12-2011

Characterization of the Human Host Gut Microbiome with an Integrated Genomics / Proteomics Approach

Alison Russell Erickson
arusse24@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Alison Russell Erickson entitled "Characterization of the Human Host Gut Microbiome with an Integrated Genomics / Proteomics Approach." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Robert L. Hettich, Major Professor

We have read this dissertation and recommend its acceptance:

Loren Hauser, Cynthia Peterson, Steven Wilhelm, Brynn Voy

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

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

Characterization of the Human Host Gut Microbiome with an Integrated Genomics / Proteomics Approach

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

Alison Russell Erickson
December 2011

DEDICATION

This dissertation is dedicated to my husband, Dr. Brian Erickson, for his patience and continual support throughout my education, career development, and our marriage. I also dedicate my dissertation work to my parents, Mr. and Mrs. Lee Russell, and my brother, Brandon Russell, for their eternal love, patience, and motivation to succeed throughout my childhood and adulthood.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Robert Hettich, for his guidance and support throughout my graduate career. I would like to acknowledge my committee members: Dr. Loren Hauser, Dr. Cynthia Peterson, Dr. Steven Wihelm, and Dr. Brynn Voy for their time, helpful comments, suggestions, and debate regarding this work, and for challenging me to think and step outside of my comfort zone. I would like to thank the Organic and Biological Mass Spectrometry group and Nathan VerBerkmoes at the Oak Ridge National Laboratory for their assistance, guidance, and patience throughout the course of this work. I would also like to thank the staff, Dr. Von Arnim, Terri Yeatts, and Kay Gardner of the Genome Science and Technology program at the University of Tennessee for their assistance and personal support throughout my graduate career. A special thanks to Dr. Kevin Lewis for the initial opportunities of exploring graduate school and scientific research at Texas State University. The education, motivation, and respect provided by Dr. Lewis provided me with the necessary desire, motivation, and tools to pursue my current PhD. Lastly, a special thanks to the love of my life, Brian Erickson, for providing a wealth of unconditional support and assistance throughout the course of this work.

ABSTRACT

The new field of ‘omics’ has spawned the development of metaproteomics, an approach that has the ability to identify and decipher the metabolic functions of a proteome derived from a microbial community that is largely uncultivable. With the development and availabilities of high throughput proteomics, high performance liquid chromatography coupled to mass spectrometry (MS) has been leading the field for metaproteomics. MS-based metaproteomics has been successful in its’ investigations of complex microbial communities from soils to the human body.

Like the environment, the human body is host to a multitude of microorganisms that reside within the skin, oral cavity, vagina, and gastrointestinal tract, referred to as the human microbiome. The human microbiome is made up of trillions of bacteria that outnumber human genes by several orders of magnitude. These microbes are essential for human survival with a significant dependence on the microbes to encode and carryout metabolic functions that humans have not evolved on their own. Recently, metaproteomics has emerged as the primary technology to understand the metabolic functional signature of the human microbiome.

Using a newly developed integrated approach that combines metagenomics and metaproteomics, we attempted to address the following questions: i) do humans share a core functional microbiome and ii) how do microbial communities change in response to disease. This resulted in a comprehensive identification and characterization of the metaproteome from two healthy human gut microbiomes. These analyses have resulted in an extended application to characterize how Crohn’s disease affects the functional signature of the microbiota.

Contrary to measuring highly complex and representative gut metaproteomes is a less complex, controlled human-derived microbial community present in the gut of gnotobiotic mice. This human gut model system enhanced the capability to directly monitor fundamental interactions between two dominant phyla, Bacteroides and Firmicutes, in gut microbiomes colonized with two or more phylotypes. These analyses revealed membership abundance and functional differences between phylotypes when

present in either a binary or 12-member consortia. This dissertation aims to characterize host microbial interactions and develop MS-based methods that can provide a better understanding of the human gut microbiota composition and function using both approaches.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1D	Single-Dimenssion
2DE	Two-Dimensional Gel Electrophoresis
2D-PAGE	Two-Dimensional Polyacrylamide Gel Electrophoresis
AMD	Acid Mine Drainage
CAZy	Carbohydrate Active Enzymes
CCD	Colonic Crohn's Disease
CFU	Colony Forming Units
CID	Collision-Induced Dissociation
DTT	Dithiothreitol
ESI	Electrospray Ionization
FDR	False Discovery Rate
FFE	Free-flow electrophoresis system
FPR	False Positive Rate
FT-ICR	Fourier Transform Ion Cyclotron Resonance
FWHM	Full Width at Half Maximum
GCF	Gingival crevicular fluid
GH	glycoside hydrolases
GO	Gene ontology
H	Healthy
HF/HS	Hight-Fat/High-Sugar
HGMI	Human gut microbiome initiative
HM	Human Microbiome
HMP	Human Microbiome Project
IBD	Inflammatory Bowel Disease
ICD	Ileal Crohn's Disease
IEF	Isoelctric Focusing
LC	Liquid Chromatography

LF/HS	Low-Fat/High-Sugar
LF/PP	Low-fat/Plant Polysaccharide
MALDI	Matrix-Assisted Laser Desorption Ionization
MRM	Multiple Reaction Monitoring
MS	Mass Spectrometry
MS/MS	Tandem (fragment) Mass Spectrum
MS1	Full (survey) Spectrum
MS2	Tandem (fragment) Mass Spectrum
M/Z	Mass to Charge
MudPIT	Multidimensional Protein Identification Tool
MW	Molecular weight
NSAF	Normalized Spectral Abundance Factor
pI	Isoelectric Point
PLs	Polysaccharide lyases
PPM	Parts per Million
PSM	Peptide-spectrum match
PTM	Post-translational Modification
PUL	Polysaccharide utlilization loci
qPCR	Quantitative Polymerase Chain Reaction
QqQ	Triple quadrupole
RP	Reverse Phase
SCX	Strong Cation Exchange
SDS	Sodium dodecyl sulfate
SRM	Selected Reaction Monitoring
TCA	Trichloroacetic Acid
TOF	Time-of-flight

Chapter One

Characterization of human host-microbiome interactions at the molecular level with metaproteomics approaches

Part of the introduction is adapted from the 'proteomics and metaproteomics' chapter in 'The Human Microbiome' book (CABI with editor Dr. Julian Marchesi) written by Alison R Erickson (2012 release date).

1.1: Introduction

In a natural ecosystem, microbes do not exist in isolation, but rather in populations and communities in which competition and cooperation are essential to shaping the composition and function of a microbial community. To understand microbial community composition, structure, function, and evolution, research has focused on development of approaches to advance beyond single pure-culture laboratory experiments to *in situ* analyses of environmental microbial populations and communities, since single microbial isolates in lab cultures do not accurately capture the complexities of microbial interactions in environmental communities. An organism cultivated in the laboratory may not represent or reflect its true activity and physiology in a natural environment where conditions such as resource competition and predation are widespread[1]. In addition, estimates suggest that ~90% of the microorganisms inhabiting the environment are not cultivable[2,3]. As a result, intensive research efforts have focused on improving methodologies for cloning, sequencing, and annotating whole genomes from heterogeneous microbial environments. For many complex environmental communities, metagenomics (genomic sequencing and analysis of uncultured microbes [4]) has provided insight into the genetic diversity, evolution, and metabolic potential of uncultivable microorganisms[5,6,7,8] otherwise not possible with traditional laboratory techniques.

With the emergence of metagenomics, extensive research has focused on sequencing and characterization of environmental microbial communities collected from extreme ecosystems, such as the Acid Mine Drainage[9] and hydrothermal systems[10],

the ocean[11,12], soils[13], and more recently the human body[14]. The human body is one example of a unique ecosystem where microbes and the human host live in symbiosis (**Figure 1.1**), in which the microbial cells outnumber human cells by 10-fold. Our collective microbial counterpart inhabits multiple body sites (e.g., skin, vagina, oral and gastrointestinal tract) and is referred to as the human microbiome. Although the human microbiome is made up of thousands of bacterial species (**Figure 1.2),** ~20-60% of the bacteria inhabiting the human-associated microbiome cannot be cultured[15,16,17,18,19]. For example, ~50% of the human oral microbiome is estimated to be non-cultivable[20]. However, with the increasing availability and quantity of human microbiota-associated metagenomic sequence data[21,22,23,24,25], we can begin to study and understand the human microbiome in both healthy and disease conditions. This sequence data has already increased our knowledge of the human microbiota gene content and variability and paved the way for systems biology ('omic') type studies, in particular making metaproteomics, a comprehensive community proteome analysis, feasible [26]. These whole community genomic sequences currently serve as the foundation for metaproteomics in the human microbiome (**Figure 1.3**) and enable the identification of hundreds to thousands of proteins. However, the dogma that DNA and RNA are equivalent is no longer considered accurate, as suggested by Li *et al.,* mandating that the primarily focus cannot only be on DNA, but RNA as a separate entity, since its alterations induce changes in the final end product, the proteins[27]. Although metagenomic sequencing unveils the collective functional 'potential' of a microbial community, this prediction is not directly related to the 'actual' host-microbial functional signature where protein information provides a deeper look at the molecular activities. Because metagenomics only captures the complete reportiore of genome capacity, it will be the integration of metaproteomics and/or metatranscriptomics with metagenomics that can serve as a powerful tool to study and collectively characterize the functional and metabolic signatures of the complex human microbiome.

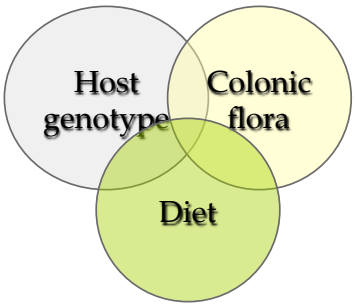


Figure 1.1: An interconnected landscape of host genetics, microbiota, and external factors such as diet regulate the stability of the unique human ecosystem where microbes and the human host live in symbiosis.

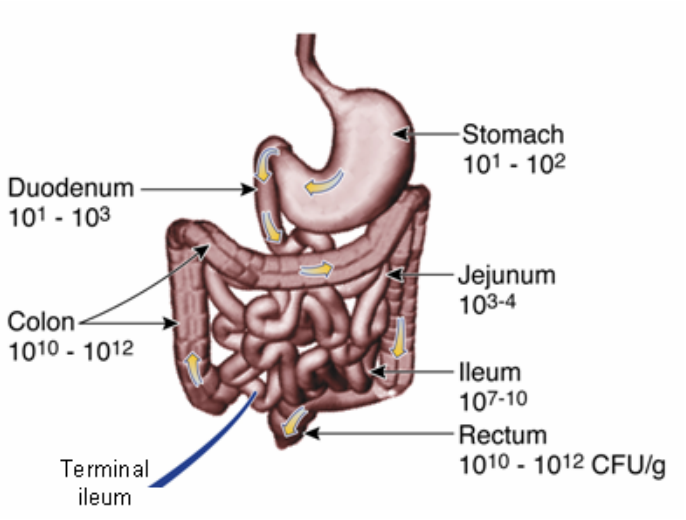


Figure 1.2: Microbial density in the gastrointestinal tract[26].

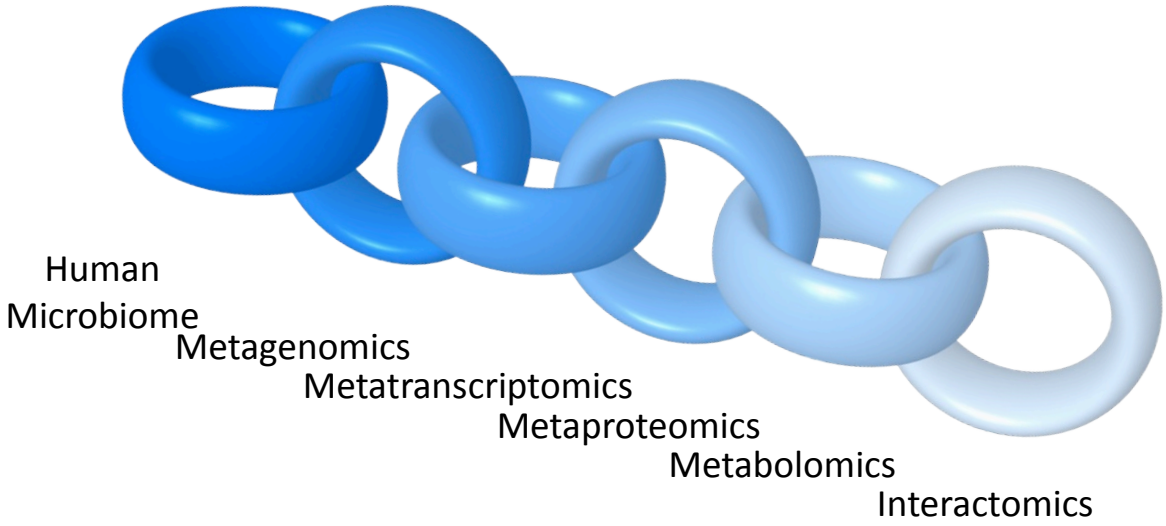


Figure 1.3: Integration and interconnectivity of metaproteomics with other 'omic' disciplines, with metagenomics as the primary foundation for all other 'omics.' Metagenomics provides DNA information, metatranscriptomics provides RNA information, metaproteomics provide protein-level information, metabolomics provides information about small-molecule metabolites, and interactomics provides information about all interactions between proteins and other molecules.

1.2: Microbial Community Functional Analysis

Proteomics, the identification and cataloguing of the entire suite of proteins translated in an organelle, organism(s), or tissue, has begun to rise in significance in the 'post-genomic era', since proteomics reveals the final gene products that are inscribed in the genome "dictionary." It is the proteins, not genes, that are the active enzymatic and metabolic players, and their complex network interactions and pathways that are responsible for the complexity of humans and their microorganisms' phenotype[28] in the human microbiome. For example, the specific order and arrangement of genes in the genome does not provide any information about the structures and functions of protein complexes. Studies have shown that proteins rarely function on their own, but rather usually exist in multi-component complexes and function with remarkable specificity[29,30]. Protein-protein interactions and post-translation modifications and are also very important and are not revealed by genomes or metagenomes. Therefore, to understand the human microbiome, one has to not only identify and characterize the gene content, but identify how its complete suite of proteins actually function *in vivo*, which can not be revealed through metagenomics.

To enable comprehensive functional analysis, high-throughput sequencing approaches (of transcripts and/or peptides) with high accuracy, sensitivity, and reproducibility are necessary to study the complex and diverse human microbiota. Microbial community functionality can be measured with either metaproteomics or metatranscriptomics, which is the sequencing of community mRNA. Recent developments in microbial sequencing technologies from microarrays to RNA-seq have yielded a tremendous increase in the throughput, accuracy, and sequence coverage (number and length of reads) of microbial transcriptomes and metatranscriptomes. Metatranscriptomics has shown to be successful in the analysis of environmental communities[31,32,33,34,35,36] and recently was applied to the analysis of the human gastrointestinal microbiome[37,38,39,40,41]. In comparison, proteomics has also advanced from two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) and single protein measurement platforms to MS-based proteomics and metaproteomics. The first large-scale metaproteome characterizations of microbial communities from a

static snapshot in time include the low-diversity Acid Mine Drainage (AMD)[42,43], wastewater sludge[44], as well as the more complex ecosystems such as the ocean[45,46] and the human gastrointestinal[26,47] and salivary[48,49] microbiome. These omics' studies have demonstrated the experimental capabilities and feasibility of applying metatranscriptomics and metaproteomics to complex environmental communities, including the human microbiome. Unlike traditional laboratory techniques, both technologies are valuable resources that can be used to characterize the functionality of uncultivable microorganisms and how it relates back to the genomic and taxonomic diversity in microbial communities derived from complex natural environments. While both 'omic' technologies are viable, there are advantages and disadvantages to selectively use one over the other to understand microbial functionality, as outlined below.

