Characterizing Early-life Microbiome Functionality in Premature Infant Gut by a Metaproteomics Approach

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Weili Xiong

May 2016

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to the following people who have always supported and encouraged me in this Ph.D. journey. Without their patience, guidance and persistent help, this dissertation would not have been completed.

First of all, I would like to thank my advisor Dr. Robert Hettich for his guidance, support and dedication to helping me complete this dissertation. In my graduate life, he was an excellent mentor who helped me develop expertise in all aspects of scientific research, and also a “father” who always display trust and encouragement that allowed me to grow whenever I got challenged. I am grateful for all things I have learnt from him, especially his enormous scientific enthusiasm and positive spirit, and all opportunities he has provided during my time working with him.

I would like to thank my committee members: Dr. Jeffrey Becker, Dr. Mircea Podar, Dr. Chongle Pan and Dr. Loren Hauser for their effort and valuable comments in this dissertation. Their inspiration and knowledge helped me think from different perspectives and broaden the scope of my project. Their time and inputs to my dissertation are greatly appreciated.

I would like to thank all my lab colleagues for their guidance, discussions and encouragement throughout my PhD life. Thank you for the time when we share joys and difficulties of being a graduate student. In particular, I would thank Dr. Richard Giannone and Dr. Paul Abraham for their substantial help in teaching me all technical aspects of mass spectrometry experiments and always providing support and suggestions whenever I need their help. I would also like to acknowledge and thank the Genome Science and Technology Program for providing me this great opportunity and environment to pursue my scientific career.

This project allowed me to collaborate with Dr. Jillian Banfield from University of Berkeley and Dr. Michael Morowitz from University of Pittsburg whom I had fantastic experience working with and would like to thank for their contributions in this project.

Lastly, I would like to thank my family and all my friends. I want to give great thanks to my parents, who made my education a priority and always supported and believed in me. I thank my beloved husband, Xi Wang, for giving me constant love, care, support and patience that helped me through difficult times and made it possible for me to complete my degree.

ABSTRACT

Microbes inhabit all parts of human body that are exposed to the environment and their interactions with human host mutually benefit each other and play significant roles in human health and diseases. The human gastrointestinal tract harbors the largest population of the microbiota and has gained broad research attention and efforts over the past decade. Colonization of the gut by microbes begins at birth and this early-life bacterial establishment can impact infants' health and even the human health and lifestyle across an entire life span. Recent studies on community structure and composition of infant gut microbiota have revealed the species shifts and variations during early bacterial colonization of the infant gut. However, little is known about functional activities of the community and how these functions change in response to different life events. Therefore, comprehensive proteomic characterization of the infant gut microbiome is needed to elucidate biological activities in this complex ecosystem. In this dissertation, we first developed a metaproteomics pipeline integrating both experimental and informatics components with careful considerations to simultaneously access microbial and human host proteins contained in infant fecal samples. The developed approach was applied in a longitudinal study of a healthy premature infant gut, revealing the overall metabolic cooperation between the human host and the gut microbiota and the temporal functional shifts in both microbiome and corresponding host response during the colonization process. To further investigate the commonalities and differences of gut microbiome between individuals, time-series metaproteomic studies were performed in three more premature infants, uncovering common core proteins/metabolic pathways established during early life microbiome establishment, as well as unique pathways that were specific to particular infants or present in certain colonization time period. In a broad perspective, the approaches and results presented in

this dissertation have provided insights into functions and activities of human gut ecosystem, and developed techniques and outlined general considerations that can be extended to proteome characterization of all complex ecosystems.

TABLE OF CONTENTS

CHAPTER 1. Introduction to mass spectrometry based metaproteomic characterization of human gut microbiome 1

1.1 Meta-omics technologies enable microbial community functional characterization 1

1.1.1 Microbial community 1

1.1.2 Metaproteomics among various –omics technologies 3

1.2 Mass spectrometry (MS) based proteomics and metaproteomics 8

1.3 Metaproteomics of human gut microbiome 12

1.3.1 Introduction to human microbiome 12

1.3.2 Current human gut metaproteome studies 14

1.4 Scope of the dissertation 23

CHAPTER 2. Experimental and computational platform for human gut microbiome research 25

2.1 General workflow for metaproteome measurements of human gut microbiome 25

2.2 Sample preparation 27

2.2.1 Sample collection 27

2.2.2 Direct versus indirect sample preparation method 28

2.2.3 Cell lysis and proteome extraction 29

2.2.4 Protein digestion 31

2.3 Liquid chromatography 31

2.4 Mass spectrometry instrumentation 35

2.4.1 Ionization sources 35

2.4.2 Mass analyzer and detector 38

2.4.3 Data acquisition in mass spectrometry 43

2.4.4 Tandem mass spectrometry 44

2.5 Bioinformatics 45

2.5.1 Database searching algorithm 45

2.5.2 Protein inference 49

2.5.3 Protein quantification 50

2.5.4 Functional groups assignment 52

CHAPTER 3. Development of an enhanced metaproteomic approach for deepening the microbiome characterization in the human infant gut 53

3.1 Introduction 53

3.2 Materials and methods 56

3.3 An enhanced strategy for infant fecal proteomics to improve the overall depth of proteome measurement 62

3.4 Microbial protein group identifications are enriched by depletion of abundant human proteins 67

3.5 Enriched microbial protein identifications facilitate more comprehensive information for microbial functional categorization 77

3.6 Conclusions 79

CHAPTER 4. Instrumental and informatics considerations for metaproteomics 81

4.1 Introduction 81

4.2 Enabling monoisotopic precursor selection for in-depth proteome measurement 82

4.3 Informatic considerations for human gut metaproteome 88

4.3.1 Construction of protein sequence database 88

4.3.2 Impact of metagenome quality and complex on peptide identifications 92

4.3.3 Quality assessment of tandem mass spectra 93

4.3.4 Protein grouping and clustering 97

4.3.5 Protein quantification 102

4.4 Conclusions 104

CHAPTER 5. Metaproteomics of a healthy premature infant gut to access early-life microbial functionality and host responses 105

5.1 Introduction 105

5.2 Materials and methods 108

5.3 General overview of metaproteomic datasets 110

5.4 Metagenome - metaproteome comparisons 117

5.5 Global metabolic pathways of human proteome and gut metaproteome 120

5.6 Microbial functional characterization 124

5.7 Human host response changing across time 132

5.8 Conclusions 136

CHAPTER 6. Characterization of temporal and inter-individual functional differences in infant gut microbiome by metaproteomics approach 140

6.1 Introduction 140

6.2 Materials and methods 141

6.3 General overview of metaproteomic datasets 143

6.4 Microbial community profile 148

6.5 Main microbial functionality in infant gut microbiome 152

6.5 Characterization of temporal and inter-individual differences in microbial functions 154

6.6 Comparison of human proteins among multiple infants 161

6.7 Discussions and conclusions 167

CHAPTER 7. Conclusions and future perspectives 170

7.1 Conclusions from the development and application of metaproteomics approach in the characterization of infant gut microbiome 170

7.2 Remaining challenges and future perspectives for human gut metaproteome research ... 174

7.2.1 The need for better assembled and annotated metagenomes 174

7.2.2 The need for high-throughput measurement campaigns 175

REFERENCES 177

VITA 192

LIST OF TABLES

Table 3.1. 21 microbial isolate reference genome database 60

Table 3.2. Overview of proteomic results from two fecal microbiomes measured by the direct and the indirect DF method..... 66

Table 3.3. Collected and assigned mass spectra results..... 68

Table 4.1. Comparisons of collected/assigned spectra, identified peptides/proteins with MIPS enabled and disabled..... 84

Table 5.1. Number of identified peptides, protein groups and MS/MS spectra 111

Table 5.2. Frequencies of human and microbial proteins identified across time 118

Table 5.3. Enriched BP GO terms among different clusters..... 128

Table 5.4. Top 20 abundant human proteins in 12 fecal samples..... 133

Table 6.1. Summary of infant medical information..... 142

LIST OF FIGURES

Figure 1.1. "Multi-omics" approaches enabling the comprehensive understanding of biological ecosystems 4

Figure 1.2. Top down and bottom up mass spectrometry 10

Figure 2.1. General workflow of human infant gut metaproteomics 26

Figure 2.2. Schematic diagram of MudPIT column setup 33

Figure 2.3. Schematic representation of the electrospray ionization process 37

Figure 2.4 Schematic of the Orbitrap Elite Hybrid Mass spectrometer 40

Figure 2.5. Peptide fragmentation ion type 46

Figure 2.6. Computational metaproteomics workflow 47

Figure 3.1. Workflow of the indirect double filtering (DF) method 58

Figure 3.2. Reproducibility of methodological (sample preparation) replicates 63

Figure 3.3. Protein group quantification reproducibility 64

Figure 3.4. Rank-abundance plots of protein groups 69

Figure 3.5. Distributions of ScanRanker scores for collected mass spectra 71

Figure 3.6. Comparison of protein group identification and quantification results by two methods 73

Figure 3.7. Microbial protein group identification 75

Figure 3.8. COG category analysis of microbial protein groups 78

Figure 4.1. Number of high and low abundance identified proteins with MIPS enabled and disabled 85

Figure 4.2. Number of MS/MS events followed by every full MS scan in 11 salt pulse steps with MIPS enabled 86

Figure 4.3. Box plot of MS1 percentages in 11 salt pulse steps with MIPS enabled (a) and disabled (b) 87

Figure 4.4. Informatics workflow for metaproteomics 89

Figure 4.5. Impact of database quality on peptide identifications 94

Figure 4.6. Impact of sample complexity on peptide identifications 95

Figure 4.7. Evaluation of ScanRank to determine unidentified high quality spectra 98

Figure 4.8. Degree of database redundancy 100

Figure 4.9. Determination of similarity threshold 101

Figure 4.10. Comparison of protein MITs with spectral counts 103

Figure 5.1. Number of human and microbial protein groups identified (a) and relative abundance of human/microbial spectra (b) over time 113

Figure 5.2. Number of total collected spectra and high quality spectra for each sample 115

Figure 5.3. Multidimensional scaling (MDS) plot of 12 fecal proteomes for microbial proteins (a) and human proteins (b) 116

Figure 5.4. Organism-specific proteome coverage 119

Figure 5.5. Pattern of changes in microbial abundance (a) and protein abundance (b) 121

Figure 5.6. KEGG pathways mapping for human proteome and human gut metaproteome 122

Figure 5.7. Microbiome GO term distributions at level 3 of biological process (BP), molecular function (MF), and cellular component (CC) 124

Figure 5.8. Hierarchical clustering of microbial proteins 126

Figure 5.9. Comparisons of relative protein abundance in two samples collected on the same day 127

Figure 5.10. Changes of human proteome across time 135

Figure 5.11. GO enrichment of human proteome over time 136

Figure 6.1. Number of identified human (blue) and microbial (red) protein groups of four infants over time 144

Figure 6.2. Relative abundance of human (blue) and microbial (red) protein groups of four infants over time 145

Figure 6.3. MDS plots of microbial proteins for infants #19, #21 and #23 146

Figure 6.4. MDS plots of human proteins for infants #19, #21 and #23 147

Figure 6.5. Pattern of changes in microbial abundance (a, b and e) and protein abundance (c, d and f) for infants #19, #21 and #23 149

Figure 6.6. Venn diagram of assigned KOs in four infants 153

Figure 6.7. KEGG pathways mapping of common microbial Kos 155

Figure 6.8. Temporal and inter-individual differences of major microbial functions 157

Figure. 6.9. Most significantly differentially expressed KOs among infants 160

Figure 6.10. Venn diagram of human proteins among infants 162

Figure 6.11. KEGG pathways mapping of common human proteins..... 163

Figure 6.12. Hierarchical clustering of human proteins among infants..... 165

CHAPTER 1

Introduction to mass spectrometry based metaproteomic characterization of human gut microbiome

Part of the text below was adapted from:

Weili Xiong, Paul Abraham, Zhou Li, Chongle Pan, Robert L. Hettich. Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. *Proteomics*, 2015, 15 (20), 3424-3438.

Weili Xiong's contributions included: literature review, manuscript writing in experimental workflow and human gut metaproteoimcs studies sections, data analysis, and manuscript editing.

1.1 Meta-omics technologies enable microbial community functional characterization

1.1.1 Microbial community

Microorganisms constitute the largest population grouping on Earth, outnumbering all other living organisms and constituting the main portion of the Earth's biomass [1]. In nature, microbes do not live in isolation, but rather function together as members of complex dynamic communities that can range from low complexity systems in extreme environments, such as acid mine drainage biofilm [2], to a moderate complexity of communities that inhabit in the human body [3], to tens of thousands of species that colonize the environment ecosystems such as soils [4] and seawaters [5]. Ubiquitous throughout the natural environment, these microorganisms represent a remarkable biodiversity and greatly contribute to all ecosystem processes, e.g. biogeochemical recycling of essential elements such as carbon and nitrogen [6]. In addition, microbial communities are associated with and also directly impact other macro-organisms like plants, animals and humans. The human body is one example of a unique ecosystem that is

continuously inhabited by a diverse collection of microbes (including bacteria, fungi, protozoa, viruses and archaea) [7]. The interaction between the microbiome and human host has a profound impact upon human health and so the disruption of this intricate balance between the microbial colonization and human host corresponding responses has been linked to a wide range of inflammatory, metabolic, and central nerves system disorders [8-11].

Despite the tremendous success achieved by classic microbiology focusing on single-protein study and pure-culture isolate investigation, the major challenge in modern microbial physiology and ecology is to globally understand microbial community composition, structure, function and most importantly, how microbes interact *in situ* and respond to their environment. The advent of comprehensive whole-genome sequencing technology [12, 13] has enormously expanded our knowledge of microbial metabolism by revealing a complete gene inventory list for single pure culture isolates. Nevertheless, estimates indicate that over 90% environmental microorganisms are not cultivatable or not yet cultured [14] and the true microbial activities and interactions in the natural environment cannot be simulated in the traditional laboratory experiment. As a result, the field of metagenomics has developed [15, 16], where genomic analysis of a mixed microbial community is carried out based on microbial DNA that is directly extracted from the environmental samples, regardless of the isolation and cultivation of microbial members. These sequence data provide the potential function of microbial communities and also pave the way for system biology studies employing multiple meta-omics approaches [17]. These large-scale technologies have allowed the investigation of all the levels of biological information (DNA, RNA, proteins and metabolites), each of which provides a different level of insight into the complex metabolic processes and interactions contributing to the microbial ecosystem.

1.1.2 Metaproteomics among various –omics technologies

"Omics" technologies offer a holistic view of all molecules in an organism or a community. Aiming to characterize complete microbial ecosystems, four major types of high-throughput measurements are commonly employed: metagenomics (DNA), metatranscriptomics (RNA), metaproteomics (proteins) and meta-metabolomics (metabolites) (Figure 1.1). Since DNA carries the instruction that builds the functional biomolecules, the metagenome provides the potential list of genes that could possibly be expressed by the communities. The metatranscriptome reveals the transcriptionally active subset of the genomes, which provides some information about the gene expression and regulation at the time of sampling. Primarily responsible for structural and enzymatic activities, proteins reflect dynamics and specific microbial activities in a given environment, making metaproteomics a particularly useful tool for the characterization of the microbial functionality. Additionally, the meta-metabolome reveals small molecules, which can be substrates, inhibitors or products that participate in the metabolic network and, together with the metagenome and the metaproteome, provides a strong insight into the microbial interaction and activities. In total, these omics technologies have revolutionized microbial ecological studies, in the way that they enable in-depth and high throughput measurement and drastically expand our understanding of community structure, function and dynamics *in situ*.

Significant advances in sequencing technologies and bioinformatics over the past decade have created new ways to characterize the genetic repertoire of many types of environmental communities, including microorganisms inhabiting the human body [10]. However, compared to single isolate genome sequencing, metagenomics studies face major challenges in obtaining deeper sequencing measurement and developing computational tools for intensive analysis [18].

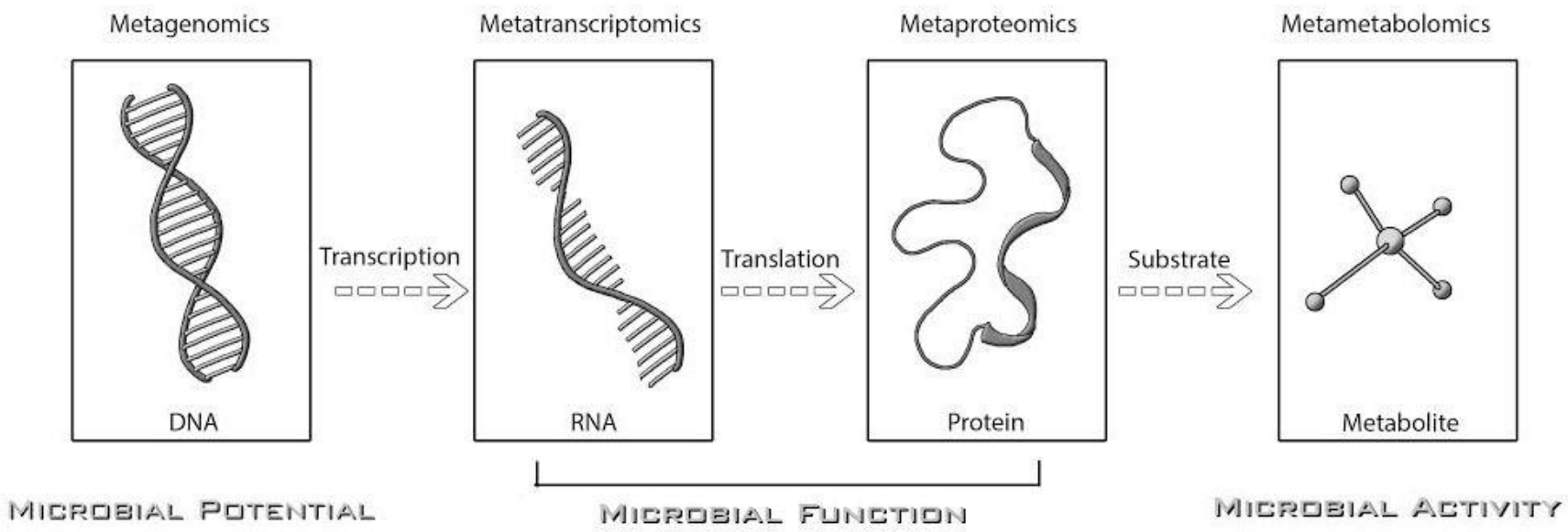


Figure 1.1. “Multi-omics” approaches enabling the comprehensive understanding of biological ecosystems. Each level of information (DNA, RNA, protein and metabolites) achieved by “multi-omics” approaches provides a different insight into the metabolic processes of microbial ecosystems.

The complexity of communities increases not only in the number of membership, but also the range of relative abundance of all organisms. Therefore, complete genome reconstruction requires higher sequence coverage and more extensive assembly/curation processes. Although different sequencing approaches have been applied to complex communities, the most commonly used technology such as Illumina sequencing, still generates short-length reads, which make the metagenome assembly particularly challenging, due to the repeated regions and homologous sequences derived from closely related species [19]. In particular, low abundance genomes and less sequenced fragments are less likely to be recovered from the metagenomics data. Although new sequencing technologies generating longer reads or combining multiple sequencing approaches begin to appear, it still remains to be seen what impact they will have on metagenomics for complex microbial communities. In order to reconstruct single genomes from metagenomics data, binning algorithms have been specifically developed for metagenomics reads or assembled contiguous sequences (contigs). For example, one of the most widely used binning approaches is ESOM (Emergent Self - Organizing Map), which bins assembled contigs based on tetra - nucleotide frequencies [20] or time series abundance profiles [21]. Further functional annotation, taxonomic classification as well as extensive curation of the genome are required, in order to access the taxonomic information and genetic functional profiles of microbial communities. Overall, metagenomics has become a powerful tool for microbial community characterization, and has also provided reference information for metatranscriptomic data mapping and metaproteomic data matching. While metagenomics reveals the potentially expressed gene in the community, metatranscriptomics can be used to examine changes in gene expression. Next generation sequencing (NGS) has facilitated sequencing of all populations of RNA (termed RNAseq [22]), including messenger RNA (mRNA) and other non-coding RNAs.

As non-coding RNAs account for 95-98% of total RNAs [23], metatranscriptomics [24] commonly involves mRNA enrichment procedures and followed by reverse transcription to cDNA, which can be sequenced using the same technologies as for metagenomics. Transcriptomic reads can be mapped to assembled genomic sequences or assembled *de novo* for transcript reconstruction and quantification. As mRNAs cannot account for post-translational regulation and indeed not all mRNAs can be transcribed to proteins, transcriptome cannot be used as a direct proxy for metabolic activities but only reflects gene expression and potential functions.

Metaproteomics [25] aims to investigate the actual gene translational products, proteins, in a community and provides additional information about post-translational modifications and localization information over transcriptome measurements. The correlation between mRNA and protein levels is generally poor at a single time snapshot, suggesting that proteome is likely more indicative of biological phenotype than transcriptome. Commonly, metaproteomic studies identify proteins via multi-dimensional chromatography – tandem mass spectrometry (MS/MS). The relative protein abundance can be determined by label free, metabolic labeling and isobaric chemical labeling approaches [26, 27]. Unlike DNA and mRNA, there is no polymerase chain reaction (PCR) amplification for proteins and therefore it is more challenging to achieve similar dynamic ranges in proteome measurement as that in genome and transcriptome measurement. Although peptide sequences can be determined *de novo*, this strategy is rarely applied to metaproteomic studies. Database searching is the more conventional way of peptide/protein identifications, in which experimental MS/MS spectra are correlated against the theoretical spectra that are derived from an available metagenome. Therefore, the success of metaproteomics is strongly dependent on the complexity and quality of the metagenomics data.

In combination with metagenomics, remarkable success in metaproteomics has been demonstrated for a variety of environmental samples [27]. Conversely, proteomic information can also be applied to improve genome annotation [28]. The advantage of metaproteomics over other omics technology is that identified proteins can be directly used for active pathway and metabolic function construction and specific functions can be further attributed to different species in the community.

Even though metaproteomics is a powerful tool to interrogate microbial function in the ecosystem, protein abundance cannot accurately predict a protein's activity or functional state. To achieve a more complete understanding of metabolic activities, meta-metabolonmics usually complements metaproteomics by providing information of small molecule metabolites (substrates, intermediates, end-products) involved in the metabolism [29]. It is recognized that metabolomics provides unique insight into metabolic dynamics and characterizes the ecosystem phenotype that results from the interplay between genomes and the environment [30]. The main challenge of metabolomics is the ability to identify and quantify the entire set of metabolites with diverse chemical properties [31]. In addition, their concentration and composition vary rapidly in response to environmental changes. Thus, various extraction methods and analytical approaches are desired, in order to capture all different compounds in the community. Unlike other omics technologies, metabolomics is not directly linked to the genome, which makes the identification of metabolites more challenging and impossible to assign specific taxa to any metabolites. Indeed, most of current meta-metabolomics studies still catalog metabolite features defined by m/z and retention time, with large amount of which are classified as unknown [32, 33].

Although above omics technologies generally measures and characterizes the gene expression event, there are dramatic temporal information differences in these omics

measurements due to different half-lives and post machinery, and thus, no exact quantitative correlation should be expected among them. In fact, each omics technology provides a unique perspective; by integrating all these information in the cascade from genes to proteins and further to metabolites, it is powerful to provide new insight into overall community metabolism at a resolution and range not previously possible.

1.2 Mass spectrometry (MS) based proteomics and metaproteomics

The first proteome measurement was achieved by using two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) [34, 35], which separates proteins by isoelectric point and protein mass. However, this technology was limited to low-complexity protein mixtures and suffered from a lack of reproducibility and methods for identifying proteins from gel spots. In attempt to obtain comprehensive protein identifications from complex samples, high throughput screening, large dynamic range and accurate protein identifications are required for the proteomic measurements. In this regard, mass spectrometry (MS) based proteomics has become the dominant and indispensable tool for the proteomics study. The implement of mass spectrometry in the proteome analysis was driven by the breakthrough of two soft ionization methods: electrospray ionization (ESI) [36] and matrix assisted laser desorption/ionization (MALDI) [37], which solved the difficulty of producing gas-phase ions from large biologically important molecules, for example, proteins and peptides. This groundbreaking invention brought a whole new dimension to proteomics and also led to the rapid development of protein separation techniques and mass spectrometer instrumentation that are compatible with either ion source. To reduce the complexity of peptide mixtures, early proteomics studies employed 2-dimensional gel electrophoresis (2DGE) for protein separation and effectively differentiate protein isoforms and

modifications [38]. However, it suffered from limited dynamic range and low-throughput, and was outperformed by liquid chromatography (LC)-based separation, which could be easily coupled on-line with ESI-MS and enabled protein characterization of complex samples [39, 40]. At the same time, tremendous advancement in mass spectrometers, for example the advent of the Orbitrap mass analyzer [41], has greatly increased the sensitivity, mass accuracy, dynamic range, resolution, and scanning speed of MS measurements, and thereby leading the field towards more advanced proteomic measurements.

As the MS platform continuously improved, two fundamental and complementary strategies for protein identification and characterization have been established in proteomics: top-down (TD) and bottom-up (BU) mass spectrometry (Figure 1.2) [42, 43]. The TD approach has merged as an essential tool for protein characterization by not only offering the ability to sequence and quantify intact proteins, but also giving the opportunity to study protein isoforms, protein structure and post-translational modification. However, TD approach is still challenged by the difficulty in measuring protein molecules with high molecular weight and the insufficient separation of complex protein mixtures prior to the MS analysis [44]. As a more favorable approach for large-scale proteomic investigations, BU proteomics, also termed as “shotgun” proteomics, measures proteolytic peptides digested from proteins. In this method, proteins are digested into small peptides very specifically by proteases (eg. trypsin, chymotrypsin, pepsin, etc). Then the peptide mixture is separated by LC and subjected to MS where they are measured by mass and further fragmented into MS/MS spectra. Bioinformatics algorithms have been developed to assign resulting MS/MS spectra into peptides and assemble the assigned peptides into proteins for identifications [45, 46]. Although the peptide-centric strategy is capable of providing massive amount of information, protein identification and quantification can be

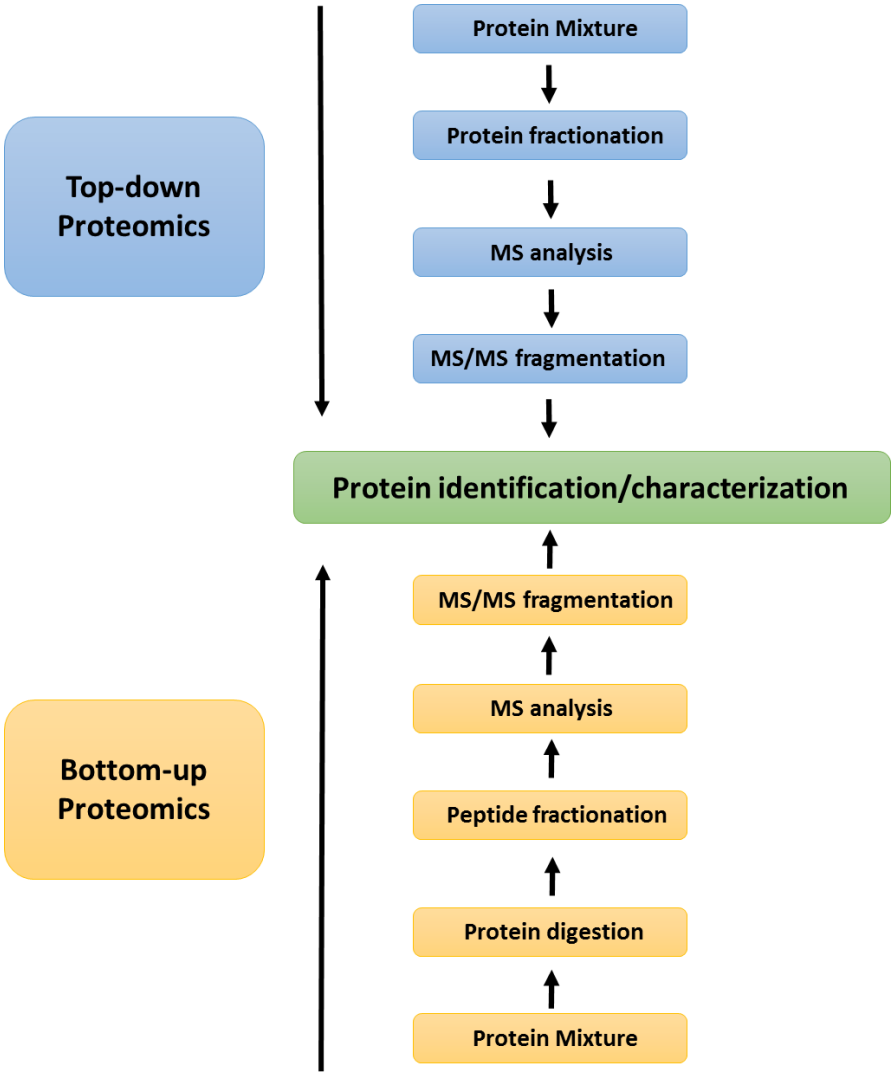


Figure 1.2. Top down and bottom up mass spectrometry. The top-down approach measures intact proteins that are separated/fractionated before MS analysis. The bottom-up approach measures proteolytic peptides digested from proteins and protein identification is achieved from peptide sequences.

ambiguous due to the genome redundancy and sequence homology among different proteins, especially for high order organisms, such as human and plant. Nevertheless, BU proteomics is currently the mainstream method for comprehensive proteomic characterization of many biological systems. Advancement in DNA sequencing methods, LC separation and high resolution mass spectrometry has allowed the transition of shotgun proteomics from profiling single organism grown in culture to metaproteomics that investigates environmental communities directly obtained in nature [47-50].

Metaproteomics, proposed by Wilmes and Bond, was first demonstrated in microbial communities collected from acid mine drainage (AMD) site [48]. As a low complexity ecosystem, greater than 2000 proteins were characterized in the five most abundant species from this extreme environment. Later, the platform has been successfully applied to much more complex ecosystems such as human gut microbiome with a thousand species of bacteria [51] and soils comprising tens of thousands to millions of microbial species [52]. Despite the remarkable progress that has been made in metaproteomics methods, notable challenges still remain in all aspects of sample preparation, mass spectrometer performance and informatics tools. While protein extraction is straightforward when applied to cultured microbial isolates, environmental samples usually face problems with limited amount of biomass and the presence of interference material, which prevents efficient cell lysis and protein extraction from environmental samples. Therefore, investigators continue to develop and optimize protein extraction protocols adapting to different types of complex samples, aiming to achieve comprehensive protein identifications [53]. Another challenge that remains is the large dynamic range within complex samples. It limits the detection of low abundance proteins and thus impedes the unbiased interrogation of the entire protein complement. In order to increase the dynamic range, improvements have been

made in the peptide separation, for example the multidimensional protein identification technology (MudPIT) [54]. Also, developments of high resolution, high speed mass spectrometers, for example, the Orbitrap-Elite, have greatly increased the sampling depth and the dynamic range of metaproteomics [55]. While experimental and technical advancements generate more comprehensive dataset, the success of metaproteomic investigation also relies on the quality of metagenome and informatics workflows that are compatible with large environmental dataset. Due to the large size of the database, metaproteomics study is hindered by computationally intensive analysis of large volume of the data, the ineffectiveness of evaluating protein identifications using false discovery rates (FDRs), the ambiguity of protein inference and quantification, and the inaccuracy of functional annotation and interpretation [56]. Overall, there is a strong need for a metaproteomics pipeline that takes into accounts both experimental and informatics complications described above, to improve the metaproteomic characterization of environmental communities. A general informatics workflow will be further discussed in Chapter 4.

1.3 Metaproteomics of human gut microbiome

1.3.1 Introduction to human microbiome

As one of the unique environmental communities, the human body is a complex and dynamic ecosystem where human host and microbes live in symbiosis and together play an important role in human physiology. There are more than 100 trillion microbial cells residing in the human body, such as skin, oral cavity, gastrointestinal tract and vagina, which outnumber human cells by 10 to 1 and collectively constitute our microbiome [57]. As technologies for

sequencing and bioinformatics continue to develop, microbial structure and functional potential of the microbiota have been revealed in different sites of the human body [58]. Substantial variations have been displayed in intra-individual microbiomes at different body sites as well as inter-individual microbiomes at the same sites across different people [59]. However, microbiomes seem to be temporally stable and highly resilient at the same site of a single human adult. In other words, body site is a more determinant factor than the temporal variation, unless there is a large perturbation such as infection, or use of antibiotics that could completely change the community structure [60]. Current research interests therefore focus on discovering different patterns of microbiomes in healthy versus diseased states and temporally how microbial composition changes during the course of the disease [61, 62].

The largest microbiome locates in our gastrointestinal (GI) tract ecosystem, which is inhabited by trillions of microbes, representing thousands of bacterial species [63]. Gut microbiota play an essential role in human health and diseases by processing and providing human host essential nutrients, acting as a barrier to prevent the pathogen invasion and aiding in the establishment and development of human immune system [64]. For example, intestinal microbes contribute to the degradation of certain food components such as planted-derived polysaccharides, production of short chain fatty acids (SCFA) such as acetate, propionate and butyrate, and synthesis of vitamins that are essential to human health. Indeed, a balance between the microbial colonization and human host responses is crucial to the maintenance of human health. However, to date, relatively little is known about the intricate details and balance of the human gut microbiota.

As studied over the past decade, intestinal microbiota can be impacted by host genetics, environmental factors, types of diets and health status [65-68]. For example, a recent study

identified microbial taxa whose abundances were more similar in monozyotic twins as compared to dizygotic twins [65]. Studies on obesity have revealed that dietary interventions could change some individual's microbiota [69]. Also, the dysfunction of microbiota has been linked with a large number of diseases such as obesity and Crohn's disease [3, 70, 71]. Given that not a single organism but rather community characteristics are more associated with diseases, the metaproteomic characterization of functional activities and interactions for gut microbiota will be central to our understanding of microbiome in relation to diseases.

1.3.2 Current human gut metaproteome studies

So far, relatively few studies have been conducted on the gut microbial metaproteome, in spite of large numbers of metagenomic interrogations. This is due at least in part by several challenges in gut proteome studies: 1) heterogeneity of bacterial species composition among different individuals; 2) wide dynamic range of protein abundances, especially dominant human proteins that mask the low abundance microbial microbiome; 3) lack of matched metagenomes or low quality metagenome assemblies/annotations that impede comprehensive MS/MS spectrum assignment; 4) informatics hurdles, such as differentiation and quantification of proteins from closely related species and characterization of diverse post-transcriptional modification events. In the following section, we will briefly highlight the range of human gut metaproteomics studies published to date, and how the information they provide is helping to shape our understanding of this unique ecosystem, and its effect on health vs. disease.

1.3.2.1 Insights into the stable microbiome of a healthy human adult gut

The first metaproteomic study of an adult human gut microbiota was performed on a healthy female monozygotic twin pair from a Swedish twin cohort [72]. Thousands of identified

proteins facilitated the first glimpse of the functional signature of the human gut microbiome, providing insight into the host-bacterial interaction in the gastrointestinal tract. Expectedly, a substantial proportion of the proteins identified in the samples (30%) were human proteins, including but not limited to the functional categories humane innate immunity, cell-to-cell adhesion, and digestion enzymes. Notably, most of the relatively abundant human proteins were similar among the two individuals, yet some differences were found in less abundant proteins, which can be expected due to the stochastic sampling nature of the approach.

A high-level overview of biological processes occurring in gut microbiota was obtained by cataloging identified proteins by Cluster of Orthologous Groups (COGs) [73]. An uneven distribution of relative abundances of each COG in the identified metaproteome relative to metagenome was revealed [72]: the metaproteome was enriched in proteins involved in translation, energy production, and carbohydrate metabolism, whereas the metagenome was dominated by proteins involved in inorganic ion metabolism, cell wall and membrane biogenesis, cell division, and secondary metabolite biosynthesis. Although there are clearly measurement depth differences between these datasets, these observational differences emphasize the important point that *in situ* functional activities (as measured by the metaproteome) can be significantly distinct from what is predicted from the metagenome information alone.

1.3.2.2 Microbial functional divergence of healthy versus disease state

The first comparison of the intestinal microbiota between healthy and diseased adults focused on Crohn's disease (CD) [10]. In brief, CD is an inflammatory bowel disease with evidence converging to suggest that imbalance in the microbiota plays a central role in chronic inflammation associated with CD. In contrast to the healthy twin pair described above, five other

twin pairs were selected here, including one concordant colonic CD (CCD) twin pair, two concordant ileum CD (ICD) twins, and two discordant ICD twins were analyzed. Due to advancements in protein sequencing technology as well as sample preparation, this study was able to achieve a more detailed investigation into the presence of microbial and human proteins, identifying 4,120 microbial protein groups and 1,646 human proteins. With a comprehensive cataloging of proteins and their relative abundances across the individuals, this study highlighted key functional signatures of CD, which were associated with alterations in bacterial metabolism (e.g. deficiency in general processes, depleted enzymes for carbohydrates and mucin degradation, and depletion of butyrate and other short-chain fatty acids), bacterial-host interactions (e.g. higher expression of bacterial outer membrane proteins that participate in inflammatory immune responses), and host corresponding response (e.g. impaired epithelial barrier and high abundance of pancreatic enzymes). Consistent with previous 16S rRNA-based phylogenetic analysis and metabolite analysis of the same cohort, the measured metaproteomes clustered according to individuals' disease status, rather than host genetics. Additionally, reduced protein abundances for butyrate production and degradation of mucin from beneficial bacteria were in agreement with the decreased abundances of these species revealed from previous 16S based phylogenetic profiling. Overall, this study revealed a catalogue of proteins exhibiting the functional signatures of the disease and therefore provided potential targets for future diagnostic and therapeutic research.

In a more targeted investigation, a recent cross-sectional study conducted on six CD patients and six healthy controls also characterized protein signals associated with CD [74]. A subset of 13 candidate proteins was selected and confirmed by selected reaction monitoring (SRM). 12 bacterial proteins mainly derived from *Bacteroides* were strongly linked to CD, as

well as one depleted human glycoprotein 2 of zymogen granule membranes (GP2), which may promote bacteria binding to host cell receptors and induce inflammatory responses. In total, this study not only discovered but also confirmed and quantified a list of CD-associated microbial proteins, which can serve as candidate targets for IBD treatment.

In effort to identify how the gut microbiota contributes to obesity, Ferrer *et al.* performed comparative metagenomics and metaproteomics of human fecal samples from one 'lean' and one 'obese' adolescent [75]. In brief, the proteins identified by shotgun proteomics revealed a drastic change in the total and functionally active microbial community; *Bacteroidetes* represented the most functional bacteria (81% of total protein) in the lean gut, whereas the obese gut had relatively equal abundances of *Firmicutes* and *Bacteroidetes* proteins. This observation is consistent other studies that have shown that the relative abundance of *Bacteroidetes* increases as obese individuals lose weight [76]. Overall, this study highlighted the importance of comparative metaproteomics approaches to further our understanding of the functional changes that occur in response to obesity.

1.3.2.3 Longitudinal changes and shifts in microbiota functionality

To date, only two studies have examined the change of adult gut metaproteomes as a function of time. In the first study, the metaproteomes of three healthy, omnivorous female subjects were characterized twice within a year [77]. As a novel finding, the fecal metaproteome of each individual was relatively stable during one year period, despite distinct inter-individual differences. In addition, approximately 1,000 proteins were observed in all subjects and likely represent core functional categories, which were also highly representative in other intestinal metaproteome studies [72]. These observations suggested a presumable common functional core

in healthy individuals, which is mainly involved in carbohydrate transport and degradation as well as a variety of surface proteins reflecting bacterial adaption to the intestinal environment. A later time-series study examined gut microbial communities over multiple time points from an individual before and after antibiotic (AB) treatment [78]. Based on integrated multi-omics data, the study proposed a presumptive model describing temporal responses of intestinal microbiota to AB therapy, from the perspective of microbial composition dynamics and metabolic activity regulation.

1.3.2.4 The mucosal luminal interface (MLI)

In general, the intestinal mucosal surface is a barrier layer that prevents the invasion of pathogens and mediates most interactions between the host and luminal intestinal microbiota. Thus far, two studies have profiled MLI metaproteomes in mucosal lavage samples. The first study analyzed 205 lavage samples from six colon regions of 38 healthy subjects [79]. The results were compared with mucosal biopsy transcriptome and showed enrichment in extracellular proteins involved in immune response. Also, metaproteomes from 6 colon regions were further compared, revealing biogeographic features of MLI metaproteome with distinct differences between the proximal colon and the distal colon. The second study investigated the bacteria and metaproteome at the MLI of CD, ulcerative colitis, and healthy subjects, and identified five bacterial phylotypes and a large number of proteins associated with the inflammatory bowel disease (IBD) [80]. Moreover, the relationship between bacteria and metaproteome provided a correlation that could be used to sort most subjects by disease type, supporting the potential role of host-microbe interaction in the etiology of IBD. Investigating the metaproteome of the MLI provides an additional dimension to the characterization of host-

microbial interaction, because the approach is capable of analyzing the biogeographic-specific metaproteomes at different locations along the gastrointestinal tract.

The two studies outlined above provide evidence that the bacteria and proteins identified in MLI are clearly involved in host-microbe interactions, which are potentially critical for disease biology. In a somewhat distinct but complementary fashion, fecal microbiota undoubtedly represent a mixture of species from various intestinal regions, thereby presenting an average but broader picture of all microbes and their functional activities along the human gut. Altogether, microbiome studies focused on both fecal and mucosal materials can be complementary to more fully characterize the functions of gut microbiota in human physiology.

1.3.2.5 Model gut microbiome systems in gnotobiotic animals

Gnotobiotic mice can be custom-designed with a defined microbial membership and therefore provide a tractable *in vivo* model to study bacterial and host dynamics. In fact, the microbiome can be 'humanized' by inoculating the germ-free gnotobiotic mice with a defined collection of human gut members. For example, to study the adaption of dietary *Lactococcus lactis* to the digestive tract, Roy et al. colonized gnotobiotic mice with a *Lactococcus lactis* strain and then analyzed the metaproteomes of fecal and cecal samples [81]. Although increased GroEL expression in fecal samples suggested that the bacteria were adapting to dehydrated environment in the colon, nearly identical protein profiles were identified between bacteria from feces and cecum. As compared to proteins from *in vitro* culture, the *in vivo* proteome showed activation of pathways involved in carbon source assimilation, pyruvate catabolism and pentose phosphate, reflecting changes in the fermentative metabolism of *L. lactis* in the digestive environment. A similar study on the proteome of commensal *E. coli* in a gnotobiotic mouse was

later performed [82]. In this case, *E. coli* appeared to express proteins/enzymes that facilitate the utilization of a variety of carbohydrates and amino acids present in the intestinal tract.

Gnotobiotic mice have also been employed to better understand colonization and microbial interactions in the host gut. For example, a model two-member human gut microbiome consisting of *E. rectale* and *B. thetaiotaomicron* was created in gnotobiotic mice to study how they interact and respond to host diet [83]. The study mainly focused on the transcriptomic changes after colonization, but proteins present in luminal contents were also analyzed by high-resolution mass spectrometry. In general, the proteome datasets were complementary to the transcriptome information, and revealed proteins abundant in both microbes as ribosomal proteins, elongation factors, chaperones, and proteins involved in energy metabolism.

Moving beyond a two-member community, a higher level of microbial complexity was evaluated by colonizing gnotobiotic mice with a model human gut microbiota comprising 12 human gut bacterial species and feeding them with high-fat vs. low-fat diets [84]. Importantly, as the complexity of the metaproteome increased, the assignment of peptides unique to proteins was affected by homologous proteins and closely related species. Furthermore, the correlation between mRNA and protein data was evaluated for *Bacteroides cellulosilyticus* WH2 genes revealed a moderate correlation (r =0.53) between overall mRNA and protein levels; yet, the correlations of genes in different functional categories were significantly different. For example, genes involved in translation showed no correlation whereas genes predicted in carbohydrate metabolism had a strong correlation between RNA and protein observations. This further emphasizes the significance of proteome measurement because proteins represent actual functional molecules that may have different temporal and stability characteristics compared with their corresponding transcripts.

The development of “humanized” gut microbiomes in gnotobiotic animals provides a unique ecosystem in which microbial membership can be carefully designed (to control complexity), controlled, and manipulated in a systematic fashion that is not possible in human subject studies. Clearly, the eukaryotic host differences are important here as well, but this system is becoming increasingly important for sorting out and simplifying the complex variables present in human systems.

1.3.2.6 Human infant gut metaproteome

Compared to the adult gut microbiota, the human infant gut has been much less studied. While the variability of the human gut microbiota is astounding, it is not unexpected, given the influences from genetic variation and diverse cultural environments. Although the human infant gut is thought to be generally sterile at birth, this theory has been recently challenged by new evidence suggesting the presence of microbes in amniotic fluid, placenta, and the infant’s meconium [85-87]. Following birth, rapid microbial colonization occurs and, for the next few years, the microbial composition continues to undergo dramatic changes until a stable microbiota is established [88]. As such, the early microbial composition of the infant gut is relatively simple and of low complexity, and therefore poses fewer analytical challenges (e.g., sampling depth) than the richer, more diverse microbial communities evident in the adult human gut. However, with increasing time, the microbial composition varies tremendously, even from week to week, and therefore a comprehensive profiling of the infant gut requires a greater number of sampling points to effectively capture this inherent variation across time.

Emerging evidence has suggested that not only does the initial colonization of the gastrointestinal tract play a critical role in the development of a stable, healthy ‘adult’ microbiota,

but also that deviations from the native early-life bacterial establishment can impact human health and lifestyle across an entire life span [89, 90]. For example, one of the most devastating conditions of premature infants is necrotizing enterocolitis (NEC), which affects 10% of premature infants and has a mortality rate of 30% - 50% [91]. NEC typically develops within the first 2 weeks after birth and once developed, the progression is rapid with significant clinical consequences. NEC is often incurable, and those who survive can suffer from long-term complications such as intestinal stricture and short-gut syndrome [92]. The specific pathogenesis of NEC is still unknown, but factors involved in prematurity, low birth weight, feeding type, use of antibiotics and intestinal bacterial colonization have been indicated in the etiology of NEC [93, 94]. Given the findings that NEC does not occur in germ – free animals [95] and outbreak of NEC could occur in multiple neonates [96], the pattern of microbial community remains to be the most important target in the pathogenesis of NEC. As compared to term infants, premature infants typically harbor delayed and less complex microbial communities. However, no single species or microbial pattern has been identified as infectious agents [97], suggesting that microbial activities, or interactions among microbes and between microbes and human host may contribute to the onset of NEC. Therefore, it is of great interests to not only capture the genetic diversity of the infant gut microbiome, but to also identify which genetic and external factors alter the molecular composition and activity of the infant gut microbiome.

Although the human infant gut microbiome is a logical place to begin metaproteome studies, to date there have been very little published in this arena. Despite having limited genome information, Klaassens *et al.* reported the first attempt to use a metaproteomics approach to functionally characterize microbial protein composition changes over time in a human infant fecal sample [98]. Although the level of protein identification was severely limited in this study,

this report revealed the need for enhanced experimental (sample preparation as well as measurement methods) and informatics (in particular, more detailed and accurate metagenomes) methodologies. Our understanding of the infant microbiome has since broadened, in part owing to the tremendous improvements in DNA and protein sequencing technologies, as well as significant advancements in the bioinformatic tools used to assemble, annotate, and analyze the data generated. In a more recent study, Young *et al.* achieved a more comprehensive metaproteome analysis of the infant gut microbiome, providing a rich dataset that has led to a better understanding of the dynamic changes in the functional signature of the infant microbiome [99]. For example, this study demonstrated that the functional signature of the microbial community increased in complexity within 2-3 weeks, stabilized relatively early, and remained remarkably conserved thereafter. Additionally, several microbial-related human proteins were concomitantly observed. In particular, several innate immunity proteins in the same fecal samples revealed a level of human host - microbiome cross-talk.

