifying Proteomic LC-MS/MS data sets with bumbershoot and IDpicker. *Current Protocols in Bioinformatics* 1–15 (2012) doi:10.1002/0471250953.bi1317s37.

166. Kong, A. T., Leprevost, F. v., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods* **14**, 513–520 (2017).

167. Jeremiah J. Faith, Philip P. Ahern, Vanessa K. Ridaura, Jiye Cheng, J. I. & Gordon. Identifying Gut Microbe-Host Phenotype Relationships Using Combinatorial Communities in Gnotobiotic Mice. **6**, (2014).

168. Ridaura, V. K. *et al.* Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science (1979)* **341**, (2013).

169. Martens, E. C. *et al.* Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biology* **9**, (2011).

170. Tuncil, Y. E. *et al.* Reciprocal prioritization to dietary glycans by gut bacteria in a competitive environment promotes stable coexistence. *mBio* **8**, (2017).

171. McNulty, N. P. *et al.* Effects of Diet on Resource Utilization by a Model Human Gut Microbiota Containing Bacteroides cellulosilyticus WH2, a Symbiont with an Extensive Glycobiome. *PLoS Biology* **11**, (2013).

172. Tauzin, A. S. *et al.* Functional characterization of a gene locus from an uncultured gut Bacteroides conferring xylo-oligosaccharides utilization to Escherichia coli. *Molecular Microbiology* **102**, 579–592 (2016).

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

174. Groer, M. W. *et al.* Development of the preterm infant gut microbiome: a research priority. *Microbiome* **2**, 1–8 (2014).

175. Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biology* **5**, 1556–1573 (2007).

176. Laforest-Lapointe, I. & Arrieta, M.-C. Microbial Eukaryotes: a Missing Link in Gut Microbiome Studies. *mSystems* **3**, (2018).

177. Greenberg, R. G. & Benjamin, D. K. Neonatal candidiasis: Diagnosis, prevention, and treatment. *Journal of Infection* **69**, (2014).

178. Tamburini, S., Shen, N., Wu, H. C. & Clemente, J. C. The microbiome in early life: Implications for health outcomes. *Nature Medicine* vol. 22 Preprint at https://doi.org/10.1038/nm.4142 (2016).

179. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45**, D353–D361 (2017).

180. Russo-Abrahão, T. *et al.* Biochemical properties of Candida parapsilosis ecto-5'-nucleotidase and the possible role of adenosine in macrophage interaction. *FEMS Microbiology Letters* **317**, (2011).

181. Polpitiya, A. D. *et al.* DAnTE: A statistical tool for quantitative analysis of -omics data. *Bioinformatics* **24**, (2008).

182. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* vol. 13 Preprint at https://doi.org/10.1038/nmeth.3901 (2016).

183. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology* **428**, 726–731 (2016).

184. Darzi, Y., Falony, G., Vieira-Silva, S. & Raes, J. Towards biome-specific analysis of meta-omics data. *ISME Journal* vol. 10 Preprint at https://doi.org/10.1038/ismej.2015.188 (2016).

185. Luo, W., Pant, G., Bhavnasi, Y. K., Blanchard, S. G. & Brouwer, C. Pathview Web: User friendly pathway visualization and data integration. *Nucleic Acids Research* **45**, (2017).

186. Darzi, Y., Letunic, I., Bork, P. & Yamada, T. IPath3.0: Interactive pathways explorer v3. *Nucleic Acids Research* **46**, (2018).

187. Croft, D. *et al.* Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Research* **39**, (2011).

188. Gensollen, T., Iyer, S. S., Kasper, D. L. & Blumberg, R. S. How colonization by microbiota in early life shapes the immune system. *Science (1979)* **352**, 539–544 (2016).

189. Hansen, C. H. F. *et al.* Patterns of early gut colonization shape future immune responses of the host. *PLoS ONE* **7**, 1–7 (2012).