With respect to metatranscriptomics, there are several challenges when working with bacterial RNA. For example, the half-life is short and fraction of obtainable bacterial mRNA is limited. Compared to eukaryotic mRNA, most bacterial mRNA do not have 3' poly-A tails[50], which complicates isolation and purification from other non-coding RNA types when bacterial RNA preps consist of ~50-80% of rRNA and tRNA[51]. Although several studies have begun to identify methods to remove or deplete these contaminating RNA species[52,53,54,55], this comes at the risk of potentially altering or disrupting the true composition and nature of the environmental community transcriptome. Similarly, technical challenges exist with for metaproteomics. Protein extraction from natural environments can be i) biased for high or low-abundance members within a community and ii) inefficient when exposed to complex sample matrices such as humic compounds commonly found in soils and sediment. The separation and resolution of thousands of proteins from a consortium of microorganisms can prove difficult with liquid chromotraphy (LC) timescales and current MS platforms. Additionally, the vast dynamic range and complexity of microorganisms and proteins within a consortium (e.g., 10-100 trillion microorganisms inhabiting the human gastrointestinal tract) can hinder the ability to comprehensively identify the proteome of all members of a microbial community. While both technologies have their limitations,

using either or both metatranscriptomics and metaproteomics, is still highly valuable in providing insight into the functional profile and physiological responses of a natural microbial community under various and extreme environmental conditions, which is not possible with metagenomics.

At this point in time, metaproteomics is better suited for understanding the actual functional signature of the microbiota, as it directly measures and determines the phenotype of the cells being studied. Metatranscriptomics (mRNA abundance) does not always provide a direct correlative link to protein activity[56], since the final gene products (proteins) can can be constitutively expressed and post-transcriptionally modified. Additionally, mRNA abundance is an unreliable estimator for the corresponding proteins' abundance[57,58,59]. Therefore, protein expression should not be directly predicted from mRNA expression data. Unlike metatranscriptomics, metaproteomics directly provides measurement information about protein abundances, turnover, post-translation modifications, and protein-protein interactions not possible with metatransciptomics. We will focus on proteomics and metaproteomics in the human microbiome from this point on.

1.3: Metaproteomics of Microbial Communities

Shotgun proteomics is a term commonly used to represent a variety of experimental (2D-PAGE and LC) and analytical methodologies (mass spectrometry) that identify the composite set of expressed gene products (proteins/proteome) collected from cells of a microorganism, organ, or tissue. Similarly but distinct, metaproteomics, as first defined by Wilmes and Bond, is identification of the suite of proteins that are derived directly from an environmental consortia that can contain a mixture of several microorganisms, and that cannot be binned into species or organism types[60]. Like shotgun DNA sequencing, shotgun proteomics consists of digesting proteins into peptides, which can be more easily separated by gel electrophoresis and/or liquid chromatography prior to analysis via mass spectrometry. Traditionally, shotgun proteomics has been used to catalogue all proteins from a single prokaryote (e.g., *E. coli*) or eukaryotic organism (e.g., *S. cerevisiae*) grown in culture under variable, but controlled, growth conditions to

evaluate the phenotypic changes as reflected by the measured proteome. Meanwhile, developments in high resolution mass spectrometry, LC-based separation, and genomic sequencing led to the viability of community proteomics of natural environmental samples (metaproteomics) as first demonstrated by the large-scale proteome measurements of the microorganisms inhabiting the AMD[42]. From this point on, shotgun proteomics was also no longer restricted to using cultivable organisms or artificially created mixtures of known proteins in culture. With the current state and adaptation of MS, metaproteomics has transitioned from low complexity communities consisting of few dominant organisms[5], to much more complex ecosystems such as soil with 10^3 – 10^6 taxa per gram of soil[28,61], human gut microbiome with 500-3,000 bacteria species[62], and the oral microbiome with 500-700 bacteria species[18,20,63,64,65].

In an attempt to identify the *entire protein complement* of a microbial community, shotgun MS-based proteomics has been the most effective and comprehensive tool to date. This involves the generation and identification of thousands of peptides in a single sample. Protein identifications are generated by matching experimental tandem mass spectra (MS/MS) against a peptide sequence database(s) using well-established programs, such as SEQUEST, Mascot, and X!Tandem, to identify peptides (peptide-spectrum matching)[66,67,68]. The accurate interpretation and assignment of MS/MS spectra is the first step in the informatics data processing pipeline, called database searching (discussed further in chapter two). Therefore, a relevant genome or metagenome (protein database) is the necessary starting point to infer biological meaning from complex environmental metaproteomes. Hence, the depth and quality of DNA sequencing have a significant impact on protein database searching. Relative to single microbial isolates, the complexity and sequence diversity (strain- and species-level variation) and reduced coverage of community metagenomic sequences can pose many challenges for metaproteomics. As additional metagenomes are acquired and the sequencing technologies and depth of coverage improve, this will correlate with an increase in protein identification and deeper proteome coverage of complex, human

microbial community metaproteomes using a metagenome(s) as the foundation for MS peptide-spectrum matching.

1.4: Human microbiome

Like other environmental microbial communities, the human microbiome is a complex and dynamic system that plays an important role in many aspects of human physiology. The human microbiome (i.e., the host of our microbial symbionts) not only consists of microbes living outside the body, but internally within the oral cavity, gastrointestinal tract, and vagina. Together, these microbes (microbiota) outnumber the human somatic and germ cells by 10:1 (**Figure 1.4**). A deep understanding of human genetic and physiologic diversity requires characterization of our microbiome by focusing on factors that influence its assembly, stability, functions, and functional variations. The Human Microbiome Project (HMP) is currently focused on generation of large datasets describing the microbial lineages and genes present in our gut communities. A central challenge will be to move beyond compositional information and develop ways of determining how these communities operate to influence human health as well as disease predisposition.

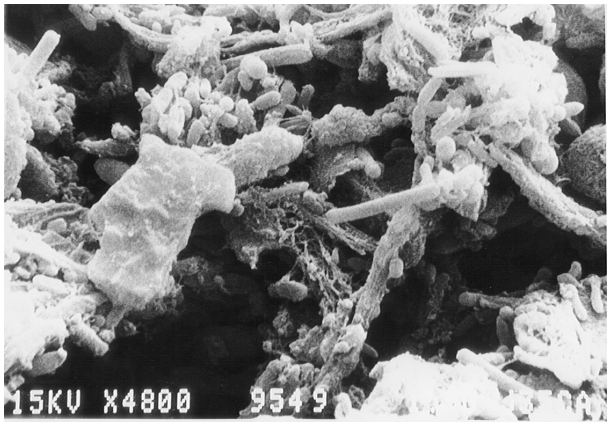


Figure 1.4: The complex healthy human gut microbiome (mixture of human, bacteria, and digested food components) and microbial diversity in human feces. Micrograph courtesy of Janet Jansson, LBNL.

Fortunately, with the advent of the HMP[14,69], a multitude of human-habitat associated microbial metagenomes have been sequenced and are publically available for researchers to use for additional investigations to help improve our understanding of the human microbiome. In addition, a number of human-derived microbial reference genomes[70] have been sequenced and are publically available; however, the vast majority of microorganisms have yet to be cultured. Currently, >1,000 microbial reference genomes collected from various sites of the human body including the mouth, skin, gut, and vagina are publically available. Therefore, when a metagenome is not available or representative for a particular sample, the human-derived microbial reference genomes could be used as a substitute for a metagenome(s). Metaproteomics can also take advantage of these metagenomes and human-derived reference genomes for protein database searches against relevant human proteome samples. While metagenomic predicted protein sequence database searches provide highly relevant proteome information, it can be difficult to unambiguously assign microbial species/strain information to many proteins. The reference genomes can be used to overcome this by providing definitive species/protein identifications, which can be used separately or as a complement to metagenomic predicted protein database searches[71]. Several common experimental, analytical, and informatics workflows suitable for MS-based proteomics will be explored in more detail with respect to their applicability and challenges with the human gut microbiome.

The objective of this dissertation research is a comprehensive and mechanistic understanding of the microbial functional signature in the gut microbiome. To achieve this goal, it was necessary to begin with a lower complexity, synthetic gut microbiome and then progress to a more realistic and complex human gut microbiome. Challenges are present in both approaches, but both approaches provide different information that will eventually contribute to an overall larger understanding of how the human gut, normal or diseased, functions with our microbial counterparts. For both approaches, liquid chromatography coupled with tandem mass spectrometry will be used to characterize the gut microbial community proteomes of human twins and gnotobiotic mice.

1.4.1: Shotgun proteomics of a model human gut microbiome

It has become clear that the human gut manages to function in symbiosis with an indefinite number of microorganisms which are necessary for normal gut function. For a less complex, systematic approach, gnotobiotic mice were used as a model system due to the ability to control what microbial flora is present in the gut. Collaborator Dr. Jeffrey Gordon (Washington University, St. Louis, MO) has sequenced a multitude of genomes from members of the two dominant phyla present in the normal distal human gut microbiota: the Firmicutes and the Bacteroidetes[62,72]. To explore the interactions between the Bacteroidetes and Firmicutes *in vivo*, adult germ-free mice were colonized by the Gordon group with either: two, seven, or twelve human-derived microorganisms. For example, to examine the fundamental interactions between these two phyla in gut biomes, germ-free mice were colonized with either *B. thetaiotaomicron* or *E. rectale*, or both (chapter 3). These gnotobiotic mice provided a novel model system in which to study not only microbial mono- versus bi-association, but up to twelve microbial inhabitants at a single point in time. The overall goal of this initial two-member community study was to demonstrate proteomic measurements on gut microbiomes from gnotobiotic mice. With increasing complexity, the goal was to (i) improve experimental and informatics applications to resolve closely-related species within a low-complexity system (chapter 4) and (ii) elucidate information about the functional activities of these low complexity microbial systems under various dietary conditions (chapter 5). The results presented in chapters 3, 4, and 5 emphasize the value of combining gnotobiotics with high resolution proteomics as a strategy for developing the experimental and computational pipeline needed to characterize gene expression in more complex, human body habitat-associated microbial communities.

1.4.2: Metaproteomics of the human gut microbiome

The composition and stability of the gut microbiota can be disrupted by external factors influencing the likelihood of developing inflammatory bowel diseases (IBD) and the propensity for obesity. IBD can be divided into two disease categories: Ulcerative Colitis and Crohn's Disease. Crohn's is a chronic, relapsing, immunologically mediated

disorder that can have severe physical consequences. The current hypothesis is that this disease is due to an overly aggressive immune response to a subset of commensal enteric bacteria. Studies to date on IBD have suggested that the disorder may be caused by a combination of bacteria and host susceptibility. Until recently, no study has reported the use of advanced integrated systems biology techniques such as metagenomics and metaproteomics, for the characterization of the natural microflora in Crohn's patients.

A non-targeted MS-based approach is ideal for studying complex communities based on its ability to directly measure expressed proteins from complex environmental matrices. This approach was applied to elucidate the differences and functional activities of commensal microbiota between monozygotic concordant (genetically identical twins with the same trait) and discordant (genetically identical twins, but differ phenotypically for a trait) human twins with and without Crohn's disease with a focus on biological inference (chapter 6 and 8) in addition to method development (chapter 7).

1.5: Metaproteomics informatics for the gut microbiome

A challenge in MS-based proteomics is "protein inference" and the ability to accurately assign a peptide to the protein from which it originated. This peptide-protein assignment is often complicated by homology between proteins of different microbial strains/species found in environmental communities. In a protein database containing multiple bacterial reference genomes, gene redundancy between multiple strains/species belonging to the same genera can make it difficult to accurately assign a unique peptide to a MS/MS spectrum. Erickson *et al.* published a method whereby i) only matched metagenomic derived protein databases were searched against the same samples' tandem mass spectra (multiple metagenomes are not concatenated into one FASTA protein database) and ii) for proteins identified across multiple samples used for comparison, a clustering approach was used to cluster identified proteins with peptides that have >80% sequence identity to reduce the level of peptide and protein redundancy[71]. For the human gut microbiome, Cantarel *et al*., Erickson *et al*. and Roojers *et al*. suggest that if a matched metagenome is available, it should be used as

the protein database for the same metaproteome sample to comprehensively identify many MS/MS spectra without massive redundancy. If a matched metagenome is not available, a synthetic metagenome can be created by concatenating available metagenomes with relevance to the body site of interest, as proposed by Verberkmoes *et al*. and/or use some combination of the publically available human-derived reference genomes from the human microbiome project (HMP)[73].

Although matched metagenomes are highly desirable for several reasons, the human-derived bacterial reference genomes have sufficient sequencing coverage where the protein sequences are full length. Metagenomic sequencing of environmental samples are not often sequenced to a sufficient depth to contain the complete gene repertoire; as a result, the predicted genes are often fragmented, resulting in incomplete protein sequences in a database. However, because many of the genomic sequences captured from an environmental sample do not map to any available reference genome, relying solely on reference genomes for metaproteomic protein identification limits the proteins identified to only those in found in sequenced organisms, which is a very small proportion of the total bacteria that is cultivable from the HM. Additionally, the human microbiome is estimated to contain trillions of bacterial cells and thousands of bacterial species, many of which are uncharacterized. Because the majority of these bacteria have not been sequenced, their proteomes are likewise undecipherable. As a result, these 'unknown' bacteria and proteins (hypotheticals) cannot be assigned to MS/MS spectra with a protein database that contains only a collection of known sequenced reference isolate genomes. Additionally, the human-derived reference genomes do not have any disease-representation or individual sequence/strain variations, whereas a matched metagenome is usually derived from the same human individual sample that may be associated with some disease (e.g., ileal Crohns disease or gingivitis). An approach that uses either a metagenome or genome/metagenome database search strategy will be able to capture the presence of these unknown microorganisms/proteins and sequence variations which can more accurately represent and reflect the entire metaproteome of the human microbiota being studied. Finally, careful attention should be given to using only reference isolate

genomes, because unidentified and uncharacterized microorganisms and their proteins are equally important in revealing the function of the human microbiome.

In conclusion, the quality of environmental protein databases, whether it be a matched or relevant metagenome(s) or a collection of reference genomes, and database searching is a critical step and in many MS-based studies it is the bottleneck in the informatics workflow and metaproteomics pipeline. With the exponential increase in size, availability, and complexity of metagenomic sequence data from the human microbiome[25], metaproteomics investigators will be severely affected and hindered by the quality of the metagenomes in terms of obtaining full length contigs (incomplete protein encoding genes) and subsequently the accuracy of annotation, degree of technological sequencing errors, and impractical size of databases to generate reverse databases and accurately predict false discovery rates (FDRs) using traditional methods. There is a strong need for a new database search engine that is compatible with large environmental protein databases that takes into account many of the complications described above. Additionally, new informatics workflows that may include a combination of both protein database searching and *de novo* peptide sequencing may prove to be highly beneficial for covering the range of both known (database predicted proteins; chapter 7) and assist in revealing the unknown proteins that may not be sequenced (ie, due to low abundance) and/or are missed in the assembly of metagenomic sequence reads.

1.6: The future of metaproteomics in the HM

While metaproteomics has advanced significantly beyond low resolution two-dimensional gel electrophoresis (2DE), many obstacles still remain. Not only is metaproteomics heavily reliant upon mass spectrometry technologies, but also on the foundation of protein database searching, the metagenomes. To improve protein identification using MS-based proteomics, we will need the quality (e.g., sequence lengths, assembly and annotation) of metagenomic-predicted protein databases to improve in parallel. However, given the error rates in genomic sequencing, technologies that rely solely on database searching with genomes and/or metagenomes

may not be as effective or desired in the future, but some alternative(s) and/or a combination of methodologies that will provide for metaproteomics to achieve a deeper and wider level of proteome coverage in the human microbiome. The initial lack of metagenomic information hindered many environmental community studies from performing comprehensive and reliable protein identification. However, investigators are now readily sequencing environmental communities (ie, cost of sequencing has decreased with the simultaneous improvement of DNA sequencing techologies such as Illumina) where they can also be used for matched and/or relevant metagenomic–predicted protein database searches. With large-scale DNA sequencing efforts, metaproteomics-based functional analyses will simultaneously improve for protein identification in microbial communities inhabiting the human microbiome.

In order to cope with the large and increasing quantities of sequence data that has been acquired by large-scale sequencing efforts in the human microbiome, metaproteomics will equally require advanced technology and vast informatics resources to manage and preserve its significance within the 'omics' field. However, given the advancement and cost of sequencing technologies, the scientific community is driving towards smaller and smaller sequence read lengths from Sanger (~1,000bp) to 454 (~500bp) and now Illumina (75-100bp), which could potentially impede the advancement of metaproteomics by limiting the ability to achieve full length protein sequences in communities. While the sequencing depth of coverage and error rate is equally important for metagenomics and metaproteomics, the lack of full-length protein sequences can be equally if not more detrimental to protein identification and biological inference. Revolutionary tools and integrated 'omic' studies are highly sought after to enable the study of the web of events rather than a static snapshot of the activities taking place in the human microbiome.