1.4 Scope of the dissertation

This dissertation focuses on the development and application of a high performance mass spectrometry and computational informatics platform for human gut microbiome research. In particular, the research presented in this dissertation aims to measure thousands of proteins from infant gut microbiome samples, in an effort to understand metabolic activities of both human host and microbial membership during early life microbial colonization of healthy and diseased premature infants. This field is still in its infancy and thus this dissertation characterizes the infant gut metaproteome with following objectives: 1) optimize the experimental methodology and MS platforms to enhance the depth of infant gut metaproteome measurement; 2) design a

bioinformatics workflow for searching large experimental datasets with sequenced metagenome and assembling assigned peptides into definitive protein information and 3) reveal functional information and temporal functionality variation in the gut metaproteome among multiple infants.

Chapter 2 will provide a metaproteomics pipeline designed for human gut metaproteome characterization with details in sample preparation methods, experimental procedures and informatics tools for protein identifications. Chapter 3 will describe the development of an optimized sample preparation method to fractionate human and microbial proteins in infant fecal samples, as a more comprehensive way to improve the infant gut microbiome characterization. Chapter 4 will focus on informatics considerations and workflow for metaproteomics, including the discussion of evaluating the quality of assembled metagenome, protein clustering and strategies for protein quantifications. Chapter 5 will perform a time-series metaproteomic analysis in a healthy premature infant and reveal temporal functionality development and host response during early life microbial colonization. Chapter 6 will include three more infants with different health status and discuss changes of functionality in both human and microbial proteins across time among different infants. Chapter 7 will conclude the research presented in this dissertation and summarize the contributions of this dissertation to the human gut microbiome field as well as future directions.

CHAPTER 2

Experimental and computational platform for human gut microbiome research

Part of the text below was adapted from:

Weili Xiong, Paul Abraham, Zhou Li, Chongle Pan, Robert L. Hettich. Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. *Proteomics*, 2015, 15 (20), 3424-3438.

Weili Xiong's contributions included: literature review, manuscript writing in experimental workflow and human gut metaproteoimcs studies sections, data analysis, and manuscript editing.

2.1 General workflow for metaproteome measurements of human gut microbiome

The shotgun proteomics approach via 2-dimensional liquid chromatography coupled with nano-electrospray tandem mass spectrometry (2D-LC-nESI-MS/MS) was employed for all metaproteome experiments discussed throughout this dissertation (Figure 2.1). Since the objective of this dissertation is to characterize the infant gut metaproteome, the general strategy begins with the sample collection of stool samples over a specific time period from multiple premature infants. Efficient protein extraction from environmental samples is not trivial, due to many different interfering materials present in the sample. Thus, collected fecal samples can be processed by direct or indirect protein extraction methods. Extracted proteins are denatured, reduced and enzymatically digested into peptide mixtures. Prior to MS measurements, the proteolytic peptide mixtures are separated by multidimensional high performance liquid chromatography (HPLC), which is directly coupled with a mass spectrometer. Peptides are then ionized and transferred into the mass analyzer and detector where their mass to charge (m/z) and

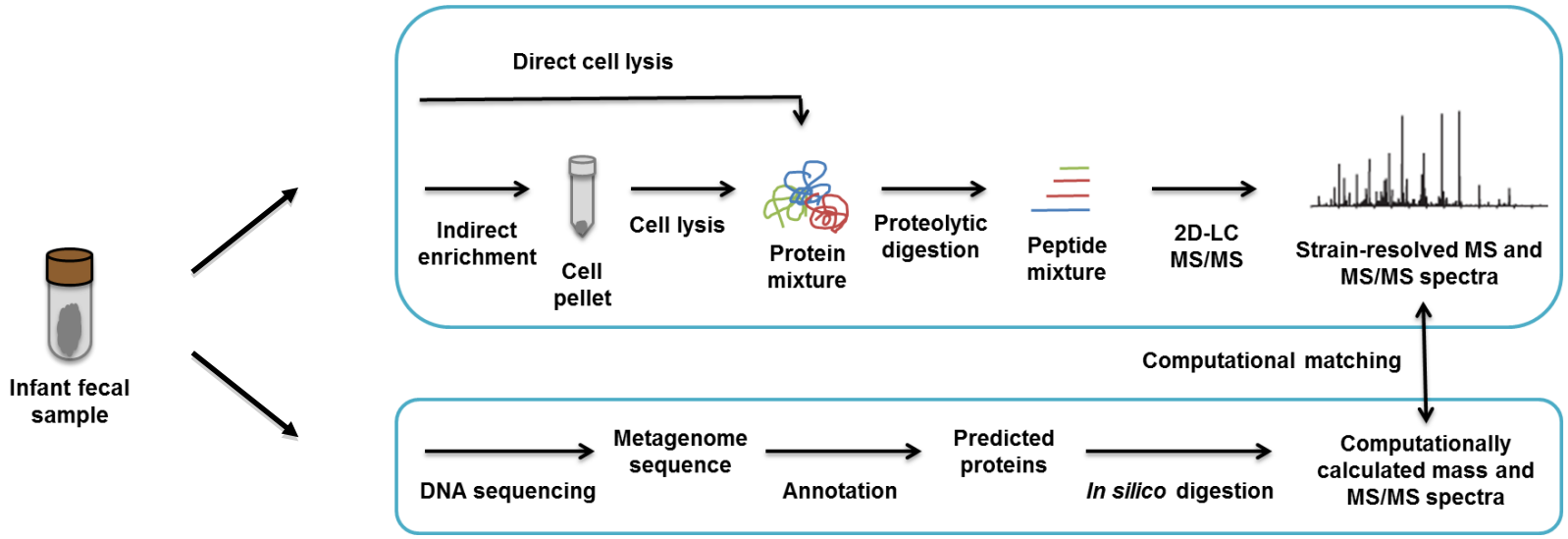


Figure 2.1. General workflow of human infant gut metaproteomics. Protein extraction of fecal samples can be performed using a direct or indirect method. Extracted proteins are digested into peptides, followed by liquid chromatography (LC) separation and mass spectrometric measurements. Protein identifications are achieved by computational matching with protein database predicted from the metagenome.

relative abundance are measured in a full mass spectrum (MS1). For peptide sequencing, top abundant peaks are subjected to fragmentations, generating tandem mass spectra (MS/MS). The identification of peptides is completed by computationally matching generated experimental MS/MS with theoretical spectra that are predicted from *in silico* tryptic digestion of the protein database. This protein database is predicted and annotated based on the metagenomic sequences. Therefore, the success of peptide identifications also relies on a high quality of metagenome construction. Identified peptides are assembled into proteins and spectral counts or matched ion intensity can be used for protein quantification. Overall, the pipeline presented above provides the ability to identify and quantify thousands of proteins in complex fecal samples, and thus characterize the infant gut microbiota functions and metabolic activities at a remarkably deep level. Specific details for each step are discussed below.

2.2 Sample preparation

2.2.1 Sample collection

Metaproteome measurements of gut microbiota are typically conducted with fecal samples, due in large part to the significant amount of microbial biomass in fecal material, and the ease of collecting temporal samples that reflect intestinal conditions under either healthy or disease-related conditions. The most common experimental challenges for this sample type include highly abundant host cells and proteins, endogenous compounds that can interfere with protein measurements, and limited sample sizes (e.g., human infants). In contrast to fecal samples, other studies have used an endoscopic saline-lavage technique to study the mucosal luminal interface (MLI) [79, 80]. Whereas fecal samples represent a mixed population of

microbiota collected from all intestinal regions, mucosal lavage sampling profiles the microbiota at specific biogeographic regions. These samples have been shown to yield robust recovery of surface microbiota and often do not require any additional preprocessing besides centrifugation to separate the cell pellet from supernatant. One potential complication of this approach is that the collection may yield low microbial biomass, so sample handling is somewhat more difficult and constrained.

In this dissertation, fecal samples were collected from multiple premature infants (#UN1 and #CA1 in Chapter 3, #3 in Chapter 5, #19, #21 and #23 in Chapter 6) over the first three months. All infants were born preterm, the gestational ages of #UN1 and #CA1 are unknown and all other infants were born at gestational age between 24 ~ 27 weeks. Detailed information about infants is discussed in each chapter below. All samples were coded for de-identifications and obtained under an IRB agreement protocol and sent to ORNL with dry ice by Dr. Michael Morowitz (University of Pittsburg). Fecal samples were stored at – 80℃ and thawed prior to cell lysis and protein extraction. In general, 0.3 ~ 0.5 g raw fecal material contains sufficient proteins for proteomic measurement. Since human fecal samples are biohazards, all samples were handled in a Biosafety Level 2 (BSL2) hood with proper personal protections.

2.2.2 Direct versus indirect sample preparation method

Depending on the focus of the study, protein extraction in fecal samples can be accomplished by either a direct or indirect enrichment protocol. Although feces are a complicated environmental matrix consisting of bacteria and other microbes, host cells, food particles, and fibrous material, it is possible to extract proteins via a direct cellular lysis of raw fecal material (typically a few grams of material), followed by protein precipitation and cleanup

procedures [77]. A unique advantage of a direct extraction is the ability to simultaneously extract and thus monitor both host and microbial proteins, facilitating the characterization of bacterial signatures as well as their interplay/communication with the host. However, the depth of microbial proteome measurement can be limited by the presence of highly abundant host proteins, especially in infant gut samples where microbial colonization is significantly reduced. To circumvent this challenge, indirect enrichment methods, in which bacterial cells are separated/enriched by differential centrifugation [10, 72, 75, 78] (low speed centrifugation to remove large fecal debris followed by high speed centrifugation to pellet bacterial cells) or high-speed centrifugation on a Nycodenz density gradient [74, 98, 100], facilitate deeper bacterial proteome measurements, though at the expense of increased sample losses and possible sample bias. In Chapter 3, an indirect method with double filtering strategy, which removes large fibrous material and human cells in the first filter and then collects microbial cells in the second filter, has developed and shown to effectively enrich microbial populations in the infant fecal samples with dominant human proteins.

2.2.3 Cell lysis and proteome extraction

With collected cell pellets from direct or indirect method, different methods can be used to lyse cells, including chemical (i.e., detergents, acids, alkalis or organic solvents) lysis, mechanical (i.e., homogenization, bead-beating, sonication) disruption, or a combined approach of both, as has been reported for proteome extraction methods from other complex media [53, 101]. While these approaches are moderately comparable in efficacy, each one has distinct advantages and disadvantages that need to be matched to demands of the instrumentation measurement technique employed for protein/peptide identifications.

Recently developed SDS-TCA method has been successfully applied to efficient protein extraction in complex environmental samples such as soil and feces [10, 53]. Sodium dodecyl sulfate (SDS) is a strong detergent that disrupts cell membrane and denatures proteins. Due to the presence of thick peptidoglycan layer, Gram-positive bacteria are typically less susceptible to the detergent based cell lysis. Therefore, boiling samples in SDS followed by sonication effectively lyse the cell and recover proteins, in particular hydrophobic proteins located in the cell membrane. However, SDS is essentially not compatible with protease activity and MS analysis and thus needs to be removed from protein samples. TCA (trichloroacetic acid) precipitating proteins followed by acetone wash is frequently used to clean the sample and has shown to help remove interfering materials such as detergents, lipids and humic acids in soil samples. Another method, FASP (filtered – aided sample preparation) [102] employs a molecular weight cut-off filter (typically 30 kDa) to capture proteins on the filter and wash away detergents. It performs on-filter protein extraction and digestion, and has shown enhanced proteome results especially for small amount of samples. However, FASP has limitations in high amount of proteins or complex environmental samples, which can easily result in the filter clogging and failure. Besides SDS, MS compatible detergents RapiGest™ SF (Waters) [103, 104] or sodium deoxycholate (SDC) [105, 106] can also be used for protein extraction and show significant advantages since they are not disruptive to protease activity and easily removed by acidification. Nevertheless, SDS has the strongest ability to disrupt cell membrane and solubilize proteins. By considering both the robustness and cost-effectiveness, SDS-based cell lysis followed by TCA precipitation cleanup procedure has been applied for protein extraction in this dissertation.

2.2.4 Protein digestion

Prior to enzymatic digestion, the total concentration of proteins was quantified using the bichinchoninic acid (BCA) assay. This step is critical in two aspects: first, the amount of proteins subjected to investigation can be standardized across all samples; second, an optimal amount of protease can be applied to proteins for effective digestion. Despite many different proteases can be used for protein digestion, trypsin is the most widely used enzyme in bottom-up proteomics. Trypsin specifically cleaves at the C-terminal side of lysine and arginine residues (if the adjacent amino acid is not proline), which generates tryptic peptides with average 10~15 amino acids and positions a basic amino acid (lysine or arginine) at C terminus. In such a way, generated peptides have a molecular weight optimal for most mass analyzers with mass range of 400 – 2,000 m/z and carry positive charges on both ends so that peptides can be ionized and detected in the mass spectrometer. In this study, all extracted proteins were first denatured and reduced to open up the protein for efficient digestion. Trypsin (Promega; Madison, WI) was added in 1:100 ratio (trypsin : protein (w/w)) according to measured protein concentration. The digestion was performed with an initial incubation for 4 hours and followed by an additional incubation overnight at room temperature.

2.3 Liquid chromatography

Proteolytic digestion of proteins extracted from fecal samples generally results in complex peptide mixtures, which must then be fractionated to simplify sample complexity prior to the mass spectrometric measurement [107]. For all proteomic experiments discussed in this dissertation, multi-dimensional liquid chromatography separations, in specific, multidimensional

protein identification technology (MudPIT, Figure 2.2), was used to separate the complex peptide mixture in a high-throughput manner. MudPIT is a non-gel-based chromatography system to resolve tens of thousands of peptides based on their charge and hydrophobicity. First described by Yates groups, the method uses two dimensional liquid chromatography consisting of a strong cation exchange (SCX, separating peptides by charge) and a reverse phase (RP, separating peptides by hydrophobicity) resin in a microcapillary chromatrographic column.

In this approach, a back column was first packed with ~4 cm SCX (Luna 5 µm particle size, 100 Å pore size, Phenomenex, Torrance, CA) followed by ~4 cm RP (C18, Aqua 5 µm particle size, 125 Å pore size, Phenomenex, Torrance, CA) material in a fused silica microcapillary column (150 µm inner-diameter, Polymicro Technologies, Phoenix, AZ). Prepared back column was pre-washed for 5 minutes with Solvent B (70% acetonitrile, 30% HPLC grade water, 0.1% formic acid (FA)) and equilibrated for 5 minutes with Solvent A (95% HPLC grade water, 5% acetonitrile, 0.1% FA). 25 ~ 50 µg peptides were pressure-loaded onto the biphasic back column and bind to RP due to the hydrophobicity. Back-columns were washed offline for 45 minutes by equilibrating with Solvent A first and eluting peptides from RP to SCX by a gradient to Solvent B. This step is critical to the clean-up of the sample by removing the salts or any residual SDS prior to MS measurements of the peptides. Next, the back column was coupled in-line with a front column, which is an in-house pulled nanospray emitter (100 µm inner-diameter, Polymicro Technologies, Phoenix, AZ). The front column was packed with ~15 cm RP material and connected to the back column via a PEEK union and 0.5 µm inline filter.

After the offline wash, the back column was coupled with a HPLC pump (U3000, Dionex, San Francisco, CA) for liquid chromatographic separations. The assembly is provided in Figure 2.2. Solvents, including A, B (both described above), and D (500 mM ammonium acetate in

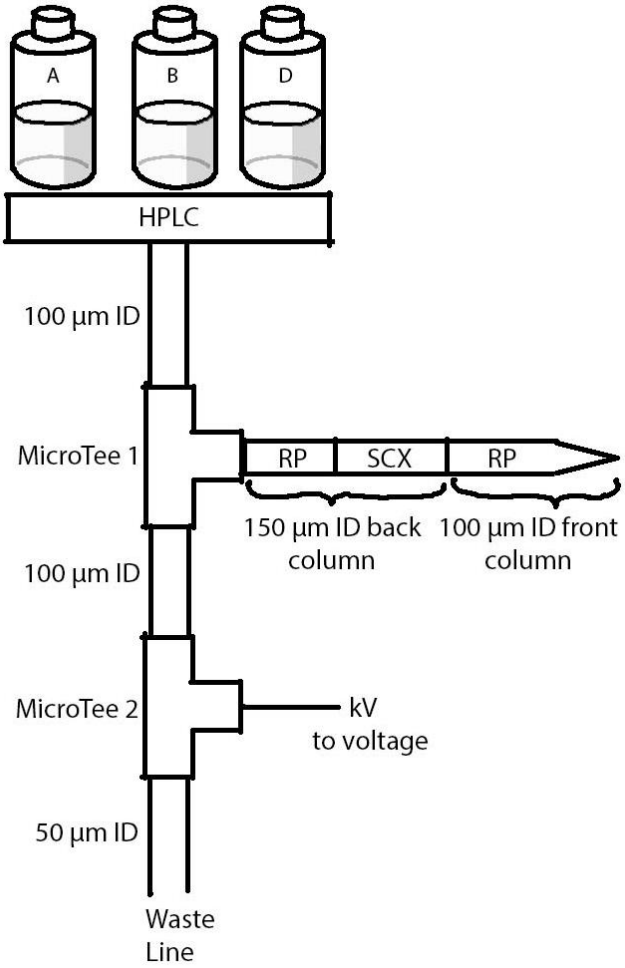


Figure 2.2. Schematic diagram of MudPIT column setup. HPLC, high performance liquid chromatography.

Solvent A), flowed from the HPLC pump (µL/min) and split to either the mass spectrometer or to the waste. This design results in a lower flow rate at the front column (nL/min) and direct the majority of salt into the waste. The back column was connected at the first MicroTee and directed the flow to the front column and then into the mass spectrometer. The other branch of the first MicroTee was connected to a second MicroTee where a voltage (2-6 kV) can be applied for the ionization process. At the end, the rest flow went into the waste through a 50 µm fused silica providing back pressure. An appropriate length (~ 75cm) of the waste line can provide a back pressure of ~75 bar and a flow rate of ~ 400 nL/min at the front column which aid in the formation of electrospray.

In a MudPIT experiment, peptides were separated/eluted in 11 steps (each lasting ~2 hour) with an increasing amount of salts (ammonium acetate) followed by the RP organic gradients in each step. During a single step, a short salt pulse (Solvent D, 5 minutes) was applied to displace a fraction of peptides from SCX of the back column onto RP in the front column. A long gradient (105 minutes) of increasing Solvent B from 0% to 50% was followed to further separate the peptides in the second dimension. The concentration of salt was increased in each iterative step from 5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20%, 25%, 35%, 50% to 100%. Only in the last step, the gradient reaches 100% Solvent B rather than 50% in the previous 10 steps for a complete elution of peptides. Thus at every step, only a portion of simplified peptides were measured, and as a result, this strategy greatly improved analytical dynamic range and protein coverage in proteomic measurements. The 11-step MudPIT was applied to proteomic measurements in Chapter 5 and 6. A modified three-step mini MudPIT, only consists of three salt pulses (10%, 25%, and 100% of 500 mM ammonium acetate) was employed in Chapter 3.

2.4 Mass spectrometry instrumentation

Over the past few decades, mass spectrometry has undergone tremendous improvements in developing and adapting various types of mass analyzers allowing for robust, accurate and high-through proteomics experiments. A mass spectrometer typically consists of three major components: the ionization source, the mass analyzer and the detector. The analytes of interest are first introduced into the ionization source to produce gaseous ions with positive or negative charges. The ions are then passed through the mass analyzer and resolved according to their mass to charge (m/z) ratio. Ions emerging from the analyzer are detected and measured by their relative abundance. After the signal is converted into an intensity value, the result is displayed graphically on the computer as a mass spectrum with the relative abundance on the y-axis and the m/z ratio on the x-axis.

2.4.1 Ionization sources

The invention of soft ionization technologies, including ESI and MALDI [36, 37], has made possible the ionization (generally no fragmentations) of large, nonvolatile and thermally labile biomolecules, for example peptides and proteins. These ionization methods have therefore made mass spectrometry a popular and powerful tool for biological studies. While both methods are widely applied in the analysis of proteins, they have distinctive but equally important features that complement each other. In MALDI, a pulsed laser is used to desorb the analyte that has been mixed with a matrix, promoting the ionization. MALDI has excellent sensitivity, broad mass range and predominantly generate singly-charged ions which simplify the interpretation of results. With all these strengths, MALDI-MS gains its high popularity in the analysis of intact proteins with accurate molecular mass achieved. However, MALDI is incompatible with LC and

less efficient in the fragmentation process due to the singly charged precursor ions. In contrast, ESI generates multiply-charged ions, allowing the analysis of large molecules on limited mass-range mass analyzers and facilitating more complete fragmentation spectra in tandem MS measurements. Another important advantage of ESI over MALDI is that ESI can be easily interfaced with LC and thus more suitable for the identification of complex peptide mixtures in a bottom-up proteomics experiment.

ESI is conducted by applying a high voltage (typically 4 kV) to the capillary emitter that transfers the peptide solution into the mass spectrometer (Figure 2.3). The solution moving through the emitter is ionized into a protonated form and forms a Taylor cone (described by Sir Geoffrey Taylor) [108] at the emitter tip where the surface tension force competes with the electric force. When the applied voltage is high enough, the Taylor cone emits a fine jet of liquid, which easily breaks into charged parent droplets that contain like charges (positive) distributing across the surface of the droplet. There are two forces counterbalance each other in the droplet: the surface tension which holds the droplet together; and the like charge repulsion (known as Coulomb force) that tries to break down its spherical shape. As these charged droplets move toward the heated capillary, their solvent begins to evaporate and their size shrinks. This causes an increase of surface charge density until it reaches the “Rayleigh stability limit” where the surface tension force no long can sustain the repulsion force. At this point, the droplets undergo the “Coulomb fission” [109] and form smaller charged progeny droplets. The fission process will occur repeatedly and eventually the gas-phase positive-charged peptide ions are formed. Produced ions will pass through the heated capillary, which can help further evaporate the remaining solvent and lead ions into the analyzer of the mass spectrometer.

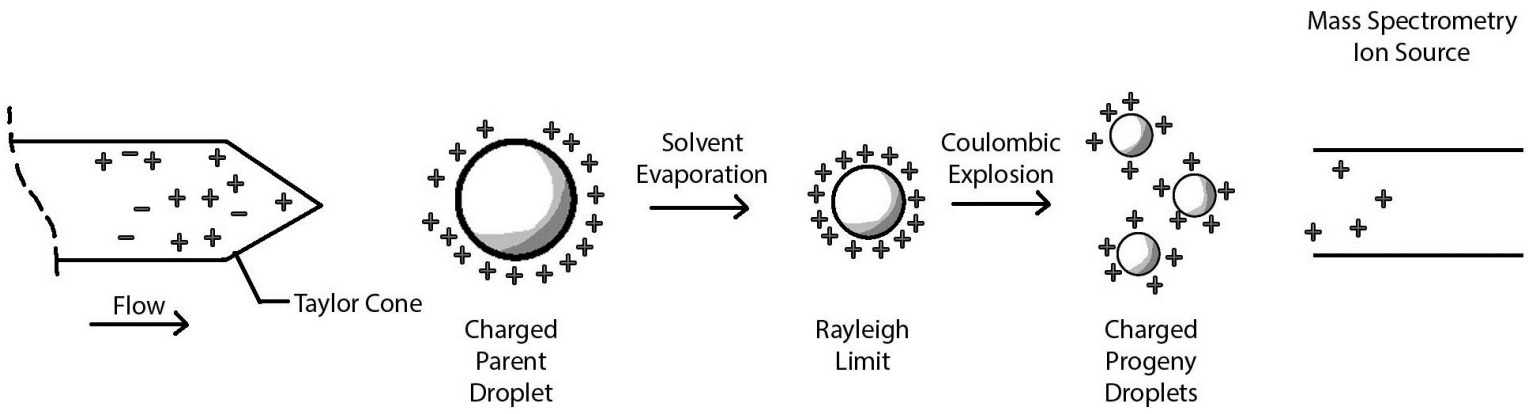


Figure 2.3. Schematic representation of the electrospray ionization process.

One drawback of ESI in comparison with MALDI is lower sensitivity, but this has been greatly improved by the introduction of nanoelectrospary ionization (nESI) [110]. nESI is operated in a more narrow spray capillary at a lower flow rate of nL/min and positioned closer to the entrance of the mass analyzer. Lower flow rate extends the analysis time and thus greatly reduces the amount of sample needed. Also, when the flow rate is dramatically reduced, smaller droplets are formed, leading to increased ionization efficiency, greater sensitivity and better tolerance to salts and other impurities as compared to conventional ESI [111]. Despite these advantages, ESI and MALDI both suffer from ion suppression due to the competition for charge in the ionization process. For example, larger molecular, more hydrophobic ions are more easily to be ionized and thus will be over-represented in the mass spectra [112]. To alleviate this problem, one solution can be coupling chromatographic separations with the mass spectrometer. For example, nESI can directly interface with 2D-LC described above (2D-LC-nESI), which further enhances the sensitivity. Currently, nESI has become the most commonly used ionization technique in peptide and protein analysis with limited amount of material available. For all measurements in this dissertation, nESI coupling 2D-LC was used as the ionization method.

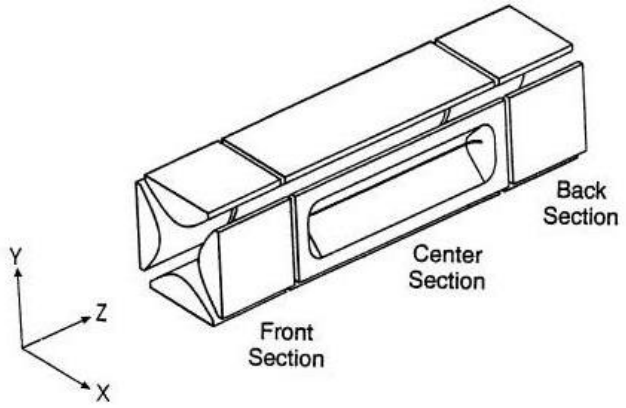
2.4.2 Mass analyzer and detector

After ions have been transferred into the gas phase and introduced into the mass spectrometer, the mass analyzer is responsible for separating ions by their m/z. There are many types of mass analyzers but not a specific one can be universal for all applications. When choosing the appropriate instrument, one should consider the type of information required for the biological problems and evaluate the instrument based on the following figures of merit: *mass resolution*, which determines the ability to differentiate two close mass spectral peaks and typically is given as FWHM (full width of at half maximum) of a peak; *mass accuracy*, which

compares the difference between the measured mass and calculated mass and is calculated as error corresponding to percentage or parts per million (ppm); *mass range*, which gives the range of largest m/z and smallest m/z that a mass analyzer can measure; *dynamic range*, which measures the ratio of datectable signal between the most abundant component and the least abundant component; *sensitivity*, which is the ability to measure small differences in concentration of an analyte and usually defined as the slope of the analytical calibration curve (plot of signal response as a function of concentration); and *scan speed*, which refers to the number of spectra can be recorded per unit time. For example, peptide sequencing are usually achieved in ion trap analyzers that provide high sensitivity and fast scan speed, but have less resolving power, while the characterization of intact proteins requires instruments with high resolution and mass accuracy, in order to determine the charge state and accurate mass of the protein. It is also noted that different types of mass analyzers can be combined together, to obtain desirable features of all linked mass analyzers, for example liner ion trap - Orbitrap hybrid mass spectrometer (discussed below) that delivers high resolution, speed, sensitivity and flexibility.

Despite various types of mass analyzers, ion trap mass analyzers such as linear trapping quadrupole (LTQ) [113] and Orbitrap [114] are most commonly used in proteomic measurements. LTQ is one type of two-dimensional linear ion trap (LIT) consisting of a square array of four hyperbolic metal rods with a longitudinal space between them (Figure 2.4 a). A combination of direct current (DC) and alternating current (AC) radio frequency (RF) voltage can be applied to the metal rods to control the ion motion. Ions are trapped radially in RF electric field and axially in a static electric field using DC voltage. Three different DC voltages are applied to the front, center, and back section of the trap aiming to trap ions axially in the central section of the trap. In addition, RF voltages with constant frequency but variable amplitude are

(a)



(b)

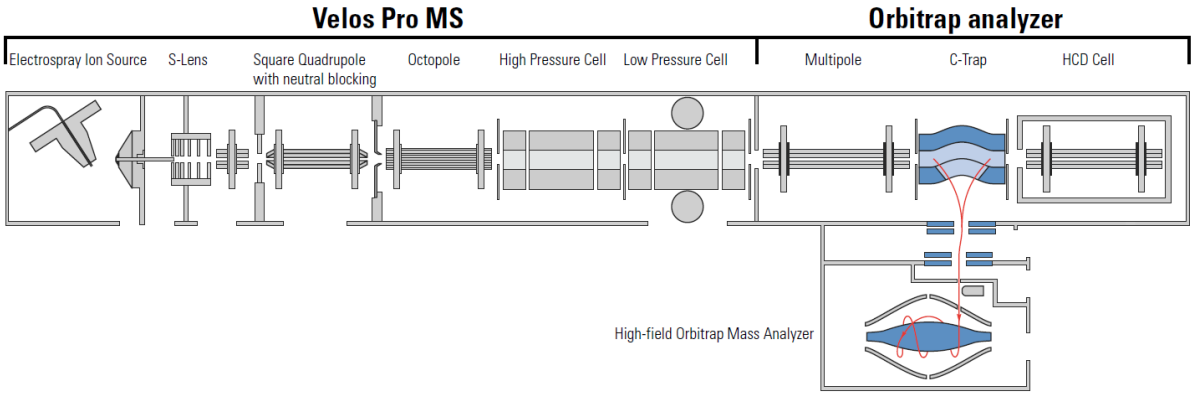


Figure 2.4 Schematic of the Orbitrap Elite Hybrid Mass spectrometer. (a) Image source: Thermo Scientific LTQ Series Hardware Manual, http://www.thermoscientific.com/en/product/ltq-xl-linear-ion-trap-mass-spectrometer.html (b) Image Source: Thermo Scientific LTQ Orbitrap Series Hardware Manual, http://www.thermoscientific.com/en/product/orbitrap-elite-hybrid-ion-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

applied to the rods and voltages are ramped during scan. When RF voltage is low, ions with a broad range of m/z values can be trapped and stabilized in the radial direction. However, as the RF voltage increases, ions of increasing m/z are sequentially unstable and ejected out of the rods and into the detector system where a mass spectrum can be acquired.

Once ions are separated in LTQ, they reach the ion detector. The most commonly used detector for LTQ is the electron multiplier [115], where ions strike the dynode and generate electric current that can be measured. Ions striking the first dynode surface result in an emission of electrons. These electrons then strike the second surface and more secondary electrons are generated. As this process repeats in a series of dynodes, the electrons are multiplied and thus initial signal is amplified by 10^6 or higher.

Recently, the dual cell differential pressure LIT [116] was developed to improve ion trapping and the mass resolution. Usually, a single ion trap uses a compromised pressure between the ion trapping/fragmentation efficiencies (can be increased in higher pressure) and mass resolution/scan speed (can be increased in lower pressure). Therefore, in order to achieve improvements in all events, rather than a single ion trap, the dual pressure cell uses a first high pressure trap to trap, isolate and fragment ions, and also a second low pressure trap to scan ions out for detection. All these improved features in sensitivity, dynamic range and acquisition speed greatly benefit the proteomic analysis of complex samples [117].

Another type mass analyzer is the Orbitrap, which consists of a spindle-shape central electrode and a barrel-shaped outer electrode (Figure 2.4 b) [118]. Without RF field as in LIT analyzers, ions are trapped in a static electric field supplied by a DC voltage on the two electrodes. The stability of ions involves a balance between an electric force (due to ions orbiting

around the central electrode) and a centrifugal force (due to the initial velocity of ions). At the same time, ions also oscillate along the z-axis and m/z values of the ions can be determined by the oscillation frequencies. The ion detection in Orbitrap is an image current from the oscillating ions and a mass spectrum is obtained by Fourier transform of this current. Such measurements in the Orbitrap can achieve very high resolving power (up to 100k) and thus high mass accuracy (<1 ppm). In addition, a new high-field Orbitrap analyzer was developed with a decreasing gap between the inner and outer electrodes. This compact, high field Orbitrap analyzer provides higher frequencies of ion oscillations and hence higher resolving power (up to 240k).

Because of the enormous sample complexity in human infant fecal samples, ultrahigh resolution LTQ-Orbitrap Elite hybrid mass spectrometer (Thermo Scientific, Waltham, MA) was used for peptide sequencing in this dissertation [55]. LTQ-Orbitrap Elite is a combination of two mass analyzers, a LTQ and an Orbitrap mass analyzer (Figure 2.4 c). It features a novel ion transmission pipeline (enhancing transfer efficiency), dual-pressure ion traps (mentioned above), discrete dynode LTQ electron multipliers (yielding six orders of magnitude dynamic range) and a high field Orbitrap mass analyzer (mentioned above). The system enables ultra-high resolution (greater than 240,000 at m/z 400), high mass accuracy (<1 ppm using internal calibration), great dynamic range (>5000), attomole-femtomole sensitivity and fast scanning speed (up to 12.5 MS/MS spectra in the rapid scan mode). All these features together enable increased proteome coverage and a greater depth of measurement in complex samples even with very low sample amounts.

In general, ions from the capillary pass a number of lens and multipoles (called ion optics) and reach the ion trap. Ions are stored, isolated, fragmented and scanned by m/z in the LTQ and ejected into the C-trap (nitrogen-filled RF-only curved ion trap). Ions trapped in the C-trap are

then ejected into the Orbitrap for high resolution and accurate mass determination. This particular hybrid instrument not only retains the high sensitivity and fast scanning speed from the LTQ, but also gains high resolution and mass accuracy from the Orbitrap. This has driven the "high-low" approach, one of the most favorite data acquisition modes in proteomics, where precursor ions and fragment ions are measured in high-resolution Orbitrap and low-resolution LTQ respectively [55]. All the studies in this dissertation employed the high-low approach on the LTQ-Orbitrap Elite.

2.4.3 Data acquisition in mass spectrometry

In a full MS1 spectrum, thousands of ions with different m/z and varying intensities can be scanned and measured. In order to obtain the peptide sequence information, precursor ions in MS1 are selected and subjected to further fragmentation. However, even with high performance instrument, not all precursor ions can be analyzed if one is handled at a time. Therefore, it is critical to operate the instrument in a mode that maximizes the sequencing efficiency of the mass analyzer. So far, two acquisition methods have been described: data-dependent approach (DDA) [119] and data-independent approach (DIA) [120]. Two methods differ on whether the events of peptide fragmentation depend on the precursor intensity information from MS1.

In a DDA scheme, top N (usually top 20) most abundant peptide ions are isolated and fragmented following a MS full scan. On the other hand, in a DIA approach, the entire m/z range is divided into a series of consecutive windows and all peaks in a given window are fragmented simultaneously. Since the DIA sequences the peptides independent of their intensity and covers the entire m/z range, complete sequencing of every available ion can be achieved. But co-fragmentations of all peptides in a range of m/z, especially in complex samples, create a

significant bioinformatics challenge in the data interpretation [121]. Therefore, despite the promising advantages of DIA, DDA is still the most widely used scheme in the metaproteomics of complex communities. Since one major drawback of the DDA is that abundant peptides can be repetitively fragmented and thus limits the measurement depth, dynamic exclusion is necessary to be applied [122]. It works by putting a precursor that has been already fragmented, onto a dynamic exclusion list for a defined duration time (30s – 60s), and so that in the next full scan, lower abundance peptides ions can be visible and have the chance to be triggered for fragmentation. In this dissertation, all mass spectrometric measurements were performed in the DDA mode with specific criteria discussed in each chapter.

2.4.4 Tandem mass spectrometry

As mentioned above, top abundant peptide ions are selected for peptide sequencing, which is obtained by the tandem mass spectrometry (MS/MS) [123]. Precursor ions are subjected to fragmentation and the m/z ratios of resulting fragment ions are measured. By reconstructing the fragment ions, peptide sequence information can be achieved. The most common type of ion fragmentation is the collisional induced dissociation (CID) [124] by which parent ions are accelerated by applying appropriate voltages and collided with a neutral inert gas such as helium, argon or nitrogen. In the collision, the kinetic energy of parent ions is converted into the internal vibrational energy which causes the fragmentation. The mechanism of protonated peptide fragmentation can be explained by the mobile proton model [125, 126]. When peptides are protonated by the electrospray ionization, protons can reside in multiple sites, normally N-terminal and side chains of basic amino acids such as lysine, arginine and histidine. Once the collision occurs, the translational energy allows protons to mobilize onto any one of the peptide amide bonds. This weakens the stability of amide bond, thus resulting in the cleavage of the

peptide bond. Mostly, b and y type ions are observed and a type ions can be generated by the sub-fragmentation of b type ions (Figure 2.5) [127]. In addition, this cleavage can occur at various amid bonds in a peptide backbone, thus generating a series of b and y type ions in a tandem mass spectrum. Theoretically, a peptide sequence can be manually interpreted using MS/MS spectra if b and y type ions can cover the complete peptide. However, part of fragment ions are often missing and massive amount of MS/MS spectra can be collected in a MS measurement, making the manual interpretation obviously impractical. Thus, computational algorithms and tools that allow for reliable and efficient identifications of peptides, have been successfully developed and are discussed below.

2.5 Bioinformatics

2.5.1 Database searching algorithm

The analysis of fragment ion spectra to decode peptide sequences has been significantly facilitated by the development of various database searching algorithms. In general, these algorithms are employed to match the collected experimental fragment ion spectra against theoretical fragment ion spectra that have been predicted for peptide sequences from the genome information (Figure 2.6). The vast array of experimental mass spectra is matched against a predicted protein sequence database with an appropriate search algorithm. This process begins by first identifying a list of candidate peptides which appear to match to the experimental spectra. Then each potential match is scored based on the level of similarity between the experimental and predicted fragmentation spectra. The algorithm selects the candidate with the highest score as the identified peptide. The identified peptides are then filtered to control the false discovery

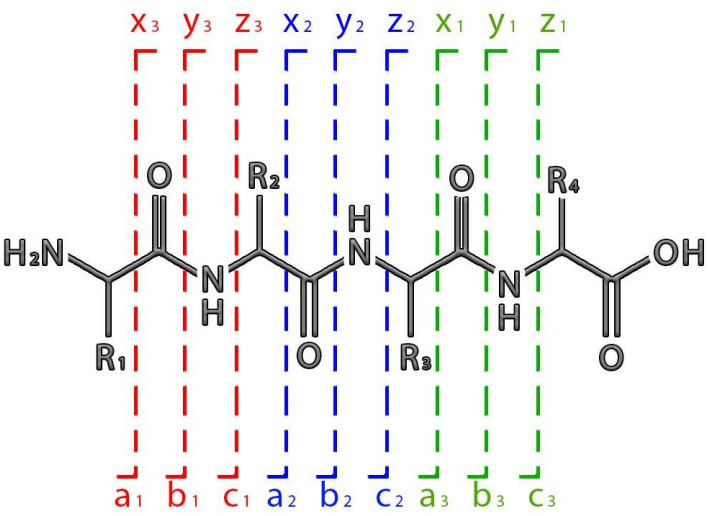


Figure 2.5. Peptide fragmentation ion type. A peptide backbone consisting of four amino acid residues (R_1, R_2, R_3 and R_4) and types of peptide fragmentation ions (a, b, c, x, y and z) generated by CID. Ion types a, b and c are generated when the charge is retained on the N-terminus while ion types x, y and c are generated when the charge is retained on the C- terminus.

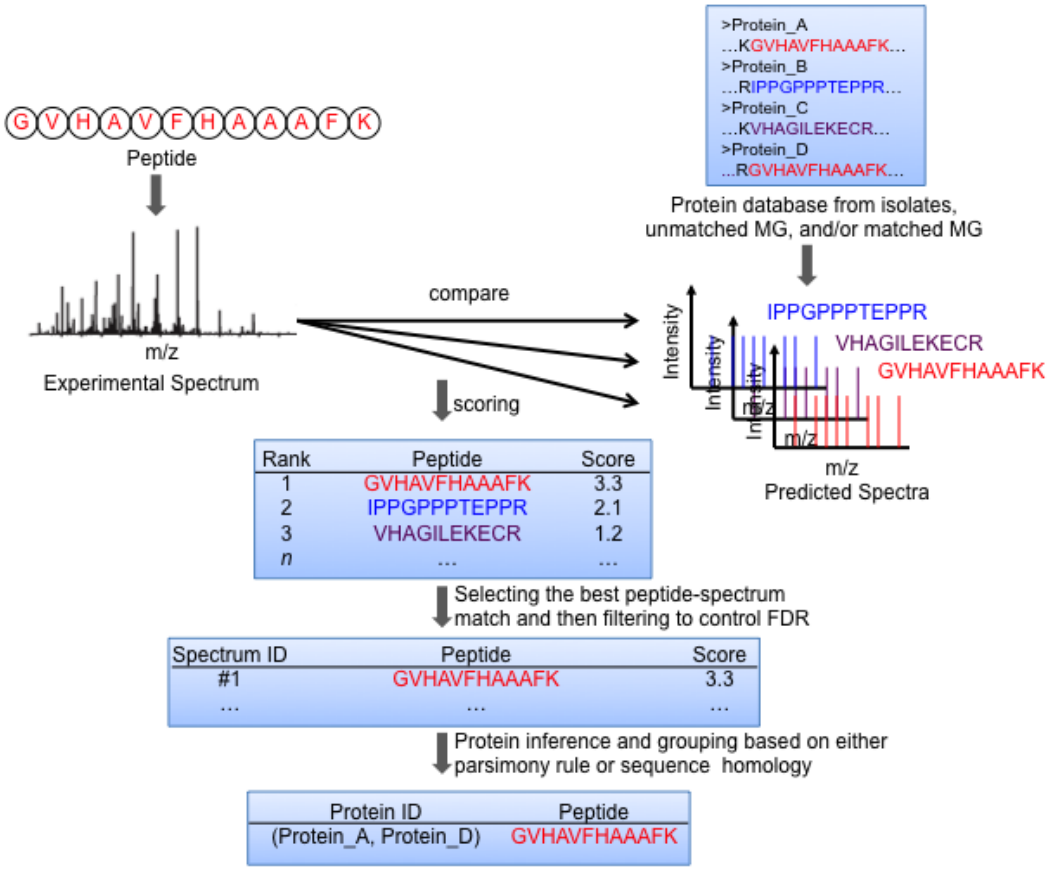


Figure 2.6. Computational metaproteomics workflow. Experimental spectra are compared with a list of candidate peptides generated from metagenome predicted database. Peptide-spectrum matches are determined based on correlation scores with a controlled false discovery rate (FDR). Peptides are assigned to protein groups in order to alleviate ambiguous protein identifications.

rate (FDR) [128]. FDR is computed by searching all MS/MS spectra against the target reference databased and also the reversed, shuffled or randomized decoy database [129]. It is calculated by doubling the number of reverse hits over the total number of hits (forward and reverse), since spurious random hits can occur in either forward or reverse searches. Those peptides that pass the scoring threshold are computationally linked to appropriate proteins using an inference approach. Due to sequence redundancies in the predicted protein sequence database, peptides often cannot be uniquely linked to specific proteins, so they are clustered into protein groups based on either parsimony [130] or sequence homology rules [40].

The most commonly employed database searching algorithms are SEQUEST [45], Mascot [131], MyriMatch [132], OMSSA [133], and X!Tandem [134]. In this dissertation, all raw files were searched with the MyriMatch algorithm. Since hundreds of thousands of fragment ion spectra can easily be acquired per day from a typical MS metagenome databases becomes computationally intensive. MyriMatch is a multi-threaded software that allows parallel computing on clusters, permitting a much faster data processing and database searching. One distinctive feature of MyriMatch is that not only the pattern of a fragment spectrum but also the intensity of each peak is considered during the scoring of matching the observed spectrum to the peptide sequence. Before the matching, data are pre-processed to remove noise and retain ions based on the intensity. Rather than removing a percentage of low intensity ions, MyriMatch retains ions by targeting a percentage (98% by default) of total ion current (TIC) for each spectrum. Retained ions are then matched with candidate peptides using the m/z tolerance (usually 1.5 m/z for average parent mass and 10 ppm for monoisotopic precursor mass). For each paring of observed spectrum and candidate sequence, MyriMatch employs the multivariate hypergeometric (MVH) distribution to compute the probability of random match and score the

matches. In this model, matching a high intensity peak contribute more to the score. Also, an mzFidelity score is employed to evaluate the proximity of observed and expected fragment m/z values. With MVH and mzFidelity sores, candidate peptides are ranked and best match is assigned.

2.5.2 Protein inference

After the database searching and peptide identifications are performed by MyriMatch, IDPicker [135] was used to assemble the proteins. In general, IDPicker extracts sequence, scan, intensity and score information of each spectrum from MyriMatch output files and determines the score threshold for confident peptide identifications according to a user defined PSM FDR cutoff. Peptides are assembled into proteins with additional peptide and protein level filters, such as minimum spectra per peptide, minimum spectra per match, maximum protein groups, maximum distinct peptides, minimum additional peptides and minimal spectra per protein. IDPicker reports a list of all identified proteins/protein groups, their corresponding detected peptides as well as spectra information including scores, mass and sequences.

To confidently identify a protein, usually one unique peptide (exclusively belonging to one proteins) and one additional peptide are required, known as “two-peptide rule” [136]. However, unlike proteins in a single microbial isolate where most peptides can be uniquely mapped to a single protein, a large amount of peptides can be shared among proteins from closely related species in metaproteomics. As a result, proteins share the same set of peptides cannot be differentiated and therefore “two-peptide rule” is difficult to be broadly applied in the metaproteomics investigations.

There are essentially two approaches to group proteins with shared peptides. The first approach uses a parsimony rule with Occam's razor constraints to identify a minimum set of proteins to explain the identified peptides [130]. While this approach has been widely used in proteomics and is able to minimize over-reporting the number of protein identifications, there is no definitive evidence to determine the presence of any particular protein within a group, and proteins in the same group may not necessarily have a similar biological function, which precludes functional analyses. An alternative approach for protein grouping is based on sequence homology [40]. Proteins with certain level of sequence similarity (e.g., 95%) are clustered together. In brief, proteins in the database are sorted by descending length and the longest protein is considered as the first "seed" protein. A pairwise comparison is performed between each protein sequence with the seed. If a protein shares a defined degree of sequence similarity with the seed, the protein is considered as a "hit" and joins the protein group. After the first group is formed, the longest protein within the rest of the proteins becomes the new seed protein and compares with subsequent proteins. By such, peptides that are not previously unique to individual proteins now can be unique to a protein group. Due to such high sequence similarity, all proteins clustered within the same group are likely to have similar biological functions. This grouping scheme allows functional analysis and relative quantification of metaproteomics at the protein group level, which helps simplify the data interpretation. Further details are discussed in Chapter 4.

2.5.3 Protein quantification

In a label-free mass spectrometric measurement, proteins are typically quantified using spectral count, which is the number of tandem mass spectra matching peptides to a particular protein [137]. The more abundant proteins and peptides are, the more times they can be observed.

Therefore, spectral count can be used to compare proteins in the relative abundance between samples. For this dissertation, proteins are quantified with spectral counts and normalized by total number of spectra collected in each measurement.

Instead of using the discrete counts, recent studies have suggested a protein quantification method based on the signal intensity, termed matched ion intensity (MIT) [138]. For every PSM, the intensity is calculated by summing up the intensities of all fragment ions matched to that peptide. One obvious advantage of MIT is that larger analytical dynamic range can be achieved as compared to the limited number of spectral counts acquired in a MS experiment. Comparisons of two methods in the quantification of gut metaproteomics are discussed in Chapter 4.