190. OLSZAK, T. *et al.* Microbial exposure during early life has persistent effects on natural killer T cell function. *Science (1979)* **336**, 489–493 (2012).

191. Belgacem Mihi & Good, M. Impact of Toll-like receptor 4 signaling in necrotizing enterocolitis: The state of the science. *Clin Perinatol* **46**, 145–157 (2019).

192. Buttó, L. F., Schaubeck, M. & Haller, D. Mechanisms of microbe-host interaction in Crohn's disease: Dysbiosis vs. Pathobiont Selection. *Frontiers in Immunology* **6**, (2015).

193. Madden, J. W. Human breast milk exosomes may protect against necrotizing enterocolitis in preterm infants. *Pediatric Research* 1–2 (2021) doi:10.1038/s41390-021-01580-w.

194. Hackam, D. & Caplan, M. Necrotizing enterocolitis: Pathophysiology from a historical context. *Seminars in Pediatric Surgery* **27**, 11–18 (2018).

195. Gopalakrishna, K. P. *et al.* Maternal IgA protects against the development of necrotizing enterocolitis in preterm infants. *Nature Medicine* **25**, 1110–1115 (2019).

196. Mirzaei, M. K. & Maurice, C. F. Ménage à trois in the human gut: Interactions between host, bacteria and phages. *Nature Reviews Microbiology* **15**, 397–408 (2017).

197. Gasparrini, A. J. *et al.* Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome. *Nature Microbiology* **4**, 2285–2297 (2019).

198. Gibson, M. K. *et al.* Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nat Microbiol* **1**, 16024 (2016).

199. Chen, Y. *et al.* Preterm infants harbour diverse klebsiella populations, including atypical species that encode and produce an array of antimicrobial resistance- and virulence-associated factors. *Microbial Genomics* **6**, 1–20 (2020).

200. Rodríguez, J. M. *et al.* The composition of the gut microbiota throughout life, with an emphasis on early life. *Microbial Ecology in Health & Disease* **26**, (2015).

201. Healy, D. B., Ryan, C. A., Ross, R. P., Stanton, C. & Dempsey, E. M. Clinical implications of preterm infant gut microbiome development. doi:10.1038/s41564-021-01025-4.

202. Berardi, A. *et al.* Are postnatal ampicillin levels actually related to the duration of intrapartum antibiotic prophylaxis prior to delivery? A pharmacokinetic study in 120 neonates. *Archives of Disease in Childhood: Fetal and Neonatal Edition* **103**, F152–F156 (2018).

203. Berardi, A. *et al.* Intrapartum beta-lactam antibiotics for preventing group B streptococcal early-onset disease: can we abandon the concept of 'inadequate' intrapartum antibiotic prophylaxis? *Expert Review of Anti-Infective Therapy* **18**, 37–46 (2020).

204. Azad, M. B. *et al.* Impact of maternal intrapartum antibiotics, method of birth and breastfeeding on gut microbiota during the first year of life: A prospective cohort study. *BJOG: An International Journal of Obstetrics and Gynaecology* **123**, 983–993 (2016).

205. Miller, J. E. *et al.* Maternal antibiotic exposure during pregnancy and hospitalization with infection in offspring: A population-based cohort study. *International Journal of Epidemiology* **47**, 561–571 (2018).

206. Dierikx, T. H. *et al.* The influence of prenatal and intrapartum antibiotics on intestinal microbiota colonisation in infants: A systematic review. *Journal of Infection* **81**, 190–204 (2020).

207. Fouhy, F. *et al.* High-throughput sequencing reveals the incomplete, short-term recovery of infant gut microbiota following parenteral antibiotic treatment with ampicillin and gentamicin. *Antimicrobial Agents and Chemotherapy* **56**, 5811–5820 (2012).

208. Cheng, G. *et al.* Selection and dissemination of antimicrobial resistance in Agri-food production. **2**, 1–13 (2019).

209. Pietsch, F. *et al.* Selection of resistance by antimicrobial coatings in the healthcare setting. *Journal of Hospital Infection* **106**, 115–125 (2020).