Future MS-specific developments include the optimization of up-front cellular lysis and protein extraction methods in addition to the development of new high-throughput hybrid mass spectrometers capable of higher resolution and mass accuracy. While protein extraction methods are straightforward when applied to culture-based microorganisms, these methods are not always efficient or non-biased for the *in situ*

extraction of proteins from complex environmental matrices. As seen with soil and sediment samples, efficient cell lysis and extraction can be technically challenging. However, as investigators continue to adapt and improve lysis and extraction protocols for environmental samples[74] (chapter 4), protein identification will benefit significantly by increasing the accessibility, depth, and coverage of proteins contained in complicated environmental matrices such as human feces. Additionally, a direct cell lysis and protein extraction method of samples collected from the human microbiome may prove to be more efficient and/or representative of the microbial community compared to indirect methods that enrich for microorganisms. To characterize the metaproteome of the human microbiota one step further would incorporate the analysis of post-translational modifications and strain-level variants in addition to protoegenomics, where we can use metaproteomics and identified peptides to refine metagenomes and gene predictions to assist in the identification of false starts/stops, misassembled contigs, gene boundaries, and incorrect protein annotations. However, in order to accurately assess and correct these critical issues with high confidence, the acquirement of and use of high mass accurate mass spectrometric data is vital. High mass accurate mass spectrometers that are capable of discriminating all amino acids will be essential to filter and control for false positives as the size and redundancy increases with higher complexity environmental communities such as those in soil and the human microbiome.

As described above, protein redundancy and peptide degeneracy is a challenge and scales substantially with environmental communities such as the human gut. Due to the level of protein redundancy found in higher complexity microbial communities, it is difficult to assign and suggest that one species/strain is uniquely responsible for a specific function within the community using MS/MS spectral abundance. With current informatics workflows, it is often challenging to identify 1) to which proteins the measured peptides originated from, 2) to which organisms the identified proteins belong to, and 3) to estimate an accurate peptide or protein false discovery rate[75,76]. Even with the use of high mass accuracy, the available informatics algorithms are not capable of differentiating and classifying peptides that are not only shared amongst multiple

proteins, but between species and strains regardless of the database type (reference genomes and/or a matched or relevant metagenome) that is used for complex environmental communities. While many algorithms assign and differentiate unique from non-unique (shared) peptide identifications, a large portion of the peptide identifications are non-unique, complicating the accuracy of biological inference in environmental communities. Adaptations of either the available informatics workflows or the generation of new algorithms may prove to be more effective, accurate, and computationally higher throughput due to the sizes of metagenomic-derived protein databases for dealing with peptide and protein redundancy on a large-scale with environmental communities.

Other challenges that remain include the ability to assess the “complete” or entire bacterial functionality of the human microbiome due to i) the biological dynamic range of low abundant proteins and on a broader level, microbial species/strains with variable abundance in the same sample and ii) dynamic range limitations of the mass spectrometer instrumentation. With improvements in peptide separations (LC) and technological developments over the past couple of years, the dynamic ranges have increased by 1-2 orders of magnitude[77]. Technological developments have increased ion transmission and speed while delivering ultra-high resolution and accurate mass data as seen in the LTQ-Velos and LTQ-Oribtrap Velos[78] that have provided new capabilities to achieve deeper coverage of less abundant proteins. More recently, the Orbitrap Elite was released with a novel high-field mass analyzer that increases the speed, sensitivity, and dynamic range of complex proteome samples. Nevertheless, in spite of these partially solved challenges, metaproteomics is already providing remarkable insight into the functional activities of the gut and oral microbiota with technological advancements that will provide unrivaled capabilities for future metaproteomics analyses of the human microbiome.

With increasing large-scale DNA sequencing efforts, metagenomics will drive the emerging field of metatranscriptomics and metaproteomics in the human microbiome. With innovation and new developments emerging in MS-based proteomics, the scientific community will have access to many more capabilities to study the metaproteome of the

human microbiome. For example, it is anticipated that proteomics will be capable of identifying all isoforms and modifications (e.g., PTMs) on a large scale in the future. To achieve this, not only will shotgun proteomics (bottom-up) continue to prevail in its applications, but also higher-throughput top-down analysis of proteins will emerge as a necessity. It is the combination of both approaches, with advancement of all areas within mass spectrometry-based proteomics including technology and informatics workflows, which may serve as the revolutionary tool to fully characterize environmental community metaproteomes in the human microbiome.

1.7: Scope of the dissertation

This dissertation encompasses a set of experimental and informatics methods and advancements that have enabled biological inference of the highly complex and diverse human gut microbiome. As described above, the human gut microbiome is highly complex and diverse with thousands of microbial species. Therefore, to comprehensively identify and characterize the metaproteome of each microbial member inhabiting the gut can be challenging as described throughout this dissertation. Chapter 2 will provide a detailed experimental and methodological overview and platform for successful applications of MS-based proteomics to microbial communities collected from complex gut-related sample matrices (ceca and feces). As described in chapter 2, different approaches have been proposed and described as viable methods to effectively characterize the proteomes of gnotobiotic mice and metaproteomes of human individuals. An approach that focuses on a less complex and carefully designed gut microbiota allows for the focused study of human-derived microbial structure, cooperation, competition and adaptation *in vivo* as described in chapters 3-5. Chapter 3 will introduce the first application of MS-based proteomics to a human-derived microbial community in gnotobiotic mice. In this chapter, we define the interactions between two members of the Firmicutes and the Bacteroidetes that are commonly represented in the human gut microbiota. The functional differences between *B. thetaiotaomicron* and *E. rectale* are revealed through an integrated approach of genomics, transcriptomics and proteomics. This chapter lays the framework for comparing and developing new methods to increase protein identifications and coverage of individual microbes within a defined consortium as described in chapter 4. Chapter 4 will describe and compare several up-front sample processes for the lysis and extraction of proteins from a higher complexity, 7-member microbial community composed of only phylotypes belonging to Bacteroidetes. Chapter 5 will use the previous two chapters as a guide for both experimental applications and biological inference of a 12-member microbial consortium. This chapter will not only focus on characterizing the community proteome as a single entity, but each individual species and their responses to diet perturbations. In contrast to chapters 3-5, an approach that directly measures the expressed

metaproteome of a highly complex gut microbiome derived from feces can more accurately represent the diversity and abundances of a human gut microbiome as described in chapters 6-8. Chapter 6 will introduce the first application of high throughput MS-based metaproteomics to human gut microbiomes. This chapter specifically describes the capabilities of the first LC-MS application to healthy gut microbiomes collected from adult feces and establishes a baseline for the biological inference of healthy metaproteomes. This work lays the groundwork for developing new integrated informatics workflows to increase the identification and coverage of metaproteomes using a variety of metagenomic sequencing and assembly strategies as described in chapter 7. Chapter 7 will deliver a new integrated metagenomic/metaproteomic approach that uses metagenomic sequence reads as a database to search against matched tandem MS/MS spectra collected from a healthy human twin pair. Using the new methodology and informatics workflow discussed in chapter 7, chapter 8 will highlight the benefits of this method with increased protein, peptide, and spectra identifications in both healthy and diseased gut microbiomes. Compared to the previous two chapters, this chapter highlights the significant improvements and capabilities of MS-based metaproteomics in the revelation of a core gut microbiome in addition to the shared and functional differences in healthy and diseased (Crohn's disease) metaproteomes. Finally, chapter 9 will summarize the experimental, technological and informatics methods that served as the platform to permit in-depth biological inference of complex human-gut microbial metaproteomes of healthy and disease.

Chapter Two

Development of an integrated experimental/computational omics platform for human gut microbiome research

Part of the introduction is adapted from the 'proteomics and metaproteomics' chapter in 'The Human Microbiome' book (CABI with editor Dr. Julian Marchesi) written by Alison R Erickson (2012 release date).

2.1: Introduction

Metaproteomics has advanced from the protein identification in low-complexity ecosystems (AMD)[42] to highly complex microbial communities inhabiting the soil, ocean, and the human microbiome (HM). The capability to identify hundreds to thousands of proteins in the HM is predicated on experimental optimization of sample collection and lysis/preparation methods, high throughput liquid chromatographic separation coupled to tandem mass spectrometry (MS/MS), and integration with genomics and metagenomics to perform sequence database searching (**Figure 2.1**).

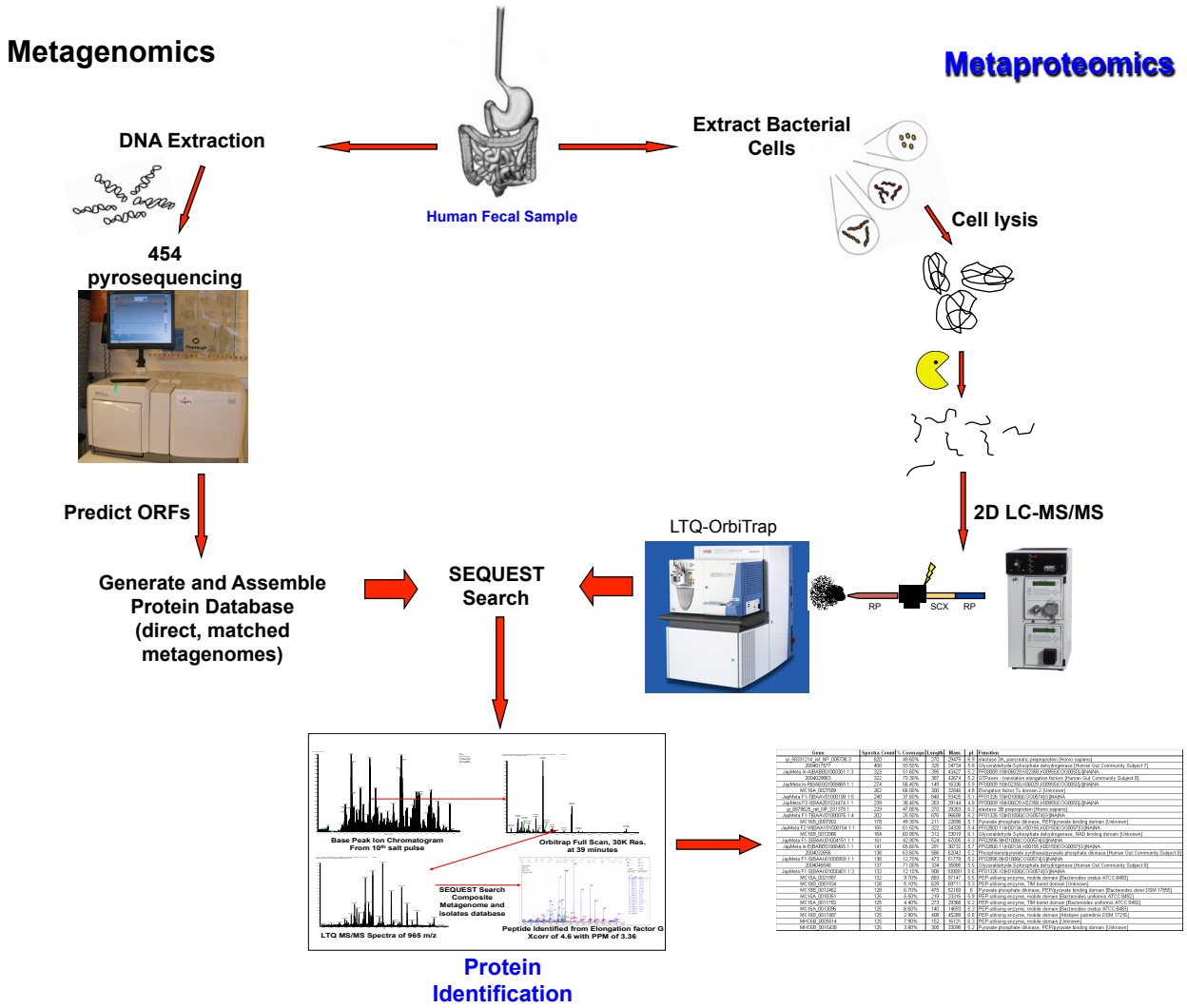


Figure 2.1: Experimental design for human gut microbiome sample collection and integrated metagenomic / metaproteomic MS characterization.

The collection of microbial cells and extraction of proteins from the human gut-associated microbiota is the one of the most important steps in the experimental design of MS-based proteomics to insure accurate representation of the collective microbiota that is sampled and sufficient biomass for all downstream processes and MS analysis. The primary goals for MS-based community proteomics are: efficient sample processing, peptide separation, high sequence coverage of proteins and the proteome, and coverage of high and low abundant organisms (dynamic range). Presently, a multitude of sample processing methods for microbial cell lysis, protein extraction,

denaturation and digestion, and purification are available to perform error-prone tandem mass spectrometry. In general, there are two general methods for microbial cell lysis where a sample is collected for proteomics and either i) derived, processed, and lysed directly (*in situ*) from the source (i.e., feces, tissue biopsy) where both microbial and human cells are included (direct approach) or ii) the collected sample is enriched for microbial cells to eliminate all human proteins and contaminants via centrifugation (ultra- or differential centrifugation; indirect approach). Currently there is no widely accepted or approved method suggesting that either the direct or indirect approach is better or worse than the other for metaproteomics of HM related samples, and this will be a key focus of part of the research under this dissertation.

Mass spectrometry is the most comprehensive tool available for large-scale proteomics[79] and metaproteomics for several reasons. MS is high-throughput, reproducible, unbiased, and highly versatile, with applications to a variety of sample types ranging from solids, gases, small molecules, to peptides and proteins (single, mixtures, or communities). MS can provide high detection sensitivity, resolution, and mass accuracy, unlike traditional methods of Western blotting. In addition, MS can be coupled with separation techniques to increase the dynamic range of higher complexity samples. A multitude of mass spectrometers are increasingly available and range based on their i) ionization source, ii) mass analyzer, and iii) data processing and ion detection source. The first component in the proteome measurement is the ionization source for proteins and peptides to be analyzed by MS. This can be accomplished with two primary ionization methods: matrix-assisted laser desorption ionization (MALDI)[80] and electrospray ionization (ESI)[81]. The second critical component is the mass analyzer, which sorts and measures ions based on their mass-to-charge ratios (*m/z*). The most common mass analyzers include 1) trapping mass spectrometers: Ion trap, Orbitrap, and Fourier transform-ion cyclotron resonance (FT-ICR) that use dynamic electrostatic or magnetic confinement, 2) ion-beam mass spectrometers: quadrupoles (Q) and time-of-flight (TOF) that utilize spatial resolution. Ion traps are generally most suited for bottom-up proteomics and LC-MS/MS of complex proteomes and mixtures. FT-ICR instruments are generally suited for top-down proteomics and PTM identification

due to its wide *m/z* range, high resolution (500,000 FWHM) and accuracy (< 1 ppm), however, the scan rate is much slower compared to the linear ion trap. The widely used triple quadrupole (QqQ) instrument contains a series of three quadrupoles (Q) that allows for selected/multiple ion reaction monitoring (SRM and MRM). Most investigators use the QqQ platform for targeted proteomics investigation rather than comprehensive proteome identification and coverage. Finally, TOF mass analyzers can be used for either top-down or bottom-up proteomics with high duty cycle, unlimited mass range, and low cost, however, low resolution limits its general application. Novel hybrid instruments integrating more than one mass analyzer (e.g., LTQ-FT-ICR, LTQ-Orbitrap, and QqTOF) have evolved from single mass analyzer instruments to combine multiple features (**Figure 2.2**) to provide more superior, faster, and robust measurement possibilities. For example, the newest revolution is the novel dual-pressure linear ion trap mass spectrometers LTQ Velos and LTQ-Orbitrap Velos with increased ion transmission, more efficient isolation and dissociation, greater resolution, and faster scan rates (compared to the LTQ-XL)[78]. For complex metaproteomes, this instrument permits higher acquisition rates while simultaneously isolating more low-abundant peptides and proteins.