Typically, a protein's spectral count or MIT is the sum of all its peptides' spectral counts or intensities. However, a number of peptides can be shared by multiple proteins/protein groups in a proteomic study. There are several ways we could have assigned these shared spectral counts or intensities, basically three strategies: 1) discarding shared spectral count or MIT and only considering unique peptides for protein abundance; 2) adding the full spectral counts or MIT of shared peptides to each of shared proteins/protein groups; or 3) distributing a portion of spectral counts or MIT to the shared proteins/protein groups, which is termed as the spectral balancing. In general, spectral balancing distributes shared spectral counts or MIT according to the number of unique peptides in each protein/protein group. Proteins/protein groups with greater proportion of unique peptides will be assigned with a greater share of the shared peptide's spectral counts or intensities. For this dissertation, the balancing strategy has been used for assigning shared spectral counts.

2.5.4 Functional groups assignment

Once thousands of proteins are identified, several approaches /databases can be used to assign proteins into functional groups. These groups summarize/group functionally similar or related proteins and aid in our understanding of functionalities in a broader scale than looking at individual proteins. These databases include the Cluster of Orthologous Genes (COG) [73], the Kyoto Encyclopedia and Genes and Genomes (KEGG) Orthology (KO) [139] and the Gene Ontology (GO) [140]. Among these three databases, COG database contains less levels of functional groups and usually provides broad categories that cannot capture enough details in the functional characterization. KO and GO are more branched and KO can be used to map proteins onto KEGG pathways while GO provides descriptions of proteins in terms of their associated biological processes (BP), cellular components (CC) and molecular functions (MF) independent of organism information. For this dissertation, all three databases have been employed for understanding high-level functions of gut metaproteome.

Overall, the methodologies presented in this chapter demonstrated the ability to conduct metaproteomics measurements and functional characterization in the infant gut microbiome. Further, in the next two chapters, we will discuss the evaluation and optimization of the pipeline and provide both experimental and informatics strategies with careful considerations, in order to address the unique challenges posed by infant gut metaproteomes.

CHAPTER 3

Development of an enhanced metaproteomic approach for deepening the microbiome characterization in the human infant gut

Text and data presented below were adapted from:

Weili Xiong, Richard J. Giannone, Michael Morowitz, Jillian F. Banfield, Robert L. Hettich. Development of an Enhanced Metaproteomic Approach for Deepening the Microbiome Characterization of the Human Infant Gut. *J Proteome Res* 2015, 14 (1), pp 133–141.

Weili Xiong's contributions included: experimental design, sample preparation of fecal samples, mass spectrometry experiments, data analysis, wrote, edited and revised the manuscript.

3.1 Introduction

Trillions of microbes, representing more than one thousand bacterial species-level phylotypes, colonize the adult human intestinal tract [63], generating a complex ecosystem that influences many aspects of human health and diseases [3, 70, 71, 141]. In particular, the gut microbiota plays a crucial role in protecting against pathogen invasion, processing nutrients, balancing energy, and regulating host immune responses [64, 142-144]. Microbial colonization in the gut is initiated immediately after birth and undergoes remarkable changes in composition and function over the first 2-3 years of life until a resilient, stable, adult-like microbiota is established [88, 145]. This early-life microbiota development requires an intricate balance between microbial colonization and the corresponding responses of human host intestinal environment [146]. Distortions in the establishment of normal gut microbiota and commensal microbes increase the risk of inflammatory diseases, such as necrotizing enterocolitis (NEC), via disruption of the mucosal barrier and subsequent impairment of immune system [93, 147]. Despite a lower complexity than adult microbiota, the gut microbial communities in infants are

highly variable between individuals and may be influenced by a number of external factors, such as delivery mode, diet, and antibiotic use [148-150].

Two recent studies based on metagenomic data from fecal samples from two healthy premature infants have detailed both the microbial species/strains present as well as their relative abundances during the first month of life [21, 151]. Both genomic analyses revealed shifts in bacterial populations, identifying discreet compositional phases during infant gut colonization. Interestingly, the gut microbiota structure was drastically different between the two infants. Specifically, *Citrobacter*, *Serratia* and *Enterococcus* species were dominant microbial members in one infant, while *Enterococcus faecalis*, *Propionibacterium carrol* and four different *Staphylococcus epidermidies* strains were dominant in the second infant. Although metagenomic information alone highlights community variation and provides a list of all possible gene proteins, metaproteomics provides insight into real-time functional signatures which help detail metabolic activity as well as host/microbe interaction during gut colonization.

Shotgun metaproteomics via nanospay two-dimensional liquid chromatography coupled with tandem mass spectrometry (nano 2D LC-MS/MS) provides a powerful platform for the large-scale characterization of metaproteomes [152]. Notwithstanding, the interrogation of the infant gut proteome is impeded by several major factors: (1) large diversity in gut microbial composition between individuals; (2) wide dynamic range of protein abundances; (3) insufficient genome information/assembly. The advancement of high performance mass spectrometry has greatly increased proteome coverage, including quantification [27], in complex samples and the use of matched metagenomes enables more confident and accurate protein identification; coupled together, remarkable success has been demonstrated for a number of diverse and complex fecal samples [10, 27, 72].

Our initial investigations to test metaproteomics for the characterization of premature infant gut microbiomes simultaneously monitored both microbial and human proteins over the first few weeks after birth. Intriguingly, the data showed dramatic variations in human to microbial protein ratios among various infants [99]. In some cases, overwhelmingly abundant human host proteins greatly overshadowed the microbial microbiome, resulting in reduced depth of measurement into the microbial proteome. These highly abundant human proteins precluded the efficient mass spectrometric detection of medium- to low-abundance microbial peptides. In particular, microbial peptides co-eluting near dominant human peptides experience ion suppression and are thus more difficult to measure. Furthermore, the reduced detection and subsequent identification of those peptides not only lowers the number of protein identified, but also the number of unique peptides. The lack of unique microbial peptides detected makes protein inference difficult, especially when considering closely related strains/species. These two issues thereby inhibit the complete characterization of diverse gut microbial communities and must be addressed. Although microbial cell enrichment in fecal samples by differential centrifugation has been previously reported [10, 72], infant fecal samples are limited by the amount of raw material available, thus precluding the differential centrifugation approach and necessitating the development of alternative enrichment strategies.

The objective of this work was to develop and demonstrate the feasibility of an enhanced metaproteomic sample preparation strategy that provides more comprehensive interrogation into the infant gut microbiome by incorporating a double filtering (DF) separation step that selectively depletes human cells and proteins while enriching microbial biomass in the fecal sample. With the inclusion of this DF, we observed a significant increase in the number of microbial protein identifications at the calculated expense of human proteins which affords a

more extensive characterization of microbial functionality in the infant gut without too much interference from high abundance human proteins.

3.2 Materials and methods

Sample Collection. Fecal samples from two premature infants (#UN1 and #CA1) were supplied by Dr. Michael Morowitz and stored at -80℃. Samples were obtained under an IRB agreement protocol, and were de-identified before sending to ORNL. A small portion of sample was excised and thawed prior to cell lysis and protein extraction.

Protein Extraction and Enzymatic Digestion. For each infant, approximately 0.5 g of raw fecal material was processed by two methods: a direct method and an indirect double filtering (DF) method. For the direct method, fecal material was boiled for 5 min in 1ml lysis buffer containing 100 mM Tris-HCl, pH 8.0, 4% w/v SDS (sodium dodecyl sulfate) and 10 mM dithiothreitol (DTT). The suspension was vortexed and sonicated with a Branson ultrasonic cell disruptor (20% amplitude for 2 min, 10 s pulse with 10 s rest). Crude protein extract was pre-cleared via centrifugation at 21,000g for 10 min, and quantified by the BCA assay (Pierce Biotechnology, Waltham, MA). An aliquot pertaining to ~1 mg of protein was collected and precipitated by 20% trichloroacetic acid (TCA) at -80℃ overnight. Protein pellets were washed with ice-cold acetone, resuspended in 8 M urea, 100 mM Tris-HCl, pH 8.0 and sonically disrupted in order to fully solubilize the protein pellet (20% amplitude for 5 min, 10 s pulse with 10 s rest). Denatured proteins were reduced with 5 mM DTT for 30 minutes. To block disulfide bond reformation, 20 mM iodoacetamide (IAA) was added to each sample; the reaction occuring in the dark at room temperature for 15 minutes. Samples were diluted to 4 M urea in 100 mM Tris-HCl, pH 8.0 and digested with one aliquot of sequencing grade trypsin (Promega, Madison, WI; 1:100 (w/w))

overnight at room temperature. Following digestion, samples were diluted to 2 M urea for a second digestion that lasted 4 hrs. Digested samples were then adjusted to 200 mM NaCl, 0.1% formic acid (FA) and filtered through a 10 kDa cutoff spin column filter (Vivaspin 2, GE Health, Pittsburgh, PA) to remove under-digested proteins. Peptides were quantified by the BCA assay and stored at -80℃ until use. For the indirect DF method, a differential filtering method was designed and optimized based on the knowledge that bacterial cells (0.2-2 µm in diameter) are typically much smaller than human cells (10-100 µm in diameter). In addition, bacterial cells usually contain chemically complex cell walls and therefore are more resistant to mechanical shear force, while human cells are much more susceptible and easier to lyse. Prior to the detergent-based cellular lysis and protein extraction described above, two different-size vacuum filter units were employed to physically separate human from microbial cells (Figure 3.1). Fecal samples (0.5 g) were suspended in 10 mL ice-cold Tris-based saline (TBS) buffer and passed first through a 20 µm vacuum filter unit to remove larger fibrous material and intact human cells. The filtrate (including microbial cells, small human cells, secreted human and microbial proteins and proteins from lysed cells) was homogenized using VDI 12 homogenizer (VWR, USA) at speed 6 (30,000 rpm, 30 s-resting 30 s-30 s) to disrupt remaining intact human cells, followed by centrifugation (4000 x g, 10 min) to pellet intact bacterial cells. The collected pellet was resuspended in 10 ml cold TBS and passed through a second 0.22µm vacuum filter unit. This permitted human proteins to be washed through while microbial cells were captured on the filter. Captured cells were washed twice with cold TBS to remove attached human proteins and lysed by the SDS-based approach as described above. The entire filtering process was performed on ice and completed within 20 min to minimize proteomic perturbations during manipulation.

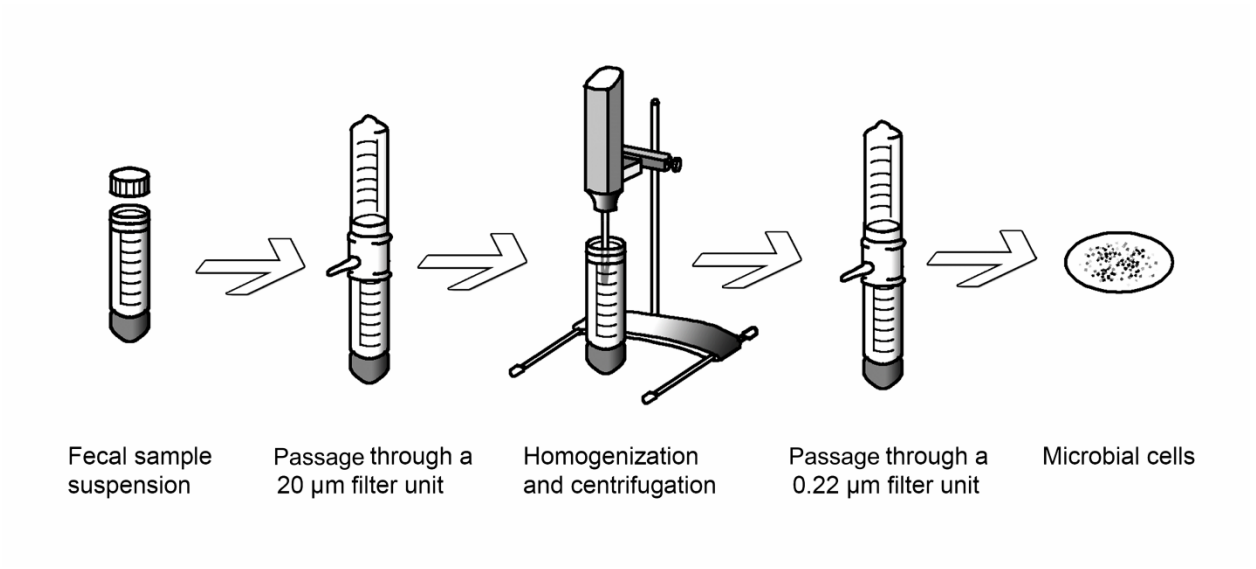


Figure 3.1. Workflow of the indirect double filtering (DF) method. Fecal raw material is suspended in cold PBS and passed through a 20µm filter to remove large particles and intact human cells. The filtrate is homogenized and centrifuged to obtain a microbial cell pellet. The pellet is resuspended and passed through a 0.22µm filter to collect microbial cells on the filter membrane. Collected cells are washed twice and subjected to SDS based cell lysis and protein purification method.

Nano 2D LC-MS/MS Measurement. Proteolytic peptide samples were analyzed via an online nano 2D LC-MS/MS system interfaced with a hybrid LTQ-Orbitrap-Elite MS (ThermoFisher Scientific). A 30 µg aliquot of peptides was loaded onto a biphasic silica back-column and analyzed by a three-step MudPIT as described in Chapter 2. The LTQ-Orbitrap-Elite was operated in a data-dependent mode with each full scan (1 microscan) collected in the Orbitrap mass analyzer at 30,000 resolution, followed by collision-induced dissociation (35% energy) of the top 20 most abundant parent ions (1 microscan). Dynamic exclusion was enabled with a mass exclusion width of 0.2 m/z and exclusion duration of 60 s.

Protein Database Construction and Searching. Due to the lack of a sample-specific metagenome-derived protein database for infant #UN1, a pseudo-metagenome was created by concatenating 21 microbial isolate reference genomes (acquired from JGI; representative organisms were chosen based on 16S rRNA information from another infant sample that was quite similar to this particular infant (Table 3.1)), human protein sequences (NCBI RefSeq_2011) and common contaminants (eg. keratin and trypsin) into a single protein database (105,671 sequences). Though not metagenomically matched, this database provides complete genome sequences of presumably present microbial species, as indicated by 16S rRNA analysis. Conversely, a matched metagenome-derived protein database (60,073 sequences) for infant #CA1 was generated by combining metagenomic sequences [21] collected on postnatal days 10, 16, 18 and 21 from the infant (provided by Dr. Jillian Banfield), along with human protein sequences and common contaminants. A decoy database consisting of reverse protein sequences was appended to the target database to calculate false discovery rates (FDR). All MS/MS spectra were searched with the Myrimatch v2.1 algorithm [132] against the appropriate database with the following configuration parameters: fully tryptic peptides with any number of miscleavages, an

Table 3.1. 21 microbial isolate reference genome database

Genome	
Acinetobacter junii SH205	*Eubacterium rectale* ATCC 33656
Bifidobacterium adolescentis ATCC 15703	*Fusobacterium* sp. 1_1_41FAA
Bacteroides fragilis NCTC 9343	*Klebsiella* sp. 1_1_55
Bifidobacterium longum infantis ATCC 15697	*Leuconostoc mesenteroides cremoris* ATCC 19254
Campylobacter concisus 13826	*Lactobacillus reuteri* 100-23
Citrobacter koseri ATCC BAA-895	*Pseudomonas aeruginosa* PAO1
Citrobacter sp. 30_2	*Staphylococcus aureus* 04-02981
Clostridium sporogenes ATCC 15579	*Serratia odorifera* 4Rx13
Enterobacter cancerogenus ATCC 35316	*Streptococcus* sp. 2_1_36FAA
Escherichia coli K12 DH10B	*Weissella paramesenteroides* ATCC 33313.
Enterococcus faecalis TX0104	

average precursor mass tolerance of 1.5 m/z, a mono precursor mass tolerance of 10 ppm, a fragment configuration parameters: fully tryptic peptides with any number of miscleavages, an average precursor mass tolerance of 1.5 m/z, a mono precursor mass tolerance of 10 ppm, a fragment mass tolerance of 0.5 m/z, a static cysteine modification (+57.02 Da), an N-terminal dynamic carbamylation modification (+43.00 Da) and a dynamic oxidation modification (+15.99). Peptides identifications were filtered with IDPicker v 3.0 [135] to < 1% peptide FDR (at the peptide level: maximum Q value < 2%, minimum one spectra per peptide and minimum one spectra per match; at the protein level: minimum two distinct peptides, minimum zero additional peptide and minimum two spectra per protein).

Protein Inference and Semi-quantification. Due to the high degree of sequence homology and redundancy in the human RefSeq database, as well as homologous proteins from different microbial species, peptides that map to multiple proteins increase the ambiguity within protein identifications and quantification. To avoid under- and over-counting protein identifications, the pseudo predicted protein database for infant #UN1 was clustered based on 90% amino acid sequence similarity using USEARCH v 5.0 software [153], as described previously [40]. This was done post-database search and is essentially a reassessment of uniqueness based on the very conservative sequence identity of 90-100%. Considering the lower level of homologous protein overlap within the constructed matched metagenome for infant #CA1, microbial proteins in the database were clustered into a group if they share 100% amino acid sequence identity (which would otherwise prevent proteomic identification based on the commonly used unique peptide criterion), and human proteins were clustered based on 90% amino acid sequence similarity. Spectral counts were balanced between shared proteins, and normalized by total numbers of

collected MS/MS of this run as previously described [40]. Data were plotted with OriginPro 8.1 graphing software (OriginLab Corporation, Northampton, MA).

Clusters of Orthologous Groups (COGs) Assignment. Protein sequences were searched against the COG database from NCBI using rpsblast [154] and the top hit was assigned with an e-value threshold of 0.00001. Assigned COGs were grouped into COG functional categories to predict functions in the gut microbiome. Abundance of each category was determined by summing normalized spectral counts of all COGs in the category.

3.3 An enhanced strategy for infant fecal proteomics to improve the overall depth of proteome measurement

To determine the feasibility and robustness of our enhanced approach for diverse samples, we selected fecal samples from two healthy premature infants (#UN1 and #CA1) that differed by human protein composition (as evident from the direct measurement method) and matched metagenome availability (See database construction in Methods and Materials for details in this study). To assess the sample preparation reproducibility and range of this method, we initially conducted replicate sample preparation processing (complete protocol) and MS measurements on an independent third preterm infant fecal sample. The overall correlation was high (R^2 = 0.85; Figure 3.2), verifying that the sample preparation approach was robust and reproducible. Technical replicates were performed for each fecal sample and were found to be highly reproducible (R^2 > 0.95) (Figure 3.3), which attests to the precision of the MS measurements as well as the enhanced search approach relative to previous methods [155].

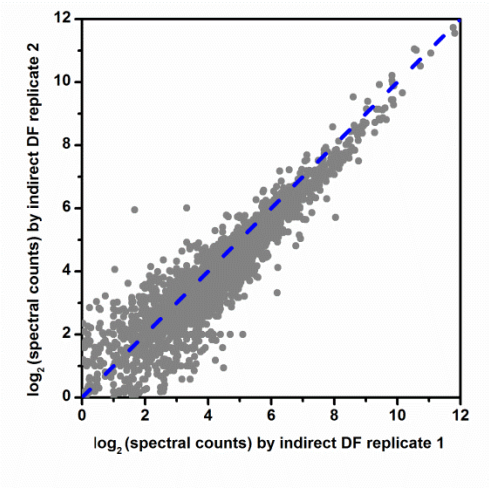


Figure 3.2. Reproducibility of methodological (sample preparation) replicates. An infant fecal sample was processed twice and measured in duplicate across two 24 h MudPIT runs. A scatterplot was generated using $\log_2$ spectral counts of protein groups in duplicate runs (R^2 = 0.85). Dashed line indicates a perfect 1:1 correlation.

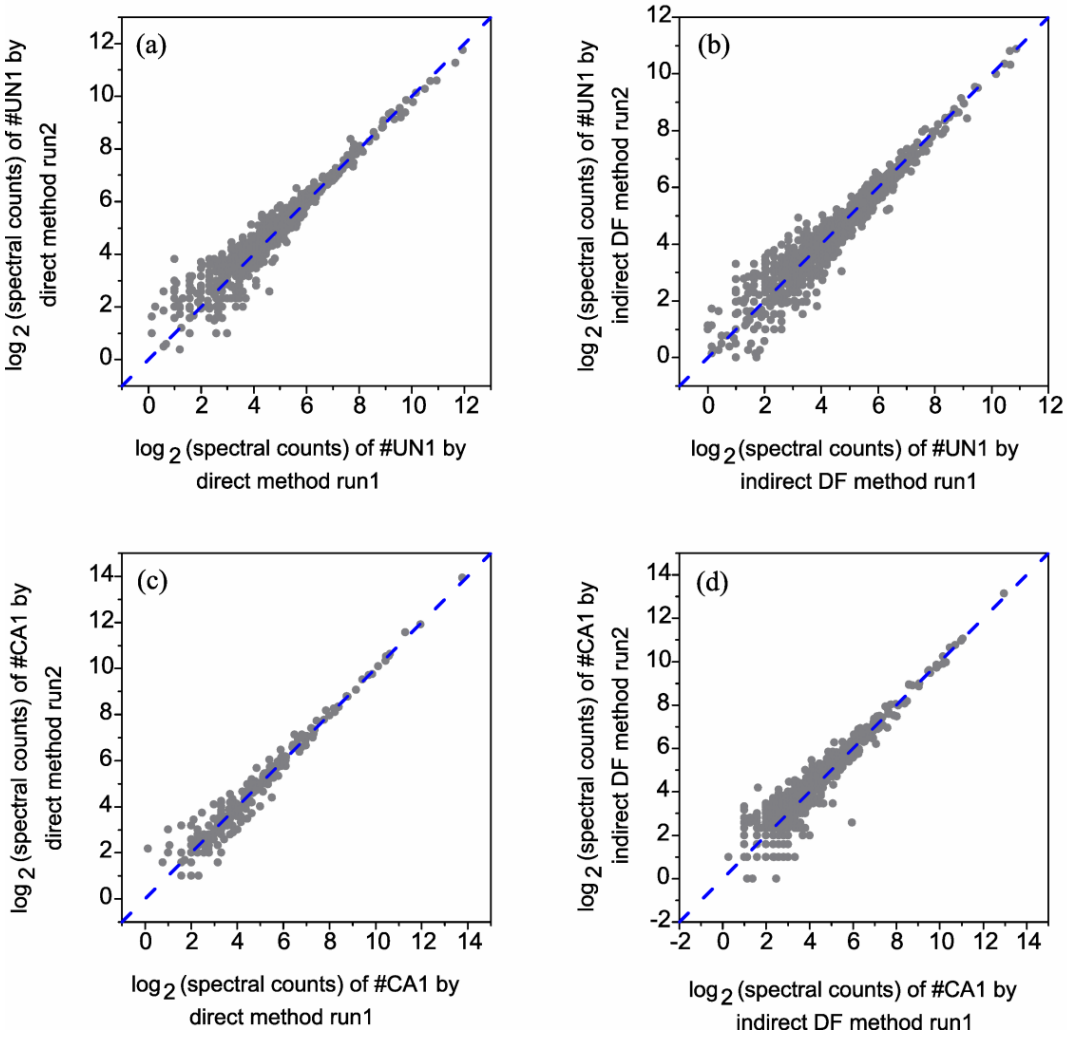


Figure 3.3. Protein group quantification reproducibility. Scatter plots are constructed using $\log_2$ spectral counts of protein groups measured in duplicate runs of infant #UN1 by the direct (a, R^2 = 0.95) and the indirect method (b, R^2 = 0.95), and of infant #CA1 by the direct (c, R^2 = 0.98) and the indirect DF method (d, R^2 = 0.94). Dash line indicates the perfect correlation.

The results of protein group identifications and spectra assignments for premature infants #UN1 and #CA1 are summarized in Tables 3.2. The task of mapping peptides to proteins for metaproteomic investigation is challenging in that peptides can be shared by multiple proteins in a reference database, which result from homologous proteins among closely related organisms and/or sequence redundancies within large databases. These shared peptides are common in infant gut databases and lead to ambiguous protein assembly, especially relative to more routine microbial isolate measurements. Previous studies proposed an effective way to deal with shared peptides by clustering proteins into groups using an algorithm based on sequence homology [40]. This protein grouping approach affords distinct advantages in data interpretation, since proteins sharing high similarity are likely to exhibit similar biological functions, allowing for a more robust interrogation of functional activities in complex communities such as the infant gut. Based on this approach, a total of 807 or 1264 (for infant #UN1) and 342 or 1012 (for infant #CA1) protein groups (non-redundant protein groups from duplicates) were generated using the direct and the indirect DF method, respectively (FDR rate < 1% at the peptide level, Table 3.2).

Having established the criteria for protein identifications, we sought to assess the overall depth of proteome coverage by this enhanced strategy. Indeed, our approach facilitated a noticeable increase in the number of overall spectra assignments and greater than 50% peptide and protein group identifications for both infant fecal microbiomes compared with the measurement using the direct approach (Table 3.2). Notably, greater improvements were observed for infant #CA1 relative to infant #UN1. This may be due to a more representative database constructed from matched metagenome of infant #CA1, providing a more complete protein inventory and thus more confident protein identifications, but more likely, the increase resulted from greater removal of human proteins, since #CA1 contained a higher abundance of

Table 3.2. Overview of proteomic results from two fecal microbiomes measured by the direct and the indirect DF method

	Infant #UN1				Infant #CA1			
Run	Direct Run1	Direct Run2	Indirect DF Run1	Indirect DF Run2	Direct Run1	Direct Run2	Indirect DF Run1	Indirect DF Run2
Spectral counts	40495	39432	42485	42544	40068	41492	45484	47221
Peptide counts	4156	4672	6688	6799	1968	1905	4475	4238
Protein counts	**2215**	**2465**	**3534**	**3697**	**691**	**542**	**1200**	**1182**
Protein group counts	**655**	**734**	**1076**	**1122**	**304**	**289**	**855**	**854**

human proteins (Table 3.3). Since the measurement of complex protein mixtures is often biased towards high abundance proteins, which generate an excess of proteolytic peptides that often occupy long periods of chromatographic space and limit the dynamic range of the measurement by precluding sampling (and identification) of co-eluting, lower abundance microbial proteins, a larger initial complement of human cells in #CA1 would have been affected more by the indirect DF method.

Once abundant proteins were removed with DF, we were able to dramatically increase identifications for those previously unmeasured microbial proteins. In fact, we observed differences in the protein abundance profiles between the two methods for both infants, with specific increases in the number low abundant proteins groups (less than 100 spectral counts; Figure 3.4). As a result, we achieved a deeper proteome characterization, primarily in the microbial membership.

3.4 Microbial protein group identifications are enriched by depletion of abundant human proteins

The success in achieving accurate protein identifications and deep proteome coverage in a complex community relies on the quality of predicted protein sequence database that is constructed from metagenomic data. Compared to the analysis of a single cell type / microbial isolate, a larger portion of high quality spectra in metaproteomic study remain unassigned due to the incompleteness of the proteomic database. To quantify this, we employed a spectral quality assessment tool, ScanRanker [156], to assign scores for all the collected spectra to evaluate the quality of the database. Using ScanRanker scores, a distribution of total collected spectra including unassigned, assigned human, and assigned microbial spectra was plotted for each

Table 3.3. Collected and assigned mass spectra results

	Infant #UN1		Infant #CA1	
Run	Direct	Indirect DF	Direct	Indirect DF
Collected	286530	286671	279174	278277
Assigned	80322	86001	81937	93728
Human	47165	6221	78260	48550
Microbial	32730	78675	3277	44021

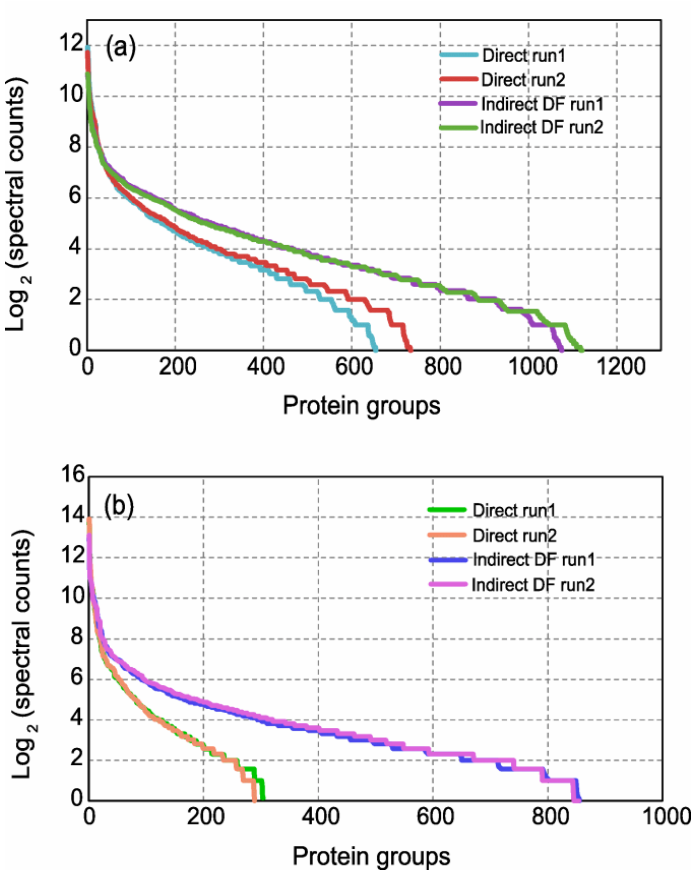


Figure 3.4. Rank-abundance plots of protein groups. Identified proteins are clustered into protein groups and their spectral counts are balanced and normalized according to the approach specified in Materials and Methods. Protein groups of (a) infant #UN1 and (b) infant #CA1 are ranked and plotted based on spectral counts. The indirect DF method facilitates an increasing number of identified protein groups. The two methods possess the same slope for top ranked groups but diverge at the group with fewer than 100 spectral counts. The indirect DF method has shallower slope and thus provides more low abundance protein group identifications.

infant, as measured by both methods (Figure 3.5). For each distribution, a total of ~280,000 spectra were represented, as measured in duplicate runs, and ~15% of those with scores below -0.6 were recognized as peptide identifications, implying lower quality spectra reside at the lower end of collected mass spectra can be assigned to peptides for an organism with a completely sequenced genome (without accounting from PTMs, sequence variants, and other unknown contaminants, of course). However, due to the increased complexity of these samples, as well as the fact that the metagenomic databases used here are incomplete, approximately 27% and 29% of total collected spectra were assigned for infant #UN1 and infant #CA1, respectively, using the direct approach (Figure 3.5 (a) and (b)), while a slightly higher percentage of 30% and 33% were achieved via the indirect DF approach (Figure 3c and 3d). Despite similar spectral assignment efficiency, one readily observable difference between the two infants is the ratio of human versus microbial assigned spectra. For infant #UN1, the microbial peptide spectral matches (PSMs) accounted for 40% of total assigned spectra with the direct method (Figure 3.5 (a)) while for infant #CA1, this value was much lower (~4%). Consequently, this suppression of microbe-derived PSMs by the presence of abundant human proteins severely impedes the interrogation of microbial functional activities in the gut, especially when considering semi-quantitation (Figure 3.5 (b)). Therefore, it is a challenge to investigate the inter-individual variability through the direct approach given the relative dearth of microbial PSMs. However, compared to the direct method, our DF strategy substantially increased microbial PSM proportions within total assigned spectra, from 40% to 93% for infant #UN1, and from 4% to 48% for infant #CA1 (Figure 3.5).

Although it depends on the experimental question being asked, the ultimate goal here was to remove abundant human proteins and peptides in order to enhance microbial protein identification depth - a process that would undoubtedly facilitate functional characterization at

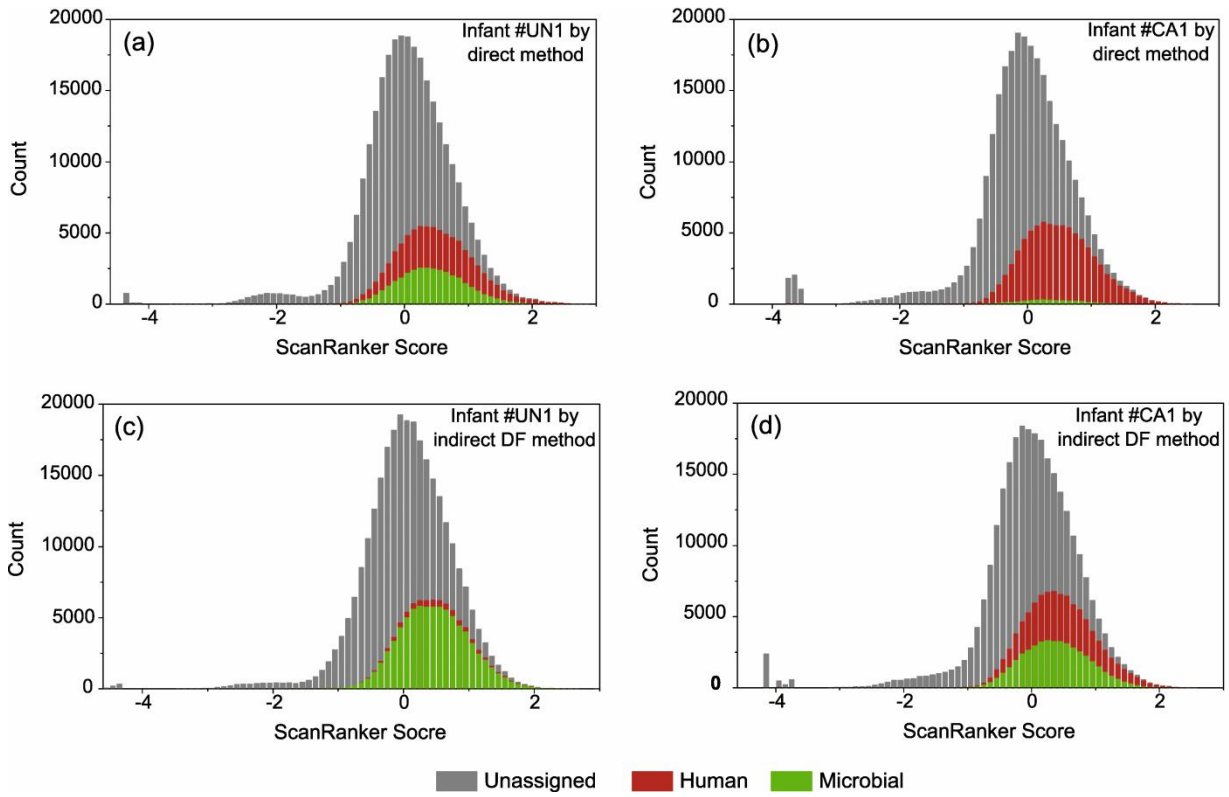


Figure 3.5. Distributions of ScanRanker scores for collected mass spectra. ScanRanker scores are used to assess spectral quality for all collected mass spectra. Stack histograms are generated for ScanRanker scores of (a) infant #UN1 measured by the direct method, (b) infant #CA1 by the direct method, (c) infant #UN1 by the indirect DF method and (d) infant #CA1 by the indirect DF method. The color encodes ScanRanker score distributions of unassigned (gray), assigned human (red) and assigned microbial (green) mass spectra in replicates. The indirect DF method enriches microbial mass spectra assignment as decreasing human mass spectra assignment.

the microbial-level. Using infant #UN1 as a test case, we found that 593 protein groups overlap between the two methods, with 214 protein groups uniquely identified by the direct method and 671 protein groups uniquely identified by the indirect DF method (Figure 3.6 (a)). We next evaluated Venn groupings at the PSM-level, specifically on the PSM partitioning between organisms (i.e. human-derived PSMs vs. microbial-derived PSMs). In this case, spectral counts were averaged between replicates. Considering protein groups specific to either method, more unique human protein groups (155 out of 214) were found by the direct method, while more unique microbial protein groups (649 out of 671) were detected in the indirect DF method. Of those commonly identified between the two methods, DF led to the identification of substantially more microbial PSMs. Collectively, our enhanced approach facilitated a 2-fold increase in the number of identified microbial protein groups and a 2.4-fold increase in the microbial spectral counts. Taken together, these observations indicate that improving the overall protein/peptide identification rate/sampling depth of the microbial complement of a fecal sample was attributed to the significant depletion of human proteins. Similar analyses of infant #CA1 further validated the enrichment of microbial protein groups (Figure 3.7 (a)).

We also examined the identification reproducibility and quantification consistency of the two methods regarding microbial protein groups. Over 90% of microbial protein groups identified by the direct method were also identified by the indirect DF method (Figure 3.7 (b) and (c)). There was a high rank correlation (r_s = 0.76 for infant #UN1; r_s = 0.77 for infant #CA1) of log2 spectral counts between the two methods, with the correlation offset likely due to microbial protein enrichment (Figure 3.6 (b) and (c)). These results demonstrate that increases in microbial PSM rate through the DF method does not bias the sample, but instead provides more confidence for microbial protein quantifications through selective enrichment.

Figure 3.6. Comparison of protein group identification and quantification results by two methods. The Venn diagram (a) shows unique and overlapped protein group identifications of infant #UN1 between the direct and the indirect DF method. Bar charts indicated human (red) versus microbial (green) protein group counts and spectral counts in the part of uniquely identified by the direct method (left), commonly identified (bottom) and uniquely identified by the indirect DF method (right). Scatter plots are constructed using $\log_2$ spectral counts of microbial protein groups measured by two methods for infant #UN1 (b, $r_s = 0.76$) and infant #CA1 (c, $r_s = 0.77$). Solid line indicates the perfect correlation and the dash line indicates the offset owing to microbial protein enrichment. Microbial protein groups are enriched with a relatively high ranked correlation.

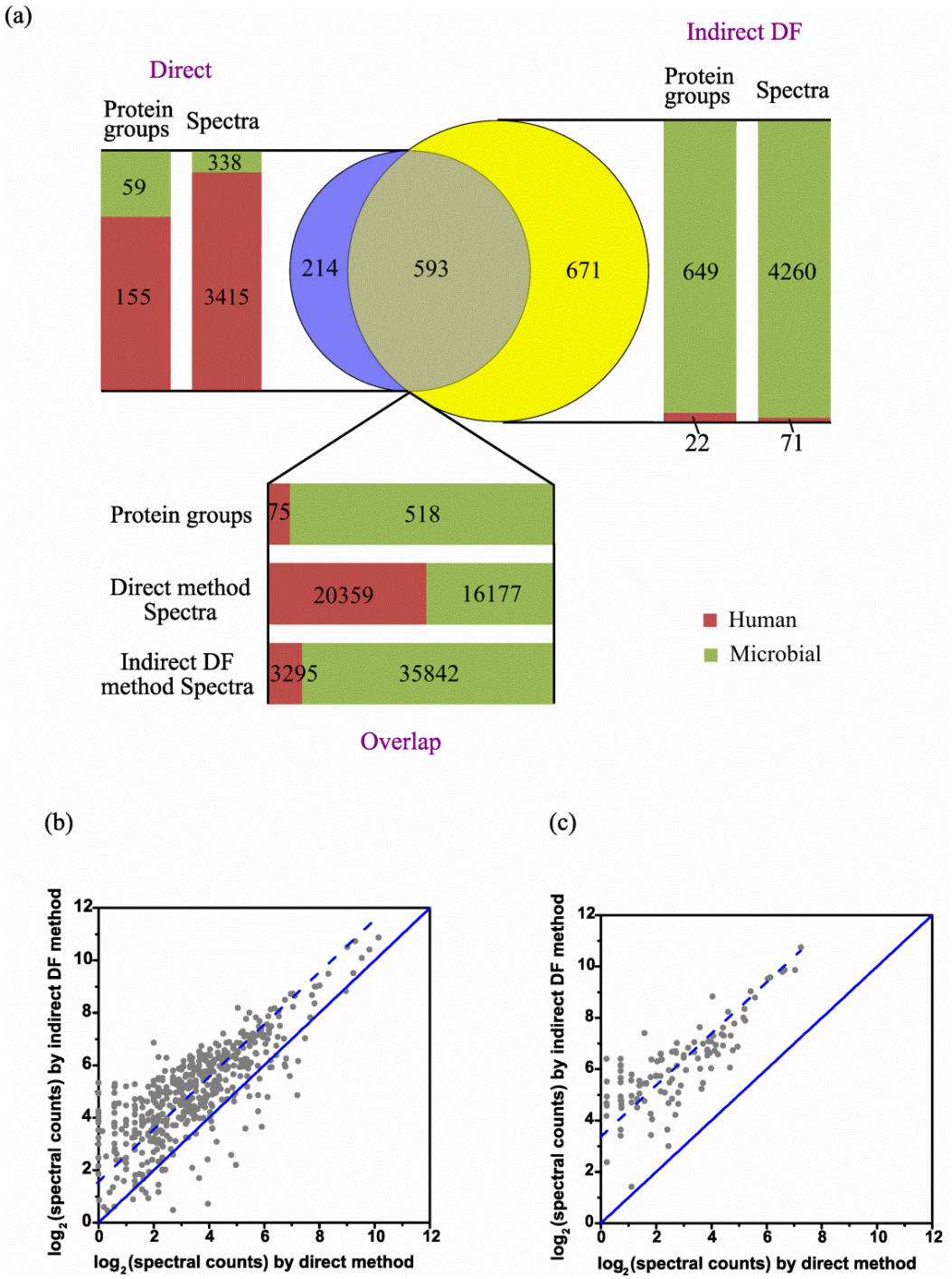
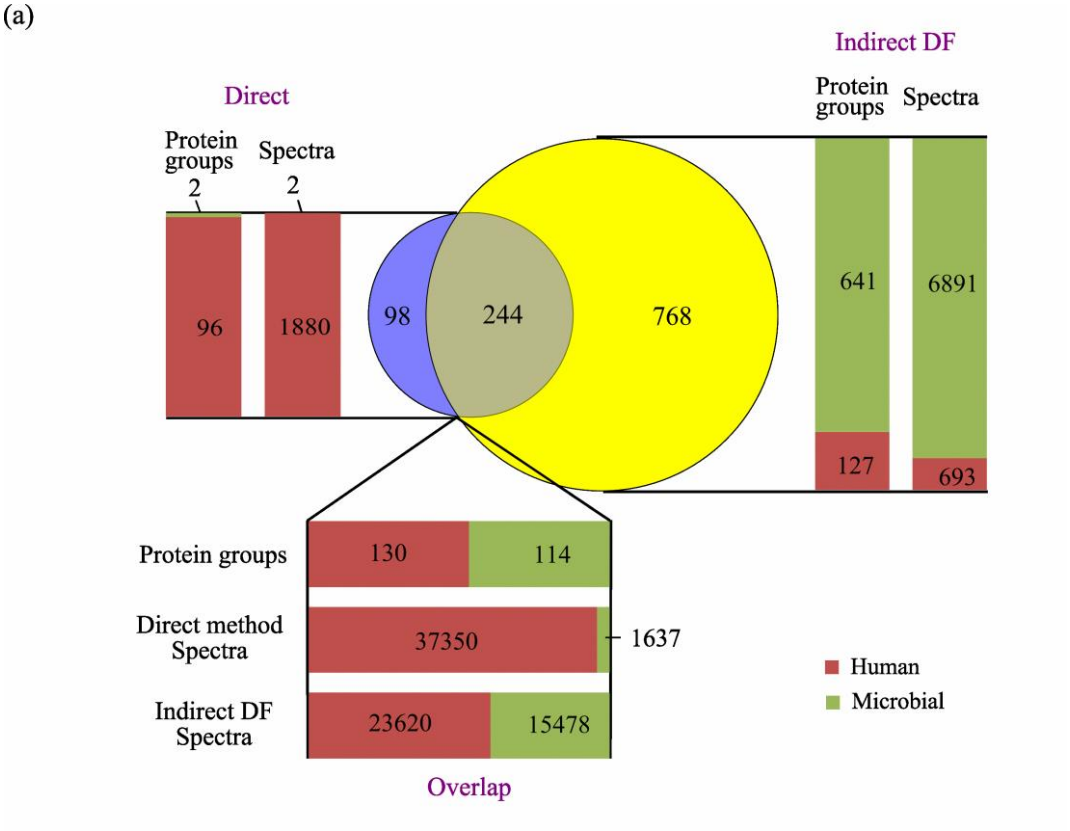
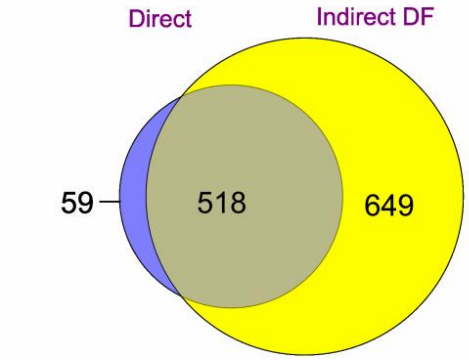

(a)
Direct
Protein groups
Spectra
59
155
338
3415
214
593
671
Indirect DF
Protein groups
Spectra
649
4260
22
71
Protein groups
75
518
Direct method Spectra
20359
16177
Indirect DF method Spectra
3295
35842
Human
Microbial
Overlap
(b)
log2(spectral counts) by indirect DF method
log2(spectral counts) by direct method
(c)
log2(spectral counts) by indirect DF method
log2(spectral counts) by direct method

Figure 3.7. Microbial protein group identification. The Venn diagram (a) shows unique and overlapped protein group identifications of infant #CA1 between the direct and the indirect DF method. Bar charts indicate human (red) versus microbial (green) protein group counts and spectral counts in the part of uniquely identified by the direct method (left), commonly identified (bottom) and uniquely identified by the indirect DF method (right). The Venn diagrams (b, infant #UN1) and (c, infant #CA1) show the overlap of microbial protein group identifications of two infants between two methods. Over 90% of microbial protein groups in the direct method are identified by the indirect DF method.

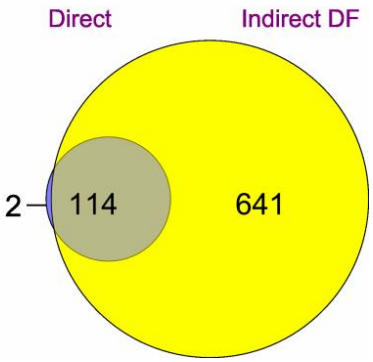
(a)


Direct
Protein groups
Spectra
2
2
96
1880
98
244
768
Indirect DF
Protein groups
Spectra
641
6891
127
693
Protein groups
130
114
Direct method Spectra
37350
1637
Indirect DF Spectra
23620
15478
Human
Microbial
Overlap

(b)


Direct
Indirect DF
59
518
649

(c)


Direct
Indirect DF
2
114
641

3.5 Enriched microbial protein identifications facilitate more comprehensive information for microbial functional categorization

To further elucidate the advantages of the DF sample preparation approach, we tabulated and analyzed the COG functions for the two infant gut microbiomes measured by these two methods. We clearly recognize that COG families are relatively broad and characterize the functionality at a lower resolution than a more specific, detailed metabolic pathway analysis. Nevertheless, they were used here only to provide a general metric for the power of microbial protein enriching approach and not necessarily to assess biological differences between the two infants. For both infants, we found several highly represented COG categories, including *Carbohydrate transport and metabolism*, *Energy production and conversion*, *Translation, ribosomal structure and biogenesis*, *Posttranslational modification, protein turnover, chaperones* and *Amino acid transport and metabolism* (Figure 3.8). A similar distribution of COG functions was reported for a healthy adult twin pair [72]. These results suggest that the establishment of microbial communities in these two infants gut environments is fairly quickly migrating towards a relatively stable and adult-like microbiota, which plays a crucial role in carbohydrate metabolism and nutrient production.