210. Touati, A., Zenati, K., Brasme, L., Benallaoua, S. & de Champs, C. Extended-spectrum β-lactamase characterisation and heavy metal resistance of Enterobacteriaceae strains isolated from hospital environmental surfaces. *Journal of Hospital Infection* **75**, 78–79 (2010).

211. Mourão, J., Novais, C., Machado, J., Peixe, L. & Antunes, P. Metal tolerance in emerging clinically relevant multidrug-resistant Salmonella enterica serotype 4,[5],12:i:- clones circulating in Europe. *International Journal of Antimicrobial Agents* **45**, 610–616 (2015).

212. Gómez-Sanz, E. *et al.* Novel erm(T)-carrying multiresistance plasmids from porcine and human isolates of methicillin-resistant Staphylococcus aureus ST398 that also harbor cadmium and copper resistance determinants. *Antimicrobial Agents and Chemotherapy* **57**, 3275–3282 (2013).

213. Hasman, H. *et al.* Copper resistance in Enterococcus faecium, mediated by the tcrB gene, is selected by supplementation of pig feed with copper sulfate. *Applied and Environmental Microbiology* **72**, 5784–5789 (2006).

214. Hasman, H. & Aarestrup, F. M. tcrb, a gene conferring transferable copper resistance in Enterococcus faecium: Occurrence, transferability, and linkage to macrolide and glycopeptide resistance. *Antimicrobial Agents and Chemotherapy* **46**, 1410–1416 (2002).

215. Yang, Q. E., Agouri, R., Tyrrell, M. & Walsh, R. Heavy Metal Resistance Genes Are Associated with blaNDM-1- and blaCTX-M-15-Carrying

Enterobacteriaceae. *Antimicrobial Agents and Chemotherapy* **62**, e02642-17 (2018).

216. Becker, K. W. & Skaar, E. P. Metal limitation and toxicity at the interface between host and pathogen. *FEMS Microbiology Reviews* **38**, 1235–1249 (2014).
217. Djoko, K. Y., Y. Ong, C. L., Walker, M. J. & McEwan, A. G. The role of copper and zinc toxicity in innate immune defense against bacterial pathogens. *Journal of Biological Chemistry* **290**, 1854–1861 (2015).
218. Waters, L. S. Bacterial manganese sensing and homeostasis. *Current Opinion in Chemical Biology* **55**, 96–102 (2020).
219. Lemire, J. A., Harrison, J. J. & Turner, R. J. Antimicrobial activity of metals: Mechanisms, molecular targets and applications. *Nature Reviews Microbiology* **11**, 371–384 (2013).
220. Brown, C. T. *et al.* Hospitalized premature infants are colonized by related bacterial strains with distinct proteomic profiles. *mBio* **9**, (2018).
221. Melville, J. M. & Moss, T. J. M. The immune consequences of preterm birth. *Frontiers in Neuroscience* **7**, 1–9 (2013).
222. Henderickx, J. G. E. *et al.* Maturation of the preterm gastrointestinal tract can be defined by host and microbial markers for digestion and barrier defense. *Scientific Reports* 1–12 (2021) doi:10.1038/s41598-021-92222-y.
223. Busi, S. B. *et al.* Persistence of birth mode-dependent effects on gut microbiome composition, immune system stimulation and antimicrobial resistance during the first year of life. *ISME Communications* **1**, 1–12 (2021).
224. Kai-Larsen, Y., Gudmundsson, G. H. & Agerberth, B. A review of the innate immune defence of the human foetus and newborn, with the emphasis on antimicrobial peptides. *Acta Paediatrica, International Journal of Paediatrics* **103**, 1000–1008 (2014).
225. Nogacka, A. *et al.* Impact of intrapartum antimicrobial prophylaxis upon the intestinal microbiota and the prevalence of antibiotic resistance genes in vaginally delivered full-term neonates. *Microbiome* **5**, 1–10 (2017).