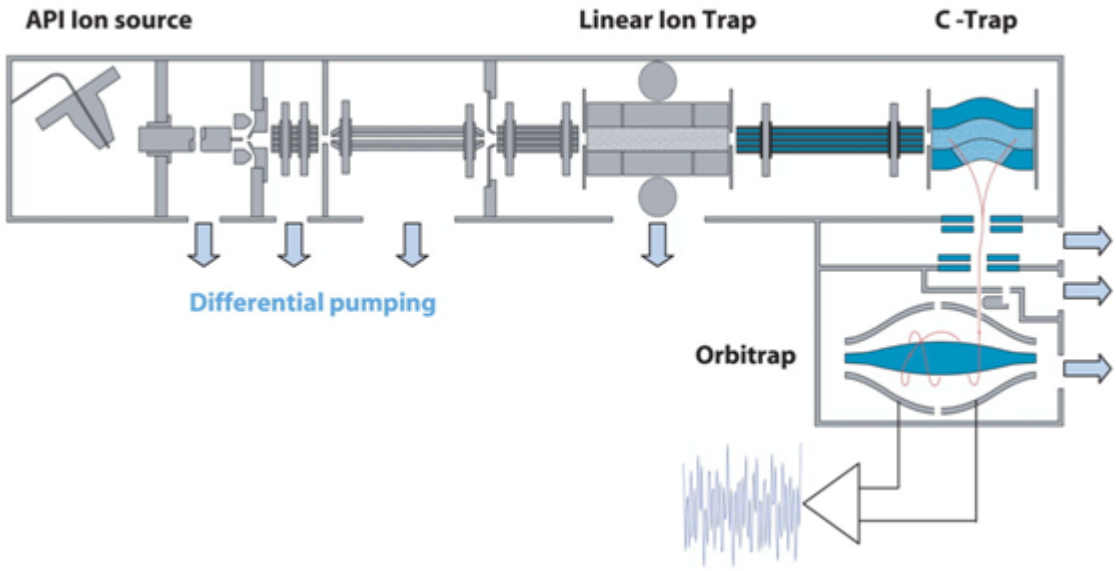


Figure 2.2: Schematic of the hybrid LTQ-Orbitrap[82]. The Orbitrap performs high resolution full MS scans and the linear trapping quadrupole performs MS/MS.

For the technical reasons described previously, mass spectrometry is an unparalleled analytical tool that has shown to be successful in its application to human microbiome-related samples. Briefly, several studies have applied a variety of MS platforms to the human gut and oral microbiome, such as the work by Klaassens et al. to use MALDI-TOF mass spectrometry to analyze the infants' gastrointestinal tryptic peptides[47]. While this study established the role of MS and metaproteomics in the human microbiome, it was very limited in protein and proteome coverage, with 55 excised protein spots and only one identified microbial protein with 91% identity to Bifidiobacterium. The lack of identifications may have been a result of the experimental technique (i.e., low resolution 2D gels), MS platform, and lack of peptide separation. VerBerkmoes *et al.* and Erickson *et al.* applied 2D-ESI-LC-MS/MS on a LTQ-Orbitrap mass spectrometer, specifically exploiting its' high mass accuracy capabilities to acquire low false discovery rates of peptides and proteins in these complex microbial communities[26,71]. Similarly, Rooijers *et al.* used an LTQ-Orbitrap mass spectrometer, but did not focus on high mass accurate identifications for their analyses of the human gut microbiota[83]. Hongwei and Rudeny *et al.* both used an LTQ mass spectrometer without the Orbitrap for all MS analysis on human whole saliva[48,84]. Grant et al. used a 7 T LTQ FT mass spectrometer for the analysis of human GCF to study the oral microbiome[49]. LC-ESI-MS/MS and hybrid mass spectrometers are the technology of choice for high-throughput peptide and protein identification[78] featuring greater sensitivity, acquisition rates, and resolution to accurately and comprehensively identify and characterize complex metaproteomes as those in the human microbiome.

While up-front sample collection, preparation, and selection of the most appropriate MS platform are important for complex human microbiota samples, the final post-MS/MS steps are critical for data interpretation and biological inference. Finally, the last step in the shotgun proteomics methodology and pipeline is informatics. Protein identifications are generated by matching MS/MS spectra against a sequence database(s) to identify the most accurate peptide-spectrum matches (**Figure 2.3**). The correct interpretation and assignment of such MS/MS spectra to peptides is the major step in the informatics data processing pipeline, called database searching. This step is

not only heavily reliant upon MS/MS quality, but the quality and accuracy of a sequence database. A ‘sequence database’ is a FASTA formatted file that contains the entire theoretical proteome and sequences that are predicted from a genome (e.g, bacterial isolate) or metagenome (e.g., microbial community). Thus, we are limited to only matching spectra to the peptides that are found in the sequence database, hence, mutated or alternatively spliced genes, and post-translationally modified peptides will not be identified. Although having a sequenced genome or metagenome for the exact same biological (protein) sample is extremely valuable for MS-based peptide-spectrum matching, a metagenome is of higher complexity compared to a single genome. The simplicity of peptide-spectrum matching with, for example, a well-characterized bacterial genome (e.g., *E. coli*), via MS database searching is not as straightforward with a metagenome(s). The traditional isolate/genome-based approaches and methodology for DNA sequencing, assembling, and predicting genes, thus, proteins for well-studied model organisms, although routine and widely accepted, may not provide the most expansive and reliable gene sequences for increasing both protein identifications and proteome coverage via MS-based proteomics for community samples.

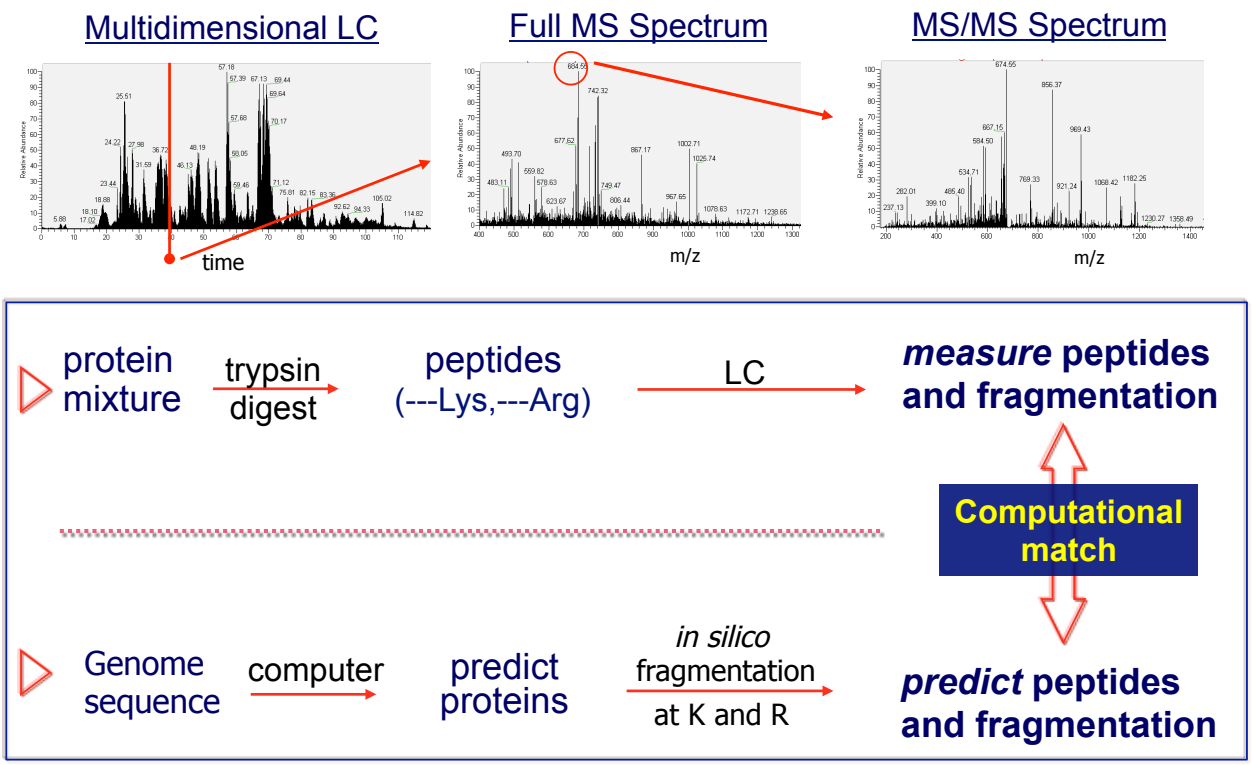


Figure 2.3: Mass spectrometry protein database searching workflow.

For each of the methodological steps and parameters described previously, this dissertation will focus on using a variety of sample processing methods (indirect and direct), hybrid ion trapping mass spectrometers (LTQ-Orbitrap), and protein database searching to optimize and examine the metaproteomes to provide an unprecedented molecular level glimpse into the complex human gut microbiome of both human individuals and gnotobiotic mice.

2.2: Reagents and solvents

All salts, chemical reagents (i.e., guanidine HCl, urea, acetic acid, dithiothreitol (DTT), sodium dodecyl sulfate (SDS)) were acquired from Sigma Chemical Co. (St. Louis, MO) and were used as supplied without further purification. Modified sequencing grade trypsin (Promega, Madison, WI) was used for all protein digestions. HPLC-grade water and acetonitrile were obtained from Burdick & Jackson (Muskegon, MI), and 99% formic acid was purchased from EM Science (Darmstadt, Germany).

2.3: Sample collection

The selection of host subjects and body sites is a critical step for representing the collective metaproteome of the human microbiota. The majority of subjects that have been selected to represent the metaproteome of the HM include human, but also gnotobiotic mice that have been colonized with human-derived microbes to control for and monitor a less complex, but representative human gastrointestinal microbiota[37]. The use of culture-independent techniques for the collection of cells representative of a specific microbial metaproteome niche(s) within the human microbiome can be challenging for several reasons. The invasiveness, quantity, quality, and preservation of sample collection is important for maintaining an intact, native proteome that has not been altered or disrupted such that biological and technical variation is minimal. For example, biological variation would occur if the proteome were not treated (lysed and denatured) immediately upon removal from freezing temperatures where the proteome would begin experience changes as a result of active endogenous proteases. Therefore, minimization of both technical (LC-MS/MS-related) and biological variation will provide for more accurate biological inference relative to the microbiota, sampling subject, and site. Sampling sites have included feces and ceca to represent the human gastrointestinal microbiome[26,37,47,71,83] where microbial cells are typically separated and enriched from the raw human fecal material to remove exfoliated human epithelial cells, interfering food debris and other contaminating compounds. Klassens *et al.* was one of the first to apply metaproteomics to processed human fecal samples (infant). Klassens and colleagues used mechanical homogenization with glass beads and centrifugation to enrich for microbial cells and remove debris. Proteins were then extracted via bead beating. A second, more comprehensive study was performed on adult human fecal samples described from Verberkmoes *et al.* where healthy human feces was processed with a five-cycle differential centrifugation method published by Apajalahti et al.[85]. Microbial cells were lysed and proteins extracted via a small-scale microbial biomass protocol[86]. This protocol was also adapted by Mahowold *et al*. for gnotobiotic mice (chapter 3) and by Erickson *et al.* for healthy and diseased human subjects (chapter 8).

2.3.1: Gnotobiotic mice

Gnotobiotic mice were used as a human model system to control the microbial diversity of the microbiota present in the gut. A collaborator, Dr. Jeffrey Gordon and his research group at the Washington University in St. Louis performed all microbial inoculations and cultivations of the gnotobiotic mice (**Figure 2.4**). Dr. Gordon has sequenced genomes from several members of the two dominant phyla present in the normal distal human gut microbiota: the Firmicutes and the Bacteroidetes. To explore the interactions between the Bacteroidetes and Firmicutes *in vivo*, adult germ-free male mice were gavaged with a mixture of sequenced human-derived microorgansims, ranging from a binary mixture of two bacteria (binary community), seven-members, and finally, a twelve-member community. qPCR analysis of both feces and cecal contents indicated that at the time of sacrifice, the microbial species had colonized the distal gut of recipient mice. Unlike the binary communities containing two evolutionarily distinct microbes (*Bacteroides thetaiotaomicron* and *Eubatcterium rectale*), a second set of adult germ-free mice were colonized with 7 sequenced human gut-derived microbes belonging to the same phyla, Bacteroidetes. After comparing the proteomes of these binary communities to the proteomes acquired from the 7-member communities, the focus shifted to sample processing optimization to better suite “cecum” and feces. The traditional method of cell lysis was not as efficient for lysing cecal material nor did it provide a thorough sampling of the individual microbes’ proteome. Thus, the initial goals for the 7-member community study were to achieve deeper and wider coverage of the ceca proteome prior to in depth biological analyses. Lastly, with the improvements and accumulation of an advanced mass spectrometer (LTQ-Orbitrap Velos), Dr. Gordon scaled up the complexity of the microbial consortium to twelve human gut-derived microorganisms consisting of the same seven phylotypes used in the 7-member consortium in addition to four Firmicutes and one Actinobacteria.

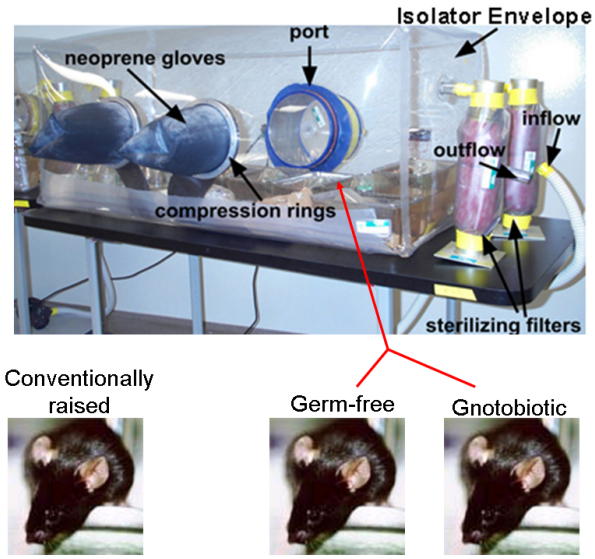


Figure 2.4: Gnotobiotic mouse isolator used to rear pups to adulthood (figure courtesty of Dr. Jeffrey Gordon at Wash. Univ.).

Adult male germ-free mice belonging to the NMRI inbred strain were colonized via gavage with of either 10^8 Colony Forming Units (CFU) of *B. thetaiotaomicron* VPI-5483 or a log-phase culture of *E. rectale*, or both. All mice were fed a standard diet rich in complex plant polysaccharides. Organisms were present in equivalent numbers in the inoculum. Distinct microbial samples were obtained from the distal gut (cecum) of eight gnotobiotic mice provided by Dr. Jeffrey Gordon and Michael Mahowald. Two mice were not colonized with any bacteria (germ-free control); two were colonized only with *B. thetaiotaomicro*n; two were colonized with a mixture of both *B. thetaiotaomicron* and *E. rectale*; and the last two were colonized only with *E. rectale*.

Dr. Gordon's group provided a total of seven cecal samples for the 7-member community proteomics experiments. The C57BL/6 mice, labeled as 2, 3, 5, and 7-10, were gavaged with an equal inoculum of the following species: *Bacteroides caccae, B. ovatus, B. uniformis, B. WH2, B. thetaiotaomicron, B. vulgatus, and Parabacteroides distasonis* and fed a standard BK diet *ab libitum*. The total microbial does was ~ 8.7 x 10^7 corresponding to 1.2-1.3x10^7 CFUs/microbe.

Dr. Gordon's group provided a 12-member community of gut-derived microorganisms consisting of the same seven phylotypes used in the 7-member consortium (*Bacteroides caccae, B. ovatus, B. uniformis, B. WH2, B. thetaiotaomicron, B. vulgatus, and Parabacteroides distasonis*) in addition to four Firmicutes (*Dorea longicatenta, Ruminococcus obeum, Clostridium spiroforme, and C. scindens*) and one Actinobacteria (*Collinsella aerofaciens*). This 12-member consortium was selected for a diet oscillation study in fourteen gnotobiotic mice where ORNL only received the ceca belonging to four mice. There are two treatment groups for which two mice consumed a high fat and simple sugar diet (termed 'western' diet) and the other two other mice consumed a standard high-protein BK diet.

2.3.2: Human gut swedish twin cohort

Human fecal samples from normal, concordant, and discordant human twins with and without Crohn's Disease were provided by a collaborator, Dr. Janet Jansson (Lawrence Berkeley National Lab). The purpose of these studies is to apply proteogenomic techniques to understand the physiology of complex microbial communities in concordant and discordant twins with Crohn's disease. Highly representative, complex gut microbiomes were extracted from bulk human fecal samples (estimated $>10^{11}$ bacteria cells/g of feces) from a total of 6 monozygotic twin pairs (**Table 2.1**) including: 1 set of healthy twins (6a and 6b), 1 set of concordant twins with Crohn's disease inflammation localized in the colon (CCD; 9a and 9b), 2 sets of concordant twins with Crohn's disease inflammation localized in the ileum (ICD; 10a and 10b, 15a and 15b) and 2 sets of ICD discordant twins (16a and 16b, 18a and 18b, ICD and healthy respectively), via differential centrifugation[85] to obtain enriched microbial pellets. The resulting bacterial cell pellets were immediately frozen at –70°C and shipped overnight to the Oak Ridge National Laboratory (ORNL).