An in-depth inquiry into infant gut microbiota establishment during early life, as well as identifying the relationship between microbiota and inflammatory disorders, requires delineation of the full range of microbial functions, especially those that are seemingly of low abundance. For infant #UN1, human proteins were only moderately abundant, and so the enrichment of microbial proteins did not change the overall pattern of COG categories but allowed both more confident protein identifications as well as an order of magnitude deeper spectra assignment for each functional category. This enhancement provided deeper protein signatures with better

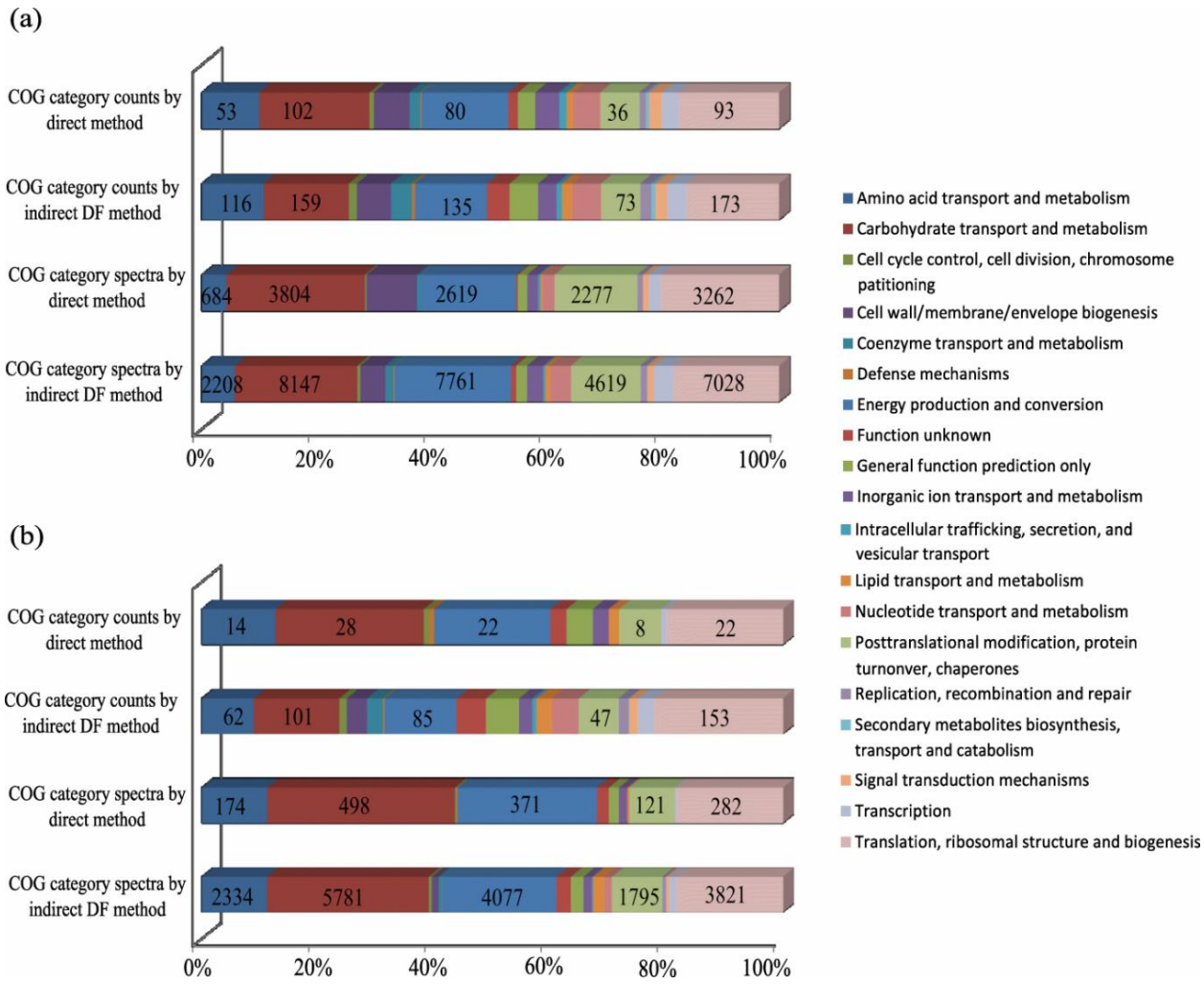


Figure 3.8. COG category analysis of microbial protein groups. Microbial protein groups are assigned into COG categories via rpsblast against the COG database from NCBI. Distributions of identified categories were constructed by category counts and spectra of infant #UN1 (a) and infant #CA1 (b). Abundant categories are numerically labeled.

statistics and better coverage of specific cellular pathways. However, for infant #CA1, human peptides dominated the identifications, leading to suboptimal microbial protein binning into COG categories for the direct method. Conversely, employing the DF strategy allowed us to significantly improve the resolution of microbial functional category determination/assignment, resulting in newly identified categories of *Cell wall/membrane/envelope biogenesis*, *Coenzyme transport and metabolism*, *Intracellular trafficking secretion and vesicular transport*, *Nucleotide transport and metabolism*, *Replication, recombination and repair*, and *Signal transduction mechanisms*. Based on the indirect DF approach, the microbiota of two infants intriguingly shared the similar COG function profiles despite tremendous taxonomic diversity, suggesting functional redundancy in the early intestinal ecosystem. Clearly, this approach sets the stage for more detailed time course measurements and expanded gene ontology / metabolic mapping analyses that should provide a higher resolution delineation of microbiome development/stabilization/functional activities during early infant life.

3.6 Conclusions

In this study, we report a novel metaproteomic method for extensively interrogating infant gut microbiome. By performing a double filtering strategy on the raw samples, we successfully enriched relatively low abundant microbial proteins from complex fecal samples containing dominant human host proteins, while preserving the relative distribution of protein abundances in each sample. This provided an in-depth microbial metaproteome measurement with greater than two-fold increase in microbial protein identification and quantification with relatively high correlated quantification, which improved our ability to confidently and comprehensively characterize microbial functional categories for complicated gut metaproteome.

Moreover, although the supernatant and filtrate generated by this approach were not examined in this study, these samples could be useful for future analyses that focus on human proteome and host responses to gut microbiome.

CHAPTER 4

Instrumental and informatics considerations for metaproteomics

Part of the text below was adapted from:

Weili Xiong, Paul Abraham, Zhou Li, Chongle Pan, Robert L. Hettich. Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. *Proteomics,* 2015, 15 (20), 3424-3438.

Weili Xiong's contributions included: literature review, manuscript writing in experimental workflow and human gut metaproteoimcs studies sections, data analysis, and manuscript editing.

Abraham, P. E., Giannone, R. J., Xiong, W., Hettich, R.L., Metaproteomics: Extracting and Mining Proteome In-formation to Characterize Metabolic Activities in Microbial Communities. *Current Protocols Bioinformatics* 2014, 13 (26), 11–14.

Weili Xiong's contributions included: figure generation and manuscript editing.

4.1 Introduction

Mass spectrometry based proteomics have enabled identifications of thousands of proteins in complex environmental samples, providing insights unachievable by classical biological approaches. However, proteomics analysis, especially metaproteomics, can be challenging as a comprehensive proteomic measurement by LC-MS/MS not only requires the ability to deal with samples properly, but also optimal instrument settings and appropriate bioinformatics analysis. In particular, the success of gut metaproteomics is affected by depth of the measurement, construction of a protein database, assessment of spectra quality, unambiguous protein inference and accurate protein quantification. In this chapter, we will discuss all above instrumental and informatics considerations and illustrate possible strategies with the aim of providing a reliable metaproteomic pipeline for the analysis of gut metaproteomes.

4.2 Enabling monoisotopic precursor selection for in-depth proteome measurement

As mentioned in Chapter 3, in-depth microbial protein identifications of infant fecal samples are inhibited by the presence of a few abundant human proteins. During MS measurements, these abundant proteins are repeatedly sampled and prevent the detection of co-eluting low abundance proteins. Although the dynamic exclusion is typically used to increase the measurement depth, the exclusion width is particularly critical when selecting peaks onto the exclusion list [122]. For example, an exclusion width of 1.5 m/z excludes the entire isotopic package from sampling, which effectively avoids re-sampling the same peptide but at the same time throws out other peptides that are co-eluted within the 1.5 m/z window. Another option that sets the exclusion width at a tighter window (0.2 m/z) doesn't exclude the whole isotopic distribution. Although protein quantifications can be improved, sampling isotopic peaks of the same peptide greatly bias the abundant peptides and limits the depth of measurement.

In the data-dependent acquisition (DDA) mode, monoisotopic precursor selection (MIPS) can be enabled or disabled [157]. When enabling this option, only the monoisotopic peak of the entire isotopic package will be selected for fragmentation. On the other hand, if this option is disabled, all isotopic peaks are accessible for fragmentation. In the MIPS-enabled method, the exclusion width can be set as tight as possible (typically 10ppm) to prevent any co-eluting peptides from being excluded due to the wide exclusion window. In addition, only monoisotopic peaks of peptides are subjected to MS/MS, and therefore, isotopic peaks of the same peptide won't be re-sampled, which avoids the repeat sampling of abundant species. A recent study has shown that enabling this option results in significant improvement in proteome coverage and depth [157]. Here, we evaluated how MIPS affected protein identifications in the infant gut metaproteomics.

One fecal sample was prepared using the direct TCA method described in Chapter 3. As a comparison, 50µg peptides were measured via 2D-LC-nESI-MS/MS on LTQ-Orbitrap Elite in data-dependent mode with MIPS enabled or disabled, and results were displayed in Table 4.1. Enabling MIPS facilitated less number of total collected and assigned spectra but gained three times of peptide and protein identifications, compared to the MIPS-disabled scheme. The increase of protein identifications was obtained in both human and microbial proteins, and can be observed in not only low abundance proteins (with spectral counts below 5), but also medium and medium high abundance proteins (with spectral counts ranging from 5-100) (Figure 4.1). This indicated that newly identified proteins were not acquired by an increasing number of random samplings of very low abundance species but by gaining the chance to measure co-eluted peptides that were overshadowed by abundant peptides.

Although top 20 abundant ions in a full MS are expected for MS/MS, it is possible that there are less than 20 available monoisotopic peaks in a full MS when enabling the option. Indeed, it was noticed in all 11 salt-pulse steps that many full scans were followed by less than 20 MS/MS events (Figure 4.2). This was observed during the salt pulse when no or very few peptides were transferred into the mass spectrometer as well as during the elution of very abundant peptides when these peptides were so dominant that overall peptides were less diverse. As a result, the instrument spent more time in MS1 scanning, which might explain why MIPS-enabled measurements collected less total MS/MS.

To evaluate the depth of measurements using two different options, MS1 percentage was calculated for every MS/MS scan in all salt-pulse steps（Figure 4.3). Basically, all peaks in a full MS were ranked from high to low intensity. For every MS/MS scan, the rank of its precursor ion in the full MS can be obtained according to the intensity. MS1 percentage was calculated by

Table 4.1. Comparisons of collected/assigned spectra, identified peptides/proteins with MIPS enabled and disabled

	Total Collected Spectra	**Total Assigned Spectra**	**No. of identified peptides**	**No. of identified proteins**	**No. of human proteins**	**No. of microbial proteins**
Monoisotopic precursor disabled	427909	123069	3352	955	883	72
Monoisotopic precursor enabled	387082	91618	9494	3105	2623	482

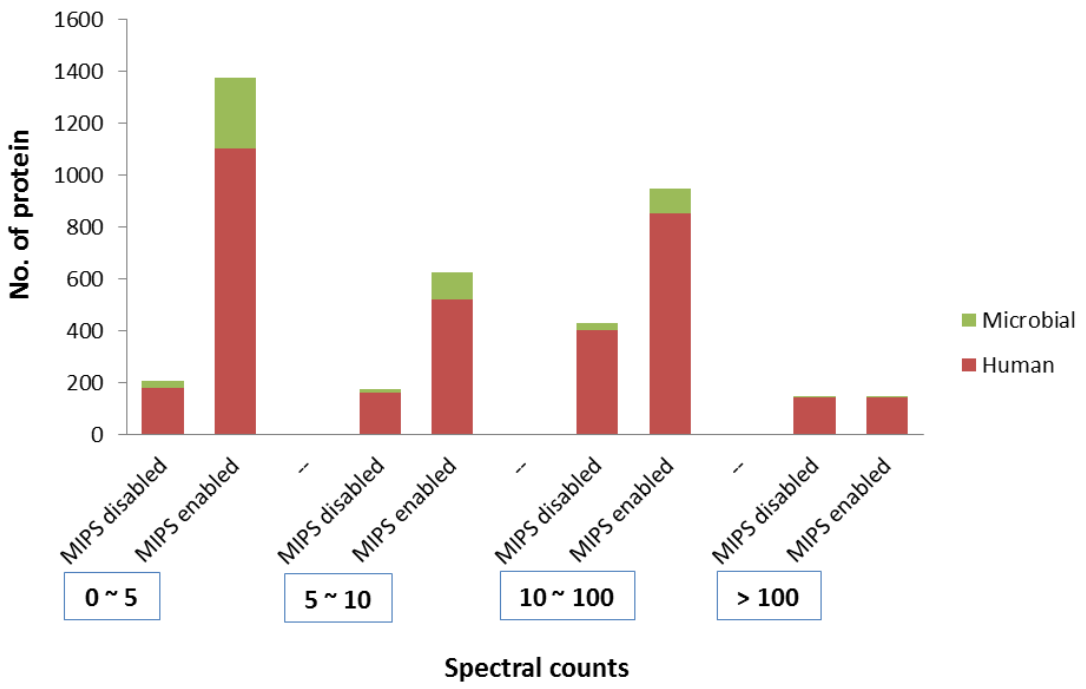


Figure 4.1. Number of high and low abundance identified proteins with MIPS enabled and disabled. Identified human (red) and microbial (green) proteins were counted in four different groups (with spectral counts larger than 100, between 10 and 100, between 5 and 10, and below 5)

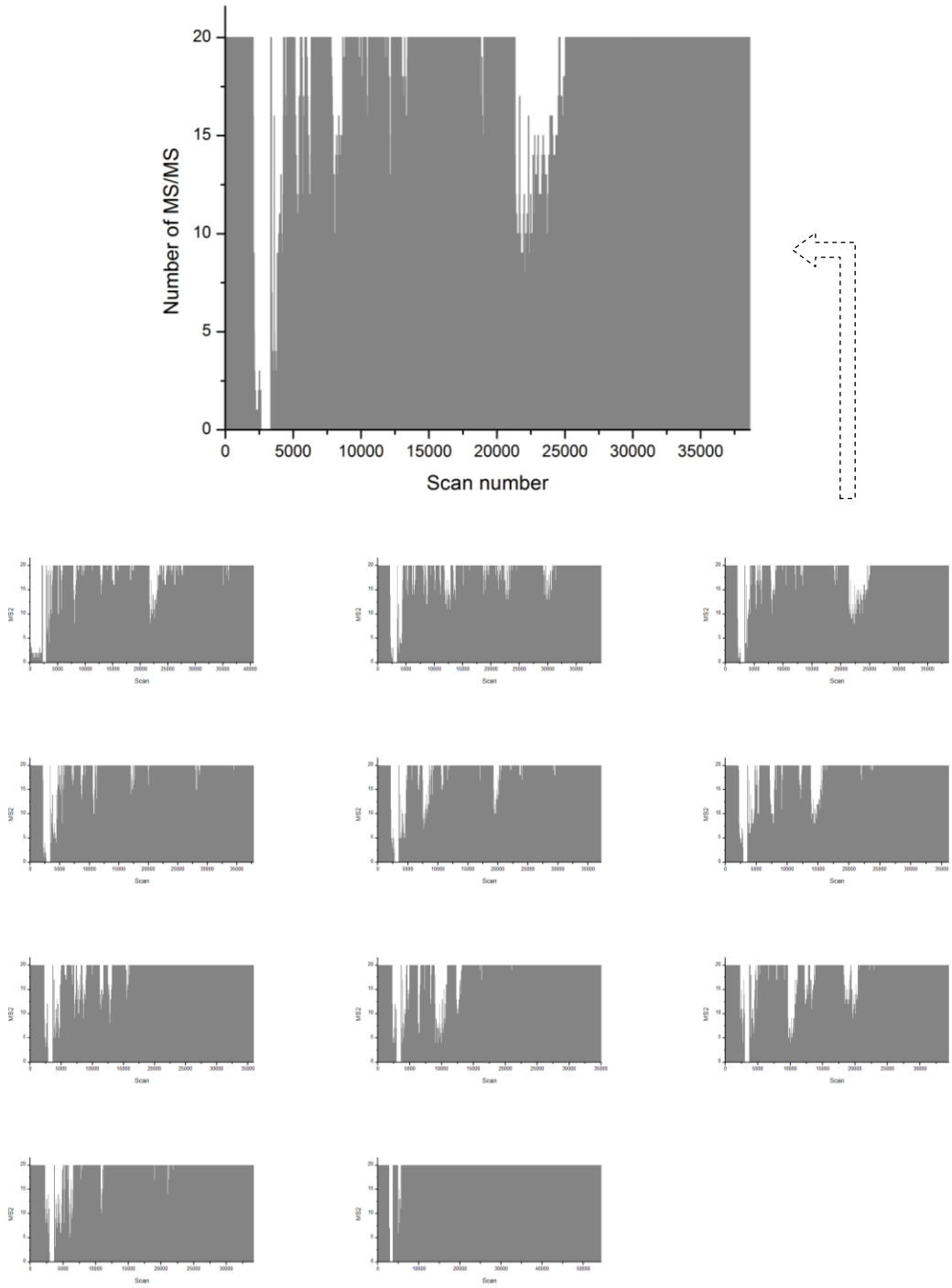


Figure 4.2. Number of MS/MS events followed by every full MS scan in 11 salt pulse steps with MIPS enabled. A zoomed in figure of salt pulse #3 was shown on the top.

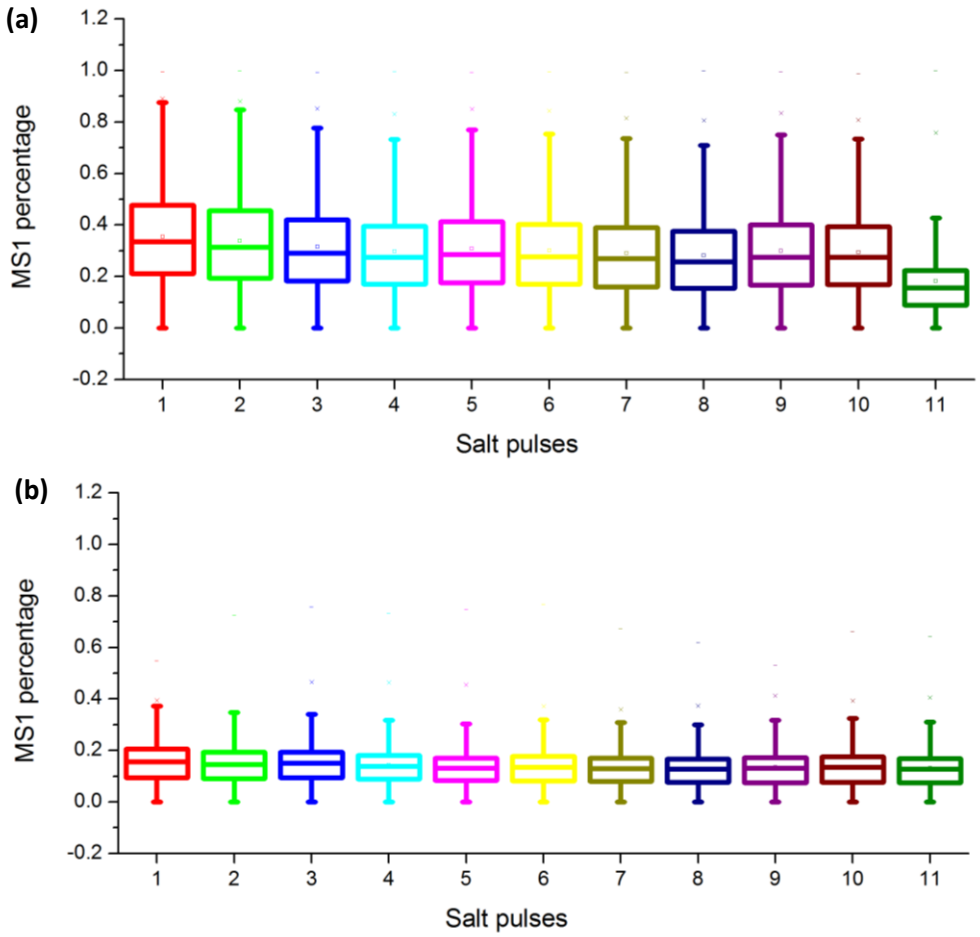


Figure 4.3. Box plot of MS1 percentages in 11 salt pulse steps with MIPS enabled (a) and disabled (b). Each color of boxes represents one individual salt pulse measurement.

the rank of a precursor ion divided by the total number of available precursor peaks in a full MS. Therefore, the higher the percentage is, the deeper measurement can be achieved since precursor ions with lower abundance have been selected for fragmentation. As shown in Figure 4.3, a medium MS1 percentage of ~30% was achieved in almost all salt-pulse steps (except ~18% in the 11th step) with MIPS enabled whereas a percentage of ~18% was obtained with MIPS disabled. The higher MS1 percentage in the first situation demonstrated that enabling MIPS allowed for a deeper proteome measurement.

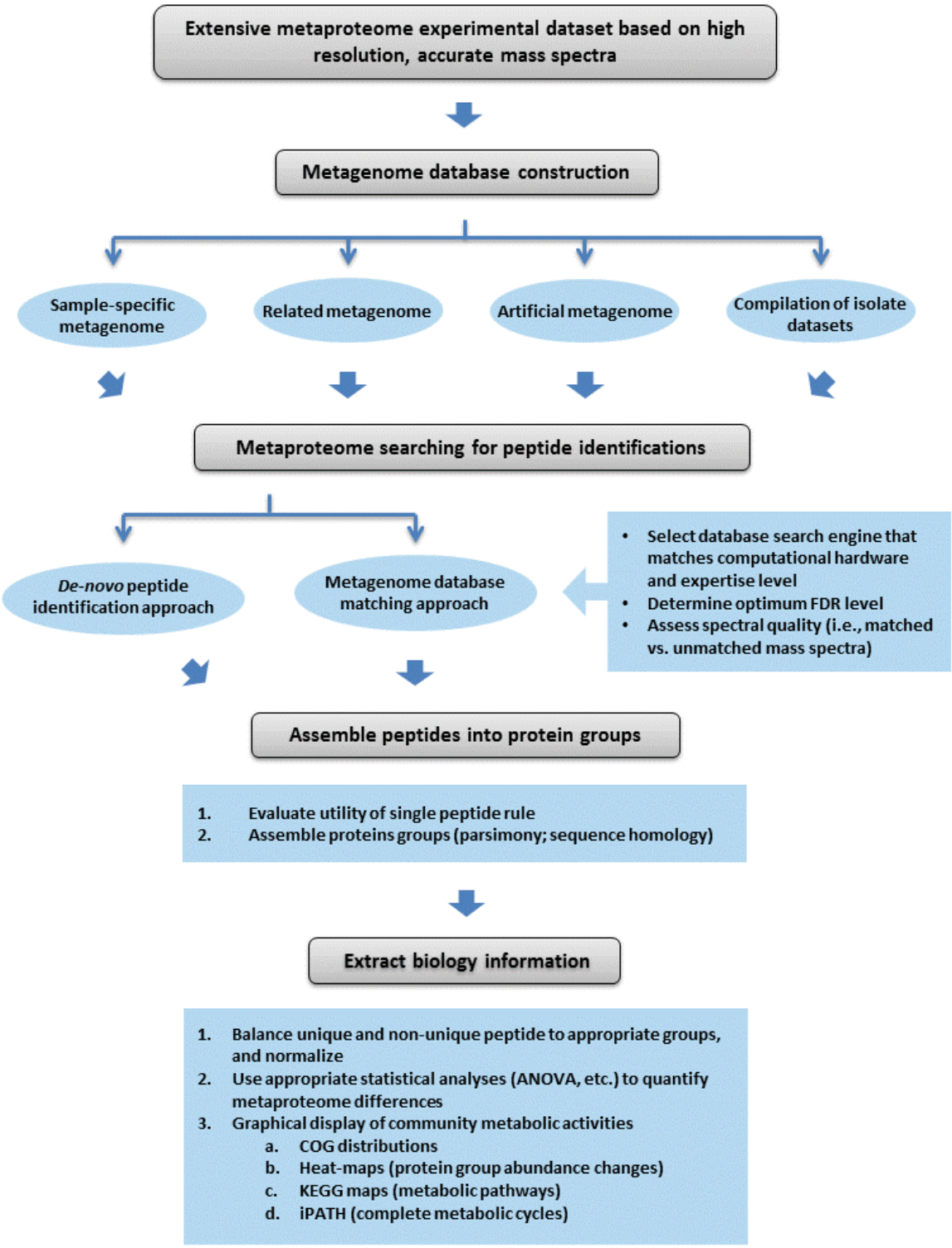
4.3 Informatic considerations for human gut metaproteome

While efforts have been made to optimize and improve the experimental and technical procedures for gut metaproteome, emphasis should also be put on the bioinformatics steps that transfer the raw MS/MS data into reliable protein identifications and meaningful biological information, as depicted in Figure 4.4. The starting point for metaproteomics data analysis is an extensive experimental dataset(s) of proteolytic peptide masses and fragmentation patterns generated by high-performance mass spectrometry. The overall informatics process consists of a number of sequential steps, including metagenome database construction, database searching for peptide identifications, protein assembly/grouping, and biological information extraction.

4.3.1 Construction of protein sequence database

Depending on the complexity of communities, the size of metaproteomic databases can range from containing tens of thousands to millions of proteins predicted from a variety of organisms. Constructing an appropriate database plays an essential role in FDR calculation and therefore determines the success of confident protein identifications [56]. At present, there are

Figure 4.4. Informatics workflow for metaproteomics. While there are various options at every step, the most common route would be preparation of a deep and well-annotated metagenome from the exact same sample targeted for metaproteome measurement, peptide identifications based on database searching/filtering/scoring against that matched metagenome, assembly into protein groups based on parsimony/sequence homology, normalization/statistical analysis, and finally graphic display of correlations for extraction of biological information.


Extensive metaproteome experimental dataset based on high resolution, accurate mass spectra
Metagenome database construction
Sample-specific metagenome
Related metagenome
Artificial metagenome
Compilation of isolate datasets
Metaproteome searching for peptide identifications
De-novo peptide identification approach
Metagenome database matching approach
• Select database search engine that matches computational hardware and expertise level
• Determine optimum FDR level
• Assess spectral quality (i.e., matched vs. unmatched mass spectra)
Assemble peptides into protein groups
1. Evaluate utility of single peptide rule
2. Assemble proteins groups (parsimony; sequence homology)
Extract biology information
1. Balance unique and non-unique peptide to appropriate groups, and normalize
2. Use appropriate statistical analyses (ANOVA, etc.) to quantify metaproteome differences
3. Graphical display of community metabolic activities
a. COG distributions
b. Heat-maps (protein group abundance changes)
c. KEGG maps (metabolic pathways)
d. iPATH (complete metabolic cycles)

essentially three different strategies for metagenome construction that are employed for metaproteome identifications: a “pseudo-metagenome” consisting of selected complete isolate genomes (the selection is typically guided by 16S rRNA information of the community) [100], a related but unmatched metagenome (which can be obtained from a similar community on the basis of the assumption that they may share most organisms) [72] and a sample-specific metagenome (which is built on the same sample as in proteomic analysis) [10]. Despite advancements in genomics leading to higher throughput and decreased costs, metaproteome database construction is still not trivial and facing significant challenges in the assembly and prediction of complete proteins. Therefore, many metaproteomics studies are often conducted on samples lack of metagenome but available for 16S rRNA information. A reference database can be constructed by concatenating a number of related sequenced isolate genomes and provides a quick and general functional characterization of the community. One advantage of the reference database is that selected isolate genomes are generally complete and accurately assembled. An example of a pseudo-metagenome approach focused on an iterative workflow for database searching, in which spectra were first searched against a synthetic metagenome comprised of over 200 intestinal species [100]. Next, a new database was created by blasting the hits from the first search against MetaHIT repository for homologous sequences. This new database was then used for a second search and permitted species-specific protein identifications. Clearly, the major disadvantage of using pseudo-metagenomes is that they do not accurately reflect the actual genome repertoire, since they lack distinct sequence information inherent to a particular microbial population. As a result, the identified metaproteome will be biased toward those organisms included in the database, leading to a skewed representation of the community being investigated.

In contrast to the pseudo-metagenome, the accuracy of protein inference increases when a closely related metagenome is available. Even without being an exact match to a particular sample, this approach improves the accuracy of protein identifications and has been used for gut microbiota studies in different humans [72]. In this scenario, the genome is much more reflective of the sample and thus a wider range of microbial membership can be evaluated. The unmatched metagenome can be also augmented with isolate genomes, which can generate even more protein identifications [72].

Of course, the most accurate means to characterize a microbial community involves employment of high quality matched metagenomes [10, 158]. In the context of the fecal proteome of two healthy human individuals, a study compared several assembly and gene finding strategies to increase microbial peptide spectral matching [158]. Overall, searching a matched metagenome facilitated a significant increase in the total number of assigned spectra, peptide identifications as well as protein identifications, as compared to the search with a concatenated database. However, as mentioned previously, there are some challenges in this approach, particularly the depth and coverage of the sequencing, as well as the accuracy of assembly and annotation. This may explain why the iterative search workflow with synthetic metagenomes showed higher spectral identifications when compared to the search with a matched metagenome.

4.3.2 Impact of metagenome quality and complex on peptide identifications

The completeness, accuracy and size of the metagenome will determine the assessment of matches between experimental spectra and databased predicted spectra by affecting the threshold for confident matches. In general, this threshold is controlled and regulated through the

calculation of FDR [129]. FDR estimate assumes that false positive PSMs are equally likely to map to either the target or decoy database. However, this assumption can be more problematic when dealing with metaproteomes. Due to the presence of many incomplete protein fragments in the metaproteome database, FDR is best and most meaningful to be evaluated at the PSM level. Moreover, FDR calculates the fractions of false positive assignments and can efficiently assess the confidence of assignment if the distributions of true and false PSMs can be well discriminated. However, this discrimination becomes increasingly blurry when the quality of the database is reduced or the size of the database is too large. Thus, in order to achieve a desirable FDR, the database search algorithms dynamically increase PSM scoring thresholds, which can result in an increase in the number of false negatives. To illustrate the importance of database size and completeness in peptide assignments, we compared the false positive and true positive PSM distributions for various database qualities and sample complexities (Figures 4.5 and 4.6). As shown in these figures, better matched databases and lower complexity biological systems facilitated better differentiation between true and false hits and thus accurately assigned a larger percentage of acquired fragment ion spectra. These results again stressed the importance of database quality and complexity with respect to identification sensitivity in metaproteomics.

4.3.3 Quality assessment of tandem mass spectra

Hundreds of thousands of tandem mass spectra are frequently collected in proteomics experiments. However, a significant number of spectra remain unidentified, especially for complex biological samples. For example, only ~30% of total collected spectra were identified in the infant gut metaproteome described in Chapter 3. Although unidentified spectra can be caused by a variety of reasons, such as poor quality spectra, incomplete database, modifications and constrained thresholds, the first and critical step of peptide identification is to evaluate the

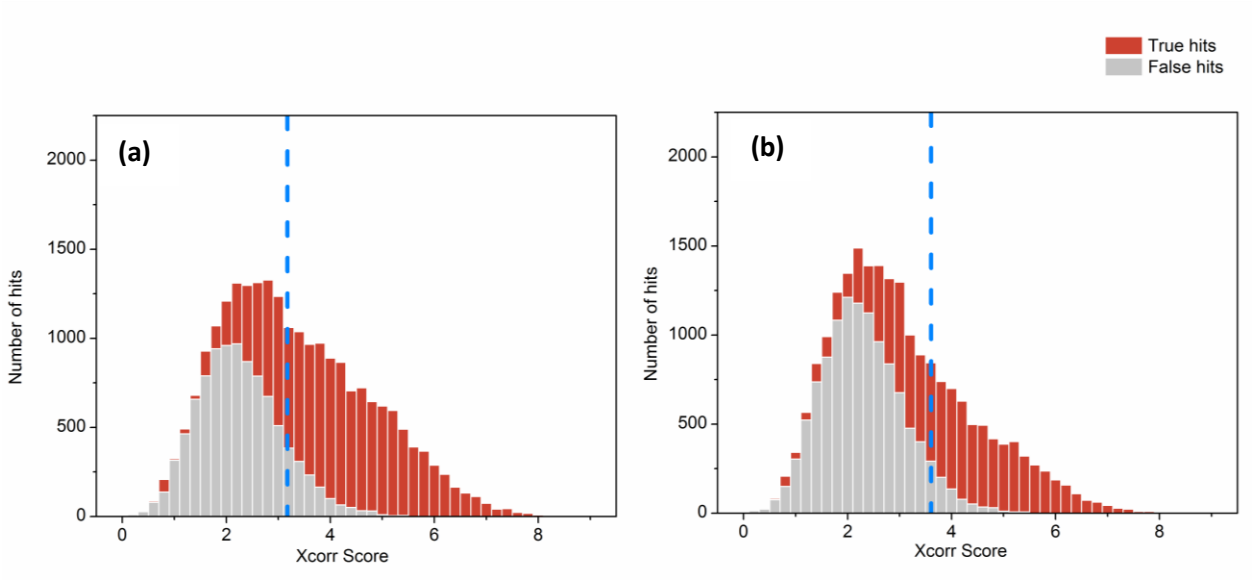
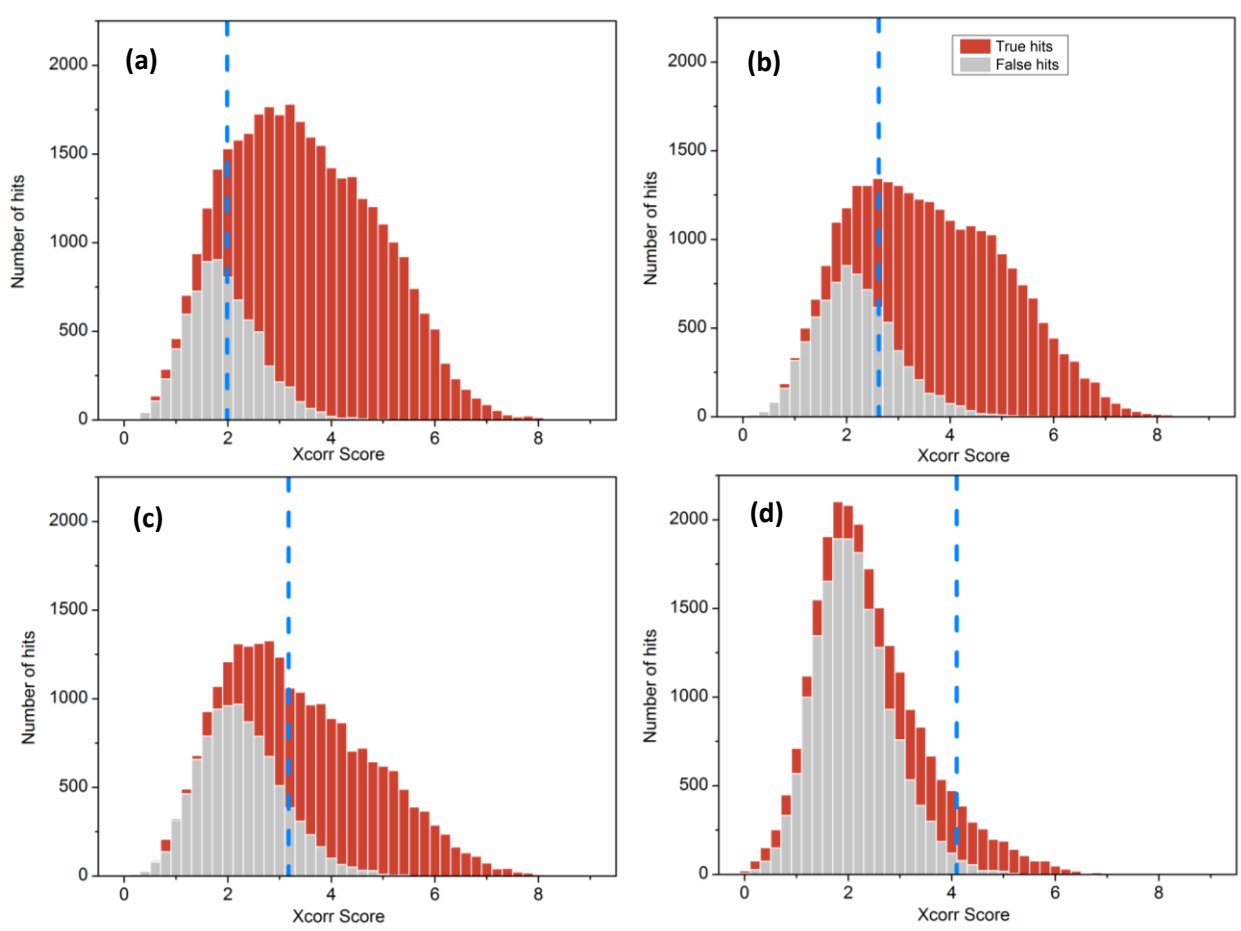


Figure 4.5. Impact of database quality on peptide identifications. Peptide spectrum matches can be ranked by MyriMatch Xcorr scores to reveal the distribution of true positive (red) vs. false positive (gray) identifications in human adult gut microbiome datasets searched with either a matched metagenome (a) or a pseudo-metagenome assembled from selected microbial isolates (b). An appropriate Xcorr score threshold (indicated by blue dashed line) is chosen to achieve a 1% PSM (peptide spectral match) FDR (false discovery rate; defined by the ratio between false PSMs and total PSMs above the score threshold). The figures reveal that the matched metagenome better differentiates true vs. false distributions, as evidenced by the higher percentage of "red identifications" to the right (i.e. higher Xcorrs) of the dashed line. Even though the pseudo-metagenome likely contains better quality, assembled microbial genomes, the matched metagenome is more closely linked to the actual environmental sample.

Figure 4.6. Impact of sample complexity on peptide identifications. Peptide spectrum matches can be ranked by MyriMatch Xcorr scores to reveal the distribution of true positive (red) vs. false positive (gray) identifications for samples of a synthetic mixture of six microbial isolates (all sequenced genomes) (a), a human *infant* gut microbiome, (b), a human *adult* gut microbiome (c), and an environmental soil (d). An appropriate Xcorr score threshold (indicated by blue dashed line) is chosen to achieve a 1% PSM (peptide spectral match) FDR. The level of true hits is greatest for the synthetic mixture, since the genomes are compete and well annotated. As the complexity of the community increases, the ability to separate true and false hits decreases, as indicated by the superior identification rates in the low complexity infant sample (b) relative to the higher complexity adult gut sample (c). For (b-d), relevant metagenomes were employed, although the metagenome of the soil sample was significantly larger (about 1.3 million genes, which was at least 2X larger than the adult gut microbiome metagenome). This metagenome could not be assembled to a satisfactory level and thus was highly fragmented, which resulted in virtually no distinction between true vs. false hits. This attests to the need for not only matched metagenomes, but well assembled and curated versions, for complex samples.


(a)
(b)
(c)
(d)
True hits
False hits
Number of hits
Xcorr Score
0
500
1000
1500
2000
0
2
4
6
8

quality of MS runs. Several spectral quality assessment tools have been developed recently. For example, ScanRanker tool assigns quality scores to every tandem mass spectra via sequence tagging [156]. In general, a high quality spectrum of a peptide is expected to contain a series of fragment ions, which can be inferred as multiple sequence tags with high scores. Spectra with higher scores/quality are more likely to be identified in a proteomic study and thus the number of high quality spectra in the dataset helps reveal the richness of identifiable spectra. This is potentially useful for metaproteomic studies, in which ScanRanker can be used to determine the number of unidentified high quality spectra (Figure 4.7). A large portion of unassigned high score spectra may suggest the incompleteness of a metagenome. Moreover, these spectra can be further reanalyzed and recovered by other approaches, such as blind modification search or *de novo* sequencing [159-161], resulting in new identifications and biological information.

4.3.4 Protein grouping and clustering

As mentioned in the previous chapter, to alleviate the ambiguity associated with shared peptides, proteins can be clustered into protein groups by sequence homology algorithms. As a result, shared peptides are found to be unique to a protein group. Before any measurements are collected, the number of unique peptides can be predicted at the database level and also the number of unambiguous proteins (likely to be detected with unique peptides) can be calculated. These numbers will provide an estimation of potentially ambiguous protein rate and to what degree, these ambiguous assignments will impact the collected data. To illustrate this, the number of unique peptides was predicted in human database and 21-isolates reference database constructed in Chapter 3. Since not all peptides are MS-friendly [162], two databases were run through PeptideSieve (Seattle Proteome Center), a software that calculates the likelihood of peptides detected by ESI-MudPIT experiments. Only those peptides that were highly possible

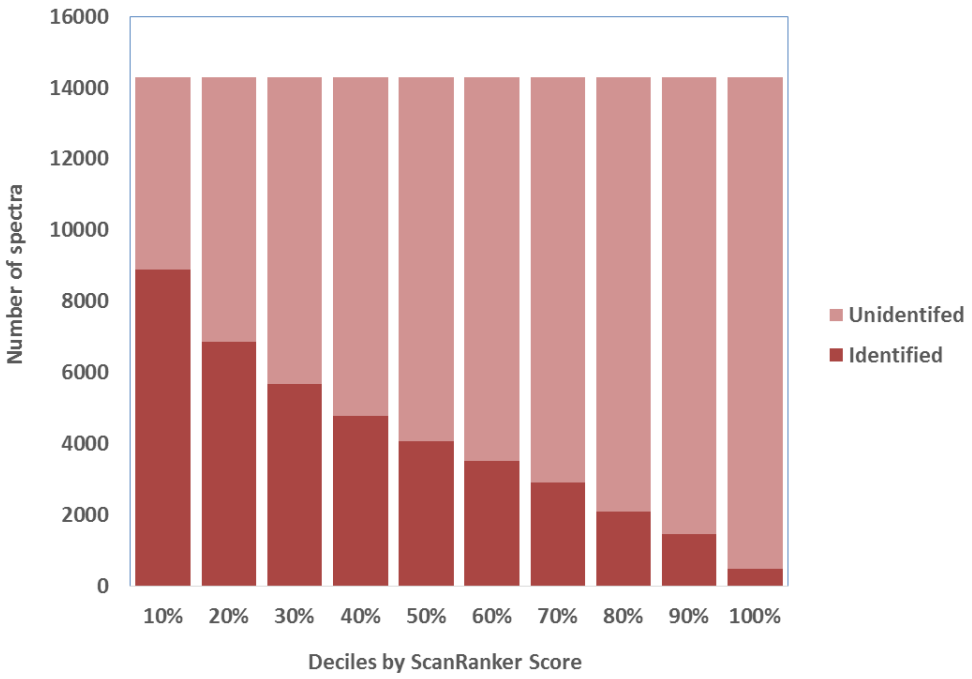
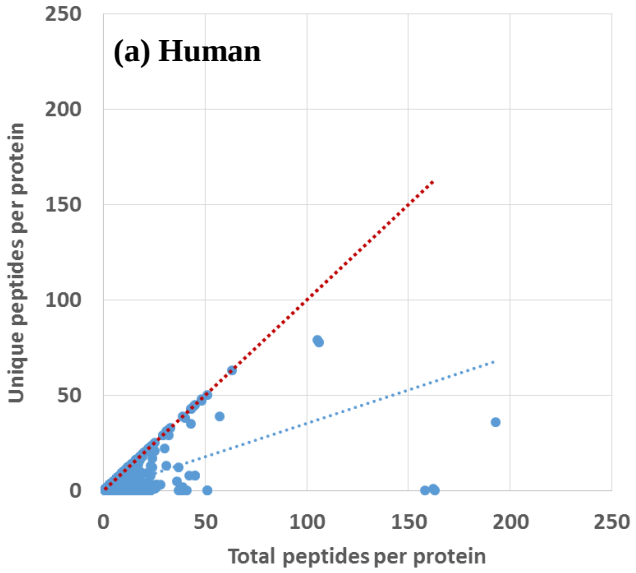


Figure 4.7. Evaluation of ScanRank to determine unidentified high quality spectra. One infant gut microbiota dataset was identified with MyriMatch. These graphs plot the distribution of identified (dark red) and unidentified (light red) spectra in deciles by ScanRanker scores. The left side represents spectra assigned with high ScanRanker quality scores and the right side represents the low quality spectra.

(p>0.9) to be detected were retained and analyzed. Scatter plots of total peptides per protein against unique peptides per protein showed the degree of redundancy for the two tested proteomes (Figure 4.8). Red dashed line along the diagonal represent the situation where all peptides are unique, while blue dashed line showed the trend line of observed data. Thus, the larger the trend line deviated from the diagonal, the higher the redundancy one proteome can have theoretically. Obviously, human database showed much higher redundancy due to paralogs present in the genome. Besides the number of unique peptides, it is also interesting to calculate the number of redundant proteins. For human proteomes, there were 73% (25173 out of 34657) detectable proteins (containing at least one peptide with p>0.9), and among those proteins, 54% had more than 95% shared peptides and thus highly redundant. In comparison, 21- isolates had 68% (48234 out of 70969) detectable proteins and only 7% of which were highly redundant.

The goal of protein clustering is to alleviate the ambiguous protein inference by grouping together those proteins that cannot be differentiated through the measurement and some shared peptides can be “rescued” and become unique to the protein group after clustering. Therefore, protein clustering has been recognized as an efficient way to study the shared functional process. However, in the process of clustering, it is possible that multiple protein identifications are combined into one identification, which reduces some level of protein resolution and loses the strain information. As the clustering threshold decreases, more peptides can be “rescued” but fewer protein groups that contain more memberships will be generated. Therefore, it’s important to test a range of thresholds to balance the tradeoff between the protein ambiguity and resolution. As shown in Figure 4.9a, a dramatic collapse (50%) of human proteins was observed when clustering at 90% while protein groups in 21 isolates were decreased much slower when lowering the similarity threshold, indicating higher redundancy in the human database and 90%



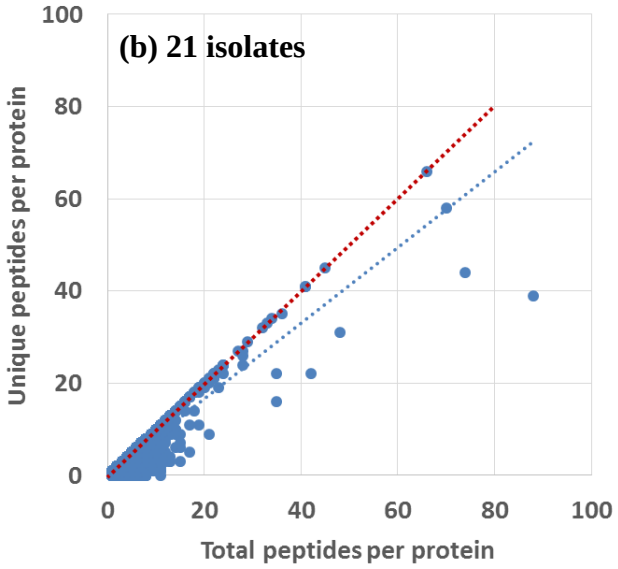


Figure 4.8. Degree of database redundancy. Peptides that were most likely to be detected in an ESI-MudPIT experiment were predicted using PeptideSieve. For each protein, total number of predicted peptides and number of unique peptide were plotted. Red dashed line represented that all predicted peptides were unique and blue dashed line represented the actual trend line of the data.