226. Gosalbes, M. J. *et al.* High frequencies of antibiotic resistance genes in infants' meconium and early fecal samples. *Journal of Developmental Origins of Health and Disease* **7**, 35–44 (2016).

227. Gerner, R. R., Nuccio, S. P. & Raffatellu, M. Iron at the host-microbe interface. *Molecular Aspects of Medicine* **75**, 100895 (2020).

228. Ho, T. T. B. *et al.* Enteric dysbiosis and fecal calprotectin expression in premature infants. *Pediatric Research* **85**, 361–368 (2019).

229. Kehl-Fie, T. E. *et al.* MntABC and MntH contribute to systemic staphylococcus aureus infection by competing with calprotectin for nutrient manganese. *Infection and Immunity* **81**, 3395–3405 (2013).

230. Turrentine, M. Intrapartum antibiotic prophylaxis for Group B Streptococcus: Has the time come to wait more than 4 hours? *American Journal of Obstetrics and Gynecology* **211**, 15–17 (2014).

231. Pajarillo, E. A. B., Lee, E. & Kang, D.-K. Trace metals and animal health: Interplay of the gut microbiota with iron, manganese, zinc, and copper. *Animal Nutrition* **7**, 750–761 (2021).

232. Casanova-Hampton, K. *et al.* A genome-wide screen reveals the involvement of enterobactin-mediated iron acquisition in Escherichia coli survival during copper stress . *Metallomics* **13**, (2021).

233. Macomber, L. & Imlay, J. A. The iron-sulfur clusters of dehydratases are primary intracellular targets of copper toxicity. *Proc Natl Acad Sci U S A* **106**, 8344–8349 (2009).

234. Djoko, K. Y. & McEwan, A. G. Antimicrobial action of copper is amplified via inhibition of heme biosynthesis. *ACS Chemical Biology* **8**, 2217–2223 (2013).

235. Ding, C. *et al.* The copper regulon of the human fungal pathogen Cryptococcus neoformans H99. *Molecular Microbiology* **81**, 1560–1576 (2011).

236. Koh, E. I. & Henderson, J. P. Microbial copper-binding siderophores at the host-pathogen interface. *Journal of Biological Chemistry* **290**, 18967–18974 (2015).

237. Festa, R. A. & Thiele, D. J. Copper at the Front Line of the Host-Pathogen Battle. **8**, 9–12 (2012).

238. Solioz, M. Copper disposition in bacteria. *Clinical and Translational Perspectives on WILSON DISEASE* 101–113 (2018) doi:10.1016/B978-0-12-810532-0.00011-2.

239. Andrei, A. *et al.* Cu homeostasis in bacteria: The ins and outs. *Membranes (Basel)* **10**, 1–45 (2020).

240. Grass, G. *et al.* Linkage between catecholate siderophores and the multicopper oxidase CueO in Escherichia coli. *Journal of Bacteriology* **186**, 5826–5833 (2004).

241. Fu, Y., Chang, F. M. J. & Giedroc, D. P. Copper transport and trafficking at the host-bacterial pathogen interface. *Accounts of Chemical Research* **47**, 3605–3613 (2014).

242. White, C., Lee, J., Kambe, T., Fritsche, K. & Petris, M. J. A role for the ATP7A copper-transporting ATPase in macrophage bactericidal activity. *Journal of Biological Chemistry* **284**, 33949–33956 (2009).

243. Miller, K. A., Vicentini, F. A., Hirota, S. A., Sharkey, K. A. & Wieser, M. E. Antibiotic treatment affects the expression levels of copper transporters and the isotopic composition of copper in the colon of mice. *Proc Natl Acad Sci U S A* **116**, 5955–5960 (2019).

244. Sluysmans, S. *et al.* PLEKHA5, PLEKHA6, and PLEKHA7 bind to PDZD11 to target the Menkes ATPase ATP7A to the cell periphery and regulate copper homeostasis. *Molecular Biology of the Cell* **32**, 1–20 (2021).