Table 2.1: Twin cohort sample descriptions and details for all subjects, healthy, ileal Crohn's disease (ICD), and colonic Crohn's disease (CCD).

Sample ID	Birth year	Phenotype	Sex	NOD2 Status	Gastro-enteritis	Age at diagnosis	Surgery (year)
6a	1951	Healthy	F	nd	Yes	-	-
6b	1951	Healthy	F	nd	No	-	-
9a	1947	CCD, Non-stricturing, Non-penetrating	M	wt	No	41	-
9b	1947	CCD, Non-stricturing, Non-penetrating	M	wt	No	40	-
10a	1962	ICD, Stricturing	F	wt	Yes	23	ileal res + right hemi (1985)
10b	1962	ICD, Stricturing	F	wt	Yes	24	ileocec res (1986)
15a	1953	ICD, Non stricturing, Non-penetrating	M	snp 8 m/w	No	23	ileal res (1980)
15b	1953	ICD, Non-stricturing, Non-penetrating	M	snp 8 m/w	No	23	ileocec res (1976)
16a	1954	ICD, Penetrating	F	wt	No	20	ileal res + right hemi (1974)
16b	1954	Healthy Co-twin	F	wt	No	-	-
18a	1953	ICD, Non-stricturing, Non-penetrating	M	wt	No	20	ileal res + right hemi (1973)
18b	1953	Healthy Co-twin	M	wt	No	-	-

Abbreviations: ileal res, ileal resection; right hemi, right sided hemicolectomy; ileocec res, ileocecal resection; nd, no data; wt, wildtype

2.4: MS-Based Sample Preparation

There are several major steps in MS-based shotgun proteomics that follow in the order of i), separation and/or *in situ* lysis of the bacteria from the environmental matrix and, ii) extraction, denaturation, and digestion of proteins into peptides and, iii) separation and fragmentation of peptides in a mass spectrometer and, iv) peptide-spectrum matching (PSM) (**Figure 2.5**). Several protocols are available and complement the delicate intricacies (e.g., quantity biomass, complexity of sample matrix, and diversity of community membership) that come with processing samples collected from the environment for metaproteomics. Widely accepted methods for the cellular lysis and extraction of proteins from complex microbial communities for metaproteomics analysis include a thermally assisted detergent-based cellular lysis (sodium dodecyl sulfate, SDS) method[74], a small-scale microbial biomass experimental approach[86], sonication[87,88], freeze-thaw cycles[89], French press[60] and published methods described above for HM-related metaproteome samples.

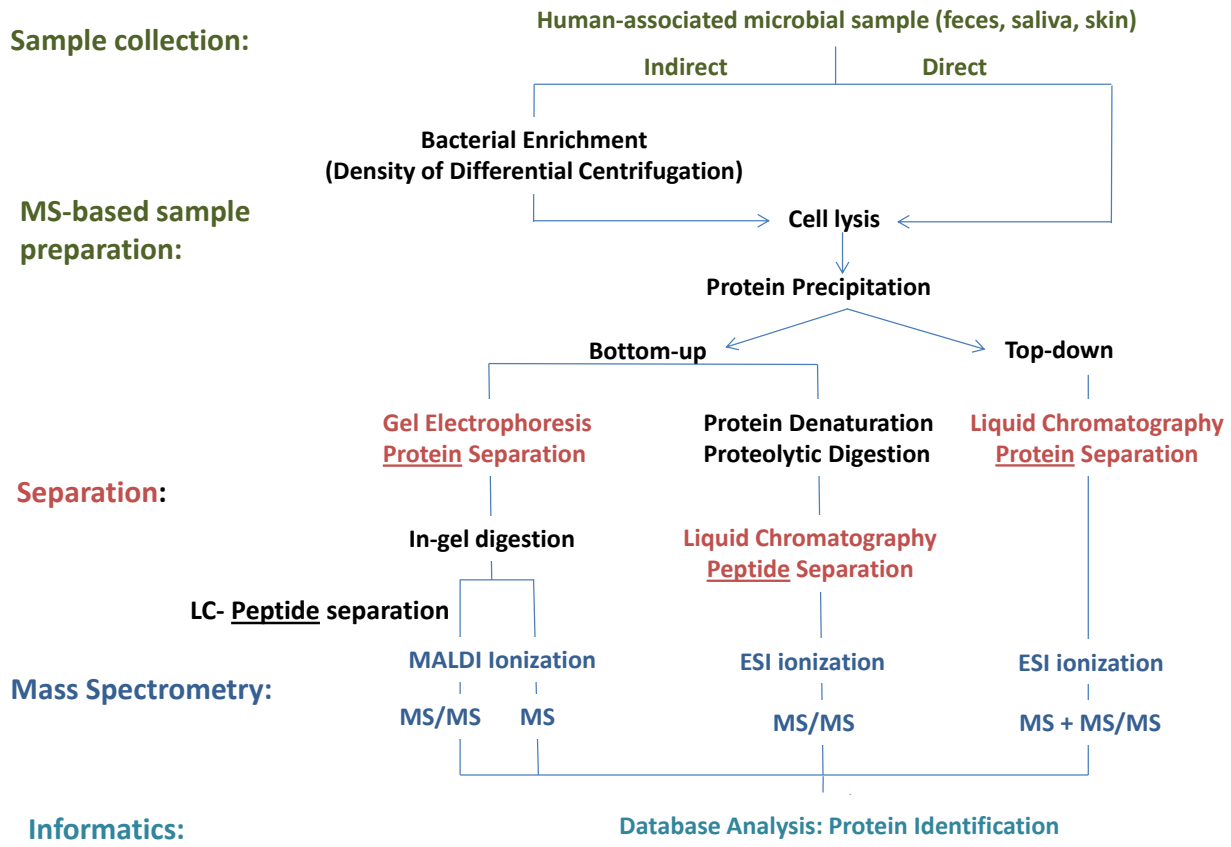


Figure 2.5: General MS-based proteomics experimental, analytical, and informatics pipeline.

Available and tested cellular lysis buffers include detergents (e.g., SDS, CHAPS, and Triton X-100), chaotropes (urea and guanidine), acid-labile surfactants (PPS silent surfactant) and many other commercially available buffers to disrupt bacterial cells with or without physical or mechanical disruption prior to protein extraction. Several precautions must be taken if a detergent is selected due to interference with binding, elution, and ionization of peptides during tandem MS experiments. To eliminate contamination and interference with detergents and mass spectrometers, several metaproteomics studies[26,42,86,90] selected to use an indirect extraction approach and chaotropes (ie, guanidine) to enrich and lyse microbial cells and denature proteins. These methods were also dependent upon the type of peptide separation that would be applied following protein denaturation and digestion, that is, an online 2D-LC separation rather than 2D-PAGE. For the third and fourth major steps, protocols for protein

denaturation and digestion of microbial environmental metaproteomes range where several methods include protein precipitation (ie, trichloroacetic acid (TCA) or ethanol) prior to denaturation and digestion to concentrate proteins away from contaminants (ie, small molecules and interfering environmental-related compounds) that were not eliminated further upstream with indirect extraction methods by centrifugation. Chourey *et al.* has shown that protein precipitation (TCA) was beneficial in eliminating humic compounds and other interfering substances commonly found in soil metaproteomes[74] and could be also improve protein purification for other complex environmental matrices such as human feces. Protein digestion is one of the most critical steps because it involves the reduction of intact proteins to peptides that are suitable for MS analysis (10-20 amino acids), which relies heavily upon sufficient lysis and solubilization of all cells in order for proteases (e.g., trypsin) to access the entire surface area of proteins for effective digestion. The final selection of one method should be based on the available starting biomass quantity and environmental sample type and quality (matrix type (ie, degree of exopolysaccharides and interfering humic and phenolic compounds) and feasibility for purification).

Although many of the described protocols are suitable for a variety of samples, there are potential biases that range with: i) the indirect (enrichment) approach and its' efficiency of bacterial extraction via density and differential centrifugation, ii) presence of other organisms (e.g., fungi, protozoa, and eukaryotes), iii) lysis of bacteria with certain properties (i.e., gram negative versus positive), iv) extraction of proteins with diverse properties (i.e., cytosolic versus membrane proteins and fractions) directly from natural environments, and v) the efficiency of digestion has not all been resolved for complex samples collected from the HM. There are many challenges that stem from processing environmental communities for efficient recovery of proteins including those mentioned above in addition to contamination with other interfering compounds imbedded in the surrounding matrix if they are not eliminated prior to MS analysis and bias in the quantitative and qualitative recovery of proteins.

2.4.1: Gnotobiotic mice and human twin cohort

The gnotobiotic mouse gut microbial communities were treated differently for each study. Similar to the human twin cohort studies, the binary community cecal contents were processed via a single tube cell lysis method[86] and proteins digested into peptides with trypsin. All eight samples were coded and mass spectrometry measurements conducted in a blinded fashion. Each of the individual cecal contents collected for the 7-member community were processed differently for method comparisons and development and is described in more detail under chapter 4. Using the results from chapter 4, the 12-member microbial community cecal contents were solubilized in SDS lysis buffer and lysed mechanically by sonication and heat. Following a TCA precipitation, the precipitates were resolubilized and reduced in 8M urea and DTT and digested with trypsin.

The bacterial cell pellets (~100mg) that were extracted from bulk human fecal samples were lysed; proteins were denatured and reduced, and digested into peptides with trypsin using the protocol developed by Thompson et al.[86]. These samples were used throughout chapters 6-8.

2.5: Liquid Chromatography

Complex biological samples often contain thousands to hundreds of thousands of proteins and can be a challenge in terms of total comprehensive intact protein identification. Therefore, proteins and proteomes are generally digested into smaller products (peptides) that are technically easier to separate, measure, and identify compared to intact proteins. However, following proteolytic digestion, a complex environmental sample can contain hundreds of thousands to millions of peptides. Therefore, complex peptide digests are fractionated by either a one or multiple chromatographic dimensions or electrophoretic steps to reduce the complexity of peptides analyzed at a single time point by mass spectrometry (**Figure 2.1**).

Since the use of and coupling of two-dimensional gel electrophoresis (2DE), SDS-PAGE, and mass spectrometry with environmental communities[60,91], the

scientific community has moved towards using other gel-less alternatives, such as, higher throughput protein and peptide separation via LC coupled to mass spectrometers. This transition has enabled the high resolution identification of a few thousands of proteins from cultured bacteria (e.g., *R. palustris*[92]) to an environmental microbial community[42]. For LC-MS, either a single-dimensional (1D) or orthogonal two-dimensional chromatographic system could be used to separate proteins or peptides online or offline[93], however, complex mixtures can overwhelm the capacity of a single dimension. Therefore, a "multi-dimensional" separation is better suited for complex mixtures and it is the current standard for large-scale proteomics consisting of two- or three-step peptide fractionation. Yates and colleagues founded the technique that is referred to as "multidimensional protein identification technology" (MudPIT) where orthogonal 2D chromatography is used to separate complex peptide mixtures prior to MS analysis[94,95,96]. The most popular MudPIT setup consists of i) strong cation exchange (SCX) chromatography followed by ii) reversed-phase (RP) chromatography to achieve peptide separation prior to MS/MS[96]. SCX chromatography serves as the primary dimension that separates peptides based on charge where it has an increased loading capacity compared to RP which separates peptides by their hydrophobicity while simultaneously eliminating any salts in the sample. Compared to 2DE and 1D-LC, the MudPIT technique has the advantage of being higher throughput, higher resolution and reproducible with both the chromatography and identified proteins, and it is unbiased to a range of proteins with variable and extreme pIs, MW, location (membrane or cytosol) and abundance.

In the human microbiome, both separation technologies have been applied in a variety of ways. Klaassens et al. demonstrated for the first time that 2D-PAGE and protein identification via matrix-assisted laser desorption (MALDI)-time of flight (TOF) mass spectrometry was applicable to study the metaproteome of the gut microbiome in human infants using infant fecal material[47]. On the contrary, LC-MS-based proteomics was first demonstrated in the adult human gastrointestinal microbiome[26,71] with the application of 2D-(SCX-RP)-LC-MS/MS to human fecal material. Rooijers et al., on the other hand, used gel electrophoresis for protein

separation followed by in gel protein digestions[83]. The tryptic peptides were then separated via 1D-(RP)-LC coupled to electrospray MS/MS on a Thermo LTQ-Orbitrap. For the oral microbiome, Hongwei et al. separated tryptic peptides via a 3-step fractionation method[84]. First, peptides were separated by isoelectric focusing (IEF) using a free-flow electrophoresis system (FFE) and fractionated into a 96-microtiter plate. Each pI fraction was subjected to preliminary MS/MS to identify fractions with the highest complexity of peptides to be characterized and separated further. For the final selected IEF fractions, the peptide fractions were purified (to remove high MW polymers) and fractionated by SCX (second fractionation step) using a step-gradient. The SCX peptide fractions were then desalted, concentrated, and loaded onto and separated on a RP column (third fractionation step) where peptides were directly eluted and analyzed by ESI-MS/MS on a thermo LTQ linear ion trap. Similarly, Rudney et al. adapted the same protocol[84] to focus on the metaproteomics analysis of the bacterial component, taxonomy and metabolic activity, of the human oral microbiome using human whole saliva. Grant et al. also investigated the oral microbiome in healthy humans using gingival crevicular fluid (GCF) and a non-invasive gingivitis model that is used to study the inflammatory response as a result of increasing bacteria over 21 days[49]. GCF samples were pooled and treated with dithiothreitol, heat, and proteolytically digested with trypsin prior to quantitative labeling with iTRAQ (discussed in further detail under 'quantitative proteomics in the HM'). iTRAQ labeled samples were separated and fractions collected offline using SCX-HPLC. The fractions were vacuum centrifuged, desalted, and peptides acidified with formic acid prior to LC-MS/MS. Lastly, tryptic labeled peptides were separated online using 1D-(RP)-LC and eluted directly into a thermo LTQ-FT mass spectrometer. In conclusion, metaproteomics can be applied to complex samples collected from the human microbiota using a variety of separation technologies as described (2DE, 1D- or 2D-LC, and/or multiple fractionation steps online or offline). In my opinion, an approach that uses a "multidimensional" separation approach (e.g., MudPIT) will provide the most comprehensive and representative coverage of peptides and proteins from the human microbiome.

2.5.1: Gnotobiotic mice and human twin cohort

For all gnotobiotic mouse (chapters 3-6) and human twin (chapters 6-8) studies, the microbial proteins were extracted and processed for 2D-LC-MS/MS using an Ultimate HPLC system (Dionex, Sunnyvale, CA) coupled to a LTQ, LTQ-Orbitrap, or LTQ-Orbitrap Velos (Thermo Fisher Scientific, San Jose, CA). At a flow rate of ~100 µL/min (set on the Ultimate pump), the peptide mixtures of all twelve samples were separated across a split-phase column (packed in-house with SCX and C_{18} reverse-phase chromatographic resins) that was connected to a 15-cm C_{18} analytical column by a 12 step, multidimensional high-pressure liquid chromatographic elution consisting of eleven salt pulses (0-500 mM ammonium acetate) followed by a 2 hour reverse-phase gradient from 100% solvent A (A: 95% H2O, 5% acetonitrile, 0.1% formic acid) to 50% solvent B (B: 30% H2O, 70% acetonitrile, 0.1% formic acid). The last salt pulse was followed with a gradient from 100% solvent A to 100% solvent B. During a single chromatographic separation (~22-24 hr run), mass spectral data acquisition was performed in data-dependent mode under the control of Xcalibur software (version 2.0.7; Thermo Fisher Scientific).

2.6: Mass spectrometric measurements

As described previously under 2.1, the selection of a mass spectrometer is important and dependent upon several factors including: (i) the proposed biological questions (ie, comprehensive characterization or targeted analysis of a subset of proteins) and (ii) the type and complexity of the biological sample (single or mixture of proteins or environmental sample; dynamic range of proteins). A linear ion trap mass spectrometer would be best suited for a comprehensive characterization of a proteome whereas a QqQ platform is preferred for targeted proteomics investigation. Due to the complexity of the samples (i.e., microbial communities) used in both the human microbiome and gnotobiotic mice, and the desire to comprehensively characterize the proteomes, ion trap mass spectrometers and hybrid mass spectrometers are ideally best suited for these bottom up studies due to their rapid scan time, resolution, and mass accuracy.