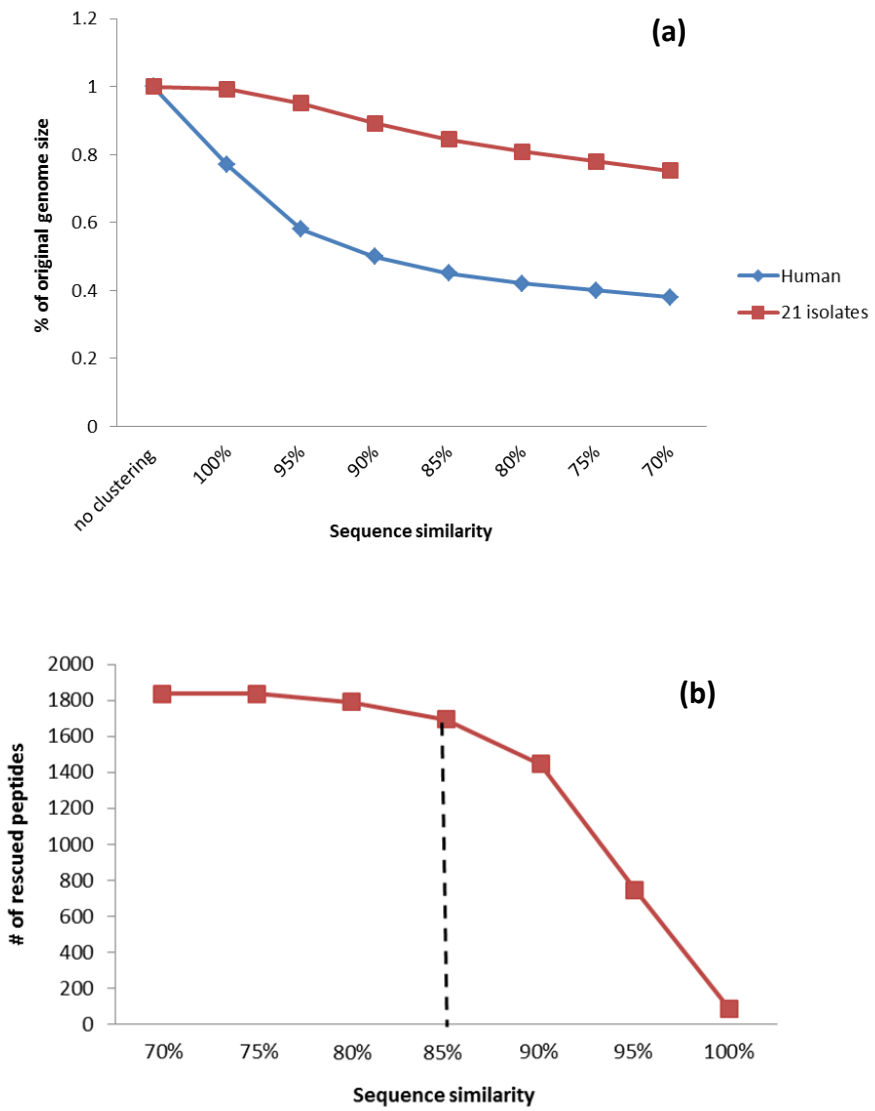


Figure 4.9. Determination of similarity threshold. (a) Human (blue line) and 21 isolates (red line) database were clustered at different sequence similarities and % of original genome sizes were calculated by total number of protein groups after clustering over total number of proteins in the original database. (b) Post database searching, number of "rescued" peptides (peptides were non-unique to the database but became unique to a protein group after clustering) were calculated at different sequence similarities.

can be an appropriate threshold. Also, we plotted the number of "rescued" peptides at multiple thresholds for 21 isolates (Figure 4.9b) and found a sharp increase when lowering the threshold to 85%, which was suggested to be the appropriate threshold. Nevertheless, there are not exact rules for the threshold, it should be as conservative as possible and more importantly, considering the objective of the study. For example, if the study focuses on the strain resolved proteome identifications, 100% may be chosen even though this threshold usually can't help unambiguous protein identifications.

4.3.5 Protein quantification

Equally as important as protein identifications, the other challenging task in proteomics is to accurately quantify and differentiate proteins from different biological samples. Although various quantification methods have been developed [27], such as metabolic and chemical labeling, label-free quantitative proteomics has been widely used to compare proteins across samples because its workflow is straightforward and much less expensive. Label-free quantification based on MS2 product ion scans is typically performed through two strategies: spectral counting estimates the number of MS/MS spectra matched to all peptides from a given protein; and intensity-based MIT that sums up the matched product ion intensities in each MS2 spectrum assigned to a given protein [163]. Here we tested two quantification methods on an infant gut metaproteome data and found that the quantification by MIT correlated very well with that using spectral counts (r_s = 0.9, Figure 4.10). But MIT presented a greater range of difference than spectral counts, particularly for those low abundance proteins. For example, a number of proteins that were quantified with two spectral counts can't be differentiated if spectral counting quantification was applied. However, the intensities of these proteins varied from 2^{15} to 2^{20}, providing the greater level of specificity and more accurate protein abundances. More recently,

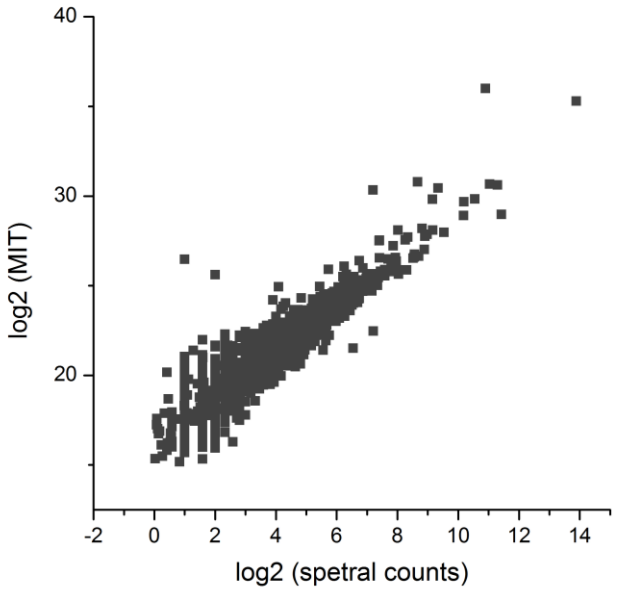


Figure 4.10. Comparison of protein MITs with spectral counts. Protein plotted here required at least one unique peptide. For MIT, a peptide MIT was achieved by summing up the intensities of all scans that matched to the peptide. MIT of shared peptides were balanced among all shared proteins. A protein MIT was calculated by summing up all unique peptide MITs and balanced portions of shared peptides. Protein spectral counts were obtained by summing all unique peptide spectral counts and balanced spectral counts from shared peptides. The spearman correlation coefficient of log2 (spectral counts) and log2 (MIT) was 0.9.

researchers have described that MIT showed good linear response to a range of protein concentrations but not spectral counts [164]. This indicates that MIT is sensitive to variations of sample loading amount and useful in the comparison of protein levels relative to units (for example, protein levels per gram raw material). On the other hand, spectral count is more tolerant to unequal loading amount, which can be a valuable feature for samples that are difficult to achieve equal loading amount. From this perspective, spectral counting quantification may be more robust in metaproteomics studies where the relative abundance of every organism can vary across samples.

4.4 Conclusions

Although metaproteomics has emerged as a valuable research tool for the investigation of metabolic activities of complex communities, a careful consideration of both experimental and informatics components is needed for the successful deployment. We have demonstrated that considerations of sample preparation and instrumental settings play an important role in improving the depth and accuracy of proteome measurement. As for informatics component, the starting point is to evaluate the metagenome quality, complexity and redundancy, since these properties of the metagenome directly impact the resulting protein identifications. Due to the redundancy in the metaproteomic data, an effective way for unambiguous protein inference is to group/cluster proteins with an appropriate threshold that alleviates the ambiguity and provides each group with meaningful biological information. Finally, an extensive inventory of protein identities and abundances can be achieved and further translated into metabolic information.

CHAPTER 5

Metaproteomics of a healthy premature infant gut to access early-life microbial functionality and host responses

5.1 Introduction

Microbes colonize all internal and external surfaces of the human body and influence all aspects of human physiology. The largest microbiome locates in the human gastrointestinal tract (GIT), which is composed of up to 100 trillion microorganisms, comprising thousands of different species and five million unique genes [57]. Microbes residing in the gut interact with each other and the host, play important roles in host nutrients through the production of vitamins, short chain fatty acids (SCFA) and amino acids, regulations of immune system by establishing immune tolerance to commensal bacteria and immune protection against pathogen, and maturation and integrity of the intestinal epithelium [165]. Dysbiosis of the gut microbiota has been linked to many diseases, such as Crohn's disease [166-168], diabetes [11, 169] and autoimmune diseases [170]. The establishment of gut microbiota begins during infancy, and emerging evidence has suggested that this initial colonization may have a life span effect on the human health [88]. While most recent studies about gut microbiome have revealed the microbial development of human adults in healthy and diseased state, fewer studies have focused on understanding the establishment of the microbiota at birth and how microbiota is associated with infant health and diseases, especially for preterm infants.

Both term and preterm infants were thought be born sterile in the gut, but this has been challenged by the presence of microbes in the placental and meconium samples [89].

Immediately after birth, the sterile/near sterile gut of the newborn infant is colonized with bacteria through the first contact with the mother and the environment. Typically the initial colonizers of the gut are facultative anaerobes. Within days or weeks, there is a shift from facultative to obligate anaerobes [171]. The establishment of the microbiota is influenced by multiple factors, including gestational age, delivery mode, birth weight, diet and exposure to antibiotics [89, 148, 172]. For example, the microbiota of infants born vaginally resembles the mother's vaginal and fecal microbiota whereas infants born by cesarean section develop their microbiota similar to skin or environmental bacteria [148]. It has also been suggested that C-section delivered infants have lower complexity gut microbial community compared to vaginally delivered infants [173]. The infant gut undergoes rapid increase in the abundance and diversity of microbial population during the first a few weeks. Large variations have been observed among different individuals and also over time within the same individual [89]. After 2.5 ~ 3 years of life, an infant's gut microbiome become a stable and adult like microbiome [174]. It remains to be determined that what factors (host genetics, environment, diet and/or interplay between host and microbiome) and how these factors determine the path of microbiome development as well as the influences of different paths on the host health and disease status. This is particularly critical for premature infants who may have a delayed and aberrant microbiota.

Infants born prematurely are at higher morbidity and mortality risk due to the immature organ systems that are not properly functional to adapt to the extrauterine life [89]. These infants are susceptible for inflammation disorders as a result of poorly developed immune system and prenatal/postnatal events that inappropriately modulate the immunity [89, 175]. Necrotizing enterocolitis (NEC) is a devastating intestinal inflammatory disease of premature infants, especially for those born with very low birth weight (500 – 1500 grams) and born prior to 28

weeks of gestation [176, 177]. NEC typically occurs in the second to third week of life in premature infants and is characterized by intestinal inflammation and damage, such as mucosal injury or necrosis. Additional potential risk factors causing NEC may involve in formula feeding, enteral feeding, blood transfusion and overall health [178]. The role of bacterial colonization in neonatal NEC has been suggested by a number of observations, including the identification of pneumatosis intestinalis (gas in the bowel wall) which is most likely produced by intestinal bacteria, occurrence of outbreaks in hospital, and resolution of inflammation after treatment with antibiotics [93, 96, 179]. Therefore, recent researches have focused on investigating the composition of microbial community associated with NEC [180]. A number of bacteria have been implicated in the pathogenesis of NEC, but none of them has been identified as the infectious agent [181]. Recently, Raveh-Sadka et al. analyzed gut communities in a number of premature infants during an outbreak of NEC but however, found no single bacterial strain was shared among all infants who developed NEC [97]. This may indicate that NEC is not attributed to a single bacterial strain. Instead, it may be caused by the introduction of a few harmful bacteria that disrupts the essential activities of commensal microbes in protecting the intestine. Therefore, better characterizing the functional activity of bacteria during colonization in both healthy and NEC infants could help understand the role of gut microbiome in the development of NEC.

Mass spectrometry based metaproteomics approach has been widely used to analyze communities samples and has emerged as an indispensable tool in investigating the gut microbiota [10, 99]. Since many challenges still remain, our efforts were made to develop and optimize the experimental and informatics pipeline for the infant gut microbiome characterization in previous chapters. It is noted that the onset of NEC usually occurs over a time

period and therefore a comprehensive interpretation requires multiple time points to capture the variation prior to the development of NEC. Before getting into the comparisons between the healthy vs diseased state, we will first focus our study on describing the dynamic functional profiles for the gut microbiota of a healthy preterm baby over time in this chapter. Metagenomics information of these samples have been previously analyzed and revealed shifts in bacterial species, strains, and phage during early colonization. It's also worth mentioning that near-complete genomes were constructed in a few highly abundant organisms, which enabled the in-depth proteome characterization.

5.2 Materials and methods

Sample collection. Fecal samples were collected from a healthy (did not develop NEC) preterm infant (#3) over the first three months (DOL (day of life) 11, 12, 13, 15, 18, 21, 25, 28, 78 and 86) after birth. This infant was born by C-section at a gestational age of 26 weeks with birth weight 822 grams. The infant was breast milk fed starting from day 4 and withhold on days 7-9 because of blood transfusions. On day 17, the breast milk was fortified to 24 kcal/oz. The infant received initial treatment with antibiotics (ampicillin/gentamicin) for the first 7 days of life, vancomycin on days 51-58 for cellulitis, cefotaxime on days 51-53 and vancomycin/claforan on days 63-65 for sepsis evaluation. Fresh sample were collected using a previously described technique, but samples 11, 13, and 15 were left in the fridge for 5, 3, and 1 days respectively, whose proteome might be changed. For days 15 and 21, a second sample (in total two samples) was collected at different time of the same day. Samples were obtained under an IRB agreement protocol, and were de-identified before sending to ORNL.

Sample preparation and measurement. ~0.5 g raw fecal material was prepared by the indirect double filtering method described in Chapter 3 with modifications. The second filter was not applied in this study. Obtained peptide samples (50 µg) were analyzed 24 hours with 11 steps via 2D LC-nESI-MS/MS system on LTQ-Orbitrap Elite (Thermo Fisher Scientific, San Jose, CA). Full scans were acquired at 30k resolution (1 microscan) in the Orbitrap, followed by CID fragmentation of the 20 most abundant ions (1 microscan). Monoisotopic precursor selection was enabled. Unassigned charge and charge state +1 were rejected. Dynamic exclusion was enabled with a mass exclusion width 10 ppm and exclusion duration 30 seconds. Technical replicates (duplicates) were performed for each sample.

Data processing. A protein database was constructed by combining matched metagenome collected on multiple days, human protein sequences and common contaminants. All MS/MS spectra were searched with the Myrimatch version 2.1 algorithm [132] against the constructed protein database and filtered with IDPicker [135] using the same parameters described in Chapter 3. Proteins were grouped base on 90% amino acid sequence similarity for human proteins and 100% similarity for microbial proteins. Spectral counts were balanced between shared proteins, and normalized by total numbers of all collected MS/MS in each run.

Data analysis. KEGG Orthology (KO) of the metagenome was assigned by KASS (KEGG Automatic Annotation Server) using single-directional best hit (SBH) method for amino acid sequence query and KEGG pathways were constructed for both human and microbiome proteomes using KEGG Mapper (http://www.genome.jp /kegg/tool/map_pathway.html) . The multidimensional scaling plots (MDS), which measure the similarities among samples, were performed by the edgeR package [182]. The dataset was normalized based on scaling factors for library sizes, which were determined using a trimmed mean of M-values (TMM) between

samples. Hierarchical clustering of microbial proteins were carried out using heatmap.2 function in R (version 3.1.1) using the moderated log-counts-per-million values calculated by the edgeR package. Hierarchical clustering of human proteins were performed by JMP Genomics (SAS, Cary, NC) using log transformed spectral counts. Blast2GO platform [183] was employed to generate gene ontology (GO) annotations of the metagenome with a BlastP E-value hit filters of 1×10^{-6}, an annotation cutoff value of 55, and GO weight of 5. GO terms enrichment was assessed by employing the Fisher's exact test and correcting for multiple testing at a cutoff of FDR<0.05. GO analysis of human proteomes were performed by ClueGO, a Cytoscape plug-in application that interpret functionally grouped gene ontology annotation networks [184, 185]. Enrichment was calculated by right-sided hypergeometric enrichment tests at a medium network specificity selection and p-value corrections using the Benjamini-Hochberg method. The selected GO tree levels were a minimum of 3 and a maximum of 8 while each cluster was set to a minimum of between 3% and 4% genes. The GO term grouping setting was selected to minimize GO term redundancy and the highest significance term enriched was used as the representative term for each functional cluster. Only p-values less than or equal to 0.05 were considered significantly enriched.

5.3 General overview of metaproteomic datasets

Fecal metaproteomes of a healthy preterm infant were examined on days 11, 12, 13, 15, 18, 21, 25, 28, 78 and 86 after birth by shotgun metaproteomics approach. Up to 111954 spectral counts, 22779 peptides and 4140 protein groups were identified per run (Table 5.1), revealing deep proteome measurement for these complex fecal samples. Each sample was measured in duplicates with high reproducibility ($R^2 > 0.95$). Although biological replicates can have multiple

Table 5.1. Number of identified peptides, protein groups and MS/MS spectra

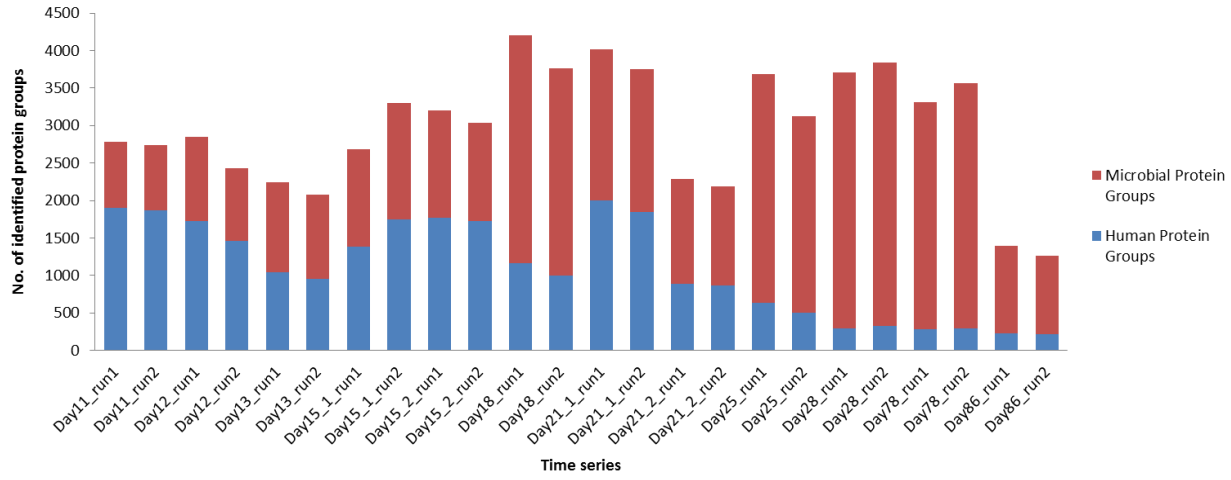
	Total Assigned SpC*	Peptides	Total Protein Groups	Human Protein Groups	Microbial Protein Groups	Total Protein Group SpC	Human Protein Group SpC	Microbial Protein Group SpC
Day11_run1	85602	17104	2643	1781	862	84778	57907	26871
Day11_run2	84056	16457	2600	1759	841	83239	572717	25968
Day12_run1	78589	15165	2666	1597	1069	77450	51425	26025
Day12_run2	74262	12490	2203	1301	902	73421	50101	23320
Day13_run1	80759	11526	2109	960	1149	79672	46986	32686
Day13_run2	77352	10272	2007	909	1098	76320	46892	29428
Day15_1_run1	90513	13237	2462	1239	1223	89844	54466	35374
Day15_1_run2	82444	17430	3138	1636	1502	82471	48805	33666
Day15_2_run1	82764	16367	3025	1660	1365	82004	57116	24888
Day15_2_run2	70144	15444	2833	1610	1223	69560	50136	19424
Day18_run1	111290	22779	4140	1116	3024	110351	35224	75128
Day18_run2	102632	19912	3611	918	2693	101437	33309	68127
Day21_1_run1	96231	22012	3829	1880	1949	95702	57939	37764
Day21_1_run2	93684	20079	3514	1707	1807	93090	55625	37445
Day21_2_run1	48000	10503	2017	767	1250	47653	25153	22500
Day21_2_run2	50563	9901	1970	766	1204	50273	26720	23553
Day25_run1	79003	17770	3495	564	2931	78965	20149	58816
Day25_run2	76006	14960	3004	470	2534	75878	19883	55995
Day28_run1	111017	21079	3683	272	3411	111436	14766	96670
Day28_run2	111954	21308	3758	278	3480	112387	15020	97367
Day78_run1	96468	18497	3253	249	3004	96148	21537	74611
Day78_run2	105730	20457	3562	265	3297	105545	22168	83377
Day86_run1	30877	6147	1212	185	1027	30268	9535	20733
Day86_run2	25235	5398	1080	159	921	24700	8557	16143

*SpC: spectral counts

meanings according to the context of the study, for example, the same organisms grown under the same conditions, it is basically impossible to obtain biological replicates for human infant fecal samples. However, it is still of interest to investigate how different samples collected on the same day correlate with each other and whether these samples can be treated as biological replicates. Fecal samples were collected twice on days 15 (samples 15_1 and 15_2) and 21 (samples 21_1 and 21_2), among which only samples 15_1 and 21_1 were analyzed with metagenomics information. Therefore, the database searching of samples 15_2 and 21_2 was conducted using unmatched but related metagenome. A Pearson correlation of r = 0.81 was found between 15_1 and 15_2 while the correlation between 21_1 and 21_2 was r = 0.52, indicating that the gut microbiome could change greatly within a day. This was also recognized by other studies that showed rapid and reproducible alteration of human gut microbiome by dietary interventions [67].

Human and microbial proteins were both monitored, providing a total of 9318 microbial and 3250 human protein groups across all time points (Figure 5.1 (a)). As the time increases, it was observed that the number of identified microbial protein groups increased while that of human protein groups decreased, mainly due to the increasing complexity of microbial composition revealed by metagenomic information. As microbial proteins became more abundant, the measuring depth of human proteins decreased and remained stable until after day 28, with ~ 200 identified human protein groups. This trend was also observed in the percentage of human/microbial relative abundance, where human proteins accounted for ~70% of total spectral counts at early time points and this percentage decreased to ~30% by day 25 (Figure 5.1 (b)). It was also noticed that less proteins and spectral counts were identified in samples 21_2 and 86. Thus, total number of collected spectra and high quality spectra were evaluated in these

(a)



(b)

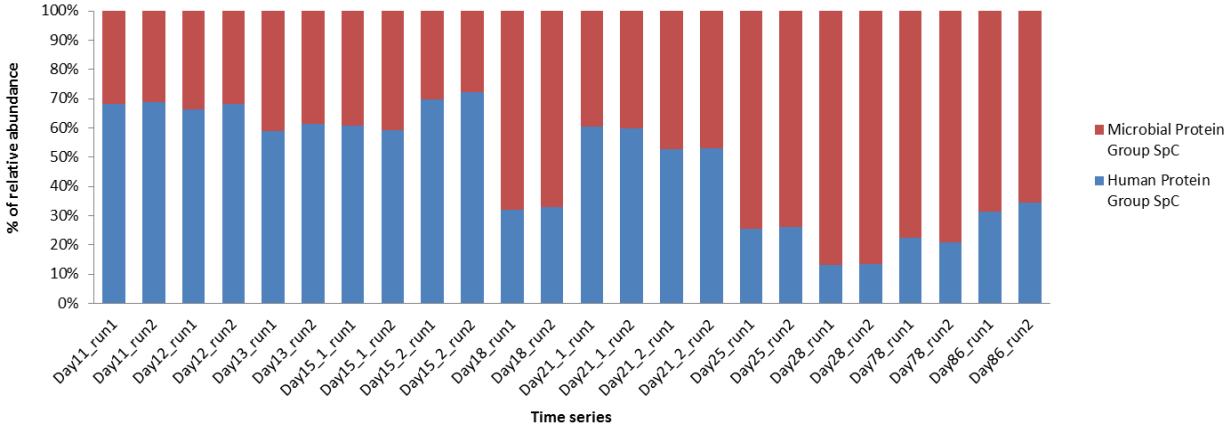


Figure 5.1. Number of human and microbial protein groups identified (a) and relative abundance of human/microbial spectra (b) over time.

two samples and found to be comparable with other samples, suggesting that the variation was less likely due to the low quality MS measurements but rather the incompleteness of the database (Figure 5.2).

Multidimensional scaling (MDS) was applied to assess similarities and differences of the protein expression across all time points for both human (Figure 5.3 (a)) and microbial proteins (Figure 5.3 (b)). To filter out low abundance proteins, total identified proteins were filtered based on having 100 counts per million (cpm) for at least one sample/library, which reduced the number of tested proteins from 12568 to 6137. After filtering, the new library size was compared with the original one and more than 95% of total assigned spectral counts were retained for each library, which indicated that half identified proteins had low number of spectral counts across all samples. MDS plotted 12 samples on a two-dimensional scatterplot and the distances approximated the protein abundance differences between the samples. It was observed that technical replicates were clustered together, which means that replicates were highly reproducible. In contrast to the high similarity in replicates, proteomes from different days were dissimilar and separated from each other. When only looking at microbial proteins, samples were separated in the order of time course and samples from adjacent days were clustered tightly. However, the pattern of human proteins was slightly different; for example, day 13 was distant from day 11 and 12, but clustered together with day 25. Although human and microbial proteins did not follow the same pattern, an obvious separation in day 25 on x-axis was observed for both of them, suggesting a possible major shift in the microbial functionality and corresponding host response during the time.

As microbial community composition shift rapidly over time, different sets of proteins were detected in each sample. The frequencies of identified proteins in 12 samples were

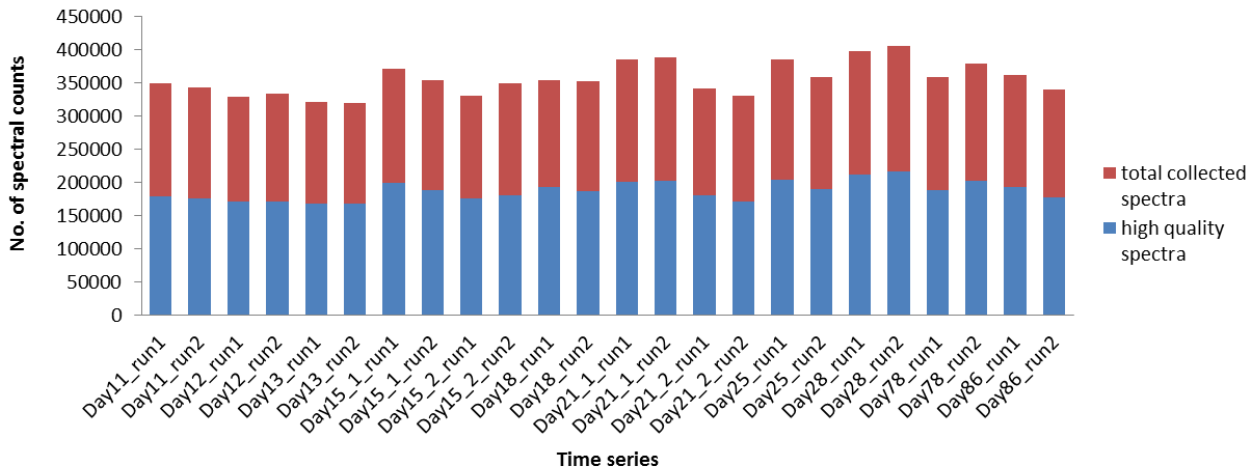


Figure 5.2. Number of total collected spectra and high quality spectra for each sample. High quality spectra were determined using ScanRanker tool.

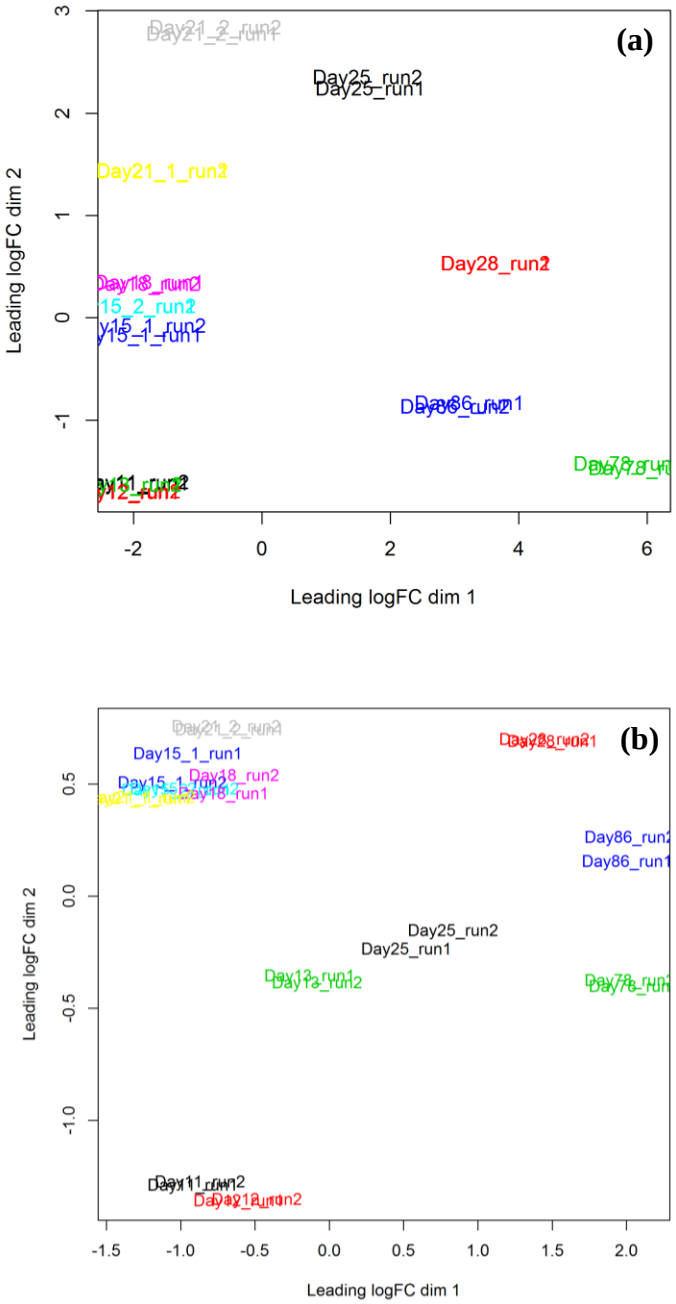


Figure 5.3. Multidimensional scaling (MDS) plot of 12 fecal proteomes for microbial proteins (a) and human proteins (b). MDS analysis was performed using the edgeR function plotMDS with log fold-change method estimated on spectral counts (normalized to library size). Color represented different fecal samples and technical replicates were labeled as the same color.

calculated in Table 5.2. Surprisingly, only 0.9% (109) of total identified proteins (12568) were identified in all 12 samples, including 94 human proteins and 15 microbial proteins. It is significant that over half total proteins were only identified once or twice and this percentage of uncommonly identified proteins was higher in microbial proteins (60%) as compared to human proteins (40%). A large number of unshared proteins made it challenging to compare changes in protein abundances across all time points.

5.4 Metagenome - metaproteome comparisons

One of the most important considerations for metaproteomic experiments is the biodiversity of the sample being analyzed. How many organisms are present and their relative abundance directly affect the proteome coverage of a single organism. A typical 24-h LC MS/MS experiment usually identifies a few thousand proteins regardless the proteome size, due to the constrained dynamic range and duty cycle of the mass spectrometer. Therefore, more complex community yields lower average proteome coverage and species with higher abundance or more active functionality tend to have a larger percentage of proteome that can be detected. Integrating strain-resolved metagenomics with deep metaproteomic measurements, we were able to characterize the proteome coverage of different species and strains across time, as shown in Figure 5.4. Since not all predicted proteins are expressed under one condition, a typical proteome of a single isolate can identify approximately 60% of the predicted proteome. Here, in total across all samples, up to 45% of the predicted proteins for an individual organism were obtained. As shown in the “TS” column in the figure, species were ranked according to the proteome coverage, with *Staphylococcus phage* the highest and *Veillonella sp.* the lowest percentage. Due to the rapid shift in the microbial composition over time, proteome specific to dominant species

Table 5.2. Frequencies of human and microbial proteins identified across time

Frequencies	# of proteins	Total	# of proteins	Human proteins (%)	# of proteins	Microbial proteins (%)
1	4308	34.3	863	26.6	3445	37
2	2551	20.3	450	13.8	2101	22.5
3	1833	14.6	330	10.2	1503	16.1
4	1000	8	245	7.5	755	8.1
5	682	5.4	265	8.2	417	4.5
6	697	5.5	220	6.8	477	5.1
7	466	3.7	212	6.5	254	2.7
8	340	2.7	210	6.5	130	1.4
9	263	2.1	160	4.9	103	1.1
10	196	1.6	114	3.5	82	0.9
11	123	1	87	2.7	36	0.4
12	109	0.9	94	2.9	15	0.2
Total	12568	100	3250	100	9318	100

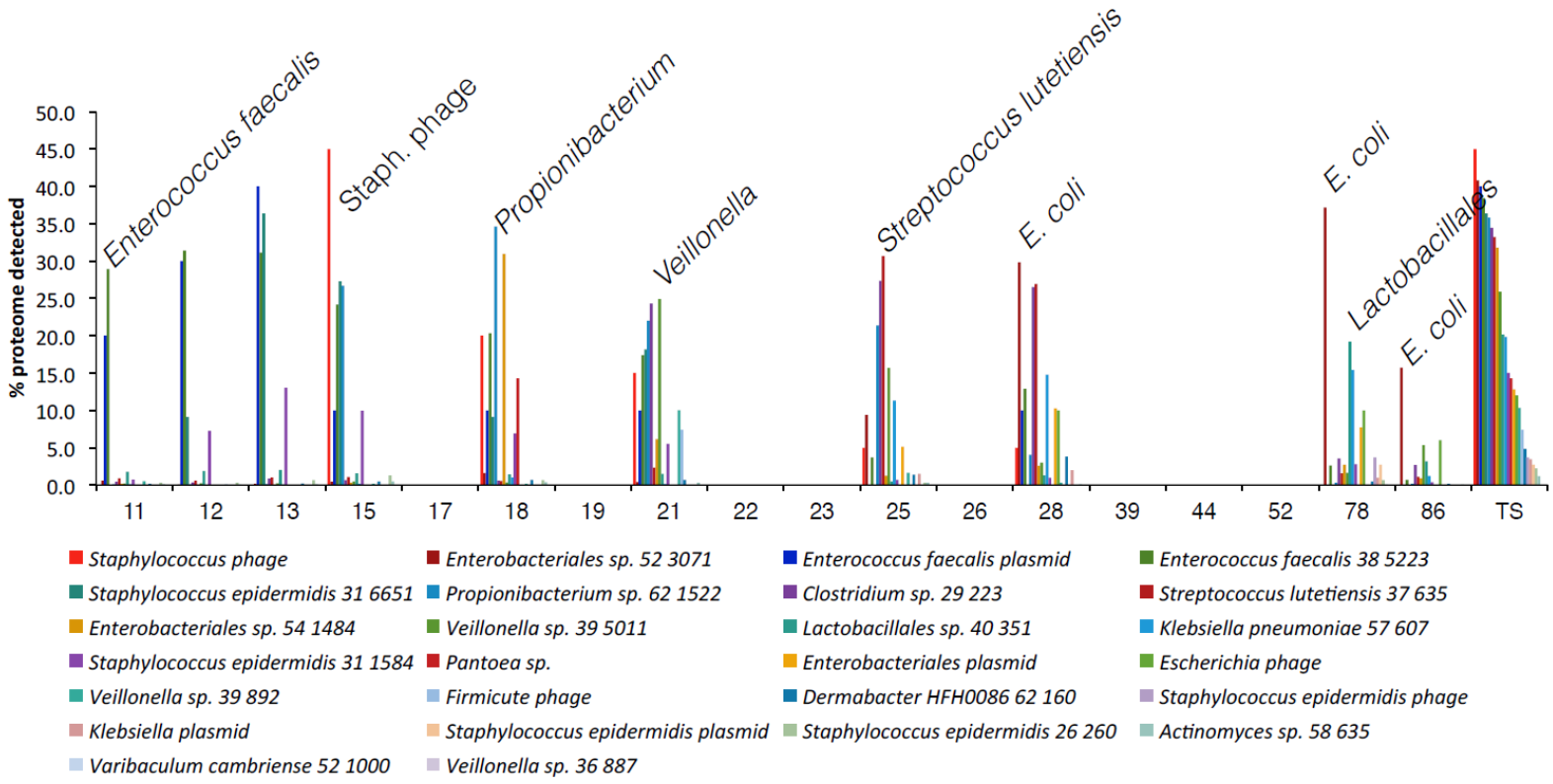


Figure 5.4. Organism-specific proteome coverage. Percentage of identified proteome was plotted for all microbial organisms at each time point being analyzed. Total proteome coverage across all samples was also shown in the last column. Different colors represented different species and figure legend was arranged in descending order of the proteome coverage. Species with highest proteome coverage within a sample was labeled on top of the figure. (Provided by Chris Brown)

and strains also shifted (as labeled on top of Figure 5.4). At early time points (days 11-13), the dominant proteome belonged to *Enterococcus faecalis* and this shifted to *Staphylococcus phage, Propionibacterium sp., Veillonella sp., Streptococcus lutetiensis* and *Escherichia coli* during the later time course.

To further investigate the activity of microbial community members, relative protein abundance was compared by assigning proteomic data to genomes (Figure 5.5 (b)). At days 11, 12, 13 and 15, proteomic data confirmed the dominance of *Enterococcus faecalis* and *Staphylococus phage* based on the relative abundance observed in the metagenomics data. Intriguingly, apparent differences were observed between the genomic and proteomic patterns on certain days, suggesting a few species that were more active in spite of low abundance, such as *Propionibacterium sp.* in day 18, *Clostridium sp.* and *Streptococcus lutetiensis* in day 25. By day 28, *Escherichia phage* began to increase their cell abundance, but however, this trend wasn't represented by their proteome abundances. The most active species in days 78 and 86 were recognized as *E. coli*, comprising 60% of relative protein abundances. These findings could have significant impacts on our understanding of dominant organism metabolic activities.

5.5 Global metabolic pathways of human proteome and gut metaproteome

To globally explore the metabolic activities in the infant gut microbiome, we next characterized and visualized microbial and human functionalities through KEGG pathways [186]. Figure 5.6 showed human proteome and gut metaproteome on a global metabolic map, where green lines indicated pathways identified by human proteins (only), red lines indicated pathways identified by microbial proteins (only) and blue lines indicated pathways identified by both. We

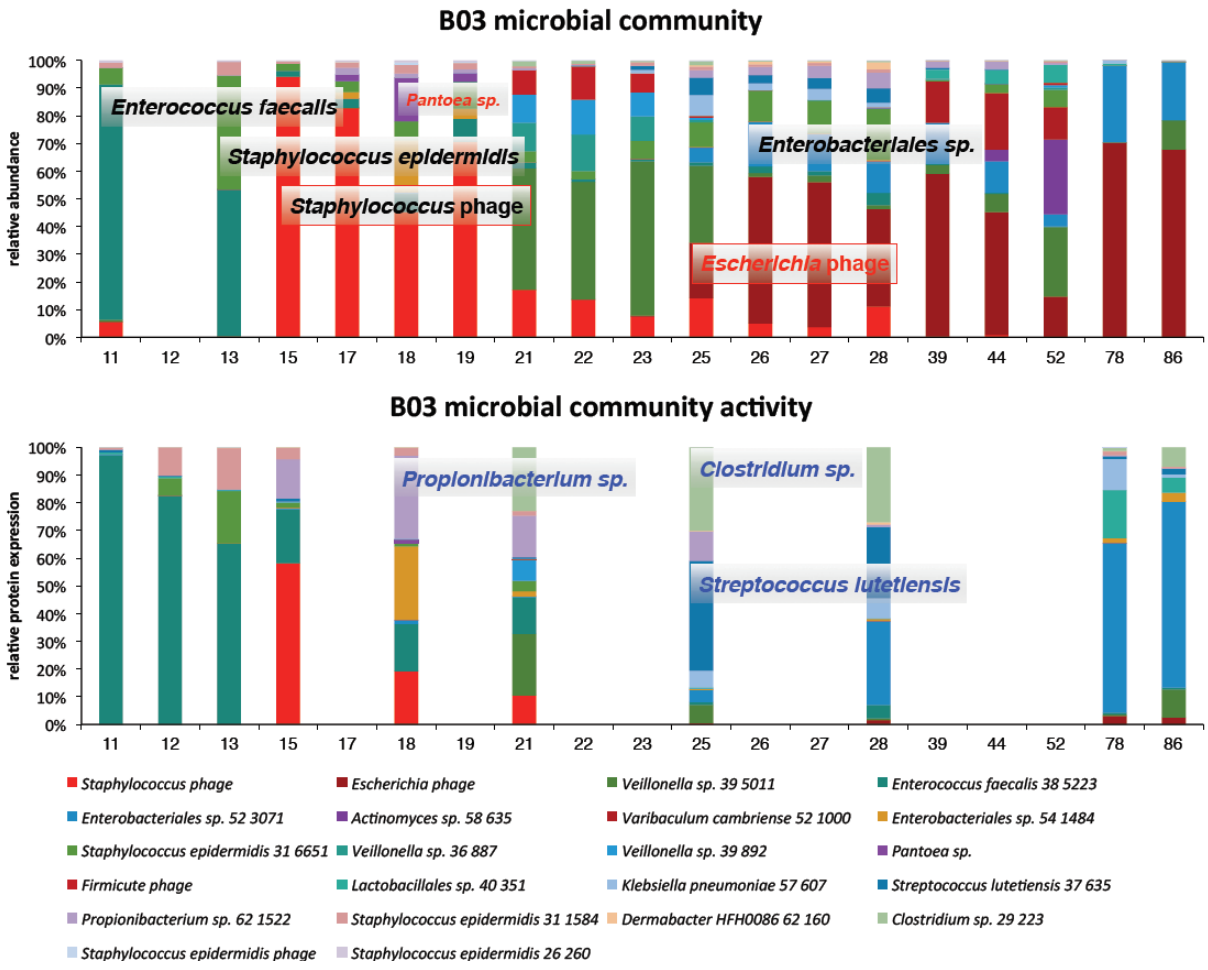


Figure 5.5. Pattern of changes in microbial abundance (a) and protein abundance (b). Relative proportion of reads (a) and summed spectral counts (b) of all proteins for all microbial organisms were plotted across time. (Provided by Chris Brown)

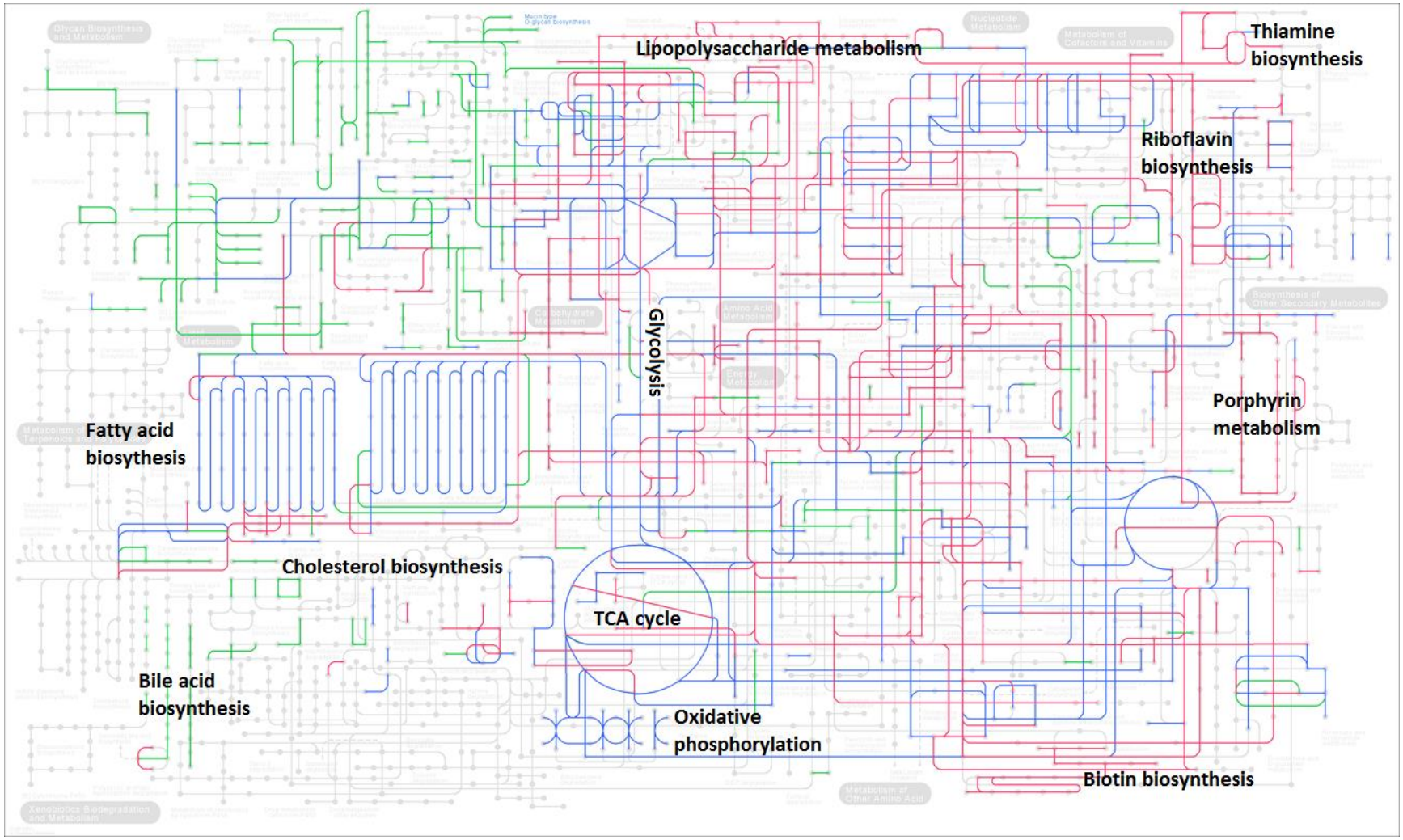


Figure 5.6. KEGG pathways mapping for human proteome and human gut metaproteome. Both human and microbial proteomes were mapped on KEGG pathways. Human proteome-only pathways were colored in green whereas microbial proteome-only pathways were colored in red. Overlapped pathways were colored in blue.

observed that pathways involving in glycolysis, citrate cycle, oxidative phosphorylation, fatty acid biosynthesis and nucleotide metabolism were commonly possessed by both human and microbiome. These are core metabolisms for both human and microbial communities, which support their respective cell growth and maintenance. However, thiamine (vitamin (V) B1), riboflavin (VB2), cobalamin (VB12) biotin (VB7), folate (VB9) and lipopolysaccharide (LPS) metabolisms were only found in the gut metaproteome, whereas pathways involving in tight junctions, regulation of actin cytoskeleton, mucin type O-glycan biosynthesis, and complement and coagulation cascades were only shown in the human proteome.

Gut bacteria synthesize B vitamins that are essential nutritional factors for human, especially the gut health [64]. Various types of vitamin B production were detected in the early bacterial colonization of the infant gut, supporting human gut nutrients. On the other hand, in response to the bacterial colonization, human host express tight junctions, actin cytoskeleton and mucins that play pivotal roles in the integrity and barrier properties of mucosal epithelial layers [187]. Dysbiosis of the intestinal mucosal barrier can lead to the pathological bacterial translocation and the initiation of an inflammatory response in the intestine. In addition, bacterial LPS and human host complement system have been suggested participating in the maturation of human innate immunity [188, 189]. Studies have also suggested that the premature infants are predisposed to intestinal inflammation due to the immature response to bacteria and therefore this initial interaction is potentially important in the development and maturation of host immune system. Overall, these observations have demonstrated that human host and gut bacteria begin to cooperate on metabolic activities during infancy, which benefit each other and initiate the establishment of a mature and healthy gut though the gut microbiome-host metabolic crosstalk.

5.6 Microbial functional characterization

To explore the establishment and changes of microbial metabolic activities over time, the metagenome was annotated with gene ontology (GO) information using the Blast2GO platform. Of 38,192 protein sequences predicted from the metagenome, 27,286 sequences were associated with at least one GO term. Of 9,318 microbial proteins identified from 12 samples, 8,307 proteins were annotated with top GO terms shown in Figure 5.7. The analysis showed that the microbiome mainly involved in the biological processes of metabolic, cellular, localization and response to stimulus processes, the molecular functions of binding, catalytic and transporter, and the components of cell and membrane part.