245. Stephenson, S. E. M., Dubach, D., Lim, C. M., Mercer, J. F. B. & la Fontaine, S. A single PDZ domain protein interacts with the menkes copper ATPase, ATP7A: A new protein implicated in copper homeostasis. *Journal of Biological Chemistry* **280**, 33270–33279 (2005).

246. Clifton, M. C. *et al.* Parsing the functional specificity of Siderocalin/Lipocalin 2/NGAL for siderophores and related small-molecule ligands. *Journal of Structural Biology: X* **2**, 100008 (2019).

247. Diaz-Ochoa, V. E. *et al.* Salmonella Mitigates Oxidative Stress and Thrives in the Inflamed Gut by Evading Calprotectin-Mediated Manganese Sequestration. *Cell Host and Microbe* **19**, 814–825 (2016).

248. Kehl-Fie, T. E. *et al.* Nutrient metal sequestration by calprotectin inhibits bacterial superoxide defense, enhancing neutrophil killing of Staphylococcus aureus. *Cell Host and Microbe* **10**, 158–164 (2011).

249. Liu, J. Z. *et al.* Zinc sequestration by the neutrophil protein calprotectin enhances salmonella growth in the inflamed gut. *Cell Host and Microbe* **11**, 227–239 (2012).

250. Kim, Y. *et al.* NDM-1, the ultimate promiscuous enzyme: Substrate recognition and catalytic mechanism. *FASEB Journal* **27**, 1917–1927 (2013).

251. Xiong, W., Giannone, R. J., Morowitz, M. J., Banfield, J. F. & Hettich, R. L. Development of an enhanced metaproteomic approach for deepening the microbiome characterization of the human infant gut. *Journal of Proteome Research* **14**, 133–141 (2015).

252. Alcock, B. P. *et al.* CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research* **48**, D517–D525 (2020).

253. Shkoporov, A. N. *et al.* Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* **6**, 68 (2018).

254. Hatfull, G. F. & Hendrix, R. W. Bacteriophages and their genomes. *Current Opinion in Virology* **1**, 298–303 (2011).

255. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature* **578**, (2020).

256. Yutin, N. *et al.* Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nature Communications* **12**, 1–11 (2021).

257. Devoto, A. E. *et al.* Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nature Microbiology* **4**, 693–700 (2019).

258. Shkoporov, A. N. & Hill, C. Bacteriophages of the Human Gut: The “Known Unknown” of the Microbiome. *Cell Host and Microbe* **25**, 195–209 (2019).

259. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nature Microbiology* **6**, 960–970 (2021).

260. Adair L. Borges, Yue Clare Lou, Rohan Sachdeva, Basem Al-Shayeb, Alexander L. Jaffe, Shufei Lei, Joanne M. Santini, J. F. B. Stop codon recoding is widespread in diverse phage lineages and has the potential to regulate translation of late stage and lytic genes. *bioRxiv* (2021).

261. Ivanova, N. N. *et al.* Stop codon reassignments in the wild. *Science (1979)* **344**, 909–913 (2014).

262. Crisci, M. A. *et al.* Closely related Lak megaphages replicate in the microbiomes of diverse animals. *iScience* **24**, 102875 (2021).

263. Hanke, A. *et al.* Recoding of the stop codon UGA to glycine by a BD1-5/SN-2 bacterium and niche partitioning between Alpha- and Gammaproteobacteria in a tidal sediment microbial community naturally selected in a laboratory chemostat. *Frontiers in Microbiology* **5**, 1–17 (2014).

264. McCutcheon, J. P., McDonald, B. R. & Moran, N. A. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genetics* **5**, (2009).

265. Polard, P., Prère, M. F., Chandler, M. & Fayet, O. Programmed translational frameshifting and initiation at an AUU codon in gene expression of bacterial insertion sequence IS911. *Journal of Molecular Biology* **222**, (1991).