2.6.1: Gnotobiotic mice and human twin cohort

All cecal and fecal samples were analyzed in technical duplicates using a two-dimensional (2D) nano-LC MS/MS system with a split-phase column (RP-SCX)[97] on a LTQ-XL, LTQ-Orbitrap (**Figure 2.2**), or LTQ-Orbitrap Velos (Thermo Fisher Scientific) with 22 hr runs per sample (LC as previously described). The mass spectrometer settings were as follows: one full MS scan was acquire in the Orbitrap (*m/z* 400-1,700) at 30k resolution followed by five or ten data-dependent MS/MS in the LTQ at 35% normalized collision energy. Two microscans were averaged for both full and MS/MS scans and centroid data were collected for all scans, with dynamic exclusion enabled at 1.

2.7: Proteome informatics

Comparative and quantitative proteomics is the evaluation of how similar and/or different environmental conditions affect protein expression and abundance. Following the acquisition of qualitative and in many studies quantitative proteome MS data, bioinformatics tools, such as DTASelect[98] and software packages, such as Scaffold[99,100,101] sort and filter through these massive MS datasets to provide the best quality peptide-spectrum matches (PSM) and their corresponding protein identifications.

As described previously, protein database searching, an informatics workflow that deduces the amino acids of a peptide sequence and assigns it to a corresponding tandem mass spectrum (MS/MS) (**Figure 2.3**), has been widely adopted by MS-based proteomics for the high throughput identification of proteins. Database search engines, such as SEQUEST[66], Mascot[67], and Xtandem![68], assign peptide sequences to MS/MS spectra by correlating the experimentally identified peptide to a theoretical peptide sequence derived *in silico* from a known FASTA formatted protein database. The peptide sequences with the highest correlation scores are reported in the final output. As a result, the protein database informatics platform is high throughput and applicable to range of protein complexity, from a single or mixture of proteins to a complex environmental community metaproteome. The alternative approach to assign

MS/MS is *de novo* peptide sequencing where no prior knowledge of proteins or a protein database is required.

2.7.1: PSM and database searching

For the first informatics platform, protein database searches, mass spectrometers first collect precursor ions (intact peptide ions; MS scan) that are selected for fragmentation by collision with inert gas (e.g., collisional induced dissociation; CID) into fragment ions (amino acids; MS/MS scan). In total, three to ten of the most abundant precursor ions (MS) are selected for MS/MS generating thousands of MS/MS during one experiment (2-24hrs) that represent fragmented peptides of a samples' protein(s). It is the quantity, quality (signal versus noise), and complexity of these MS/MS that necessitates the need for informatics workflows that can sort, filter, and confidently assign PSMs with high accuracy at a computationally reasonable speed. Currently, there are several open-source and commercial tandem mass spectrometry database search engines that are widely available. The freely open-source search engines include X!Tandem[68], OMSSA[102], Myrimatch[103]. The commercially available search engines Mascot[67], from Matrix Science, and SEQUEST[66], from Thermo Fisher Scientific, have become the most widely used and are referred to as the golden standards. However, each database search algorithm has its advantages and disadvantages in terms of spectral filtering, scoring, configuration, compatible formats, and speed/performance. Several of these factors are dependent on the quality and relevance of the protein database to the measured sample and its' proteome. For example, if a high quality experimentally identified PSM cannot be assigned to a protein because it is missing in the protein database, these high quality PSMs will go unidentified. As demonstrated by Cantarel *et al*. with the human gut microbiome, up to several thousands of high-quality MS/MS may not be identified as a result of the database even with having a matched metagenome(s)[76]. Therefore, the quality and selection of a protein database and/or metagenomes(s) is a critical component of the protein database search informatics workflow. Additionally, the larger and all-inclusive protein sequence database(s) will often take longer to search hindering the performance of several search engines and increase the number of false positive-identifications.

Currently, there are many ways of estimating error associated with peptide identifications. Until the field of proteomics comes to a conclusion on the proper way of reporting proteomic data, different versions will exist. For these large-scale studies, false discovery rates were used in order to differentiate between true and false peptide identifications. The overall false discovery rate (FDR) was estimated using the formula: $FDR= 2[n_{rev}/(n_{rev} + n_{real})]*100$ where n_{rev} is the number of peptides identified from the reverse database and n_{real} is the number of peptides identified from the real database[96].

2.7.1.1: Gnotobiotic mice and human twin cohort

All MS/MS spectra were searched with the SEQUEST algorithm[66] [(enzyme type, trypsin; Parent Mass Tolerance, 3.0; Fragment Ion Tolerance, 0.5; up to 4 missed cleavages allowed (internal lysine and arginine residues), and fully tryptic peptides only (both ends of the peptide must have arisen from a trypsin specific cut, except N and C-termini of proteins)] and filtered with DTASelect/Contrast[98] at the peptide level [Xcorrs of at least 1.8 (+1), 2.5 (+2) 3.5 (+3) and deltCN of either 0.08 or 0.0]. Only proteins identified with two fully tryptic peptides from the MS runs were considered for further biological study. Monoisotopic theoretical masses for all peptides identified by SEQUEST were generated and compared to observed masses. Observed high resolution masses were extracted from .raw files from the full scan preceding best identified spectra; parts per million (ppm) calculations were made comparing each identified peptides' observed and theoretical mass. When quality MS/MS spectra didn't have an observed mass (low intensity) due to an unassignable charge state for the precursor ion, a mass of 0 was reported and ppm was calculated as infinity.

Detailed database descriptions and search strategies can be found for each of the following chapters under their "experimental methods." The FDRs were also calculated for the majority of experiments and are described in more detail for each chapter.

2.7.2: *De novo* peptide sequencing in the microbiome

As mentioned, the second alternative workflow for protein identification from tandem mass spectra is *de novo* peptide sequencing[104,105,106]. Protein identification via database searching will not be able to identify peptides that are not in the database. Additionally, with the use of high mass resolution and mass accuracy MS data, *de novo* peptide sequencing can be used for the identification of amino acid polymorphisms and post-translation modifications (PTMs). *De novo* sequencing calculates the mass difference between two peaks in a single mass spectrum and if the difference corresponds to an amino acids' mass, the algorithm is able to assume that the two peaks are adjacent fragment ions in the peptide sequence. However, to assess this mass difference, differentiate signal versus noise, and identify a specific amino acid with high accuracy, investigators need high mass accuracy MS/MS data to eliminate the degree of interference. Several *de novo* sequencing algorithms are currently available and include PepNovo[105,107], DirecTag[108], PEAKS[109], MSNovo[110] and Vonode[111]. As described by Cantarel *et al*., the first *de novo* peptide sequencing application (high confidence sequence tags found by both PEAKS and PepNovo) to the human gut microbiome, conservative *de novo* sequencing can be highly beneficial for its revelation of novel peptides that were not identified using a database search engine (SEQUEST) and increase in protein discovery[76]. Additional details and results are described in chapter 7.

2.8 Summary

The experimental and analytical methods described above provides a robust, high-throughput, and highly reproducible platform for the application of MS-based proteomics to characterize complex human gut microbiomes collected from gnotobiotic mice ceca and human feces. Each step of the general MS-based proteomics workflow: i), sample collection, ii) MS-based sample preparation, iii) liquid chromatography, iv) MS/MS measurement and, v) peptide-spectrum matching is important and intended to provide a solid foundation for optimum identification and coverage of human gut metaproteomes. This workflow results in the assignment of hundreds of thousands of spectra and the

identification of thousands of proteins and peptides with a deeper characterization and understanding of the gut microbiota as highlighted in the following analyses.

Chapter Three

Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla

Portions of the included text are adapted from:

> *Michael A. Mahowald, Federico E. Rey, Henning Seedorf, Peter J. Turnbaugh, Robert S. Fulton, Aye Wollam, Neha Shah, Chunyan Wang, Vincent Magrini, Richard K. Wilson, Brandi L. Cantarel, Pedro M. Coutinho, Bernard Henrissat, Lara W. Crock, Alison Russell, Nathan C. Verberkmoes, Robert L. Hettich, and Jeffrey I. Gordon. "Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla." PNAS, 2009, volume 106, issue 14, pages 5859-5864.*

Alison R. Erickson's contributions include experimental preparation of ceca samples for proteomics and all experimental LC-MS/MS measurements and analysis.

3.1: Introduction

The adult human gut houses a bacterial community containing trillions of members comprising thousands of species-level phylogenetic types (phylotypes). Culture-independent surveys of this community have revealed remarkable interpersonal variations in strain- and species-level phylotypes, and two commonly abundant bacterial phyla, the Firmicutes and the Bacteroidetes[112]. This phylum-level composition is not a unique feature of humans: a global survey of the guts of 59 other mammalian species showed a similar phylum level pattern[113].

Comparative analysis of five sequenced human gut Bacteroidetes revealed that each genome contains a large repertoire of genes involved in acquisition and metabolism of polysaccharides: this repertoire includes (i) up to hundreds of glycoside hydrolases (GHs) and polysaccharide lyases (PLs); (ii) myriad paralogs of SusC and SusD, outer membrane proteins involved in recognition and import of specific carbohydrate structures[114]; and (iii) a large array of environmental sensors and regulators[115]. These genes are assembled in similarly organized, selectively regulated polysaccharide utilization loci (PULs) that encode functions necessary to detect, bind, degrade and import carbohydrate species encountered in the gut habitat –

either from the diet or from host glycans associated with mucus and the surfaces of epithelial cells[116,117,118]. Studies of gnotobiotic colonized with human gut-derived Bacteroides thetaiotaomicron alone have demonstrated that this organism can vary its pattern of expression of PULs as a function of diet: e.g., during the transition from mother's milk to a polysaccharide-rich chow consumed when mice are weaned[116], or when adult mice are switched from a diet rich in plant polysaccharides to a diet devoid of these glycans and replete with simple sugars (under the latter conditions, the organism forages on host glycans)[117,118].

Our previous functional genomic studies of the responses of *B. thetaiotaomicron* to co-colonization of the guts of gnotobiotic mice with *Bifidobacterium longum*, an Actinobacterium found in the guts of adults and infants, or with *Lactobacillus casei*, a Firmicute present in a number of fermented dairy products, have shown that *B. thetaiotaomicron* responds to the presence of these other microbes by modifying expression of its PULs in ways that expand the breadth of its carbohydrate foraging activities[119].

These observations underscore the notion that gut microbes may live at the intersection of two forms of selective pressure: bottom-up selection, where fierce competition between members of a community that approaches a population density of 10^{11}-10^{12}organisms/ml of colonic contents drives phylotypes to assume distinct functional roles; and top-down selection, where the host selects for functional redundancy to insure against the failure of bioreactor functions that could prove highly deleterious[120,121].

The content, genomic arrangement and functional properties of PULs in sequenced gut Bacteroidetes illustrate the specialization and functional redundancy within members of this phylum. They also emphasize how the combined metabolic activities of members of the microbiota undoubtedly result in interactions that are both very dynamic and overwhelmingly complex (at least to the human observer), involving multiple potential pathways for the processing of substrates (including the order of substrate processing), varying patterns of physical partitioning of microbes relative to

substrates within the ecosystem, plus various schemes for utilization of products of bacterial metabolism. Such a system likely provides multiple options for processing of a given metabolite, and for the types of bacteria that can be involved in these activities.

All of this means that the task of defining the interactions of members of the human gut microbiota is daunting, as is the task of identifying general principles that govern the operation of this system. In the present study, we have taken a reductionist approach to begin to define interactions between members of the Firmicutes and the Bacteroidetes that are commonly represented in the human gut microbiota. In the human colon, Clostridium cluster XIVa is one of two abundantly represented clusters of Firmicutes. Therefore, we have generated the initial two complete genome sequences for members of the genus Eubacterium in Clostridium cluster XIVa, (the human gut-derived *E. rectale* strain ATCC 33656 and *E. eligens* strain ATCC 27750) and compared them with the draft sequences of 25 other sequenced human gut bacteria belonging to the Firmicutes and the Bacteroidetes. The interactions between *E. rectale* and *B. thetaiotaomicron* were then characterized by performing whole genome transcriptional profiling of each species after colonization of gnotobiotic mice with each organism alone, or in combination under three dietary conditions. Transcriptional data collected by Wash. Univ. were verified by mass spectrometry of cecal proteins collected by ORNL, plus biochemical assays of carbohydrate metabolism. Lastly, we examined colonization and interactions between these microbes from a host perspective; to do so, we performed whole genome transcriptional analysis of colonic RNA prepared from mice that were germ-free or colonized with one or both species. Our results illustrate how members of the dominant gut bacterial phyla are able to adapt their substrate utilization in response to one another and to host dietary changes, and how host physiology can be affected by changes in microbiota composition.

3.2: Experimental Methods

3.2.1: Genome comparisons

All nucleotide sequences from all contigs of completed genome assemblies containing both capillary sequencing and pyrosequencer data, produced as part of the HGMI, were

downloaded from the Washington University Genome Sequencing Center's website (http://genome.wustl.edu/pub/organism/Microbes/Human_Gut_Microbiome/) on September 27, 2007. The finished genome sequences of *B. thetaiotaomicron* VPI-5482, *Bacteroides vulgatus* ATCC 8482, and *B. fragilis* NCTC9343 were obtained from GenBank.

For comparison purposes, protein-coding genes were identified in all genomes using YACOP[122]. Each proteome was assigned InterPro numbers and GO terms using InterProScan release 16.1. Statistical comparisons between genomes were carried out, as described previously[115] using perl scripts that are available upon request from the authors.

3.2.2: GeneChip analysis

Previously described methods were used to isolate RNA from a 100-300 mg aliquot of frozen cecal contents, synthesize cDNA, and to biotinylate and hybridize the cDNAs to a custom bacterial GeneChip[123]. The only modification was that in RNA isolation protocol 0.1mm zirconia/silica beads (Biospec Products, Bartlesville, OK) were used for lysis of bacterial cells in a bead beater (Biospec; 4 min run at highest speed). Genes in a given bacterial species that were differentially expressed in mono- versus biassociation experiments were identified using CyberT (default parameters) following probe masking and scaling with the MAS5 algorithm (Affymetrix; for details about the methods used to create the mask, see the Methods section of Supplementary Information).

RNA was purified from proximal colon using Mini RNeasy kit (Qiagen) with on-column DNase digestion. Biotinylated cRNA targets were prepared from each sample (n=4/treatment group). cRNA was hybridized to Affymetrix Mouse Genome Mo430 2 GeneChips, and the resulting data sets analyzed using Probe Logarithmic Error Intensity Estimate method (PLIER+16). Fold-changes and p-values were calculated using Cyber-t. Significance was defined by maintaining a FDR <1% using Benjamini-Hochberg correction[124].

3.2.3: Proteomic methods

Cecal contents were processed via a single tube cell lysis and protein digestion method as follows. Briefly, the cell pellet was re-suspended in 6M Guanidine/10 mM DTT, heated at 60°C for 1h, followed by an overnight incubation at 37°C to lyse cells and denature proteins. The guanidine concentration was diluted to 1 M with 50mM Tris/10mM CaCl2 (pH 7.8), and sequencing grade trypsin (Promega, Madison, WI) was added (1:100; wt/wt). Digestions were run overnight at 37°C. Fresh trypsin was then added followed by additional 4h incubation at 37°C. The complex peptide solution was subsequently de-salted (Sep-Pak C_{18} solid phase extraction; Waters, Milford, MA), concentrated, filtered, aliquoted and frozen at -80°C. All eight samples were coded and mass spectrometry measurements conducted in a blinded fashion.