To further investigate the pattern of microbial proteins over the time course, protein abundances were analyzed using hierarchical clustering, revealing four clusters with similar expression patterns (Figure 5.8): cluster I (days 11, 12 and 13), cluster II (days 15_1, 15_2, 18 and 21_1), cluster III (days 21_1 and 25), and cluster IV (days 28, 78 and 86). It was noted that technical replicates showed high similarity and samples from adjacent time points were closely clustered. However, two samples collected on day 21 showed notably different expression profiles, suggesting major changes occurred between the two samples and also emphasizing the need for high-resolution sampling. Therefore, we compared the relative protein abundance of two samples and found out that the introduction of a new dominant colonizer – *Clostridium sp.* in the second sample was mainly responsible for the changes (Figure 5.9), although the cause of the shift wasn't clear. It was possible that the colonization of *Clostridium sp.* disrupted the metabolism of other bacteria and therefore significantly impacted their protein expression profiles. An obvious separation observed at day 25 in previous MDS results might be also due to this dominant colonization of *Clostridium sp.*

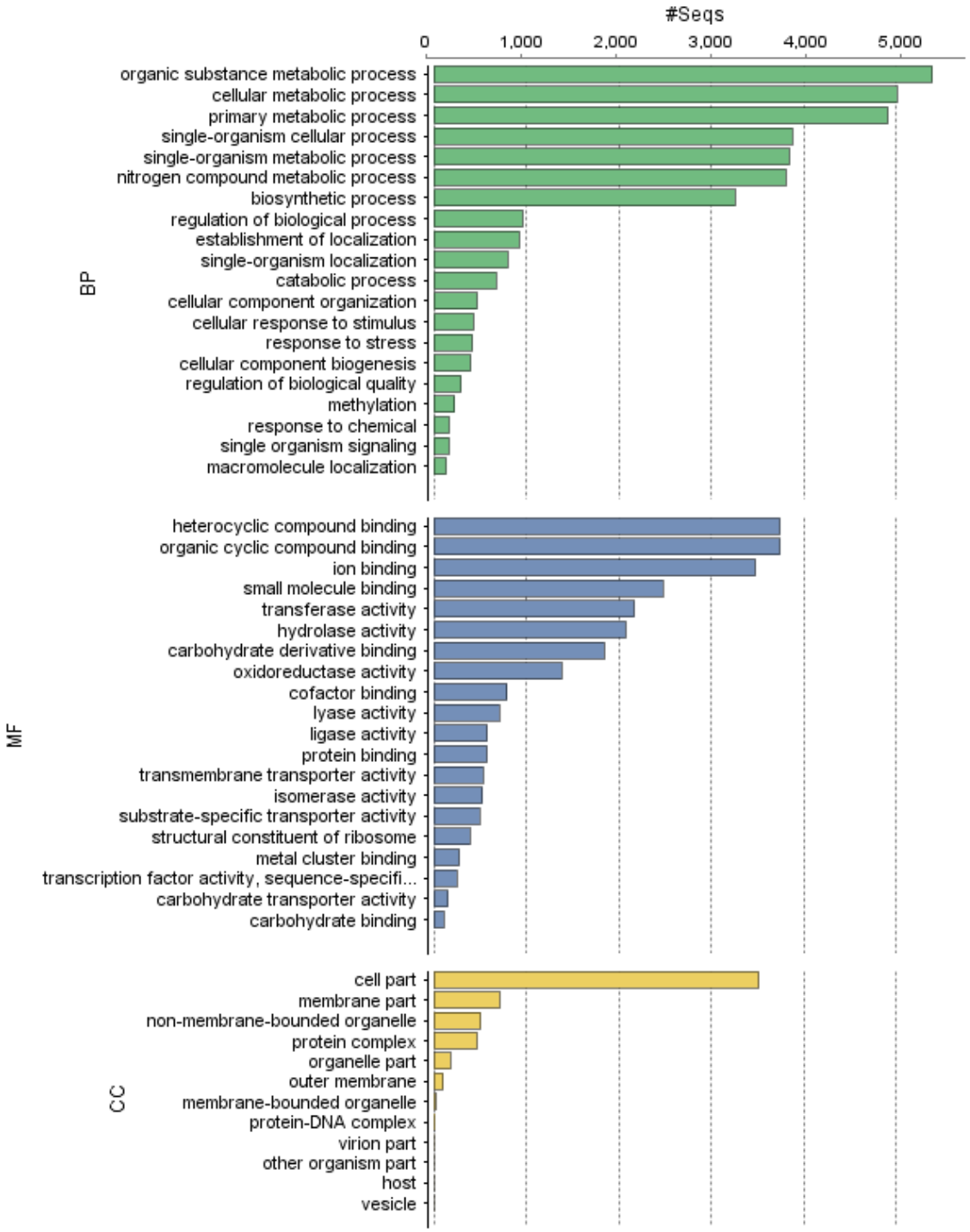


Figure 5.7. Microbiome GO term distributions at level 3 of biological process (BP), molecular function (MF), and cellular component (CC).

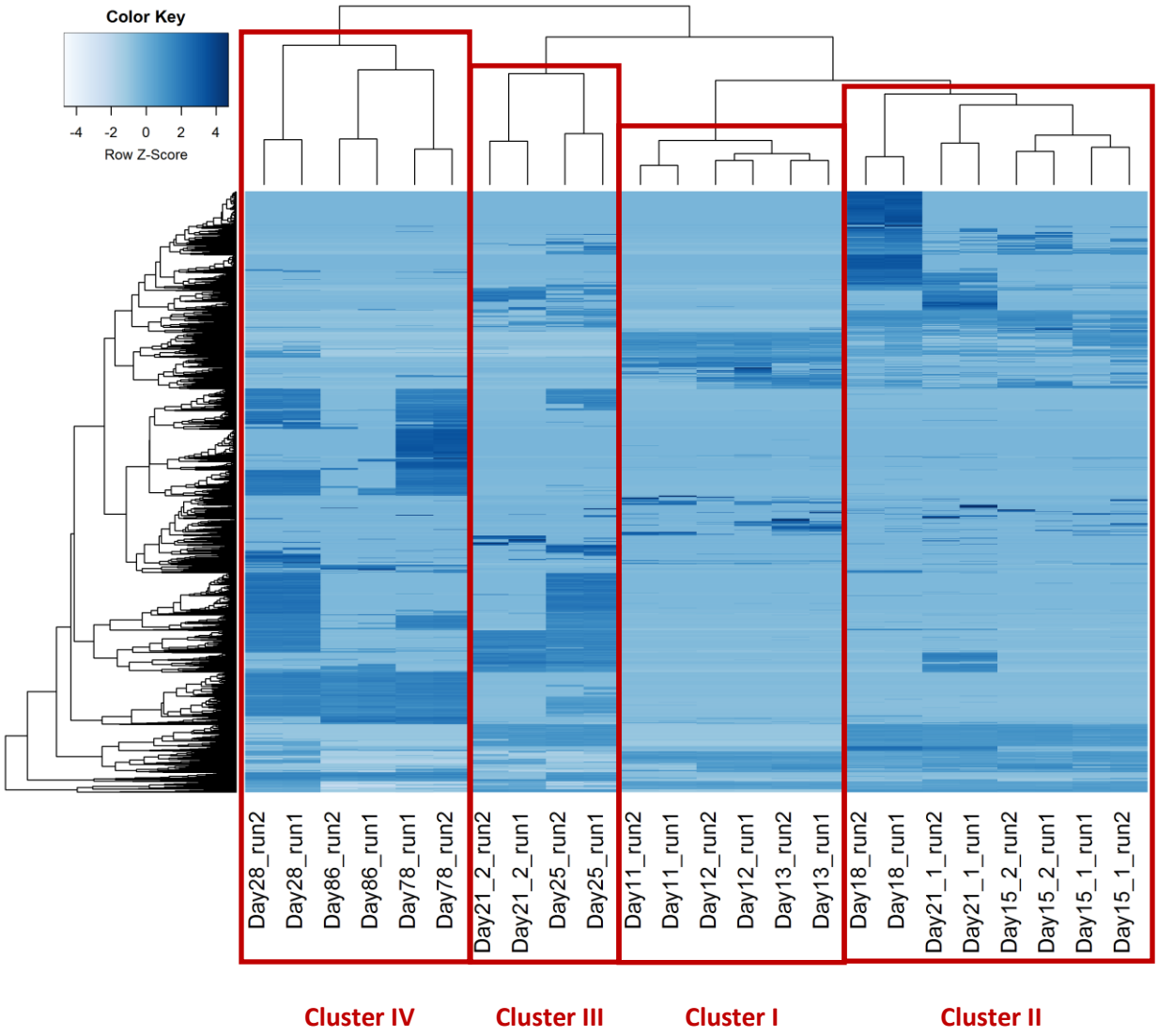


Figure 5.8. Hierarchical clustering of microbial proteins. 6137/12568 total detected microbial proteins were included based on having 100 counts per million (cpm) for at least one sample/library. The dataset was normalized based on scaling factors for library sizes. Heatmap was generated using moderated log-counts-per-million (log2 counts-per-million).

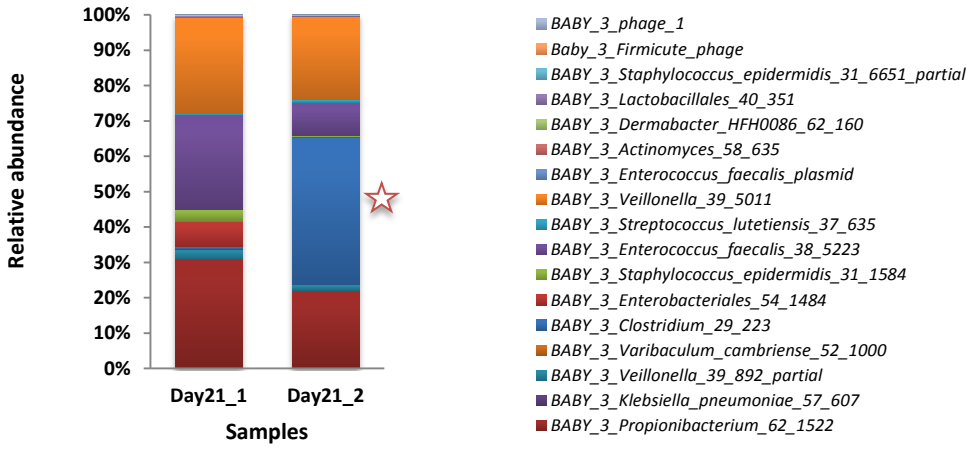


Figure 5.9. Comparisons of relative protein abundance in two samples collected on the same day. Relative protein abundance was calculated for two samples from day 21. The relative abundance of *Clostridium sp.* (as labeled in the figure) was found dramatically different between two samples.

GO enrichment analysis of identified proteins was performed for each cluster by using Fisher's Exact Test with Multiple Testing Correction of FDR (Benjamini and Hochberg) at a cutoff of 0.05 (Table 5.3). Total microbial proteins identified from all 12 samples were used as reference and enriched GO terms were reduced to most specific terms. Glycolytic process and translation were enriched in all four clusters as energy production and protein synthesis are fundamentally important in supporting bacterial cell growth. Phosphotransferase system (PTS) including glucose/fructose/mannose/lactose specific components, was significantly enriched in cluster I. PTS is a bacterial-specific method for transporting sugars into cell using the energy source from phosphoenolpyruvate. It was found to be increased in infants fed with exclusive breast milk as compared to non-exclusively breastfed infants [190]. Enrichment of PTS may indicate an increasing processing of carbohydrates in the early time course.

Enrichments in cobalamin, purine (particularly GMP) and pyrimidine (particularly UMP) biosynthetic process were observed in cluster III. Vitamin B12 (cobalamin) production is only found in bacteria and archaea, and is an essential cofactor for anti-inflammation and neurological function [191]. Previous studies have observed a decrease in the cobalamin synthesis in patients with inflammatory bowel disease (IBD) [192] and therefore, VB12 production may play key roles in the early community colonization. Glutamine metabolic process was also enriched in cluster III, providing nitrogen donor for the synthesis of purine and pyrimidine nucleotides, which play important roles in cell signaling, energy metabolism and nucleic acids (DNA and RNA) formation.

Anaerobic respiration was significantly enriched in the late time course (cluster IV). Although aerobic respiration wasn't enriched in the entire period, proteins involved in TCA cycle, electron transport chain and terminal oxidase, such as NADH: ubiquinone oxidoreductase,

Table 5.3. Enriched BP GO terms among different clusters

Cluster	GO-ID	Term	FDR	P-Value	# of identified	# of total
Cluster 1	GO:0006096	glycolytic process	9.09E-05	1.59E-07	51	145
	GO:0009401	phosphoenolpyruvate-dependent sugar phosphotransferase system	2.21E-04	1.07E-06	50	149
	GO:0006412	translation	2.41E-04	1.22E-06	190	804
	GO:0034219	carbohydrate transmembrane transport	5.47E-04	3.68E-06	42	122
	GO:0043335	protein unfolding	3.03E-02	5.53E-04	6	8
Cluster 2	GO:0006412	translation	1.62E-06	3.49E-10	445	804
	GO:0006096	glycolytic process	5.39E-03	4.54E-05	89	145
Cluster 3	GO:0006418	tRNA aminoacylation for protein translation	5.54E-11	2.46E-13	135	215
	GO:0006096	glycolytic process	3.29E-09	2.80E-11	95	145
	GO:0009236	cobalamin biosynthetic process	8.30E-04	1.81E-05	32	46
	GO:0046129	purine ribonucleoside biosynthetic process	2.38E-03	6.00E-05	99	189
	GO:0015991	ATP hydrolysis coupled proton transport	5.77E-03	1.59E-04	25	36
	GO:0006177	GMP biosynthetic process	8.36E-03	2.42E-04	17	22
	GO:0006222	UMP biosynthetic process	9.94E-03	3.05E-04	43	73
	GO:0006541	glutamine metabolic process	2.27E-02	7.80E-04	48	86
	GO:0006006	glucose metabolic process	2.27E-02	7.98E-04	54	99
	GO:0006450	regulation of translational fidelity	2.88E-02	1.03E-03	29	47
Cluster 4	GO:0006412	translation	1.37E-06	1.24E-08	554	804
	GO:0006096	glycolytic process	2.56E-05	3.81E-07	115	145
	GO:0009408	response to heat	5.89E-05	9.67E-07	47	52
	GO:0044262	cellular carbohydrate metabolic process	1.17E-04	1.97E-06	236	328
	GO:0042402	cellular biogenic amine catabolic process	9.58E-04	2.01E-05	21	21
	GO:0042710	biofilm formation	5.54E-03	1.58E-04	17	17

Table 5.3. Continued

Cluster	GO-ID	Term	FDR	P-Value	# of identified	# of total
Cluster 4	GO:0070887	cellular response to chemical stimulus	5.90E-03	1.74E-04	45	54
	GO:0006979	response to oxidative stress	7.46E-03	2.28E-04	70	90
	GO:0046677	response to antibiotic	7.97E-03	2.45E-04	44	53
	GO:0051172	negative regulation of nitrogen compound metabolic process	8.35E-03	2.63E-04	76	99
	GO:0019541	propionate metabolic process	8.35E-03	2.65E-04	16	16
	GO:0009062	fatty acid catabolic process	8.35E-03	2.65E-04	16	16
	GO:0019320	hexose catabolic process	1.31E-02	4.32E-04	26	29
	GO:0015949	nucleobase-containing small molecule interconversion	1.34E-02	4.44E-04	15	15
	GO:0009061	anaerobic respiration	1.45E-02	4.89E-04	19	20
	GO:0006066	alcohol metabolic process	1.49E-02	5.06E-04	86	115
	GO:0033036	macromolecule localization	1.84E-02	6.51E-04	99	135
	GO:1901616	organic hydroxy compound catabolic process	1.88E-02	6.69E-04	41	50
	GO:0034310	primary alcohol catabolic process	2.07E-02	7.43E-04	14	14
	GO:0006457	protein folding	2.15E-02	7.83E-04	106	146
	GO:0045892	negative regulation of transcription, DNA-templated	2.58E-02	9.55E-04	56	72
	GO:0009267	cellular response to starvation	3.27E-02	1.24E-03	13	13
	GO:0009893	positive regulation of metabolic process	3.47E-02	1.33E-03	34	41
	GO:0034622	cellular macromolecular complex assembly	4.01E-02	1.56E-03	31	37
	GO:0006970	response to osmotic stress	4.97E-02	1.97E-03	16	17

succinate dehydrogenase, cytochrome c oxidase, were identified throughout the dataset, indicating aerobic growth pathway was continuously active during the colonization. Proteins participating in the anaerobic respiration were identified in cluster IV, for example fumarate, dimethyl sulfoxide and nitrate reductase, suggesting that the gut microbiome shifted to facultative anaerobic state in the late time period as the availability of oxygen decreased. Enriched propionate metabolic process, particularly the anaerobic degradation of L-threonine to propionate by L-threonine dehydratase catabolic TdcB and pyruvate formate-lyase (PFL) – like enzyme TdcE, was also strong evidence for the anaerobic respiration of the community.

Another GO term enriched in cluster IV was hexose catabolic process, resulting in the breakdown of hexose, for example fucose, rhamnose, mannose and galactose. In particular, proteins such as fucose/rhamnose isomerase, fuculokinase, and fucose/rhammulose phosphate aldolase were detected for fucose and rhamnose degradation, which might be further converted to propionate. Propionate is a SCFA that has potential anti-inflammatory function and maintains human gut health [193]. Oligosaccharides are major components of human milk and can be digested into SCFA by intestinal bacteria. SFCA provides energy for intestinal epithelial cells and lower intestinal pH thus reducing potential pathogens. However, overproduction of SCFA has also been suggested in the pathogenesis of NEC in premature infants due to the excessively produced SCFA that can’t be absorbed promptly and injure the vulnerable intestinal mucosa [194]. Besides propionate, acetate and butyrate are two other principal SCFAs produced by gut microbiota. The fermentation processes of these two SCFAs were not enriched in our study, but acetate kinase, butyryl-CoA dehydrogenase, butyrate kinase were detected in days 25-86, and days 21-28 respectively.

Interestingly, a series of responses to stimulus, heat, oxidative stress, antibiotics, starvation and osmotic stress were enriched in cluster IV, suggesting the community was under various environmental stresses and adapted to the changing conditions. As observed in the metagenomics information, there was an *Escherichia phage* dominating the gut microbiota during this time period. Phage shock protein was detected in response to the phage infection and might play significant roles in bacterial survival for the competition of limited nutrients and energy [195]. Additionally, heat stress and reactive oxygen species (ROS) generated in the host intestine during inflammation can activate bacterial stress response and induce the expression of protective antioxidant proteins [196]. Indeed, many antioxidant proteins such as superoxide dismutase, alkyl hydroperoxide reductase, catalase, thioredoxin reductase, glutathione peroxidase and peroxiredoxin were abundantly identified. Studies have also shown that bacteria can form biofilm in response to environmental stresses [196]. The formation of biofilm allows bacteria to grow in close association with host cells and more importantly, provides bacteria resistance to antimicrobials. Therefore, enrichment of biofilm formation process in this cluster might also be raised by the bacterial stress response.

5.7 Human host response changing across time

Although a small number of proteins were shared for all samples, 63 human proteins were identified in all samples across time. David Bioinformatics Resources (http://david.abcc.ncifcrf.gov/home.jsp) was employed to investigate the functionality of these core proteins. Activities involved in defense response, inflammatory response and acute inflammatory response were enriched, suggesting that early life gut maintains a tuned inflammation that is important to the interaction between human host and gut microbes. Top 20

abundant proteins were shown in Table 5.4. Lactotransferrin (LTF) was the most abundant protein throughout all samples. As the major whey in human milk, LTF exhibits antibacterial activity and may provide benefits in the prevention of NEC in premature neonates [197]. Proteins related to digestion, such as chymotrypsin-C (CTRC) and chymotrypsin-like elastase family (CELA3A/3B) were found among the most abundant proteins. Other abundant proteins were involved in gut mucosal barrier protection, such as intelectins (ITLN1, ITLN2) that recognize pathogen-associated glycans [198] and intestinal alkaline phosphatase (ALPI) dephosphorylating bacterial LPS [199]. Secretory IgA present in mucosal surface potentially contribute to gut mucosal defense and its transport across epithelial cells is dependent on the polymeric immunoglobulin receptor (PIGR), which was also abundant in these fecal samples [200]. In addition, IgG Fc-binding protein (FCGBP) and calcium-activated chloride channel regulator (CLCA1) are significantly related to the production and maintenance of the mucosal structure [201]. Also, complement C3, S100 calcium binding protein A9 (S100A9) and glycoprotein deleted in malignant brain tumors 1 (DMBT1), also known as glycoprotein 340 (gp-340) may participate in the innate immune response in the intestine [202, 203].

To further investigate the expression pattern of human proteins over the time course, human proteins were hierarchically clustered using protein abundance and 10 clusters were identified (Figure 5.10). Each cluster was enriched for Gene Ontology using David Bioinformatics Resources [204]. In general, human proteins were more abundant in the first month (before day21) than in later time points due to the increasing microbial colonization. No terms were enriched in Cluster #1 and #9 probably due to a small number of proteins in these two clusters. Cytoskeleton organization was enriched in Cluster #10 (containing proteins mainly predominant in days 11 and 12), suggesting the early development of the intestinal epithelial

Table 5.4. Top 20 abundant human proteins in 12 fecal samples

Gene symbol	Protein description	Summed Spectral counts
LTF	Lactotransferrin	95440.46
FCGBP	IgGFc-binding protein	57425.85
CLCA1	Calcium-activated chloride channel regulator	56090.18
CTRC	Chymotrypsin-C	23883.97
DMBT1	Deleted in malignant brain tumors-1	21064.67
CELA3B	Chymotrypsin-like elastase family	18633.93
PIGR	Polymeric immunoglobulin receptor	17731.18
IGHA1	Immunoglobulin heavy constant alpha-1	15774.32
ALPI	Intestinal alkaline phosphatase	15151.16
C3	Complement C3	12527.15
ITLN1	Intelectin-1	10702.16
ALB	Serum albumin	10544.29
MME	Membrane metallo-endopeptidase	8914.46
XDH	Xanthine dehydrogenase	8525.57
VDAC1	Voltage-dependent anion channel-1	7391.13
ACTA1	actin, alpha-1, skeletal muscle	7300.23
SERPINA1	alpha-1 antitrypsin	7096.47
MTTP	Microsomal triglyceride transfer protein	6378.62
S100A9	S100 calcium binding protein A9	6172.59
ITLN2	Intelectin-2	5735.22

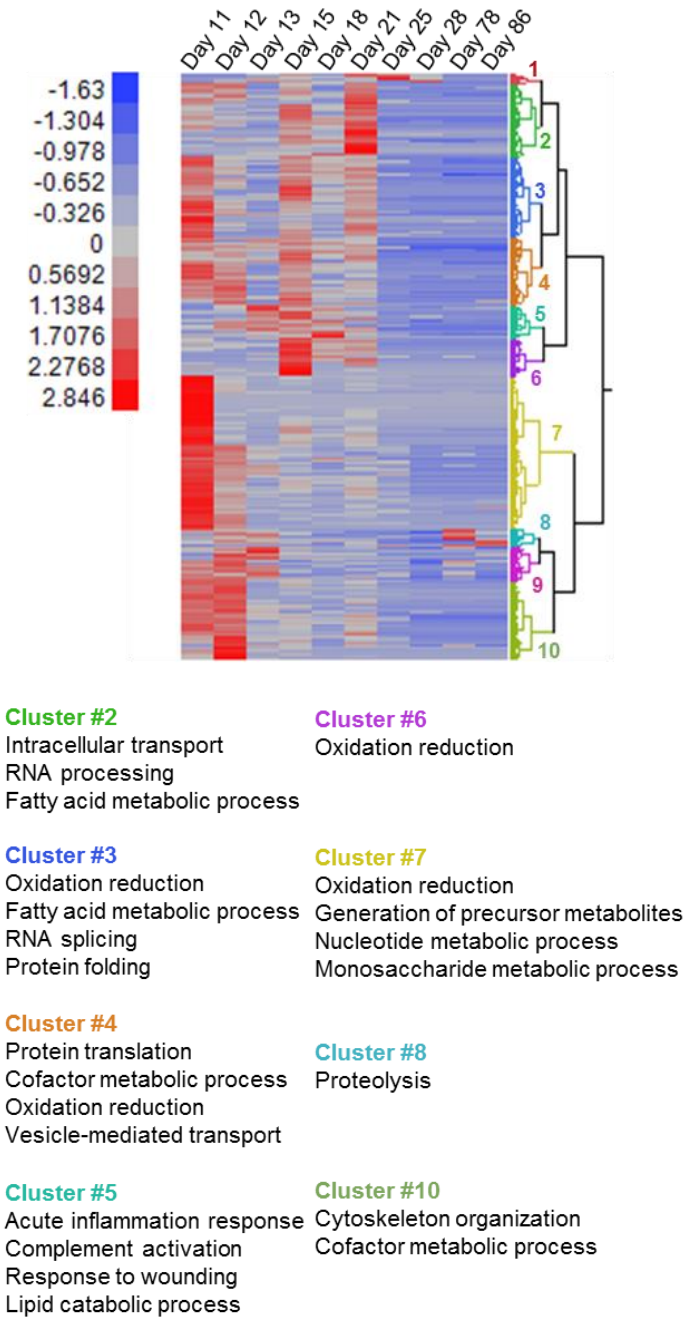


Figure 5.10. Changes of human proteome across time. Proteins were hierarchically clustered based on protein abundance changes. Each cluster was enriched for GO functions with p value less than 0.01.

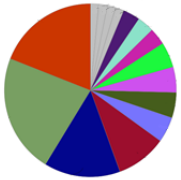
barrier, which prevents penetration of pathogenic microbes into the mucosa and submucosa. Complement activation and acute inflammation response were observed in Cluster #5 (containing proteins more abundant in days 13, 15 and 18), suggesting infection might occur during these days.

We employed ClueGO, a Cytoscape plug-in that provides representative GO terms for a large set of genes, to explore primary functions of human proteins in response to microbial colonization across the time course. Top 200 abundant human proteins of each sample were analyzed and enriched for GO terms and top 5 significantly enriched terms were listed in Figure 5.11. Most of terms contributed to human cell growth and activities such as carbohydrate metabolism and oxidation-reduction process, especially at early time points. Interestingly, almost all terms enriched in days 28, 78 and 86 were involved in inflammatory responses to microbes, which might give an explanation to bacterial stress responses observed in the late phase. The inflammation produces antibacterials, elevated ROS and heat that help the host compete with and defend pathogens.

5.8 Conclusions

By employing previously established high-performance mass spectrometry based metaproteomics approach, we achieved deep proteome measurement for both human and microbial proteins in a longitudinal study of a healthy infant gut microbiome. We identified a total of 9318 microbial and 3250 human protein groups from 12 fecal samples across all time points. It was demonstrated that human proteins were relatively abundant in early time points, and then reduced as microbes colonized rapidly and the complexity of microbial composition

Figure 5.11. GO enrichment of human proteome over time. Top 200 abundant human proteins were subjected for GO enrichment. Pie charts showed all significantly representative terms and the size/area of the slice represented the significance of a GO term that was enriched. TOP 5 significantly enriched GO terms were listed out.



Day 11

- Carboxylic acid metabolic process
- Oxidation-reduction process
- Respiratory electron transport chain
- Glycolysis
- Maintenance of location in cell

Day 12

- Oxidation-reduction process
- Monocarboxylic acid metabolic process
- Cell respiration
- Digestion
- Maintenance of location in cell

Day 13

- Acute inflammatory responses
- Regulation of coagulation
- Defense response to fungus
- Digestion
- Coagulation

Day 15

- Generation of precursor metabolites and energy
- Oxidation-reduction process
- COPI coating of Golgi vesicle
- Acute inflammatory responses
- Negative regulation of myeloid cell apoptotic process

Day 18

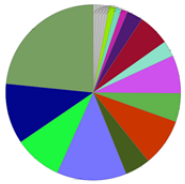
- Oxidation-reduction process
- Regulation of protein activation cascade
- Response to inorganic substance
- Lipid digestion
- Digestion

Day 21

- Oxidation-reduction process
- COPI coating of Golgi vesicle
- Monocarboxylic acid metabolic process
- Protein folding
- Citruline biosynthetic process

Day 25

- Monocarboxylic acid metabolic process
- Defense response to fungus
- Regulation of coagulation
- Oxidation-reduction process
- Digestion



Day 28

- Anion transport
- Acute inflammatory response
- Defense response to inorganic substance
- Response to bacterium
- Oxidation-reduction process

Day 78

- Defense response to fungus
- Defense response to bacterium
- Acute inflammatory responses
- Regulation of coagulation
- Regulation of endopeptidase activity

Day 86

- Protein processing
- Acute inflammatory responses
- Cellular respiration
- Response to inorganic substance
- Response to zinc ion

increased. Until about one month, the number of human proteins remained stable. The organism-specific pattern of microbial protein abundance was different from that of their genome abundance, indicating that a few species were functionally active despite low abundance and thus emphasizing the role of proteomic measurements in the functional characterization.

Starting from the infancy, human and microbes cooperated on metabolic activities in the development of gut and the maturation of immune system. Four clusters/phases were identified based on the protein expression pattern of the microbiome. Major functional shift might be related to the transition of community respiratory mode from aerobic to facultative anaerobic. Core metabolism of microbial community established early in support of microbial cell growth and maintenance. As the complexity increased, more activities such as vitamin production and short chain fatty acid metabolism were observed. It seemed that the community went through environmental stresses during the late phase, possibly raised by the phage infection or host inflammatory responses. On the other hand, in response to the microbial colonization, the first (epithelial barrier) and the second (innate immunity) line of human immune defense was under development. The complement activation and inflammatory responses indicated in certain days suggested that a possible infection occurred.

CHAPTER 6

Characterization of temporal and inter-individual functional differences in infant gut microbiome by metaproteomics approach

6.1 Introduction

Premature infants may face a number of health problems including breathing and respiratory difficulties, feeding and digestive problems and neurological and psychiatric problems, due to underdeveloped organs and systems. As compared to term infants, premature infants typically harbor delayed and less complex microbial communities [89]. Although not fully understood, the initial colonization has been related to a number of factors, for example, gestational age, birth weight, delivery mode, feeding, use of antibiotics, and host health status [89]. Therefore, huge variations of microbial composition have been observed among premature infants. It has been reported that NEC and sepsis, conditions primarily seen in premature infants, can result in differential microbiome development in premature infants. However, no single species or microbial pattern has been identified as causative agent [176].

In Chapter 5, we have demonstrated that microbial metaproteomics provided the ability to characterize metabolic activities for both the community and human host at a remarkably deep level in a healthy premature infant. To further explore inter-individual viabilities and possible functions that are associated with NEC, in this chapter, we will include three more premature infants, including two infants from a triplet set (one developed NEC and didn't survive, one didn't develop NEC but had severe sepsis) and one healthy infant co-hospitalized with the two

above infants. We aimed to explore the longitudinal and individual functional variations in the gut microbiome and host response among infants.

6.2 Materials and methods

Sample collection. In addition to the infant #3, fecal samples were collected from three more preterm infants (#19, #21 and #23) over the first three months after birth. Infants #19 and #21 were two infants from triplets, among which, #19 developed severe sepsis but not NEC while #21 developed NEC and died from NEC totalis. Infant #23 was co-hospitalized with infants #19 and #21, who was healthy aside from some mild lung disease. Additional medical details of four infants were shown in Table 6.1. Fecal samples from infant #19 (on days 12, 16, 20, 26, 31, 38 and 56), infant #21 (on days 13, 18, 21, 24, 27 and 30), and infant #23 (on days 15, 18, 21, 34, 50) were collected for metaporteomics analysis.

Sample preparation and measurement. Sample preparation and proteomic measurements were performed using the same methods described in Chapter 5.

Data processing and analysis. Protein database were constructed for each individual infant by combining proteins predicted from sequenced metagenome, human proteins and contaminants. All database searching, peptide matching and protein inference were processed by the same methods described in Chapter 5. Data were further analyzed using edgeR package [182], KEGG database [186], and Blast2GO [183] with parameters also described in the previous chapter. The Venn diagram was generated using the Venny tool. (http://bioinfogp.cnb.csic.es/tools/venny/index.html)

Table 6.1. Summary of infant medical information

Infants	***#3***	***#19***	***#21***	***#23***
Gestational age (weeks)	26	24	24	27
Gender	Female	Female	Female	Female
Delivery mode	C-section	C-section	C-section	Vaginal
Birth weight (g)	822	731	697	875
Feeding	Breast milk	Combination	Breast milk	Breast milk
Health status	Healthy	Sepsis	NEC and died	Healthy
Antibiotics use	Initial 7-day treatment with Ampicillin/ Gentamici; Day 51-63 with Vancomycin and Cefotaxime	Initial 7-day treatment with Ampicillin/ Gentamici; Day 23-31 with Vancomycin, Claforan, Nafcillin and Gentamycin	Initial 7-day treatment with Ampicillin/ Gentamici; Day 24-32 with Gentamycin	Initial 7-day treatment with Ampicillin/ Gentamici

6.3 General overview of metaproteomic datasets

Fecal metaproteomes of three more preterm infants (#19, #21 and #23) were measured on multiple time points by metaproteomics approach (Figure 6.1). A total of 9665 (7397 microbial and 2268 human), 7091 (6349 microbial and 742 human), and 11649 (10330 microbial and 1319 human) protein groups were identified for infant 19, 21 and 23 respectively. In the proteomes of baby 19, we observed an increasing microbial load and a decrease of identified human proteins over time, a similar trend as shown in baby 3, with the exception of a dramatic decrease of microbial proteins at day 26. This was coincided with the antibiotics use at day 26, suggesting that the antibiotics effectively suppressed or removed the microbiota. However, proteomes of baby 21 didn't follow this pattern but showed relatively stable number of identified human and microbial proteins. In addition, the number of human proteins was much lower than microbial proteins across the time. Although the number of human proteins was also less than microbial proteins in baby 23, it decreased as the community complexity increased. We also examined the relative abundance between the human and microbiome in all infants across the time course (Figure 6.2) and similar trends were observed in the relative abundance as in the number of identified proteins.

To discern the temporal development of human and microbial proteins among infants, we plotted samples in two dimensions using the multidimensional scaling (MDS) (Figure 6.3 and 6.4). Technical replicates were overlapped with each other, indicating the high reproducibility. For microbial proteins, in general, samples were separated from each other according to the temporal order. In the infant #19, days 16, 20, 31 and 38 were closely plotted but distant from days 12 and 56, which were the first and last day of the time course. However, day 26 was far away from any other days, might be related to the use of antibiotics. Interestingly, after the

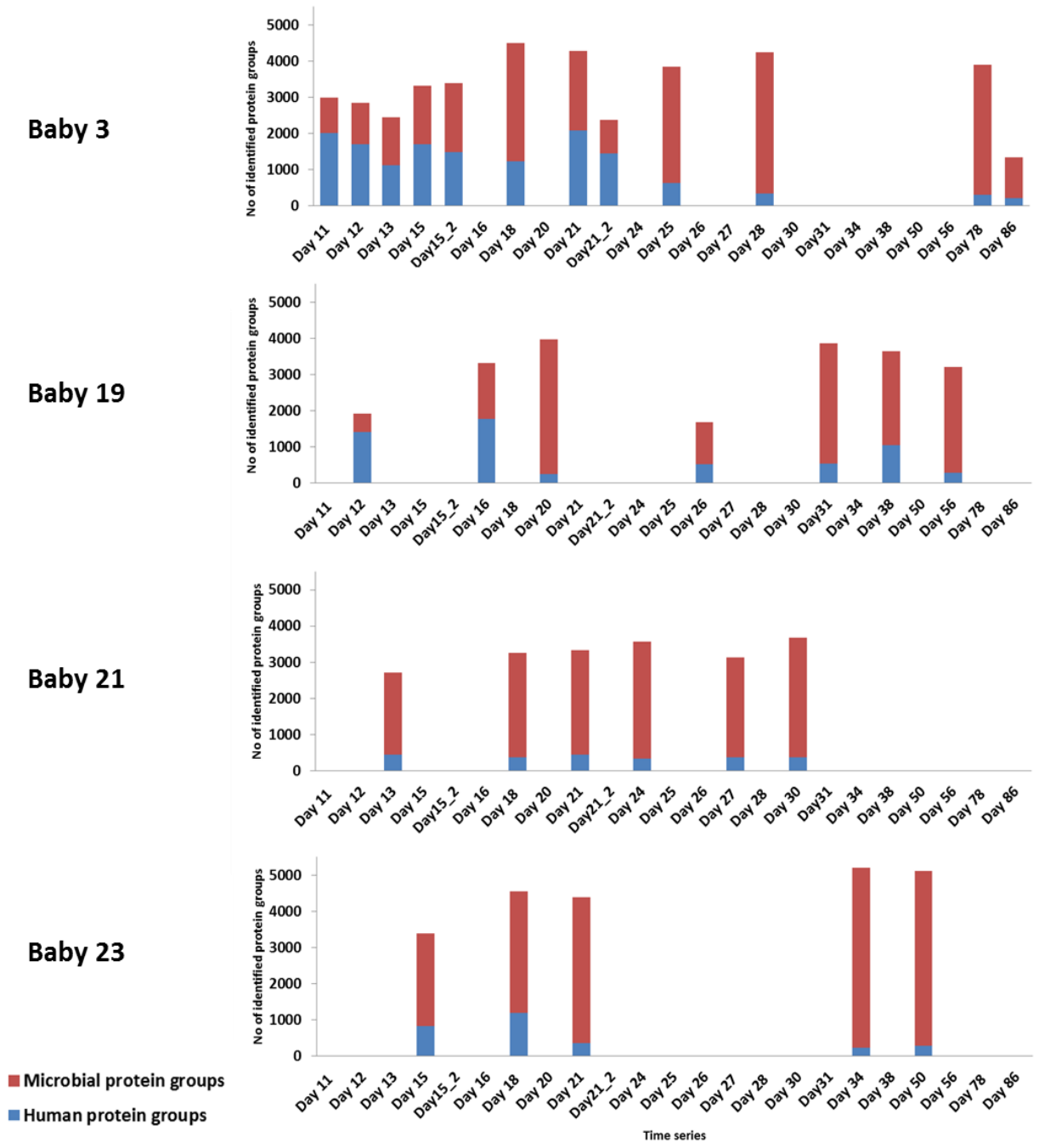


Figure 6.1. Number of identified human (blue) and microbial (red) protein groups of four infants over time. The number indicated in the figure included protein groups detected from both duplicate runs.

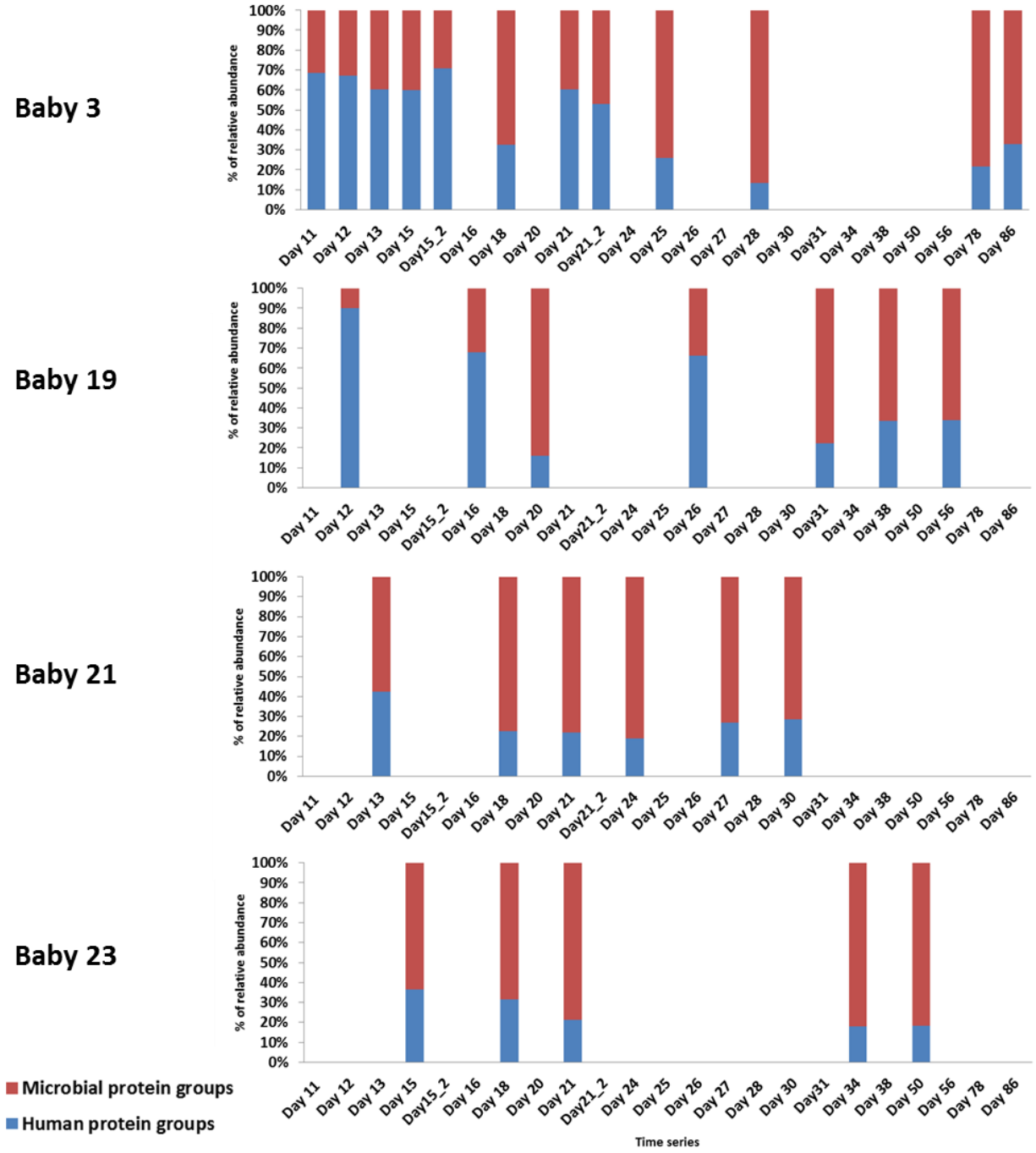


Figure 6.2. Relative abundance of human (blue) and microbial (red) protein groups of four infants over time. Spectral counts were normalized by the number of total collected spectra and averaged between duplicate runs.

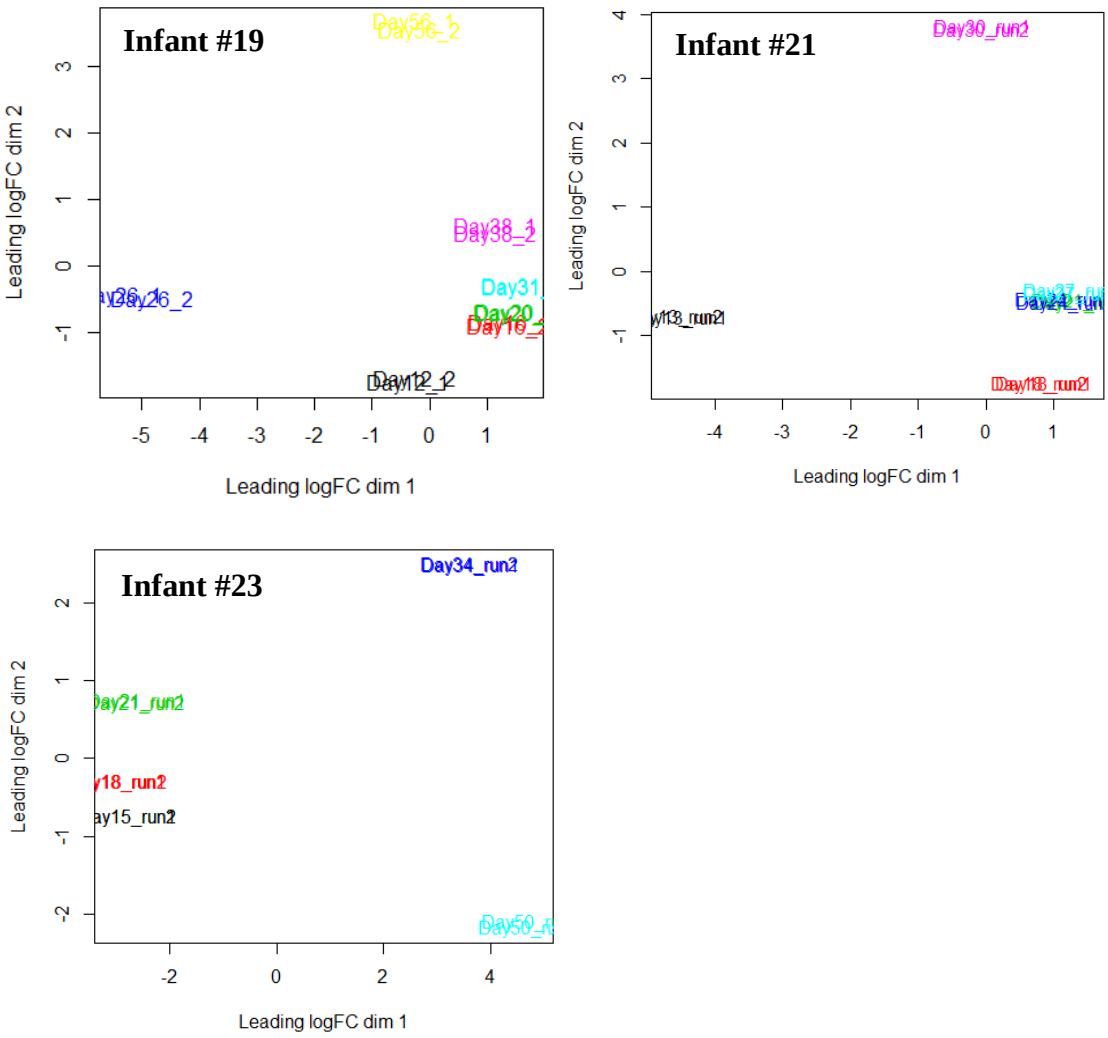


Figure 6.3. MDS plots of microbial proteins for infants #19, #21 and #23. MDS analysis was performed using the edgeR function plotMDS with log fold-change method estimated on spectral counts (normalized to library size). Different colors represented different fecal samples and technical replicates were labeled as the same color.