266. Kim, W. *et al.* Proteomic detection of non-annotated protein-coding genes in Pseudomonas fluorescens Pf0-1. *PLoS ONE* **4**, (2009).

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

268. Baudet, M. *et al.* Proteomics-based refinement of Deinococcus deserti genome annotation reveals an unwonted use of non-canonical translation initiation codons. *Molecular and Cellular Proteomics* **9**, (2010).

269. Pánek, T. *et al.* Nuclear genetic codes with a different meaning of the UAG and the UAA codon. *BMC Biology* **15**, 1–18 (2017).

270. Lou, Y. C. *et al.* Infant gut strain persistence is associated with maternal origin, phylogeny, and traits including surface adhesion and iron acquisition. *Cell Reports Medicine* **2**, 100393 (2021).

271. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host and Microbe* **26**, 527-541.e5 (2019).

272. Pannaraj, P. S. *et al.* Shared and distinct features of human milk and infant stool viromes. *Frontiers in Microbiology* **9**, 1–13 (2018).

273. John, S. G. *et al.* A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environmental Microbiology Reports* **3**, 195–202 (2011).

274. Cui, J., Schlub, T. E. & Holmes, E. C. An Allometric Relationship between the Genome Length and Virion Volume of Viruses. *Journal of Virology* **88**, 6403–6410 (2014).

275. Chaudhari, H. v., Inamdar, M. M. & Kondabagil, K. Scaling relation between genome length and particle size of viruses provides insights into viral life history. *iScience* **24**, 102452 (2021).

276. Michalski, A., Neuhauser, N., Cox, J. & Mann, M. A systematic investigation into the nature of tryptic HCD spectra. *Journal of Proteome Research* **11**, 5479–5491 (2012).

277. Hong, S. H., Kwon, Y. C. & Jewett, M. C. Non-standard amino acid incorporation into proteins using Escherichia coli cell-free protein synthesis. *Frontiers in Chemistry* **2**, 1–7 (2014).

278. Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, , Frank W Larimer, L. J. H. Integrated nr Database in Protein Annotation System and Its Localization. *Nature Communications* **6**, 1–8 (2010).

279. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960 (2005).

280. Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology* **430**, 2237–2243 (2018).

281. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. Trabajo práctico N° 13 . Varianzas en función de variable independiente categórica. *Nature Protocols* **10**, 845–858 (2016).

282. Chan, P. P. & Lowe, T. M. Structural and Functional Annotation of Eukaryotic Genomes with GenSAS in Gene Prediction - Methods and Protocols. *Gene Prediction: Methods and Protocols, Methods in Molecular Biology* **1962**, 1–29 (2019).

283. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. *Nature Biotechnology* **34**, 1256–1263 (2016).

284. Chen, L. X., Anantharaman, K., Shaiber, A., Murat Eren, A. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Research* **30**, 315–333 (2020).

285. Delannoy-Bruno, O. *et al.* Evaluating microbiome-directed fibre snacks in gnotobiotic mice and humans. *Nature* **595**, 91–95 (2021).

286. West, P. T., Probst, A. J., Grigoriev, I. v., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Research* **28**, 569–580 (2018).

VITA

Samantha Peters was born in Littleton, Colorado and raised throughout the United States, from the Great Plains of the Midwest to the Atlantic and Pacific coasts. In 2014, Samantha graduated with a Bachelor of Science degree in Microbiology with a specialization in Infectious Diseases from South Dakota State University in Brookings, SD. After completing her undergraduate degree, she spent three years working for SomaLogic, Inc. in Boulder, CO designing DNA-based high-affinity protein binding reagents and methodologies with applications in protein biomarker discovery and characterization. In 2017, she began her Ph.D. studies in the Genome Science and Technology program at University of Tennessee-Knoxville, where she completed PhD dissertation work in the Bioanalytical Mass Spectrometry group at Oak Ridge National Laboratory under the supervision of Dr. Robert L. Hettich.