Cecal samples were analyzed in technical triplicates using a two-dimensional (2D) nano-LC MS/MS system with a split-phase column (SCX-RP)[97] on a linear ion trap (Thermo Fisher Scientific) with each sample consuming a 22 hr run as detailed elsewhere[92,125]. The linear ion trap (LTQ) settings were as follows: dynamic exclusion set at one; and five data-dependent MS/MS. Two microscans were averaged for both full and MS/MS scans and centroid data were collected for all scans. All MS/MS spectra were searched with the SEQUEST algorithm[66] against a database containing the entire mouse genome, plus the *B. thetaiotaomicron*, *E. rectale*, rice, and yeast genomes (common contaminants such as keratin and trypsin were also included). To find potential food proteins, yeast and rice databases were included. The breakdown of each database component can be found in **Table 3.1**. The SEQUEST settings were as follows: enzyme type, trypsin; Parent Mass Tolerance, 3.0; Fragment Ion Tolerance, 0.5; up to 4 missed cleavages allowed (internal lysine and arginine residues), and fully tryptic peptides only (i.e., both ends of the peptide must have arisen from a trypsin-specific cut, except the N- and C-termini of proteins). All datasets were filtered at the individual run level with DTASelect (22) [Xcorrs of at least 1.8 (+1 ions), 2.5 (+2 ions) 3.5 (+3 ions)]. Only proteins identified with two fully tryptic peptides were considered.

Table 3.1: Protein sequence database components for binary microbial community SEQUEST database searches.

Database	Proteins	Size (MB)
B. thetaiotaomicron	4,958	2.3
E. rectale	3,188	1.36
M. musculus	34,966	19.2
Rice	66,710	36
Yeast	6,345	3.33
contaminants	36	0.02

For this study, false-positive rates (FPR) were used to estimate the error associated with peptide identifications. The overall FPR was estimated using the formula: FPR = $2[n_{rev}/(n_{rev} + n_{real})]*100$ where n_{rev} is the number of peptides identified from the reverse database and n_{real} is the number of peptides identified from the real database[96]. Reverse and shuffled databases were created in order to calculate FPRs[96,126]. A reverse database was created by precisely reversing each protein entry (i.e., N-terminus became C-terminus in each case) and then appended these reversed sequences onto the original database. Two runs - samples 705, Run 1 and 710, Run 2 - were randomly selected for estimating a FPR. The observed FPR rates were 0.55% and 0.31% respectively for these two runs. An additional database was created by randomly shuffling the amino acids of each protein rather than simply reversing the N-terminus and C-terminus. A FPR was estimated using a similar formula as that described above except that the number of identified reverse peptides was replaced with the number of shuffled peptides. A FPR was estimated for both samples, 705, Run 1 (0.45%) and 710, Run2 (0.31%) and was similar to the rate determined by the reverse database method. Datasets for calculating FPR rates are available on the website mentioned above.

In addition to differentiating between true and false peptide identifications with FPRs, label-free quantitation methods were used to estimate relative protein abundance. Several protein quantitation methods are currently available and routinely performed for shotgun proteomics analyses. To estimate relative protein abundance in complex protein mixtures and communities, spectral counts and normalized spectral

abundance factors (NSAF)[127] are commonly used. Spectral counting is based on the theory that the more abundant peptides are typically sampled more frequently, resulting in higher spectral counts. Liu *et al.* has shown that spectral copy number provides a more accurate correlation to protein abundance than peptide count and % coverage[128]. NSAF, on the other hand, is based on spectral counts, but takes into account protein size and the total number of spectra from a run, thus normalizing the relative protein abundance between samples[127].

3.3: Results and Discussion

3.3.1: Comparative genomic studies of human gut-associated Firmicutes and Bacteroidetes

Wash. Univ. produced finished genome sequences for *Eubacterium rectale*, which contains a single 3,449,685 bp chromosome encoding 3,627 predicted proteins, and *Eubacterium eligens* which contains a 2,144,190 bp chromosome specifying 2,071 predicted proteins, plus two plasmids. We also analyzed 25 recently sequenced gut genomes, including (i) 9 sequenced human gut-derived Bacteroidetes [includes the finished genomes of *B. thetaiotaomicron*, *B. fragilis*, *B. vulgatus*, *and Parabacteroides distasonis*, plus deep draft assemblies of the *B. caccae*, *B. ovatus*, *B. uniformis*, *B. stercoris* and *P. merdae* genomes generated as part of the human gut microbiome initiative (HGMI; http://genome.wustl.edu/hgm/HGM_frontpage.cgi], and (ii) 16 other human gut Firmicutes where deep draft assemblies were available through the HGMI. We classified the predicted proteins in these two genomes using Gene Ontology (GO) terms generated via Interproscan, as well as according to the scheme incorporated into the Carbohydrate Active Enzymes (CAZy) database [www.cazy.org;[129], and then applied a binomial test to identify functional categories of genes that are either over- or under-represented between the Firmicutes and Bacteroidetes phyla. This analysis emphasized among other things that the Firmicutes, including *E. rectale* and *E. eligens*, have significantly fewer polysaccharide-degrading enzymes and more ABC transporters and PTS systems than the Bacteroidetes[130]. We subsequently chose *E. rectale* and *B. thetaiotaomicron* as representatives of these two phyla for further characterization of

their niches *in vivo*, because of their prominence in culture-independent surveys of the distal human gut microbiota[62,131], the pattern of representation of carbohydrate active enzymes in their glycobiomes and *E. rectale*'s ability to generate butyrate as a major end product of fermentation[132,133]. These choices set the stage for an 'arranged marriage' between a Firmicute and a Bacteroidetes, hosted by formerly germ-free mice.

3.3.2: Functional genomic analyses of the minimal human gut microbiome

3.3.2.1: *Creating a "minimal human gut microbiota" in gnotobiotic mice* - Young adult male germ-free mice belonging to the NMRI inbred strain were colonized with *B. thetaiotaomicron* or *E. rectale* alone (monoassociations) or co-colonized with both species (biassociation). 10-14 d after inoculation by gavage, both species colonized the ceca of recipient mice, fed a standard chow diet rich in complex plant polysaccharides, to high levels (n=4-5 mice/treatment group in each of 3 independent experiments). Moreover, cecal levels of colonization for both organisms were not significantly different between mono- and biassociated animals.

3.3.2.2: *B. thetaiotaomicron's response to E. rectale* - A custom, multispecies, human gut microbiome Affymetrix GeneChip was designed and used to compare the transcriptional profile of each bacterial species when it was the sole inhabitant of the cecum, and when it co-existed together with the other species. A significant number of *B. thetaiotaomicron* genes located in PULs exhibited differences in their expression upon *E. rectale* colonization [55 of 106; $p<10^{-15}$ (cumulative hypergeometric test). Of these 55 genes, 51 (93%) were upregulated.

As noted in the Introduction, two previous studies from our lab examined changes in *B. thetaiotaomicron*'s transcriptome in the ceca of monoassociated gnotobiotic mice when they were switched from a diet rich in plant polysaccharides to a glucose-sucrose chow[117], or in suckling mice consuming mother's milk as they transitioned to a standard chow diet[116]. In both situations, in the absence of dietary plant polysaccharides, *B. thetaiotaomicron* adaptively forages on host glycans. The genes upregulated in *B. thetaiotaomicron* upon co-colonization with *E. rectale* have a

significant overlap with those noted in these two previous datasets ($p<10^{-14}$, cumulative hypergeometric test). In addition, they involve several of the genes upregulated during growth on minimal medium containing porcine mucosal glycans as the sole carbon source[118]. For example, in co-colonized mice and *in vitro*, *B. thetaiotaomicron* upregulates several genes (BT3787-BT3792; BT3774-BT3777) used in degrading α-mannosidic linkages, a component of host N-glycans as well as the diet. (Note that *E. rectale* is unable to grow in defined medium containing α -mannan or mannose as the sole carbon sources). *B. thetaiotaomicron* also upregulates expression of its starch utilization system (Sus) PUL in the presence of *E. rectale* (BT3698-3704). This well-characterized PUL is essential for degradation of starch molecules containing ≥6 glucose units[134].

Thus, it appears that *B. thetaiotaomicron* adapts to the presence of *E. rectale* by upregulating expression of a variety of PULs so that it can broaden its niche and degrade an increased variety of glycan substrates, including those derived from the host that *E. rectale* is unable to access. There are a number of reasons why the capacity to access host glycans likely represents an important trait underpinning microbiota function and stability: (i) glycans in the mucus gel are abundant and are a consistently represented source of nutrients; (ii) mucus could serve as a microhabitat for Bacteroidetes spp. to embed in (and adhere to via SusD paralogs), thereby avoiding washout from the ecosystem; and (iii) the products of polysaccharide digestion/fermentation generated by Bacteroidetes spp. could be shared with other members of the microbiota that are also embedded in mucus[118].

3.3.2.3: *E. rectale's response to B. thetaiotaomicron* - *E. rectale*'s response to *B. thetaiotaomicron* in the mouse cecum stands in marked contrast to *B. thetaiotaomicron*'s response to *E. rectale*. Carbohydrate metabolism genes, particularly GHs, are significantly overrepresented among the *E. rectale* genes that are downregulated in the presence of *B. thetaiotaomicron* compared to monoassociation; i.e., 12 of *E. rectale*'s predicted 51 GHs have significantly reduced expression while only two are upregulated. The two upregulated GH genes (EUBREC_1072, a 6-P-b-glucosidase and EUBREC_3687, a cellobiose phosphorylase) are predicted to break

down cellobiose. Three simple sugar transport systems with predicted specificity for cellobiose, galactoside, and arabinose/lactose (EUBREC_3689, EUBREC_0479, and EUBREC_1075-6, respectively) are among the most strongly upregulated genes. Phosphoenolpyruvate carboxykinase (EUBREC_2002) is also induced with co-colonization (GeneChip data verified by qRT-PCR assays in 2 independent experiments involving 3-4 mice/treatment group). This enzyme catalyzes an energy conserving reaction that produces oxaloacetate from phosphoenolpyruvate. In a subsequent transaminase reaction, oxaloacetate can be converted to aspartate, linking this branching of the glycolytic pathway with amino acid biosynthesis.

Additional data support the notion that *E. rectale* is better able to access nutrients in the presence of *B. thetaiotaomicron*. For example, a number of peptide and amino acid transporters in *E. rectale* are upregulated, as are the central carbon and nitrogen regulatory genes CodY (EUBREC_1812), glutamate synthase (EUBREC_1829) and glutamine synthetase (EUBREC_2543) (note that these genes are also upregulated during growth in tryptone glucose medium).

3.3.2.4: Changes in *E. rectale's* fermentative pathways - *E. rectale* possesses genes (EUBEC733-737; EUBEC1017) for the production of butyrate that show high similarity to genes from other Clostridia. This pathway involves condensation of two molecules of acetylCoA to form butyrate and is accompanied by oxidation of NADH to NAD+. Transcriptional and high resolution proteomic analyses (see below) disclosed that the enzymes involved in production of butyrate are among the most highly expressed in cecal extracts prepared from mono- and biassociated mice containing *E. rectale*.

In vitro studies have shown that in the presence of carbohydrates, *E. rectale* consumes large amounts of acetate for butyrate production[133]. Several observations indicate that *E. rectale* utilizes *B. thetaiotaomicron*-derived acetate to generate increased amounts of butyrate in the ceca of our gnotobiotic mice. First, *E. rectale* upregulates a phosphate acetyltransferase (EUBREC_1443; EC 2.3.1.8) - one of two enzymes involved in the interconversion of acetyl-CoA and acetate. Second, cecal acetate levels are significantly lower in co-colonized mice compared to *B.*

thetaiotaomicron monoassociated animals. Third, although cecal butyrate levels are similar in *E. rectale* mono- and biassociated animals, expression of mouse Mct-1, encoding a monocarboxylate transporter whose inducer and preferred substrate is butyrate[135], is significantly higher in the distal gut of mice containing both *E. rectale* and *B. thetaiotaomicron* versus *E. rectale* alone ($p<0.05$). The cecal concentrations of butyrate we observed are similar to those known to upregulate Mct-1 in colonic epithelial cell lines[135]. Higher levels of acetate (i.e. those encountered in *B. thetaiotaomicron* monoassociated mice) were insufficient to induce any change in Mct-1 expression compared to germ-free controls.

The last enzyme in *E. rectale*'s butyrate production pathway, butyrylCoA dehydrogenase/electron transfer flavoprotein (Bcd/Etf) complex (EUBREC_0735-0737; EC 1.3.99.2), offers a recently discovered additional pathway for energy conservation, via a bifurcation of electrons from NADH to crotonylCoA and ferredoxin[136]. Reduced ferredoxin, in turn, can be reoxidized via hydrogenases, or via the membrane-bound oxidoreductase, Rnf, which generates sodium-motive force. The upregulation and high level of expression of these key metabolic genes when *E. rectale* encounters *B. thetaiotomicron* indicates that *E. rectale* not only employs this pathway to generate energy, but to also accommodate the increased demand for NAD+ in the glycolytic pathway. Consistent with these observations, we found that the NAD+/NADH ratio in cecal contents was significantly increased with co-colonization. A high NAD+/NADH ratio promotes high rates of glycolysis, since NAD+ is a required cofactor and may represent an adaptation by *E. rectale* to increased nutrient uptake.

The pathway for acetate metabolism observed in this simplified model human gut community composed of *B. thetaiotaomicron* and E. rectale differs markedly from what is seen in mice that harbor *B. thetaiotaomicron* and the principal human gut methanogenic archaeon, *Methanobrevibacter smithii*. When *B. thetaiotaomicron* encounters *M. smithii* in the ceca of gnotobiotic mice, there is increased production of acetate by *B. thetaiotaomicron*, no diversion to butyrate (and no induction of Mct-1; [123] and B. Samuel and J. Gordon, unpublished observations), increased serum acetate levels, and increased adiposity compared to *B. thetaiotaomicron* mono-

associated controls. In contrast, serum acetate levels and host adiposity (as measured by fat pad to body weight ratios) are not significantly different between *B. thetaiotaomicron* monoassociated and *B. thetaiotaomicron-E. rectale* co-colonized animals (n=4-5 animals/group; n=3 independent experiments; data not shown).

3.3.2.5: Colonic transcriptional changes evoked by *E. rectale-B. thetaiotaomicron* co-colonization – We subsequently used Affymetrix Mouse 430 2 GeneChips to compare patterns of gene expression in the proximal colons of mice that were either germ-free, monoassociated with *E.rectale* or *B. thetaiotaomicron*, or co-colonized with both organisms (n=4 mice per group; total of 16 GeneChip datasets). In contrast to the small number of genes whose expression was significantly changed (1.5-fold cut off, <1%FDR) after colonization with either bacterium alone relative to germ-free controls, co-colonization produced significant alterations in the expression of 508 host genes. Expression of many of these genes also changed with monoassociation with either organism, and in the same direction as seen after co-colonization, but in most cases the changes evoked by *B. thetaiotaomicron* or *E. rectale* alone did not achieve statistical significance. Unsupervised hierarchical clustering of average expression intensity values derived from each of the four sets of GeneChips, revealed that the *E.rectale* monoassociation and *E.rectale-B.thetaiotaomicron* bi-association profiles clustered separate from the germ-free and *B. thetaiotaomicron* monoassociation datasets.

Ingenuity Pathway Analysis (www.ingenuity.com) disclosed that the list of 508 host genes affected by co-colonization was significantly enriched in functions related to cellular growth and proliferation (156 genes), as well as cell death (142 genes). A number of components of the canonical wnt/β catenin pathway known to be critically involved in controlling self-renewal of the colonic epithelium were present in this list. Many of the changes observed in biassociated mice are likely to be related to the increased influx of butyrate, generated by *E. rectale*, into colonic cells. Butyrate, a histone deacetylase inhibitor that evokes pronounced transcriptional changes in different types of cultured epithelial cell line[137,138,139,140], is the preferred energy substrate for colonic enterocytes[141]. While transcriptional changes caused by butyrate differ depending upon the cell lineage, state of cellular differentiation, and

cellular energy status[139,140,142,143] *in vitro* and *in vivo* studies have shown that it affects expression of genes involved in proliferation, differentiation and apoptosis[137,142].

As mentioned above, as part of its adaptation to the presence of *E. rectale*, *B. thetaiotaomicron* upregulates a number of genes involved in the harvest of host glycans. Included among these *B. thetaiotaomicron* genes are components of a fucose utilization operon linked to the production of a bacterial signal that induces synthesis of intestinal mucosal fucosylated glycans, and also catabolism of fucose from O-glycans[144]. GeneChip profiling of colonic gene expression disclosed that co-colonization results in increased expression of *Fut2* (α-1,2 fucosyltransferase), *Fut4* (α-1,3-fucosyltransferase), plus nine other genes involved in the synthesis of mucosal glycans (glycosphingolipids and O-glycans). By increasing host production of glycans, *B. thetaiotaomicron* can benefit itself, and through its metabolic products, *E. rectale*.