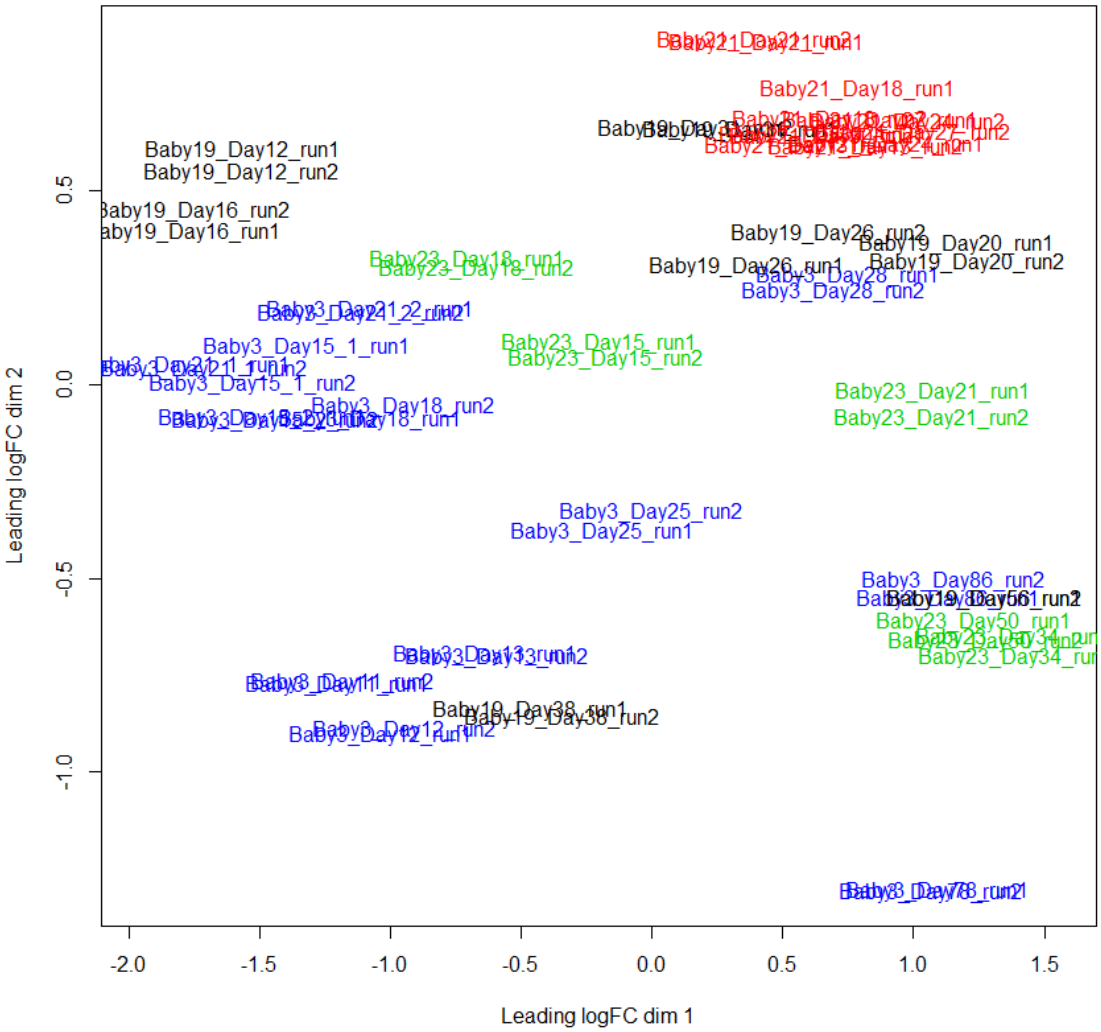


Figure 6.4. MDS plots of human proteins for infants #19, #21 and #23. MDS analysis was performed using the edgeR function plotMDS with log fold-change method estimated on spectral counts (normalized to library size).

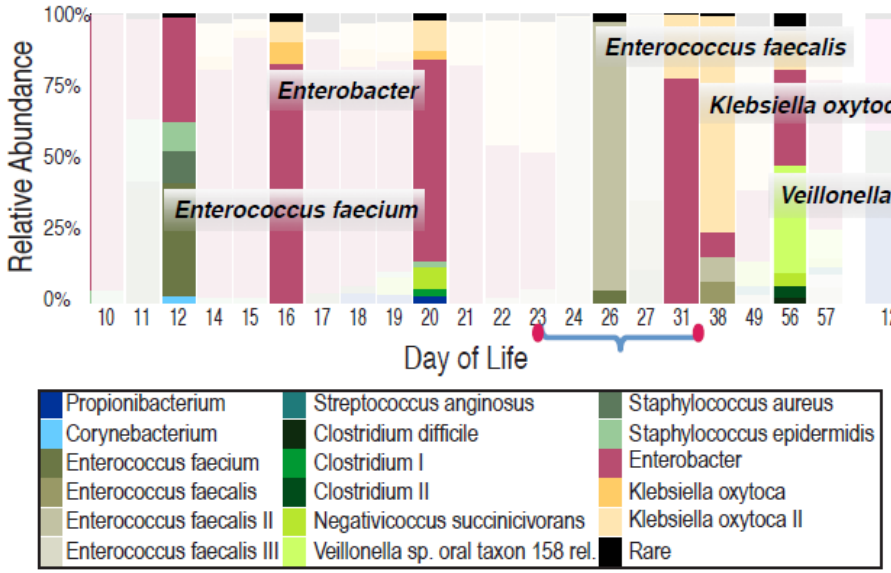
antibiotics treatment, the microbiome seemed to be restored, as day 31 was back to where was close to day 20. In the infant #21, days 21, 24 and 27 were clustered together and separated from rest days, while in the infant #23, days 15, 18 and 21 were more similar. Also, we examined the similarities/differences of human proteins for all infants, including infant #3 who was analyzed in the previous chapter. For infants #3, #19 and #23, days were scattered but did not follow the pattern seen in microbial proteins. For example, day 26 of infant #19 was dramatically different from day 20 of infant #19 in the microbiome, but human proteomes of these two samples were very similar. For these three infants, human proteomes from early time points were different from each other but relatively similar to samples from the same infant. For example, days 15, 18 and 21 of infant #3, days 12 and 16 of infant #19, and days 15 and 18 of infant #23 were closely clustered. It was also very interesting to observe that human proteomes of late time points in these three infants were almost overlapped, including day 86 of infant #3, day 56 of infant #19 and days 34 and 50 of infant #23. Intriguingly, all human proteomes of infant #21 were clustered together and separated from all other samples despite changes observed in the microbiome. Noted that infant #21 was the only infant who developed NEC.

6.4 Microbial community profile

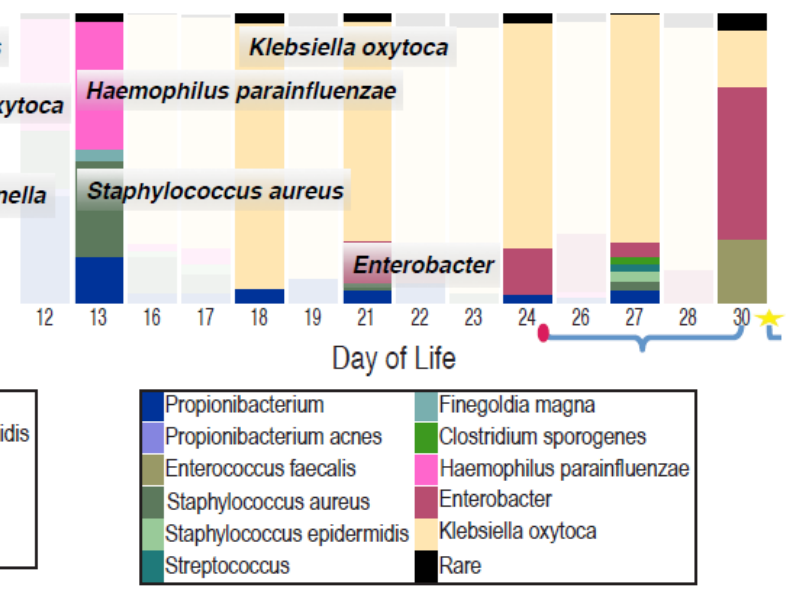
To further survey the microbial community, we performed phylogenic assignments and compared the relative abundance of the community based on metagenomics reads and protein abundance (Figure 6.5). Proteins were identified in 25, 18, 12, and 29 different species/strains for infants #3, #19, #21 and #23 respectively, showing that microbiomes of infant #19 and #21 were less complex. Microbial composition was largely different and *Enterococcus faecalis* was the only species that was shared by all infants. However, a number of species were shared between

Figure 6.5. Pattern of changes in microbial abundance (a, b and e) and protein abundance (c, d and f) for infants #19, #21 and #23. Relative abundance of microbial community was based on mapping metagenome sequencing reads to reconstructed genomes. Activity of microbial community members was based on assignment of proteomic data to genomes. Most abundant species/strains based on metagenomes or metaproteomes were highlighted for every sample. The blue brackets indicate that antibiotics were administered.

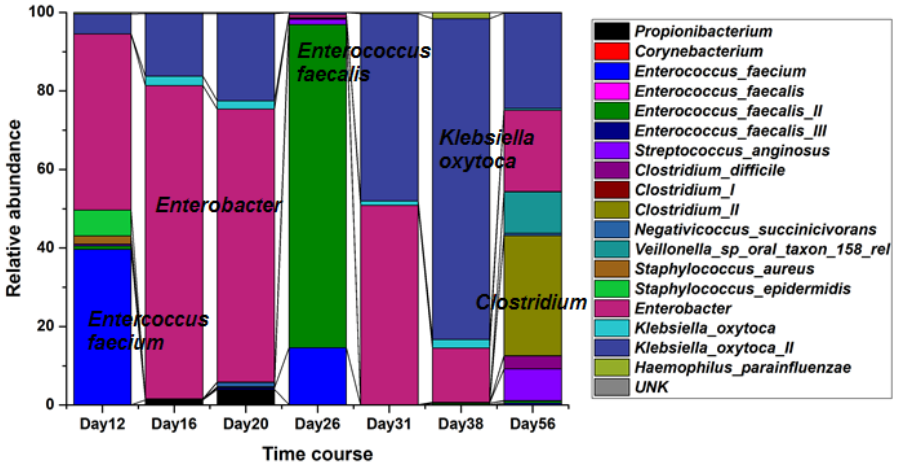
(a) Infant #19 microbial abundance



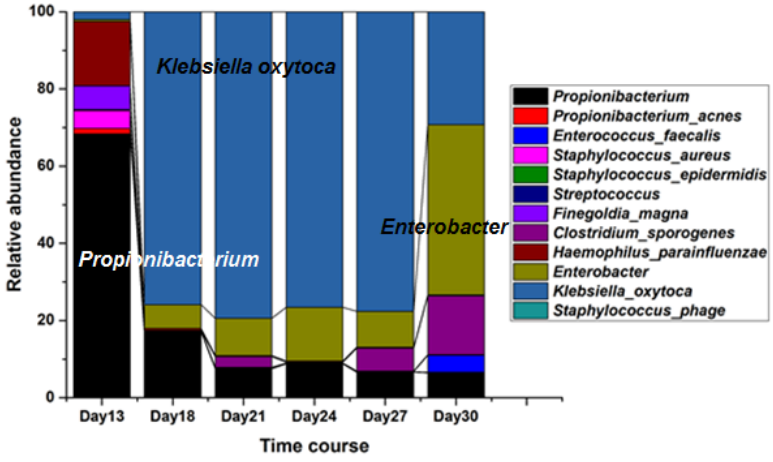
(b) Infant #21 microbial abundance



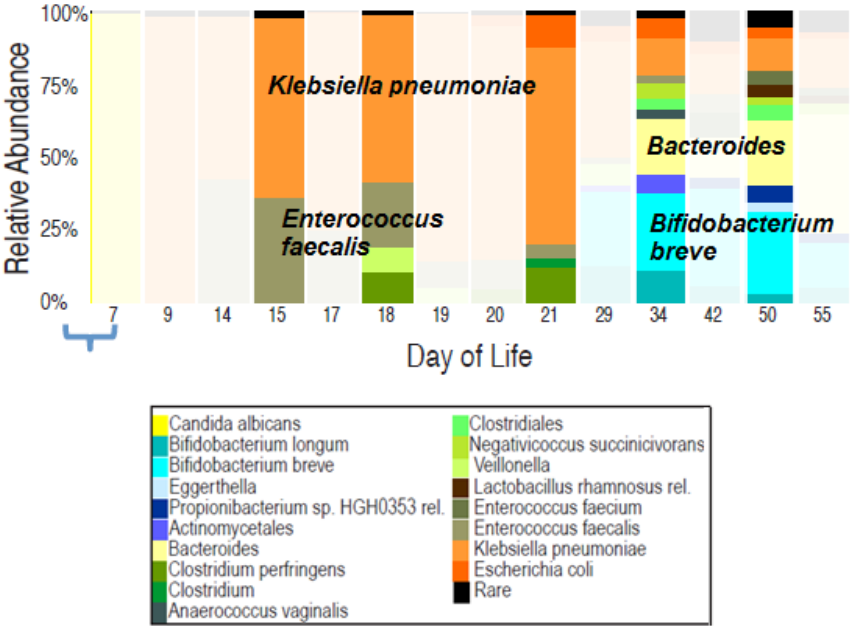
(c) Infant #19 microbial activity



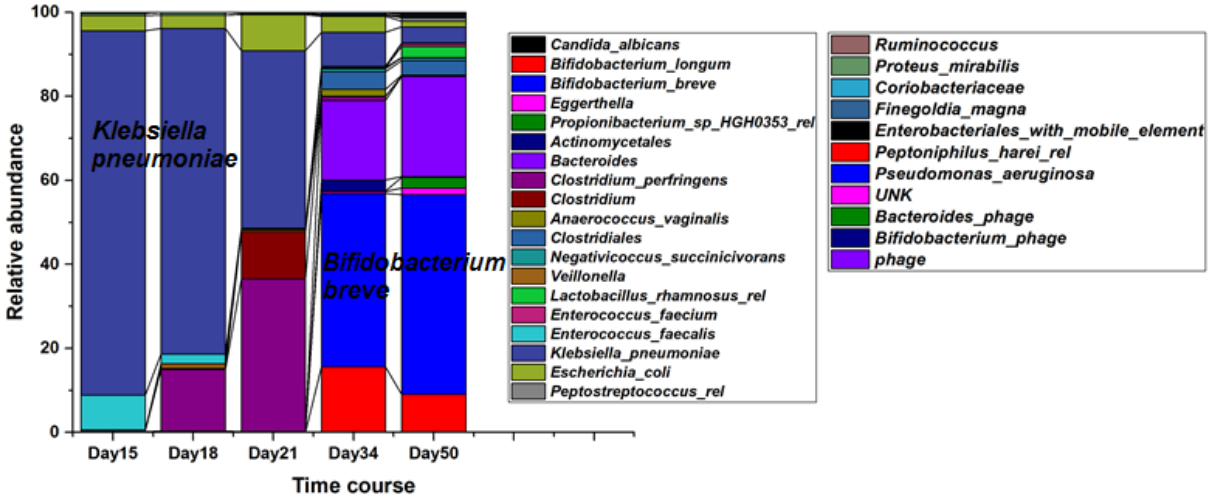
(d) Infant #21 microbial activity



(e) Infant #23 microbial abundance



(f) Infant #23 microbial activity



the twin baby 19 and 21, such as *Staphyloccoccus aureus*, *Enterobacter cloacae*, *Klebsiella oxytoca* and *Haemophilus parainfluenzae*. In baby 19, *Enterococcus faecium* and *Enterobacter cloacae* were the most active organisms in day 12 while *Enterobacter cloacae* and *Klebsiella oxytoca* comprised of the largest proportion of the metaproteome in days 16, 20 and 31. In day 26, the pattern was completely different from other samples, probably resulting from the use of antibiotics. Until day 56, *Clostridium II* accounted for 30% of the community proteome but its reads only comprised of less than 5% of the community. In baby 21, apparent difference between cell abundance and activity was observed for *Propionibacterium* in day 13, whose proteome was dominant but proportion of genome abundance was only 15%. For rest samples, *Enterobacter cloacae* and *Klebsiella oxytoca* were most active organisms, which were consistent with the relative abundance shown in genomic information. Baby 23 was colonized with most complex microbiome among four infants. The metaproteome of baby 23 was dominated by *Klebsiella pneumoniae* in days 15, 18 and 21 and shifted to *Bifidobacterium breve* in days 34 and 56.

6.5 Main microbial functionality in infant gut microbiome

As discussed above, the microbial composition and proportions not only vary dramatically during the early colonization phase but also can be remarkably different between infants, and therefore, comparing abundances of identified proteins across samples is less trivial. Hence, we employed the strategy of annotating identified proteins by orthologous groups to make comparisons between samples possible. By using KEGG orthological database, annotations were obtained for over 80% of identified proteins, with 2236, 2071, 2029 and 2230 KOs assigned for baby 3, 19, 21 and 23 respectively (Figure 6.6). Among all annotated KOs, 1468 KOs (larger than 65%) were commonly identified in four infants and 111 KOs were

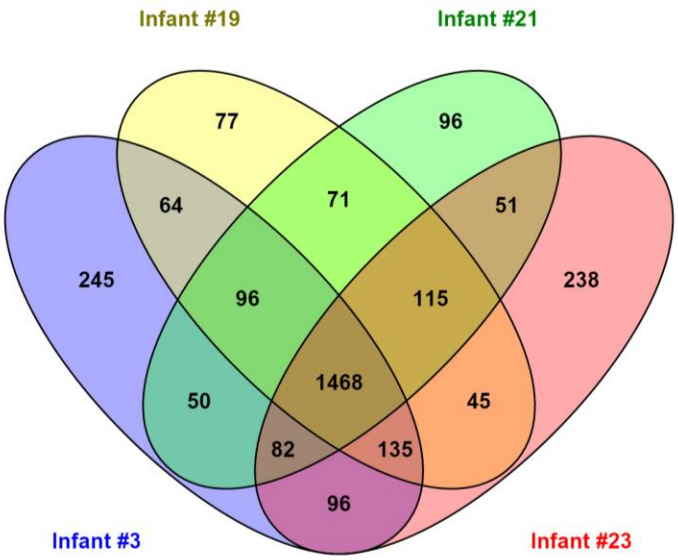


Figure 6.6. Venn diagram of assigned KOs in four infants. A total number of 2236, 2071, 2029 and 2230 KOs were assigned for infants 3, 19, 21 and 23 respectively.

commonly detected in all 30 fecal samples. Mapping these 1468 common KOs on KEGG pathways highlighted pathways specific to gut microbiome, including core central metabolisms of carbohydrates, lipid, nucleotides and amino acid, LPS biosynthetic process and cobalamin production. These pathways represented the functional core metabolism in the infant gut (Figure 6.7). In addition, many high abundance proteins with important roles were also identified, such as ATP synthase involved in the oxidative phosphorylation, chaperonin GroEL mediating the protein folding, PTS-glucose/fructose-specific component transporting sugars into bacterial cells, RNA polymerase regulating transcription, and peroxiredoxin and superoxide dismutase defending against the oxidative damage. We also examined the pathways identified in all 30 samples, (blue lines shown in Figure 6.7), which further emphasized the importance of generating energy (especially central carbon metabolism) in the microbial metabolism.

6.5 Characterization of temporal and inter-individual differences in microbial functions

The main microbial functionalities described above as well as details of longitudinal microbial functions discussed in Chapter 5 clearly depicted characteristics essential and specific to the infant gut microbiome. To identify temporal and individual differences in these core functions, we employed Blast2GO platform to annotate all identified microbial proteins with GO terms, as GO has a high annotation coverage and therefore provides functional characterization with different levels of resolution. Approximately 88% of identified proteins (6540/7390 in baby 19, 5619/6347 in baby 21, and 9130/10322 in baby 23) were annotated with at least one GO term. At a low resolution, major GO term distributions remained stable across infants (similar patterns as seen in Figure 5.7), with metabolic process including carbohydrate, lipid, protein metabolism, localization, response to stimulus and methylation among the top 20 terms. Hence, we next

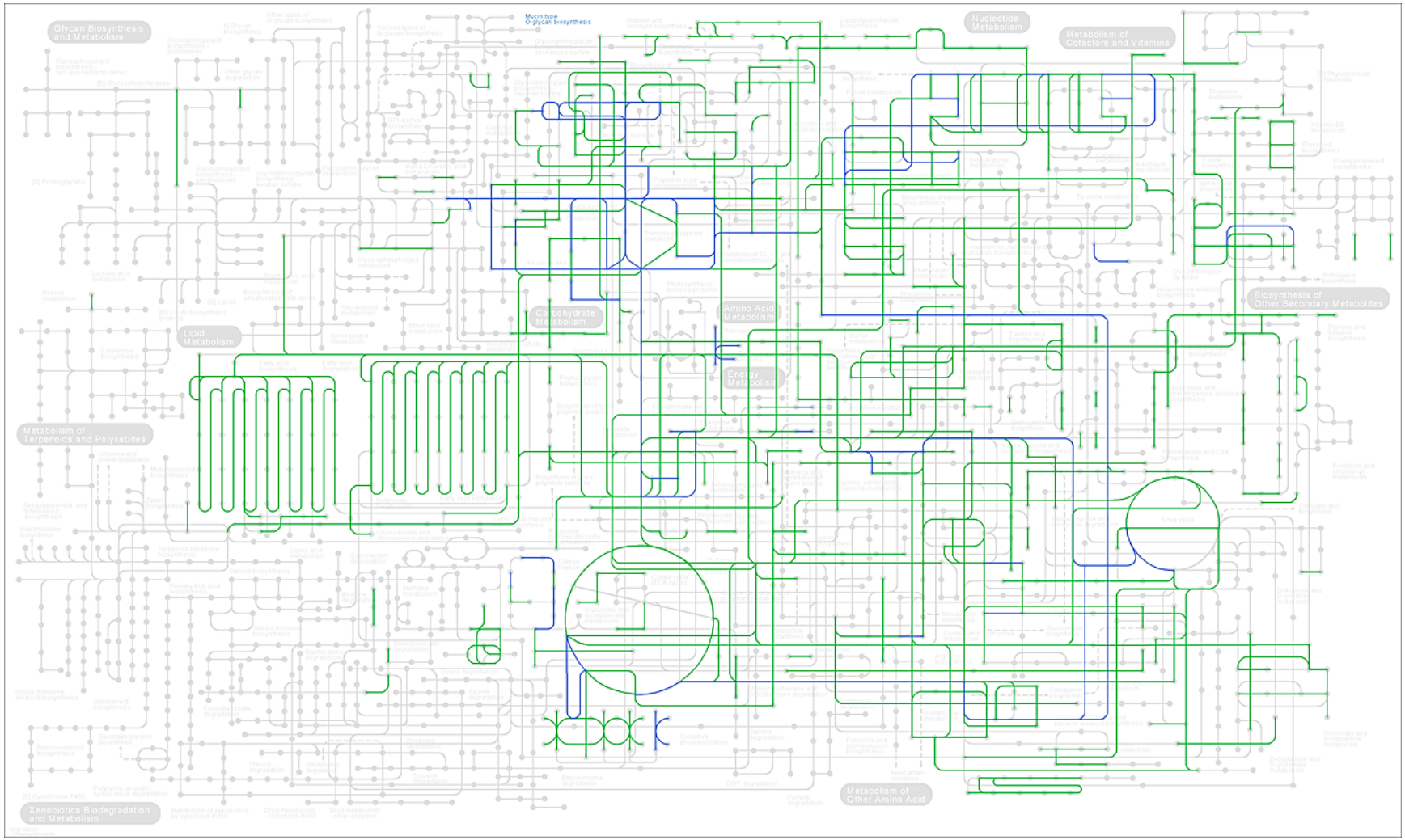
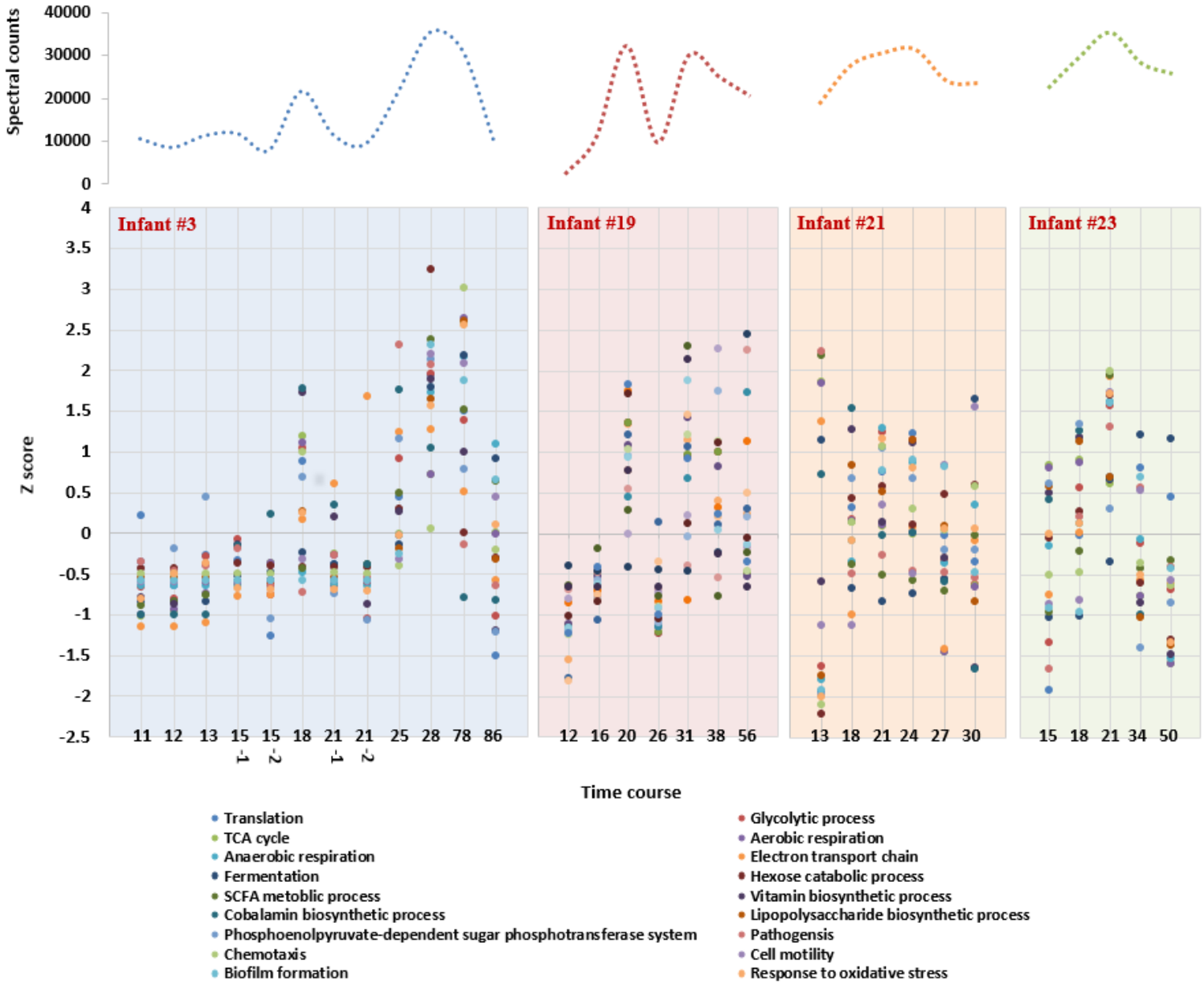


Figure 6.7. KEGG pathways mapping of common microbial KOs. 1468 common microbial KOs assigned in four infants were mapped onto KEGG pathways (shown in highlighted lines), among which 111 KOs detected in all samples were highlighted in blue.

examined the functional differences in a more detailed level consisting of 18 major biological processes, such as translation, aerobic/anaerobic respiration, SCFA metabolic process, chemotaxis and *et al* as listed in Figure 6.8. In general, all these gut microbiome functions were developed in four infants at certain time points and there was a rapid increase in functionalities during the early time course, but later, these functions fluctuated over time (Figure 6.8 upper panel) as a result of all possible factors, for example environment, diet and health status.

For the infant 3, we noticed a great increase in day 28 and 78 for functions related to pathogenesis, chemotaxis, LPS biosynthesis, response to oxidative stress and biofilm formation, suggesting a possible bacterial infection. Anaerobic respiration and SCFA were also increased in days 28 and 78 whereas vitamin production was relatively abundant in the middle time course. For the infant 19, we observed that most functionalities were decreased for days 12, 16 and 26 due to very low complexity and number of proteins identified. However, cobalamin production was increased in day 38 whereas pathogenesis and SCFA metabolic process were increased in day 56. In day 56, a high abundance of autotransporter adhesion was detected. Autotransporter adhensins are outer membrane proteins of Gram-negative bacteria that are crucial for bacteria to infect host cells via cell adhesion [205]. Also, an increasing abundance of key enzymes (butyrate kinase and phosphate butyryltransferase) involved in the production of butyrate were identified, contributing to the production of SCFA. As compared to infants 3 and 19, variations of functions in infants 21 and 23 were relatively smaller. Some interesting findings included highly increased pathogenesis in day 13 of infant 21 with abundant autotransporter adhesins detected and increased fermentation in days 34 and 50 of infant 23. In total, the gut microbiome exhibit temporal variations in biological activities described here and the variation patterns were largely different among infants.

Figure 6.8. Temporal and inter-individual differences of major microbial functions. Variations of 18 GO terms were analyzed for four infants (separated in four panels) over time. Spectral counts were summed up for all proteins involved in one GO term and z-score of each GO term was calculated and plotted. Every dot in the figure represented the z-score of a GO term characterized in a sample from an infant. A trend showing summed spectral counts of all 18 GO terms for a sample was displayed on the upper panel of the figure.


Spectral counts
40000
30000
20000
10000
0
Z score
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Infant #3
Infant #19
Infant #21
Infant #23
11 12 13 15 15 18 21 21 25 28 78 86
-1 -2 -1 -2
12 16 20 26 31 38 56
13 18 21 24 27 30
15 18 21 34 50
Time course
Translation
TCA cycle
Anaerobic respiration
Fermentation
SCFA metoblic process
Cobalamin biosynthetic process
Phosphoenolpyruvate-dependent sugar phosphotransferase system
Chemotaxis
Biofilm formation
Glycolytic process
Aerobic respiration
Electron transport chain
Hexose catabolic process
Vitamin biosynthetic process
Lipopolysaccharide biosynthetic process
Pathogensis
Cell motility
Response to oxidative stress

Besides the above major functions, we also compared all KOs assigned among four infants to visualize the significantly differential expressed functionalities. Among total 2929 KOs, we identified 973 KOs were significantly different between infants and top 100 significant KOs were plotted in the heatmap shown in Figure 6.9. A total of 7 clusters were assigned, according to the similar expression pattern of proteins. Interestingly, proteins in Cluster I were mainly abundant in the infant 21, including proteins involved in biosynthesis of siderophore group nonribosomal peptides and CRISPR system cascade. Siderophores are small, high iron affinity molecules secreted by microbial cells into the environment to scavenge for iron. Enzymes responsible for the synthesis of siderophores, for example 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase (entA), bifunctional isochorismate lyase / aryl carrier protein (entB) and nonribosomal peptide synthetase (dhbF) were found only expressed in the infant 21. Several studies have demonstrated the role of siderophores for the survival of pathogenic bacteria and the development of virulence [206]. The activated siderophore production pathway may be indicative of bacterial pathogenesis in the competition of irons with the host. We also identified two CRISPR-associated proteins Cse 4 and Cas 5 in this cluster, suggesting a possible bacterial defense against invaders. CRISPR/Cas system is a prokaryotic defense mechanism found in bacteria that provides microorganisms immunity against invading genetic elements, for example, phages and plasmids [207].

In Cluster II, we detected proteins mainly abundant in the infant 23, including ABC transporter proteins transporting lactose and other sugars, L-fuconate dehydratase participating in the fructose metabolism, alpha-L-fucosidase 2 aiding in the degradation of fucosylated glycan and xylulose-5-phosphate/fructose-6-phosphate phosphoketolase playing key roles in the pentose phosphate pathway. All these identified proteins belonged to the genus *Bifidobacterium* and

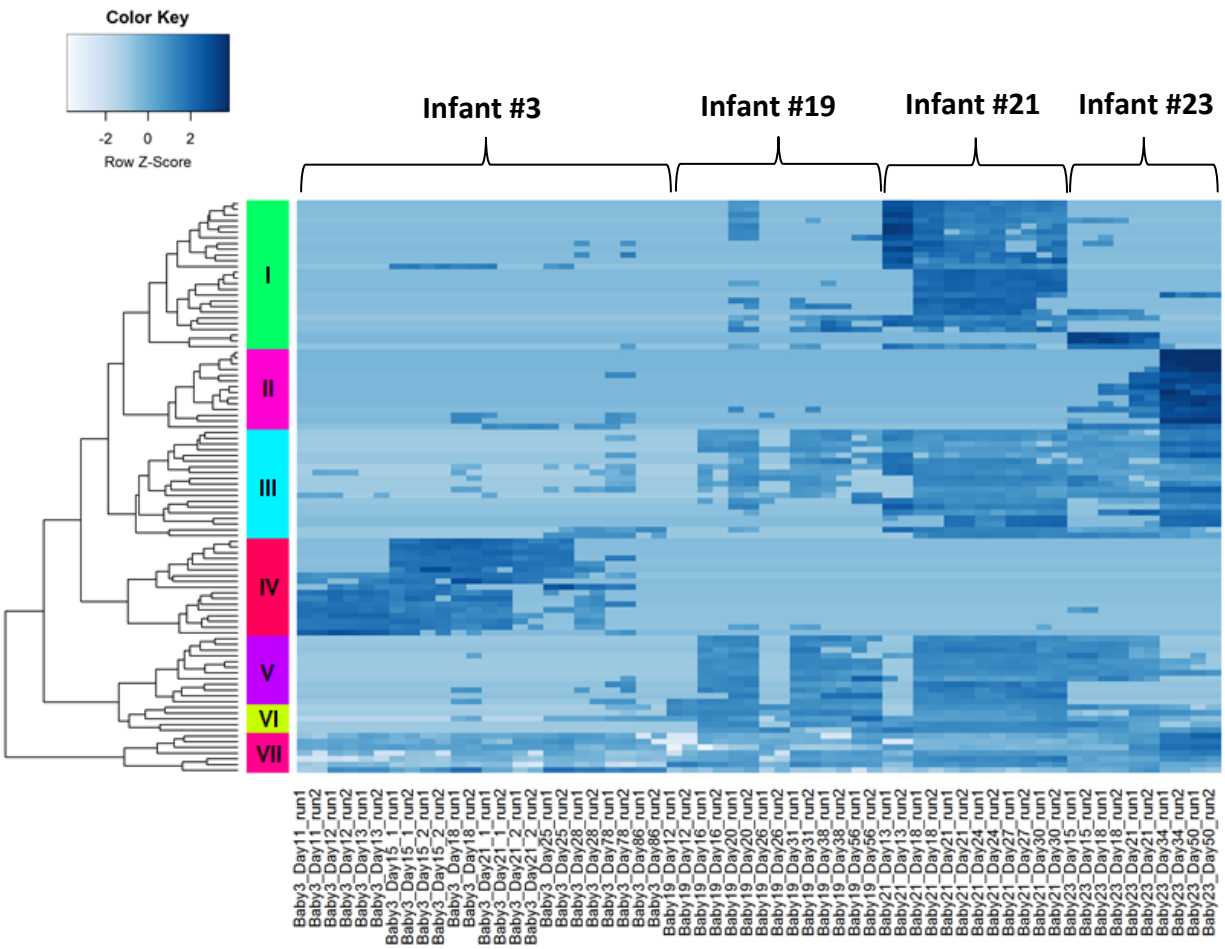


Figure. 6.9. Most significantly differentially expressed KOs among infants. KO abundance was determined by summing up spectral counts of all proteins assigned in a KO and normalized based on scaling factors for library sizes (edgeR). Significance testing for differential expression was performed using a negative binomial generalized linear model (GLM) and ANOVA like analysis in edge R package. Top 100 significantly differentially KOs were hierarchically clustered based on z-score of log-counts-per-million (log2 counts-per-million).

.

participated in the carbohydrate utilization. In particular, fermenting carbohydrates via a phosphoketolase pathway is a unique process in the *Bifidobacterium* bacteria [208]. The microbiome in the infant 23 may utilize the carbohydrates via pathways different from other infants.

In Cluster III, V and VI, we identified proteins depleted in the infant 3. An interesting finding was that all enzymes (ascorbate-specific PTS transporting system, L-ribulose-5-phosphate 4-epimerase, 3-dehydro-L-gulonate-6-phosphate decarboxylase, and L-ribulose-5-phosphate 3-epimerase) involved in the ascorbate degradation [209] were identified in infants other than the infant 3, suggesting that the microbiomes from those three infants can use the ascorbate as the carbon source under the anaerobic condition.

6.6 Comparison of human proteins among multiple infants

Our developed approach not only allowed the monitoring of microbial functionalities, but also enabled the simultaneous analysis of human host functions. In total, 3250, 2268, 742, and 1319 human proteins were identified in infants #3, #19, #21 and #23 respectively. Venn diagram showed that 547 proteins were commonly identified in all infants (Figure 6.10). Mapping these common proteins onto the KEGG pathway and GO terms highlighted biological processes responsible for energy generation, including carbohydrate metabolism (mainly glycolysis, TCA cycle), energy metabolism (mainly oxidative phosphorylation) and lipid metabolism (mainly fatty acid degradation) (Figure 6.11). These common proteins also included a number of immune defense proteins, for example, S100 calcium binding protein A8/A9, complement component 3/5/9, lactotransferrin, peroxiredoxin 1/2, serpin peptidase inhibitor which are important in providing the interaction between the human host and gut microbes [94]. We also identified

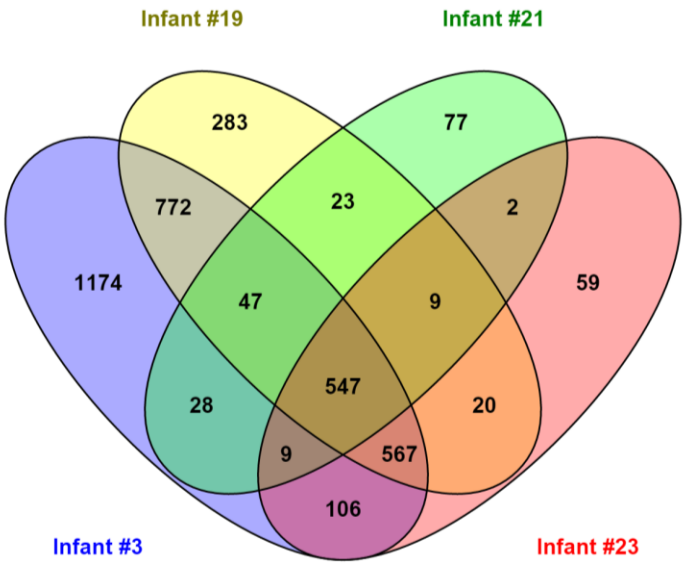


Figure 6.10. Venn diagram of human proteins among infants.

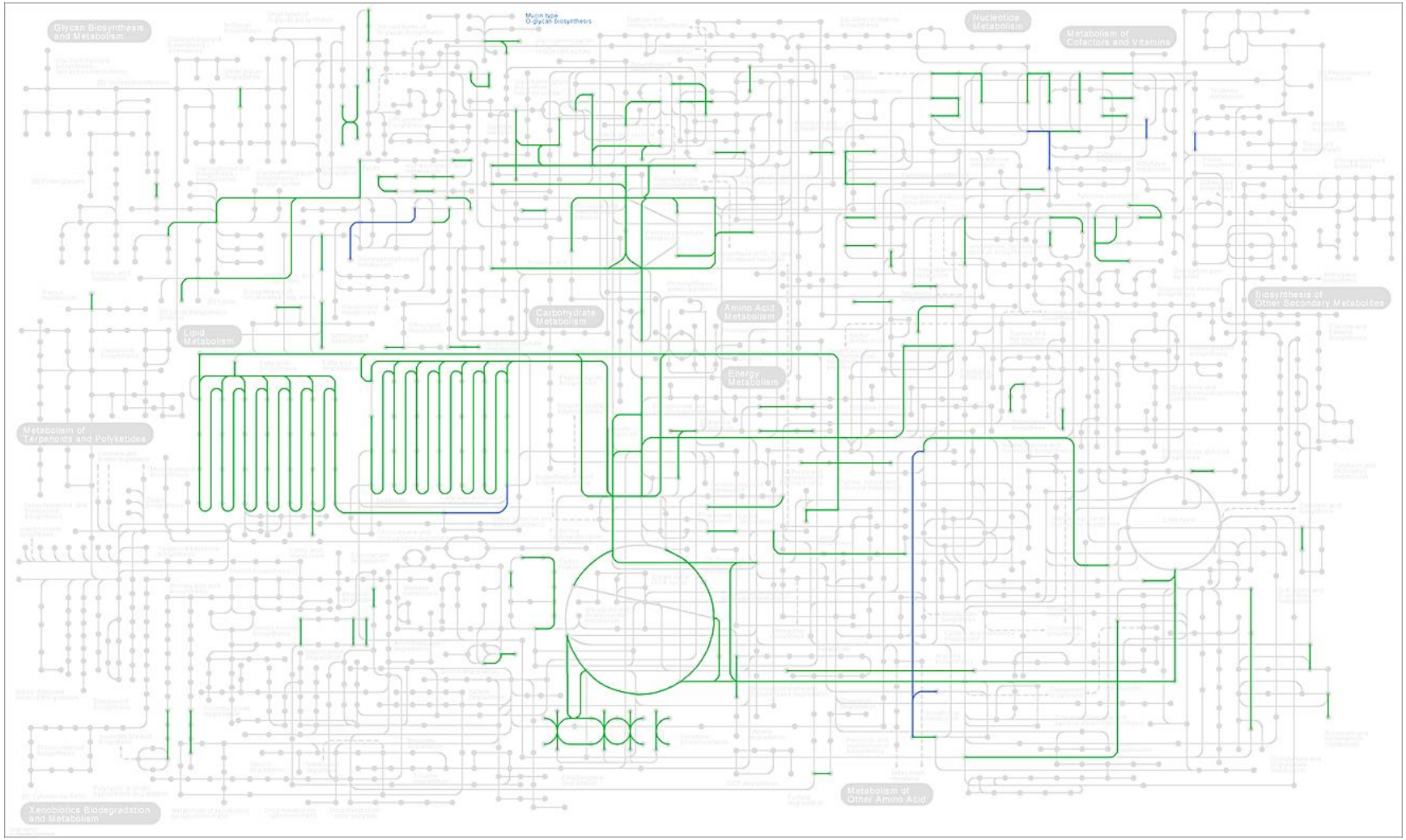
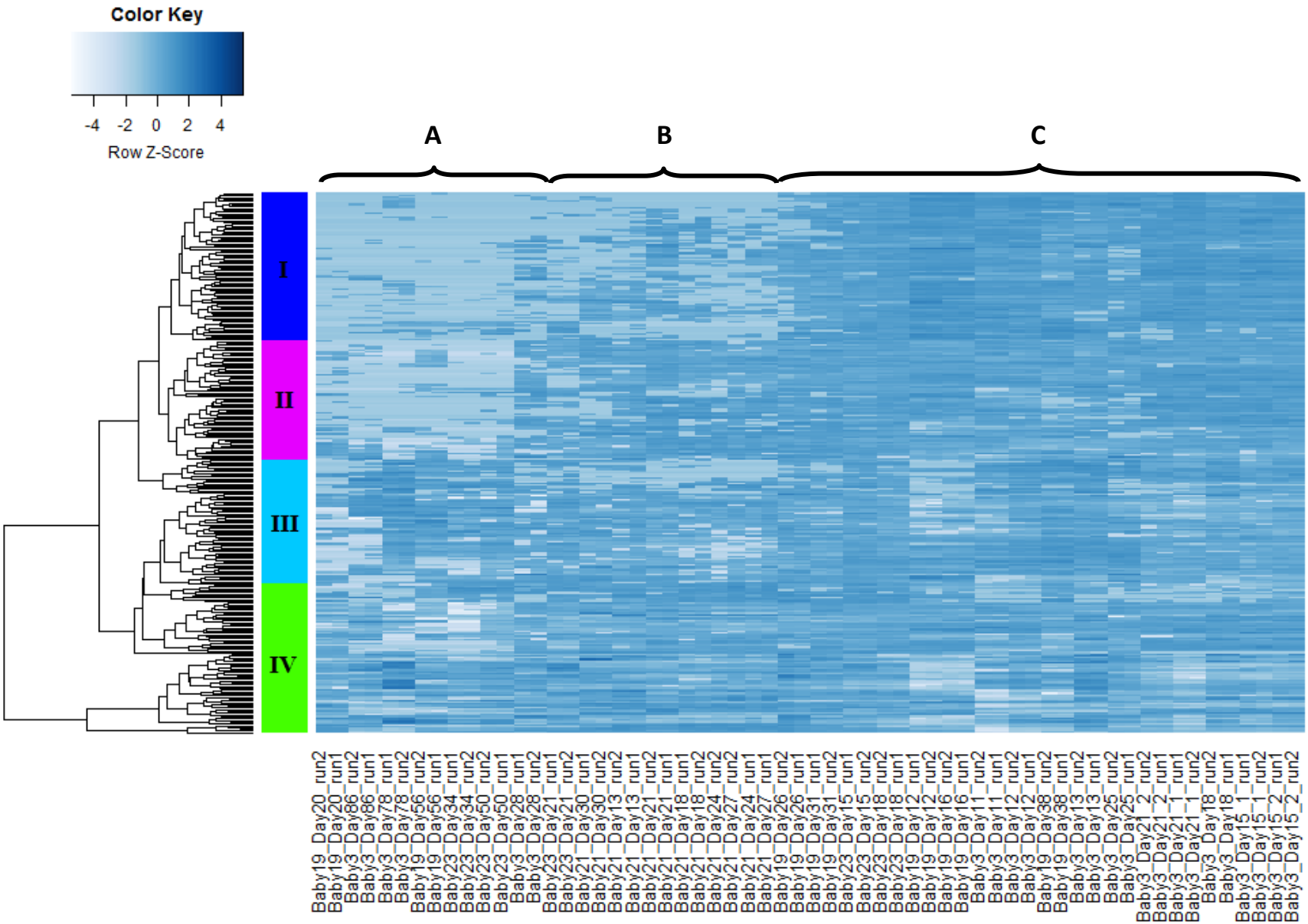


Figure 6.11. KEGG pathways mapping of common human proteins. 547 human proteins commonly identified in four infants were mapped onto KEGG pathways (shown in highlighted lines), among which 38 proteins detected in all samples were highlighted in blue.

trefoil factor 3 (TFF3) in all infants but with relatively low abundance (~ 10 spectral counts). TFF3 is typically secreted abundantly at the mucosal surface by goblet cells in the intestine and plays important roles in the maintenance and repair of the intestinal mucosal barrier. Deficiency of TFF3 has been suggested in premature infants and the pathogenesis of NEC [210]. In addition, proteins participating in the host intestinal mucus layer development were also identified in all infants, including MUC2 mucin, CLCA1 and FCGBP [210]. The intestinal mucus layer play critical roles in providing a barrier preventing bacterial invasion into the epithelium and researchers have suggested that the composition of the gut microbiota can shape the mucus structure [211]. Another important component in preventing the intestinal permeability is the production of tight junctions between epithelial cells, for example zonula occluding proteins ZO-1, ZO-2, and ZO-3, occluding, and the claudin family [212]. However, we only detected low abundance (less than 5 spectral counts) ZO-1, ZO-2, ZO-3, claudin-3 and claudin-7 proteins in one or two samples among the total 30 samples, suggesting the immature development of tight junctions for these premature infants.

We further dissected the expression pattern of human proteins across four infants over the time course. Identified human proteins were hierarchically clustered using protein abundance and four clusters were determined (Figure 6.12). GO enrichment using David Bioinformatics Resources revealed enrichment of fatty acid beta-oxidation, generation of precursor metabolites and energy, glucose metabolic process, oxidative phosphorylation, cation transport, protein transport, regulation of apoptosis, and cytoskeleton organization in Cluster I and II. We noticed that most proteins in these two clusters were generally depleted in samples collected from later time points (days 78, 86 of infant 3, day 56 of infant 19, and days 34 and 50 of infant 23), indicating essential roles of above mentioned pathways in the early development of infant

Figure 6.12. Hierarchical clustering of human proteins among infants. Identified human proteins were filtered with the threshold of copy per million (cpm) > 100 in at least half samples (edgeR). Protein abundance was normalized using trimmed mean of m-values (TMM) based on scaling factors for library sizes. Heatmap was generated using moderated log-counts-per-million (log2 counts-per-million).


Color Key
-4 -2 0 2 4
Row Z-Score
A
B
C
I
II
III
IV
Baby19_Day20_run2
Baby19_Day20_run1
Baby3_Day86_run2
Baby3_Day86_run1
Baby3_Day78_run1
Baby3_Day78_run2
Baby19_Day56_run2
Baby19_Day56_run1
Baby23_Day34_run1
Baby23_Day34_run2
Baby23_Day50_run2
Baby23_Day50_run1
Baby3_Day28_run1
Baby3_Day28_run2
Baby23_Day21_run1
Baby23_Day21_run2
Baby21_Day30_run1
Baby21_Day30_run2
Baby21_Day13_run2
Baby21_Day13_run1
Baby21_Day21_run2
Baby21_Day21_run1
Baby21_Day18_run1
Baby21_Day18_run2
Baby21_Day24_run2
Baby21_Day27_run2
Baby21_Day24_run1
Baby21_Day27_run1
Baby19_Day26_run2
Baby19_Day26_run1
Baby19_Day31_run1
Baby19_Day31_run2
Baby23_Day15_run1
Baby23_Day15_run2
Baby23_Day18_run2
Baby23_Day18_run1
Baby19_Day12_run1
Baby19_Day12_run2
Baby19_Day16_run2
Baby19_Day16_run1
Baby3_Day11_run2
Baby3_Day11_run1
Baby3_Day12_run2
Baby3_Day12_run1
Baby19_Day38_run2
Baby19_Day38_run1
Baby3_Day13_run1
Baby3_Day13_run2
Baby3_Day25_run2
Baby3_Day25_run1
Baby3_Day21_2_run2
Baby3_Day21_2_run1
Baby3_Day21_1_run1
Baby3_Day21_1_run2
Baby3_Day18_run2
Baby3_Day18_run1
Baby3_Day15_1_run1
Baby3_Day15_1_run2
Baby3_Day15_2_run1
Baby3_Day15_2_run2

intestine. In Cluster III and IV, we observed proteins almost identified across samples with enriched GO terms including defense response, response to extracellular/endogenous stimulus and anti-apoptosis. Bacterial colonization during infancy plays a critical role in the maturation of host immune system and the establishment of life-time tolerance to the commensals [213]. Colonized bacteria activate the innate immune pathways in epithelial cells and induce human immune maturation and tolerance. On the other hand, the host produces intestinal barrier and immune factors to control activities of the microbiome. Therefore, a constantly tuned immune response in the host gut is essential to keep the balance between the microbiome and human host.

6.7 Discussion and conclusions

In this study, we employed a metaproteomic pipeline to characterize the temporal microbial functions and interactions with human host in three more premature infant gut (in addition to the infant studied in the last chapter), aiming to examine the inter-individual variations among multiple infants.

An obviously increasing complexity/population of gut microbial community was observed over the time course in two infants but the trend was not clear in the other two, suggesting that the degree of infant gut microbiome diversity/density could be different despite the same/close day of life (DOL). Thus, it might be possible that the microbiomes of the latter two infants developed quickly and thus already achieved a moderate abundance of microbial population at the time when we collected the samples.

We identified temporal variations in both microbial and human proteins. When only looking at microbial proteins, in general, samples from adjacent sampling days were more

similar unless there was a drastic change, for example the complete different microbial community composition detected in day 26 of infant 19, probably relating to the antibiotics use. However, human host proteins did not respond in the same manor but still separated early time points from late time points in all infants with the exception of infant 21. Interestingly, very small differences were detected in human proteins of infant 21 although differences were observed in microbial proteins. We suspect that the low complexity of microbiome in this infant may contribute to the relatively stable expression of human proteins as a result of less diverse host-microbial interactions.

Characteristics essential and specific to the infant gut microbiome were characterized among infants, including all central energy metabolism in carbohydrates, lipid, nucleotides and amino acids as well as microbiome-specific processes such as vitamin synthesis, SCFA production and LPS biosynthesis. However, these processes can be established at different time of the colonization and varied greatly over time in a personalized pattern with impacting factors unclear. In addition to these core metabolisms, several pathways/proteins were found unique to a particular infant, for example the siderophore production in the infant 21 and the phosphoketolase pathway in the infant 23, indicating types of unique microbial functions. Core metabolisms were also examined in the human host, which mainly involved in energy production, epithelial barrier construction and immune response. However, trefoil factor 3 and tight junction proteins were detected with low abundance or not detected, suggesting that the epithelial barrier was immature or still under development in these premature infants. In addition, proteins participating in the host immunity were constantly abundant over time as essential components in the balance and control of gut microbiome and the maturation of human immune system.

The infants in this study are different in many aspects including delivery mode, feeding type, the use of antibiotics and health status, all of which can have influenced microbial colonization. Thus, more proteomic data collected from premature infants between groups, e.g. healthy versus diseased (NEC), may help the determination of the proteomic baseline for healthy individuals and the investigation of proteomic differences between healthy and diseased groups, which are important questions required to be answered for our better understanding of NEC onset.

.

CHAPTER 7

Conclusions and future perspectives

7.1 Conclusions from the development and application of metaproteomics approach in the characterization of infant gut microbiome

The advent of whole-genome sequencing technology has largely changed microbiology from analyzing cultured isolates to discovering a variety of environmental communities and understanding their interactions and responses to the environment. Our understandings of microbial community functionalities are further enhanced by the continuous development in proteomic platforms including the improved analytical capabilities with high speed, resolution and accuracy mass spectrometers and the development of available peptide/protein identification algorithms and informatics pipelines. The human gut microbiome is the most studied ecosystem due to its critical role in human physiology and association with a number of inflammatory and metabolic disorders. The beginning of the intestinal community occurs during infancy and develops as human body grows and develops. The establishment process can be shaped by many factors and have an impact on infants' health and health later in life potentially [89]. Despite increasing studies have been conducted on infants gut microbiome and revealed microbial composition changes associated with early life events, the characterization of functional shift at the protein level is much less addressed, which is required to answer "what does happen".

With regard to the sample preparation, a double filtering strategy aiming in the in-depth measurement of microbial proteins was designed in Chapter 3 as abundant human proteins preventing the detection of low abundance microbial proteins. We have demonstrated that our

approach facilitated greater than 50% increase in the overall peptide/protein identifications for two infant fecal microbiomes. Indeed, the depth of both microbial and human protein measurement was improved as a result of removing those abundant human proteins. In this study, we also addressed that less than 30% of total collected spectra were assigned as peptides and a large proportion of unassigned spectra were assigned with high quality scores, highlighting the importance of metagenome completeness and the need for new data interpretation solutions, for example *de novo* sequencing. Although functional characterization of identified proteins wasn't the focus of this study, the employment of COG categories in the analysis further demonstrated that the improved depth in microbial proteins revealed newly identified functional categories.

Not only the sample preparation, but also the instrumental and informatics methodologies affect the success of metaprotomics studies, and therefore careful considerations are need and have been discussed in Chapter 4. We demonstrated that optimized data acquisition settings by enabling isotopic precursor selection improved the depth and accuracy of proteome measurement. We summarized available approaches for the construction of a metagenome database and illustrated how the quality and complexity of the metagenome impacted confident protein identifications. As the complexity of ecosystem increased, the differentiation between true and false identifications became difficult and thus the threshold for confident identifications largely increased, which might suggest the need for new solutions other than the target-decoy search estimating FDR strategy, to efficiently determine the confidence for very large database. Redundancy of the metaproteomic data is one difficulty hindering the unambiguous protein inference in the metaproteomic analysis. One efficient way to alleviate this problem is to cluster proteins based on sequence homology. To apply the clustering approach in the infant gut metaproteomics, we evaluated the degree of redundancy for both human and microbiome

databases and determined the appropriate similarity threshold used for protein clustering, via balancing the tradeoff between the protein ambiguity and resolution. Considering the important role of protein structure in the determination of protein function, the sequence similarity is only partly reliable to group proteins with similar functions. Therefore, future clustering strategies need to be developed based on information other than protein sequences, for example protein domains/models. Lastly, we compared two label free quantification methods (spectral counting and MIT) on the infant gut metaproteome and found a high correlation between the two methods despite respective advantages and disadvantages.

With the pipeline established in above two chapters, we measured and investigated longitudinal microbial functionalities and interactions with the human host in a healthy premature infant gut. We demonstrated the feasibility and robustness of our metaproteomic approach in 12 infant fecal samples by detecting a total of 9318 microbial and 3250 human protein groups and revealing rapid microbial colonization as well as changes in the relative abundance of human and microbial proteins. Also observed in other proteomics studies, community activities were not perfectly consistent with community abundances in this study. Although the main goal of metaproteomics is not to identify the presence of microbial organisms but rather their functions, assigning proteins to specific species was helpful in the assessment of common and dominant microbial metabolisms. KEGG mapping of all identified proteins revealed metabolic pathways/activities operated both commonly and separately by human host and gut microbiome. In particular, pathways involving in the epithelial barrier establishment and immune responses were exclusively detected in human proteomes whereas vitamin and SCFA production were only seen in the microbiome, which are essential characteristics also shown in the human adult gut microbiome. Based on protein expression patterns of microbiome over the

time course, we identified four clusters/phases revealing functional shifts involving in glycolytic process, translation, aerobic/anaerobic respiration, vitamin synthesis, hexose catabolic process, LPS biosynthesis, PTS system, biofilm formation, and responses to various stimuli. At the same time, functions of human proteomes were monitored and mainly associated with immune defense and inflammatory response. Proteins associated with digestion, antibacterial activity, mucosal barrier proteins and innate immune responses were found among the most abundant proteins.

While a variety of factors can influence the microbial colonization during infancy, it is also interesting to investigate the commonalities and variations between individuals. In addition to the infant studied in the last chapter, we further measured gut metaproteomes from three more premature infants, aiming to investigating the differences of functional development among multiple infants. With the metaproteomic pipeline previously developed, we achieved in-depth metaproteomic measurements in all three infants with a total of 9666, 7092, and 11649 protein groups detected, which also revealed the intra- and inter- variations in the number of identified human versus microbial proteins as well as their relative abundances. Core metabolic pathways were identified in both human and microbial proteins, representing the establishment of the mutualistic relationship between the microbiome and human host during infancy, in which gut bacteria processed essential vitamins and activated host immune response while human host initiated the mucosal construction and immune defense. However, unlike human adult, infant gut exhibit large temporal variations in the abundance of these core metabolisms as a result of infant gut microbiome ecosystem being immature and unstable. Although most functionalities were shared among infants, pathways specific to particular infants were also identified, which might be strongly related to factors that were different among infants. Nevertheless, more proteomic data and significant number of individuals are needed to further address these observations.

7.2 Remaining challenges and future perspectives for human gut metaproteome research

The metaproteomic approach has emerged as an indispensable tool in order to explore global biological functions of complex ecosystems. The metaproteomic approach developed in this dissertation monitored both human and microbial proteins over time and revealed host and microbiome specific metabolic activities as well as temporal and inter-individual variations in these functions. Nevertheless, we are still at the beginning of understanding the gut microbiome by employing "omics" approaches generating large amounts of complex datasets, especially for metaproteomics. Metaproteomics is behind metagenomics in terms of the number of analyzed samples and available informatics platforms. However, progress in advanced metaproteomics methodologies is fast and the throughput of metaproteomic measurement is increasing. In this section, we will discuss remaining questions and future directions in the gut metaproteomics research.

7.2.1 The need for better assembled and annotated metagenomes

Assembling a high-quality metagenome is a major hurdle for extensive protein identifications in metaproteomics experiments, as current identification strategy and algorithm mainly rely on the completeness and accuracy of protein database predicted from the metagenome. Due to the complexity of gut microbiome, constructions of metagenomes from the human gut microbiome are commonly incomplete. In this dissertation, we have identified a large amount of unassigned high quality spectra which could be peptides belonging to proteins that are not included in the constructed database. Therefore, limitations in protein identifications enforce the need for new sequencing technologies and advanced assembly algorithms providing more reliable and complete metagenomes.

Functional annotation of proteins through orthologous protein database COG or KO, or GO database provide a broad overview of functions in the complex gut ecosystem and enable the functional comparison between microbial communities with different composition. Nevertheless, these databases have limitations in the sequence coverage and may not be specific to the human intestinal microbiome. Therefore, further improvements in the metagenome annotation will require a higher sequence coverage database and more preferable, an annotation database specific for the human gut microbiome.

7.2.2 The need for high-throughput measurement campaigns

The increasing genomic information of human genomes as well as the intestinal microbiomes has greatly improved our knowledge of the variety and complexity in the human gut microbiome. Large-scale comparative analyses of metagenomes have begun to uncover a large number of factors impacting human host physiology and shaping the composition and structure of microbes in the gut. Metaproteomes analyzed in this dissertation and also other related studies have also revealed that every individual harbors a unique/personalized gut metaproteome and their functions can vary temporally. In particular, dramatic variation has been observed in infants during early microbial colonization. Therefore, in order to identify all possible influencing factors and further determine which factor leading to what changes and if so, to what extent, future large-scale comparative metaproteomic investigations would be required to conduct longitudinal studies and analyze sufficient number of samples. This is critical for researchers to acknowledge what functional characteristics are essential for the definition of “a healthy microbiota” and further to determine functional differences between diseased and healthy individuals, specifically those associated with diseases. Therefore, there is a strong need

for significant improvements in throughput of mass spectrometry-based metaproteomic approaches in the coming years.

Recently, multiplexing quantification technologies using isobaric chemical tags such as TMT and iTRAQ, have allowed increasing throughput of proteomic experiments without increasing analysis complexity. Paulo et al. has demonstrated the feasibility of using 10-plex TMT to compare 10 different samples in a single mass spectrometric experiment [214]. Theoretically, multiplexing approach, for example 10-plex, is 10 times faster than the label free method and the sensitivity can be increased since multiple samples are combined. However, to achieve comparable depth of measurement as single one-sample experiment is not trivial and the technology still face its own issues, for example co-fragmented containing ions and co-alesced reporter ions. Future efforts to improve the multiplexing technology, particularly the optimization and robustness for complex fecal samples are needed and will advance gut metapreoteomics investigations in a high throughput manner.

This and related work has ignited a flurry of metaproteomic research on the human microbiome. Continued improvements in sample preparation methods, high performance mass spectrometers and bioinformatics platforms will allow thorough insights into biological information of the human-microbiome ecosystem. We anticipate that in the next five to ten years that we may be able to harness the power of our knowledge to guide the use of pre- and pro-biotics nurturing a healthy microbiota and interventions that allow for prevention and cure for microbiome-mediated inflammatory and immune diseases.

REFERENCES

1. Whitman, W.B., D.C. Coleman, and W.J. Wiebe, *Prokaryotes: the unseen majority.* Proc Natl Acad Sci U S A, 1998. **95**(12): p. 6578-83.

2. Baker, B.J. and J.F. Banfield, *Microbial communities in acid mine drainage.* FEMS Microbiol Ecol, 2003. **44**(2): p. 139-52.

3. Yatsunenko, T., et al., *Human gut microbiome viewed across age and geography.* Nature, 2012. **486**(7402): p. 222-7.

4. Wagg, C., et al., *Soil biodiversity and soil community composition determine ecosystem multifunctionality.* Proc Natl Acad Sci U S A, 2014. **111**(14): p. 5266-70.

5. Venter, J.C., et al., *Environmental genome shotgun sequencing of the Sargasso Sea.* Science, 2004. **304**(5667): p. 66-74.

6. Garbeva, P., J.A. van Veen, and J.D. van Elsas, *Microbial diversity in soil: selection microbial populations by plant and soil type and implications for disease suppressiveness.* Annu Rev Phytopathol, 2004. **42**: p. 243-70.

7. Human Microbiome Project, C., *A framework for human microbiome research.* Nature, 2012. **486**(7402): p. 215-21.

8. Fukuda, S. and H. Ohno, *Gut microbiome and metabolic diseases.* Seminars in Immunopathology, 2014. **36**(1): p. 103-114.

9. Moloney, R.D., et al., *The microbiome: stress, health and disease.* Mammalian Genome, 2014. **25**(1-2): p. 49-74.

10. Erickson, A.R., et al., *Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease.* PLoS One, 2012. **7**(11): p. e49138.

11. Qin, J.J., et al., *A metagenome-wide association study of gut microbiota in type 2 diabetes.* Nature, 2012. **490**(7418): p. 55-60.

12. Rappe, M.S. and S.J. Giovannoni, *The uncultured microbial majority.* Annu Rev Microbiol, 2003. **57**: p. 369-94.

13. Riesenfeld, C.S., P.D. Schloss, and J. Handelsman, *Metagenomics: Genomic analysis of microbial communities.* Annual Review of Genetics, 2004. **38**: p. 525-552.

14. Amann, R.I., W. Ludwig, and K.H. Schleifer, *Phylogenetic identification and in situ detection of individual microbial cells without cultivation.* Microbiol Rev, 1995. **59**(1): p. 143-69.

15. Pennisi, E., *Metagenomics. Massive microbial sequence project proposed.* Science, 2007. **315**(5820): p. 1781.

16. Faust, K., et al., *Metagenomics meets time series analysis: unraveling microbial community dynamics.* Curr Opin Microbiol, 2015. **25**: p. 56-66.

17. Abram, F., *Systems-based approaches to unravel multi-species microbial community functioning.* Comput Struct Biotechnol J, 2015. **13**: p. 24-32.

18. Scholz, M.B., C.C. Lo, and P.S. Chain, *Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis.* Curr Opin Biotechnol, 2012. **23**(1): p. 9-15.

19. Ji, Y., et al., *A new strategy for better genome assembly from very short reads.* BMC Bioinformatics, 2011. **12**: p. 493.

20. Dick, G.J., et al., *Community-wide analysis of microbial genome sequence signatures.* Genome Biol, 2009. **10**(8): p. R85.

21. Sharon, I., et al., *Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization.* Genome Res, 2012. **23**(1): p. 111-20.

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

23. Peano, C., et al., *An efficient rRNA removal method for RNA sequencing in GC-rich bacteria.* Microb Inform Exp, 2013. **3**(1): p. 1.

24. Gilbert, J.A. and M. Hughes, *Gene expression profiling: metatranscriptomics.* Methods Mol Biol, 2011. **733**: p. 195-205.

25. Siggins, A., E. Gunnigle, and F. Abram, *Exploring mixed microbial community functioning: recent advances in metaproteomics.* FEMS Microbiol Ecol, 2012. **80**(2): p. 265-80.

26. Pan, C. and J.F. Banfield, *Quantitative metaproteomics: functional insights into microbial communities.* Methods Mol Biol, 2014. **1096**: p. 231-40.

27. Li, Z., et al., *Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on LTQ Orbitrap Velos.* J Proteome Res, 2012. **11**(3): p. 1582-90.

28. Gupta, N., et al., *Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation.* Genome Res, 2007. **17**(9): p. 1362-77.

29. Patti, G.J., O. Yanes, and G. Siuzdak, *Innovation: Metabolomics: the apogee of the omics trilogy.* Nat Rev Mol Cell Biol, 2012. **13**(4): p. 263-9.

30. Fiehn, O., *Metabolomics--the link between genotypes and phenotypes.* Plant Mol Biol, 2002. **48**(1-2): p. 155-71.

31. Tang, J., *Microbial metabolomics.* Curr Genomics, 2011. **12**(6): p. 391-403.

32. Chen, J., et al., *Practical approach for the identification and isomer elucidation of biomarkers detected in a metabonomic study for the discovery of individuals at risk for*

diabetes by integrating the chromatographic and mass spectrometric information. Anal Chem, 2008. **80**(4): p. 1280-9.

33. Huang, H.J., et al., *Metabolomic analyses of faeces reveals malabsorption in cirrhotic patients.* Dig Liver Dis, 2013. **45**(8): p. 677-82.

34. O'Farrell, P.H., *High resolution two-dimensional electrophoresis of proteins.* J Biol Chem, 1975. **250**(10): p. 4007-21.

35. Van den Bergh, G. and L. Arckens, *Fluorescent two-dimensional difference gel electrophoresis unveils the potential of gel-based proteomics.* Curr Opin Biotechnol, 2004. **15**(1): p. 38-43.

36. Fenn, J.B., et al., *Electrospray ionization for mass spectrometry of large biomolecules.* Science, 1989. **246**(4926): p. 64-71.

37. Karas, M. and F. Hillenkamp, *Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.* Anal Chem, 1988. **60**(20): p. 2299-301.

38. Gygi, S.P., B. Rist, and R. Aebersold, *Measuring gene expression by quantitative proteome analysis.* Curr Opin Biotechnol, 2000. **11**(4): p. 396-401.

39. Peng, J. and S.P. Gygi, *Proteomics: the move to mixtures.* J Mass Spectrom, 2001. **36**(10): p. 1083-91.

40. Abraham, P., et al., *Defining the boundaries and characterizing the landscape of functional genome expression in vascular tissues of Populus using shotgun proteomics.* J Proteome Res, 2012. **11**(1): p. 449-60.

41. Scigelova, M. and A. Makarov, *Orbitrap mass analyzer--overview and applications in proteomics.* Proteomics, 2006. **6 Suppl 2**: p. 16-21.

42. Tipton, J.D., et al., *Analysis of intact protein isoforms by mass spectrometry.* J Biol Chem, 2011. **286**(29): p. 25451-8.

43. Yates, J.R., 3rd, *Mass spectral analysis in proteomics.* Annu Rev Biophys Biomol Struct, 2004. **33**: p. 297-316.

44. Compton, P.D., et al., *On the scalability and requirements of whole protein mass spectrometry.* Anal Chem, 2011. **83**(17): p. 6868-74.

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

46. Matthiesen, R., *Algorithms for database-dependent search of MS/MS data.* Methods Mol Biol, 2013. **1007**: p. 119-38.

47. Giannone, R.J., et al., *Proteomic characterization of cellular and molecular processes that enable the Nanoarchaeum equitans--Ignicoccus hospitalis relationship.* PLoS One, 2011. **6**(8): p. e22942.

48. Ram, R.J., et al., *Community proteomics of a natural microbial biofilm.* Science, 2005. **308**(5730): p. 1915-20.

49. Denef, V.J., et al., *Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities.* Proc Natl Acad Sci U S A, 2010. **107**(6): p. 2383-90.

50. Mueller, R.S., et al., *Ecological distribution and population physiology defined by proteomics in a natural microbial community.* Mol Syst Biol, 2010. **6**: p. 374.

51. Eckburg, P.B., et al., *Diversity of the human intestinal microbial flora.* Science, 2005. **308**(5728): p. 1635-8.

52. Torsvik, V. and L. Ovreas, *Microbial diversity and function in soil: from genes to ecosystems.* Curr Opin Microbiol, 2002. **5**(3): p. 240-5.

53. Chourey, K., et al., *Direct cellular lysis/protein extraction protocol for soil metaproteomics.* Journal of proteome research, 2010. **9**(12): p. 6615-6622.

54. Wolters, D.A., M.P. Washburn, and J.R. Yates, 3rd, *An automated multidimensional protein identification technology for shotgun proteomics.* Anal Chem, 2001. **73**(23): p. 5683-90.

55. Michalski, A., et al., *Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes.* Mol Cell Proteomics, 2012. **11**(3): p. O111 013698.

56. Abraham, P.E., et al., *Metaproteomics: extracting and mining proteome information to characterize metabolic activities in microbial communities.* Curr Protoc Bioinformatics, 2014. **46**: p. 13 26 1-13 26 14.

57. Turnbaugh, P.J., et al., *The human microbiome project.* Nature, 2007. **449**(7164): p. 804-10.

58. Human Microbiome Project, C., *Structure, function and diversity of the healthy human microbiome.* Nature, 2012. **486**(7402): p. 207-14.

59. Jalanka-Tuovinen, J., et al., *Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms.* PLoS One, 2011. **6**(7): p. e23035.

60. Raymond, F., et al., *The initial state of the human gut microbiome determines its reshaping by antibiotics.* ISME J, 2015.

61. Greenblum, S., P.J. Turnbaugh, and E. Borenstein, *Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease.* Proc Natl Acad Sci U S A, 2012. **109**(2): p. 594-9.

62. Segre, J.A., *MICROBIOME. Microbial growth dynamics and human disease.* Science, 2015. **349**(6252): p. 1058-9.

63. Lozupone, C.A., et al., *Diversity, stability and resilience of the human gut microbiota.* Nature, 2012. **489**(7415): p. 220-30.

64. Kau, A.L., et al., *Human nutrition, the gut microbiome and the immune system.* Nature, 2011. **474**(7351): p. 327-36.

65. Goodrich, J.K., et al., *Human genetics shape the gut microbiome.* Cell, 2014. **159**(4): p. 789-99.

66. Davenport, E.R., et al., *Seasonal variation in human gut microbiome composition.* PLoS One, 2014. **9**(3): p. e90731.

67. David, L.A., et al., *Diet rapidly and reproducibly alters the human gut microbiome.* Nature, 2014. **505**(7484): p. 559-63.

68. Dominianni, C., et al., *Sex, body mass index, and dietary fiber intake influence the human gut microbiome.* PLoS One, 2015. **10**(4): p. e0124599.

69. Korpela, K., et al., *Gut microbiota signatures predict host and microbiota responses to dietary interventions in obese individuals.* PLoS One, 2014. **9**(6): p. e90702.

70. Nell, S., S. Suerbaum, and C. Josenhans, *The impact of the microbiota on the pathogenesis of IBD: lessons from mouse infection models.* Nat Rev Microbiol, 2010. **8**(8): p. 564-77.

71. Berer, K., et al., *Commensal microbiota and myelin autoantigen cooperate to trigger autoimmune demyelination.* Nature, 2011. **479**(7374): p. 538-41.

72. Verberkmoes, N.C., et al., *Shotgun metaproteomics of the human distal gut microbiota.* ISME J, 2009. **3**(2): p. 179-89.

73. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic acids research, 2000. **28**(1): p. 33-36.

74. Juste, C., et al., *Bacterial protein signals are associated with Crohn's disease.* Gut, 2014: p. gutjnl-2012-303786.

75. Ferrer, M., et al., *Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure.* Environ Microbiol, 2013. **15**(1): p. 211-26.

76. Ley, R.E., et al., *Microbial ecology: human gut microbes associated with obesity.* Nature, 2006. **444**(7122): p. 1022-3.

77. Kolmeder, C.A., et al., *Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions.* PLoS One, 2012. **7**(1): p. e29913.

78. Ferrer, M., et al., *Gut microbiota disturbance during antibiotic therapy: a multi-omic approach.* Gut Microbes, 2014. **5**(1): p. 64-70.

79. Li, X., et al., *A metaproteomic approach to study human-microbial ecosystems at the mucosal luminal interface.* PLoS One, 2011. **6**(11): p. e26542.

80. Presley, L.L., et al., *Host-microbe relationships in inflammatory bowel disease detected by bacterial and metaproteomic analysis of the mucosal-luminal interface.* Inflamm Bowel Dis, 2012. **18**(3): p. 409-17.

81. Roy, K., et al., *Proteomic investigation of the adaptation of Lactococcus lactis to the mouse digestive tract.* Proteomics, 2008. **8**(8): p. 1661-76.

82. Alpert, C., et al., *Adaptation of protein expression by Escherichia coli in the gastrointestinal tract of gnotobiotic mice.* Environ Microbiol, 2009. **11**(4): p. 751-61.

83. Mahowald, M.A., et al., *Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla.* Proc Natl Acad Sci U S A, 2009. **106**(14): p. 5859-64.

84. 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 Biol, 2013. **11**(8): p. e1001637.

85. Wang, X., et al., *Comparative microbial analysis of paired amniotic fluid and cord blood from pregnancies complicated by preterm birth and early-onset neonatal sepsis.* PLoS One, 2013. **8**(2): p. e56131.

86. Aagaard, K., et al., *The placenta harbors a unique microbiome.* Sci Transl Med, 2014. **6**(237): p. 237ra65.

87. Ardissone, A.N., et al., *Meconium microbiome analysis identifies bacteria correlated with premature birth.* PLoS One, 2014. **9**(3): p. e90784.

88. Matamoros, S., et al., *Development of intestinal microbiota in infants and its impact on health.* Trends Microbiol, 2013. **21**(4): p. 167-73.

89. Groer, M.W., et al., *Development of the preterm infant gut microbiome: a research priority.* Microbiome, 2014. **2**: p. 38.

90. Morrow, A.L., et al., *Early microbial and metabolomic signatures predict later onset of necrotizing enterocolitis in preterm infants.* Microbiome, 2013. **1**(1): p. 13.

91. Stewart, C.J., et al., *Development of the preterm gut microbiome in twins at risk of necrotising enterocolitis and sepsis.* PLoS One, 2013. **8**(8): p. e73465.

92. Schnabl, K.L., et al., *Necrotizing enterocolitis: a multifactorial disease with no cure.* World J Gastroenterol, 2008. **14**(14): p. 2142-61.

93. Morowitz, M.J., et al., *Redefining the role of intestinal microbes in the pathogenesis of necrotizing enterocolitis.* Pediatrics, 2010. **125**(4): p. 777-85.

94. Gritz, E.C. and V. Bhandari, *The human neonatal gut microbiome: a brief review.* Front Pediatr, 2015. **3**: p. 17.

95. Afrazi, A., et al., *New insights into the pathogenesis and treatment of necrotizing enterocolitis: Toll-like receptors and beyond.* Pediatr Res, 2011. **69**(3): p. 183-8.

96. Boccia, D., et al., *Nosocomial necrotising enterocolitis outbreaks: epidemiology and control measures.* Eur J Pediatr, 2001. **160**(6): p. 385-91.

97. Raveh-Sadka, T., et al., *Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development.* Elife, 2015. **4**.

98. Klaassens, E.S., W.M. De Vos, and E.E. Vaughan, *Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract.* Applied and environmental microbiology, 2007. **73**(4): p. 1388-1392.

99. Young, J.C., et al., *Metaproteomics reveals functional shifts in microbial and human proteins during a preterm infant gut colonization case.* Proteomics, 2015. **15**(20): p. 3463-73.

100. Rooijers, K., et al., *An iterative workflow for mining the human intestinal metaproteome.* BMC Genomics, 2011. **12**: p. 6.

101. Sharma, R., et al., *Coupling a detergent lysis/cleanup methodology with intact protein fractionation for enhanced proteome characterization.* Journal of proteome research, 2012. **11**(12): p. 6008-6018.

102. Wisniewski, J.R., et al., *Universal sample preparation method for proteome analysis.* Nat Methods, 2009. **6**(5): p. 359-62.

103. Wu, F., et al., *Comparison of surfactant-assisted shotgun methods using acid-labile surfactants and sodium dodecyl sulfate for membrane proteome analysis.* Anal Chim Acta, 2011. **698**(1-2): p. 36-43.

104. Yu, Y.Q., et al., *Enzyme-friendly, mass spectrometry-compatible surfactant for in-solution enzymatic digestion of proteins.* Anal Chem, 2003. **75**(21): p. 6023-8.

105. Masuda, T., M. Tomita, and Y. Ishihama, *Phase transfer surfactant-aided trypsin digestion for membrane proteome analysis.* J Proteome Res, 2008. **7**(2): p. 731-40.

106. Masuda, T., et al., *Unbiased quantitation of Escherichia coli membrane proteome using phase transfer surfactants.* Mol Cell Proteomics, 2009. **8**(12): p. 2770-7.

107. Motoyama, A. and J.R. Yates III, *Multidimensional LC separations in shotgun proteomics.* Analytical chemistry, 2008. **80**(19): p. 7187-7193.

108. Taylor, G., *Disintegration of Water Drops in an Electric Field.* Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 1964. **280**: p. 383-397.

109. Rayleigh, L., *On the equilibrium of liquid conducting masses charged with electricity.* Philosophical Magazine, 1882. **14**: p. 184-186.

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

111. Wilm, M. and M. Mann, *Analytical properties of the nanoelectrospray ion source.* Anal Chem, 1996. **68**(1): p. 1-8.

112. Wilm, M., *Principles of electrospray ionization.* Mol Cell Proteomics, 2011. **10**(7): p. M111 009407.

113. Schwartz, J.C., M.W. Senko, and J.E. Syka, *A two-dimensional quadrupole ion trap mass spectrometer.* J Am Soc Mass Spectrom, 2002. **13**(6): p. 659-69.

114. Makarov, A., *Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis.* Anal Chem, 2000. **72**(6): p. 1156-62.

115. Park, C.H., *Further study of electron multiplication in conventional continuous dynode electron multiplier.* 2003 Ieee Nuclear Science Symposium, Conference Record, Vols 1-5, 2004: p. 1398-1400.

116. Olsen, J.V., et al., *A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed.* Mol Cell Proteomics, 2009. **8**(12): p. 2759-69.

117. Second, T.P., et al., *Dual-pressure linear ion trap mass spectrometer improving the analysis of complex protein mixtures.* Anal Chem, 2009. **81**(18): p. 7757-65.

118. Hu, Q., et al., *The Orbitrap: a new mass spectrometer.* J Mass Spectrom, 2005. **40**(4): p. 430-43.

119. Graumann, J., et al., *A framework for intelligent data acquisition and real-time database searching for shotgun proteomics.* Mol Cell Proteomics, 2012. **11**(3): p. M111 013185.

120. Blackburn, K., et al., *Improving protein and proteome coverage through data-independent multiplexed peptide fragmentation.* J Proteome Res, 2010. **9**(7): p. 3621-37.

121. Bilbao, A., et al., *Processing strategies and software solutions for data-independent acquisition in mass spectrometry.* Proteomics, 2015. **15**(5-6): p. 964-80.

122. Zhang, Y., et al., *Effect of dynamic exclusion duration on spectral count based quantitative proteomics.* Anal Chem, 2009. **81**(15): p. 6317-26.

123. Shukla, A.K. and J.H. Futrell, *Tandem mass spectrometry: dissociation of ions by collisional activation.* J Mass Spectrom, 2000. **35**(9): p. 1069-90.

124. Wells, J.M. and S.A. McLuckey, *Collision-induced dissociation (CID) of peptides and proteins.* Methods Enzymol, 2005. **402**: p. 148-85.

125. Wysocki, V.H., et al., *Mobile and localized protons: a framework for understanding peptide dissociation.* J Mass Spectrom, 2000. **35**(12): p. 1399-406.

126. Boyd, R. and A. Somogyi, *The mobile proton hypothesis in fragmentation of protonated peptides: a perspective.* J Am Soc Mass Spectrom, 2010. **21**(8): p. 1275-8.

127. Roepstorff, P. and J. Fohlman, *Proposal for a common nomenclature for sequence ions in mass spectra of peptides.* Biomed Mass Spectrom, 1984. **11**(11): p. 601.

128. Benjamini, Y., et al., *Controlling the false discovery rate in behavior genetics research.* Behav Brain Res, 2001. **125**(1-2): p. 279-84.

129. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nat Methods, 2007. **4**(3): p. 207-14.

130. Nesvizhskii, A.I. and R. Aebersold, *Interpretation of shotgun proteomic data: the protein inference problem.* Mol Cell Proteomics, 2005. **4**(10): p. 1419-40.

131. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-67.

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

133. Geer, L.Y., et al., *Open mass spectrometry search algorithm.* J Proteome Res, 2004. **3**(5): p. 958-64.

134. Craig, R. and R.C. Beavis, *TANDEM: matching proteins with tandem mass spectra.* Bioinformatics, 2004. **20**(9): p. 1466-7.

135. Ma, Z.Q., et al., *IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering.* J Proteome Res, 2009. **8**(8): p. 3872-81.

136. Wilkins, M.R., et al., *Guidelines for the next 10 years of proteomics.* Proteomics, 2006. **6**(1): p. 4-8.

137. Zybailov, B., et al., *Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling.* Anal Chem, 2005. **77**(19): p. 6218-24.

138. Sardiu, M.E. and M.P. Washburn, *Enriching quantitative proteomics with SI(N).* Nat Biotechnol, 2010. **28**(1): p. 40-2.

139. Kanehisa, M., *The KEGG database.* Novartis Found Symp, 2002. **247**: p. 91-101; discussion 101-3, 119-28, 244-52.

140. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.

141. Sanz, Y., A. Santacruz, and P. Gauffin, *Gut microbiota in obesity and metabolic disorders.* Proc Nutr Soc, 2010. **69**(3): p. 434-41.

142. Tremaroli, V. and F. Backhed, *Functional interactions between the gut microbiota and host metabolism.* Nature, 2012. **489**(7415): p. 242-9.

143. Nicholson, J.K., et al., *Host-gut microbiota metabolic interactions.* Science, 2012. **336**(6086): p. 1262-7.

144. Round, J.L. and S.K. Mazmanian, *The gut microbiota shapes intestinal immune responses during health and disease.* Nat Rev Immunol, 2009. **9**(5): p. 313-23.

145. Trosvik, P., N.C. Stenseth, and K. Rudi, *Convergent temporal dynamics of the human infant gut microbiota.* ISME J, 2010. **4**(2): p. 151-8.

146. Salzman, N.H., *Microbiota-immune system interaction: an uneasy alliance.* Curr Opin Microbiol, 2011. **14**(1): p. 99-105.

147. Cilieborg, M.S., M. Boye, and P.T. Sangild, *Bacterial colonization and gut development in preterm neonates.* Early Hum Dev, 2012. **88 Suppl 1**: p. S41-9.

148. Dominguez-Bello, M.G., et al., *Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns.* Proc Natl Acad Sci U S A, 2010. **107**(26): p. 11971-5.

149. Zeissig, S. and R.S. Blumberg, *Life at the beginning: perturbation of the microbiota by antibiotics in early life and its role in health and disease.* Nat Immunol, 2014. **15**(4): p. 307-10.

150. De Filippo, C., et al., *Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa.* Proc Natl Acad Sci U S A, 2010. **107**(33): p. 14691-6.

151. Morowitz, M.J., et al., *Strain-resolved community genomic analysis of gut microbial colonization in a premature infant.* Proc Natl Acad Sci U S A, 2011. **108**(3): p. 1128-33.

152. Hettich, R.L., et al., *Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities.* Anal Chem, 2013. **85**(9): p. 4203-14.

153. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST.* Bioinformatics, 2010. **26**(19): p. 2460-1.

154. Wu, S., et al., *WebMGA: a customizable web server for fast metagenomic sequence analysis.* BMC Genomics, 2011. **12**: p. 444.

155. Razumovskaya, J., et al., *A computational method for assessing peptide- identification reliability in tandem mass spectrometry analysis with SEQUEST.* Proteomics, 2004. **4**(4): p. 961-9.

156. Ma, Z.Q., et al., *ScanRanker: Quality assessment of tandem mass spectra via sequence tagging.* J Proteome Res, 2011. **10**(7): p. 2896-904.

157. Kalli, A., et al., *Evaluation and optimization of mass spectrometric settings during data-dependent acquisition mode: focus on LTQ-Orbitrap mass analyzers.* J Proteome Res, 2013. **12**(7): p. 3071-86.

158. Cantarel, B.L., et al., *Strategies for metagenomic-guided whole-community proteomics of complex microbial environments.* PLoS One, 2011. **6**(11): p. e27173.

159. Ma, B., et al., *PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry.* Rapid Commun Mass Spectrom, 2003. **17**(20): p. 2337-42.

160. Frank, A. and P. Pevzner, *PepNovo: de novo peptide sequencing via probabilistic network modeling.* Anal Chem, 2005. **77**(4): p. 964-73.

161. Mo, L., et al., *MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry.* Anal Chem, 2007. **79**(13): p. 4870-8.

162. Mallick, P., et al., *Computational prediction of proteotypic peptides for quantitative proteomics.* Nat Biotechnol, 2007. **25**(1): p. 125-31.

163. Krey, J.F., et al., *Accurate label-free protein quantitation with high- and low-resolution mass spectrometers.* J Proteome Res, 2014. **13**(2): p. 1034-44.

164. Tu, C., et al., *Systematic assessment of survey scan and MS2-based abundance strategies for label-free quantitative proteomics using high-resolution MS data.* J Proteome Res, 2014. **13**(4): p. 2069-79.

165. Walter, J. and R. Ley, *The human gut microbiome: ecology and recent evolutionary changes.* Annu Rev Microbiol, 2011. **65**: p. 411-29.

166. Hall, L.J., J. Walshaw, and A.J. Watson, *Gut microbiome in new-onset Crohn's disease.* Gastroenterology, 2014. **147**(4): p. 932-4.

167. Hofer, U., *Microbiome: bacterial imbalance in Crohn's disease.* Nat Rev Microbiol, 2014. **12**(5): p. 312.

168. Wright, E.K., et al., *Recent advances in characterizing the gastrointestinal microbiome in Crohn's disease: a systematic review.* Inflamm Bowel Dis, 2015. **21**(6): p. 1219-28.

169. Upadhyaya, S. and G. Banerjee, *Type 2 diabetes and gut microbiome: at the intersection of known and unknown.* Gut Microbes, 2015. **6**(2): p. 85-92.

170. Giongo, A., et al., *Toward defining the autoimmune microbiome for type 1 diabetes.* ISME J, 2011. **5**(1): p. 82-91.

171. Brooks, B., et al., *Strain-resolved microbial community proteomics reveals simultaneous aerobic and anaerobic function during gastrointestinal tract colonization of a preterm infant.* Front Microbiol, 2015. **6**: p. 654.

172. Guaraldi, F. and G. Salvatori, *Effect of breast and formula feeding on gut microbiota shaping in newborns.* Front Cell Infect Microbiol, 2012. **2**: p. 94.

173. Song, S.J., M.G. Dominguez-Bello, and R. Knight, *How delivery mode and feeding can shape the bacterial community in the infant gut.* Canadian Medical Association Journal, 2013. **185**(5): p. 373-374.

174. Koenig, J.E., et al., *Succession of microbial consortia in the developing infant gut microbiome.* Proc Natl Acad Sci U S A, 2011. **108 Suppl 1**: p. 4578-85.

175. Melville, J.M. and T.J. Moss, *The immune consequences of preterm birth.* Front Neurosci, 2013. **7**: p. 79.

176. Hunter, C.J., et al., *Understanding the susceptibility of the premature infant to necrotizing enterocolitis (NEC).* Pediatric Research, 2008. **63**(2): p. 117-123.

177. Jacquot, A., et al., *Dynamics and clinical evolution of bacterial gut microflora in extremely premature patients.* J Pediatr, 2011. **158**(3): p. 390-6.

178. Neu, J. and W.A. Walker, *Necrotizing enterocolitis.* N Engl J Med, 2011. **364**(3): p. 255-64.

179. Grave, G.D., et al., *New therapies and preventive approaches for necrotizing enterocolitis: report of a research planning workshop.* Pediatr Res, 2007. **62**(4): p. 510-4.

180. Zhou, Y., et al., *Longitudinal analysis of the premature infant intestinal microbiome prior to necrotizing enterocolitis: a case-control study.* PLoS One, 2015. **10**(3): p. e0118632.

181. Torrazza, R.M. and J. Neu, *The altered gut microbiome and necrotizing enterocolitis.* Clin Perinatol, 2013. **40**(1): p. 93-108.

182. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.

183. Conesa, A., et al., *Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.* Bioinformatics, 2005. **21**(18): p. 3674-6.

184. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.

185. Bindea, G., et al., *ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks.* Bioinformatics, 2009. **25**(8): p. 1091-3.

186. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.

187. Linden, S.K., et al., *Mucins in the mucosal barrier to infection.* Mucosal Immunology, 2008. **1**(3): p. 183-197.

188. Sommer, F. and F. Backhed, *The gut microbiota--masters of host development and physiology.* Nat Rev Microbiol, 2013. **11**(4): p. 227-38.

189. Wells, J.M., et al., *Epithelial crosstalk at the microbiota-mucosal interface.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**: p. 4607-4614.

190. Thompson, A.L., et al., *Milk- and solid-feeding practices and daycare attendance are associated with differences in bacterial diversity, predominant communities, and metabolic and immune function of the infant gut microbiome.* Front Cell Infect Microbiol, 2015. **5**: p. 3.

191. Hollister, E.B., et al., *Structure and function of the healthy pre-adolescent pediatric gut microbiome.* Microbiome, 2015. **3**: p. 36.

192. Morgan, X.C., et al., *Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment.* Genome Biol, 2012. **13**(9): p. R79.

193. Reichardt, N., et al., *Phylogenetic distribution of three pathways for propionate production within the human gut microbiota.* Isme Journal, 2014. **8**(6): p. 1323-1335.

194. Nafday, S.M., et al., *Short-chain fatty acids induce colonic mucosal injury in rats with various postnatal ages.* Pediatric Research, 2005. **57**(2): p. 201-204.

195. Darwin, A.J., *The phage-shock-protein response.* Mol Microbiol, 2005. **57**(3): p. 621-8.

196. Poole, K., *Bacterial stress responses as determinants of antimicrobial resistance.* Journal of Antimicrobial Chemotherapy, 2012. **67**(9): p. 2069-2089.

197. Sherman, M.P., *Lactoferrin and necrotizing enterocolitis.* Clin Perinatol, 2013. **40**(1): p. 79-91.

198. Tsuji, S., et al., *Human intelectin is a novel soluble lectin that recognizes galactofuranose in carbohydrate chains of bacterial cell wall.* Journal of Biological Chemistry, 2001. **276**(26): p. 23456-23463.

199. Bates, J.M., et al., *Intestinal alkaline phosphatase detoxifies lipopolysaccharide and prevents inflammation in zebrafish in response to the gut microbiota.* Cell Host Microbe, 2007. **2**(6): p. 371-82.

200. Johansen, F.E. and C.S. Kaetzel, *Regulation of the polymeric immunoglobulin receptor and IgA transport: new advances in environmental factors that stimulate pIgR expression and its role in mucosal immunity.* Mucosal Immunology, 2011. **4**(6): p. 598-602.

201. Kobayashi, K., et al., *Distribution and partial characterisation of IgG Fc binding protein in various mucin producing cells and body fluids.* Gut, 2002. **51**(2): p. 169-176.

202. Nacken, W., et al., *S100A9/S100A8: Myeloid representatives of the S100 protein family as prominent players in innate immunity.* Microsc Res Tech, 2003. **60**(6): p. 569-80.

203. Madsen, J., J. Mollenhauer, and U. Holmskov, *Review: Gp-340/DMBT1 in mucosal innate immunity.* Innate Immun, 2010. **16**(3): p. 160-7.

204. Huang, D.W., et al., *The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.* Genome Biol, 2007. **8**(9): p. R183.

205. Linke, D., et al., *Trimeric autotransporter adhesins: variable structure, common function.* Trends Microbiol, 2006. **14**(6): p. 264-70.

206. Buckling, A., et al., *Siderophore-mediated cooperation and virulence in Pseudomonas aeruginosa.* FEMS Microbiol Ecol, 2007. **62**(2): p. 135-41.

207. Horvath, P. and R. Barrangou, *CRISPR/Cas, the immune system of bacteria and archaea.* Science, 2010. **327**(5962): p. 167-70.

208. Gonzalez-Rodriguez, I., et al., *Role of extracellular transaldolase from Bifidobacterium bifidum in mucin adhesion and aggregation.* Appl Environ Microbiol, 2012. **78**(11): p. 3992-8.

209. Campos, E., et al., *Regulation of expression of the divergent ulaG and ulaABCDEF operons involved in LaAscorbate dissimilation in Escherichia coli.* J Bacteriol, 2004. **186**(6): p. 1720-8.

210. Claud, E.C., *Neonatal Necrotizing Enterocolitis -Inflammation and Intestinal Immaturity.* Antiinflamm Antiallergy Agents Med Chem, 2009. **8**(3): p. 248-259.

211. Jakobsson, H.E., et al., *The composition of the gut microbiota shapes the colon mucus barrier.* EMBO Rep, 2015. **16**(2): p. 164-77.

212. Berkes, J., et al., *Intestinal epithelial responses to enteric pathogens: effects on the tight junction barrier, ion transport, and inflammation.* Gut, 2003. **52**(3): p. 439-51.

213. Martin, R., et al., *Early life: gut microbiota and immune development in infancy.* Benef Microbes, 2010. **1**(4): p. 367-82.

214. Paulo, J.A., et al., *Effects of MEK inhibitors GSK1120212 and PD0325901 in vivo using 10-plex quantitative proteomics and phosphoproteomics.* Proteomics, 2015. **15**(2-3): p. 462-73.

VITA

Weili Xiong was born and raised in Nanchang, Jiangxi, China. She completed her Bachelor of Engineering degree in Biomedical Engineering, Bachelor of Laws degree in Medical Law, and Master of Science degree in Biochemistry and Molecular Biology in Tianjin Medical University, Tianjin, China. After completing her Master degree, she began her Ph.D. studies in the UTK-ORNL Graduate School of Genome Science and Technology in August 2010. She expects to receive her Ph.D. degree in May 2016.