3.3.2.6*: E. rectale's* colonization levels and production of butyrate are affected by host diet - In a final series of experiments, we assessed how *E. rectale* and *B. thetaiotaomicron* were affected by changes in host diet. Groups of age- and gender-matched co-colonized mice were fed one of three diets that varied primarily in their carbohydrate and fat content: (i) the standard low-fat, plant polysaccharide-rich diet used for the experiments described above (abbreviated 'LF/PP' for low-fat/plant polysaccharide), (ii) a high-fat, 'high-sugar' Western-type diet (abbreviated HF/HS) that contained sucrose, maltodextrin, corn starch as well as complex polysaccharides (primarily cellulose) that were not digestible by *B. thetaiotaomicron* or *E. rectale*, and (iii) a control diet that was similar to (ii) except that the fat content was 4-fold lower ('LF/HS' for low-fat, high-sugar; n=5 mice per group). Whereas *B. thetaiotaomicron*'s colonization levels were similar in all three diets, colonization of *E. rectale* was significantly reduced (five-fold) in mice fed either the LF/HS or HF/HS diets ($p<0.01$, heteroscedastic t-test).

Whole-genome transcriptional profiling of both organisms showed that relative to the standard polysaccharide-rich chow diet (LF/PP), both the Western style HF/HS diet and its LF/HS control produced a significant upregulation of *B. thetaiotaomicron* PULs

involved in harvesting and degrading host polysaccharides, and a downregulation of several PULs involved in the degradation of dietary plant polysaccharides. *E. rectale*'s response to the HF/HS and LF/HS diets was to downregulate several of its GHs as well as a number of its sugar transporters. Moreover, levels of butyrate were five-fold lower in co-colonized mice fed these compared to the standard chow LF/PS diet [0.496 ±0.0051 μmol/g wet weight cecal contents; (LF/PP) vs. 0.095±0.002 (HF/HS) vs 0.080±0.008 (LF/HS) ($p<0.05$ ANOVA)].

These dietary manipulations lend further support to the view that *B. thetaiotaomicron* functions in this model two-member human microbiota to process complex dietary plant polysaccharides and to distribute to the products of digestion to *E. rectale* which, in turn, synthesizes butyrate. The response of *E. rectale* to the HF/HS and LF/HS diets can be explained by the fact that this Firmicute does not have predicted GHs and PLs that can process host glycans. In addition, it could not utilize most of the sugars we tested that are derived from mucosal polysaccharides. Finally, the host possesses enzymes in its glycobiome that can directly process the simple sugars present in these two diets. Indeed, human subjects that are fed diets deficient in complex polysaccharides harbor lower levels of butyrate-producing gut bacteria, including members of the *E. rectale*-containing clade[145]. Our simplified gnotobiotic model of the microbiota underscores the functional implications of diet-associated changes in the representation of this clade, not only as they relate to the operations of the microbiota itself but also potentially as they relate to butyrate-mediated changes in gut epithelial homeostasis.

3.3.3: Proteomic studies of this simplified two-component model of the human gut microbiome

Model communities such as the one described above, constructed in gnotobiotic mice, where microbiome gene content is precisely known and transcriptional data are obtained under conditions where potentially confounding host variables such as diet and host genotype can be constrained, provide a way to test the efficacy of mass spectrometric methods for characterizing gut microbial community proteomes.

Therefore, we assayed luminal contents, collected from the ceca of 8 gnotobiotic mice fed the standard polysaccharide-rich LF/PP diet: (germ-free, monoassociated, and co-colonized; n=2 mice/treatment group representing two independent biological experiments).

The measured proteomes had high reproducibility in terms of total number of proteins observed and spectra matching to each species. A total of ~6,300-21,000 spectra were identified per sample and differ based on inoculations. For a complete list of the total number of identified spectra, peptides and proteins per sample and run, see **Table 3.2**. Interestingly, the total number of identified spectra was, for the most part, distinct and unique to each bacterial species. Unlike *B. thetaiotaomicron* and *E. rectale*, the number of identified spectra belonging to mouse was redundant: thus, a higher number of spectra were non-unique spectra. The difference is evident when the total spectra counts are compared to unique spectra counts only. The total average spectra count identified in the control (germ-free) mouse was 10,767 for sample 700 and 11, 221 for sample 799. The total average unique spectra count, however, decreased to 4,394 and 4,168. Therefore, the majority of identified mouse peptides are not unique within the database. The total number of unique spectra counts per species and run can be found in **Table 3.3**. The two co-colonized mice (710 and 810) had a total of ~ 77% unique spectra belonging to *B. thetaiotaomicron*, 20% unique spectra belong to *E. rectale*, and only 3% of the two species' combined spectra counts were non-unique. This suggests that the majority of identified proteins belonging to *B. thetaiotaomicron* and *E. rectale* are true unique identifications and these species can be easily differentiated by proteomics. These values were calculated by summing the total number of unique spectra per species per run, followed by an average per species across all runs (**Table 3.3**).

Table 3.2: High resolution proteomic analyses of cecal contents from gnotobiotic mice – total proteins, peptides, and spectra for each sample.

Sample	Protein IDs	Peptide IDs	Spectra	Species Inoculation
700 Run 1	716	3505	7561	None (control)
700 Run 2	702	3448	7067	
700 Run 4	596	2965	7019	
705 Run 1	1526	11515	21228	*B. thetaiotaomicron*
705 Run 2	1538	11534	20051	
705 Run 3	1513	9577	17119	
710 Run 2	1335	9270	17910	*B. thetaiotaomicron* and *E. rectale*
710 Run 3	1482	9256	16839	
710 Run 4	1612	10484	17635	
715 Run 1	914	6388	12243	*E. rectale*
715 Run 2	894	6241	12257	
715 Run 5	945	6040	11358	
799 Run 1	571	2809	6338	None (control)
799 Run 2	471	2355	6575	
799 Run 3	449	2213	6995	
806 Run 1	1407	9366	18071	*B. thetaiotaomicron*
806 Run 2	1424	9400	18915	
806 Run 3	1358	8867	15864	
810 Run 1	1409	7798	14102	*B. thetaiotaomicron* and *E. rectale*
810 Run 2	1509	8505	14659	
810 Run 3	1431	7658	14217	
817 Run 1	837	4779	10294	*E. rectale*
817 Run 2	791	4519	10346	
817 Run 3	881	4880	10829	

Table 3.3: High resolution proteomic analyses of cecal contents from gnotobiotic mice – breakdown of unique spectral counts for all species in the database for Sample Set 1 and 2.

Sample Set 1

Sample ID:	700 Run1	700 Run2	700 Run4	705 Run1	705 Run2	705 Run3	710 Run2	710 Run 3	710 Run4	715 Run1	715 Run2	715 Run5
B. thetaiotaomicron	9	8	6	17329	16455	14149	10723	10104	10596	7	23	2
E. rectale	36	23	31	22	15	6	3596	3494	3843	6817	6790	6294
M. musculus	4591	4377	4216	2324	2122	1852	1911	1827	1847	3426	3438	3317
Rice	232	254	269	171	142	105	69	46	45	121	112	98
Yeast	13	8	13	2	4	5	9	4	16	23	12	8
contams	40	32	30	37	33	19	30	24	26	30	35	27
Total:	4921	4702	4565	19885	18771	16136	16338	15499	16373	10424	10410	9746

Sample Set 2

Sample ID:	799 Run1	799 Run2	799 Run3	806 Run1	806 Run2	806 Run3	810 Run1	810 Run2	810 Run3	817 Run1	817 Run2	817 Run3
B. thetaiotaomicron	0	0	0	14403	15034	12693	9115	9562	8933	4	4	3
E. rectale	4	1	2	6	11	5	1600	1736	1658	4749	4686	5068
M. musculus	3949	4250	4306	2318	2384	1932	2260	2183	2366	3582	3739	3743
Rice	272	226	243	56	113	34	62	114	107	140	113	135
Yeast	14	12	7	4	0	4	5	9	5	16	11	3
contams	15	13	15	5	0	4	7	6	8	3	7	5
Total:	4254	4502	4573	16792	17542	14672	13049	13610	13077	8494	8560	8957

Table 3.4 provides a summary of our analyses, including the percentage of mRNAs called 'Present' in the GeneChip datasets for which there was an identified protein product. These data suggest that RNA and protein identifications are not always correlative. While both datasets provide valuable insight into the two microbes function in the gut, there are differences for which many mRNAs were identified, but were not present or identified in the final protein product. The most abundant identified products from both microbes included ribosomal proteins, elongation factors, chaperones, and proteins involved in energy metabolism. Many conserved hypothetical and pure hypothetical proteins were identified, as well as 10 genes in *B. thetaiotaomicron* whose presence had not been predicted in our initial annotation of the finished genome. Together, the results provide validation of experimental and computational procedures used for proteomic assays of a model gut microbiota, and also illustrate some of the benefits in obtaining this type of information.

Table 3.4: Summary of proteins detected by mass spectrometry of the cecal contents of gnotobiotic mice.

	E. rectale			**B. thetaiotaomicron**		
	Mono-association	Bi-association	Total	Mono-association	Bi-association	Total
Detected by MS/MS	661	453	680	1608	1367	1687
Detected by GeneChip	2139	2010	2150	3798	3865	3995
GeneChip-/ MS/MS+	7	7	8	40	21	23
MS/MS- /GeneChip+ [a]	1608	1638	1603	2280	2569	2357

[a]Positive is defined as having a 'Present' call in ≥75% of GeneChips.

3.4: Prospectus

These studies of a model two component human gut microbiota created in gnotobiotic mice support a view of the Bacteroidetes, whose genomes contain a disproportionately large number of glycan-degrading enzymes compared to sequenced Firmicutes, as responding to increasing diversity by modulating expression of their vast array of polysaccharide utilization loci. *B. thetaiotaomicron* responds to the presence of *E. rectale* by upregulating a variety of loci specific for host-derived mucin glycans that *E. rectale* is unable to utilize. *E. rectale*, which like other Firmicutes has a more specialized capacity for glycan degradation, broadly downregulates its available GHs in the

presence of *B. thetaiotaomicron*, even though it does not grow efficiently in the absence of carbohydrates. It also becomes more selective in its harvest of sugars and its transcriptional profile suggests improved access to other nutrients (e.g. there is a generalized upregulation of amino acid biosynthetic genes as well as a set of nutrient transporters that can harvest peptides).

We have previously used gnotobiotic mice to show that the efficiency of fermentation of dietary polysaccharides to short chain fatty acids by *B. thetaiotaomicron* increases in the presence of *M. smithii* [123]. Co-colonization increases the density of colonization of the distal gut by both organisms, increases production of formate and acetate by *B. thetaiotaomicron* and allows *M. smithii* to use H_2 and formate to produce methane, thereby preventing the build-up of these fermentation end-products (and NADH) in the gut bioreactor, and improving the efficiency of carbohydrate metabolism[123]. Removal of H_2 by this methanogenic archaeon allows *B. thetaiotaomicron* to regenerate NAD+, which can then be used for glycolysis. This situation constitutes a mutualism, in which both members show a clear benefit. The present study, characterizing the co-colonization with *B. thetaiotaomicron* and *E. rectale*, describes a more nuanced interaction where both species colonize to similar levels if carbohydrate substrates are readily available. Moreover, certain aspects of bacterial-host mutualism become more apparent with co-colonization, including increased microbial production and host transport of butyrate, and increased host production and microbial consumption of mucosal glycans: this mutualism is likely vital for the co-existence of these species.

It seems likely that as the complexity of the gut community increases, interactions between *B. thetaiotaomicron* and *E. rectale* will either be subsumed or magnified by other 'similar' phylogenetic types (as defined by their 16S rRNA sequence and/or by their glycobiomes). Synthesizing model human gut microbiotas of increasing complexity in gnotobiotic mice using sequenced members of our intestinal communities should be very useful for exploring two ecologic concepts: (i) the neutral theory of community assembly which posits that most species will share the same general niche (profession), and thus are likely to be functionally redundant[146], and (ii) the idea that

both bottom-up selection, where fierce competition between members of the microbiota drives phylotypes to assume distinct functional roles, and top-down selection, where the host selects for functional redundancy to insure against failure of bioreactor functions, operate in our guts.

Chapter 4

Optimization of a cellular lysis/mass spectrometric proteome characterization approach for a model 7-member gut microbial community in gnotobiotic mice

Alison R. Erickson, Nathan P. McNulty, Nathan C. VerBerkmoes, Jeffrey I. Gordon, and Robert L. Hettich

4.1: Introduction

Mass-spectrometry (MS)-based proteomics has become very powerful in providing comprehensive and unbiased characterization of proteins and proteomes. With the onset of multidimensional protein separations interfaced to high-performance tandem mass spectrometry, this experimental approach can handle substantial protein or peptide complexity and simultaneously achieve protein identification[95]. However, the complexity and 'dynamic range' of microbial proteomes containing thousands of bacterial species have hindered the ability to identify whole community proteomes, as compared to traditional single bacterial isolates (e.g., *E. coli*).

The 'standard' shotgun proteomics strategy includes cellular lysis (with or without fractionation), protein denaturation and digestion, peptide separation via LC and identification via tandem mass spectrometry (MS/MS). While much optimization has been invested in LC-MS/MS, experimental methods involving sample preparation (i.e., cellular fractionation and lysis), peptide separation (i.e., gel electrophoresis and LC), and MS/MS are equally important to enhance overall protein identifications in MS-based shotgun proteomics of complex microbial samples (i.e., soil, ocean, feces). Environmental microbial samples pose several challenges not characteristic of laboratory –based systems, such as increased dynamic range (abundance) of microbial species and proteins, and interferences derived from the environmental matrix. In general, there are two options available where a microbial community sample is either i) derived, processed, and lysed directly (*in situ*) from the source (i.e., feces, cecum, tissue) with both bacterial and host cells included ('direct approach') or ii) enriched for bacterial cells to eliminate all host proteins and contaminants via centrifugation ('indirect' approach). Environmental matrices can be problematic for the 'direct' lysis

and protein extraction approach[89] without any pre-fractionation or enrichment (i.e., differential centrifugation) of bacterial cells due to inference with downstream processes and analysis (i.e., peptide signal suppression)[74]. Widely accepted protocols for cell lysis and extraction of proteins from complex microbial communities for proteomic analysis include a thermally assisted detergent-based cellular lysis (sodium dodecyl sulfate, SDS) method[74], a small-scale microbial biomass experimental approach[86], sonication[87,88], freeze-thaw cycles[89], and French press[60]. Available and tested lysis buffers include detergents (e.g., SDS, CHAPS, and Triton X-100), chaotropes (urea and guanidine), acid-labile surfactants (PPS silent surfactant) and many other commercially available buffers to disrupt bacterial cells with or without physical or mechanical disruption prior to protein extraction. Several precautions should be taken if a detergent is selected due to their interference with binding, elution, and ionization of peptides during tandem MS experiments. To eliminate contamination and interference of detergents with mass spectrometers, several proteomics studies[26,42,43,86] opted to use an 'indirect' approach and chaotropes (i.e., guanidine) to enrich and lyse bacterial cells and denature proteins. In this study, we focused on comparing methods that are used prior to proteolysis and LC-MS/MS with emphasis on identifying an efficient *in situ* lysis and protein extraction method.

Due to the growing interest and desire to understand the human microbiome[14,69], MS-based proteomics has begun to emerge as a key player in understanding the functional signatures of the human gut and oral microbiome[26,37,47,48,49,83]. Therefore, to achieve comprehensive proteome coverage of samples collected from the human gut microbiota, optimization of experimental MS-based methods for their direct application to complex community samples (i.e., feces or ceca) would help increase protein identification and our understanding of the host-microbiota functional signatures. A variety of methods including one protein fractionation (ultracentrifugation) method, five bacterial lysis methods, and protein (TCA) precipitation were evaluated to identify the best performing method for MS-based analysis of human-derived gut microbiota in germ-free (gnotobiotic) mice (**Figure 4.1**). A model human gut microbiota of seven bacterial

species (*B. WH2*, *B. ovatus, B. vulgatus, B. thetaiotaomicron, B. caccae,* and *Parabacteroides distasonis*) belonging to one of two dominant gut phyla, Bacteroides, was inoculated in gnotobiotic mice to measure the proteomes of the microbial community, in addition to each individual species. The method that provided the most efficient lysis and higher peptide recovery, thus, increased protein identification would be identified as the best overall performing method that could be applied *in situ* to any future fecal or cecal samples.

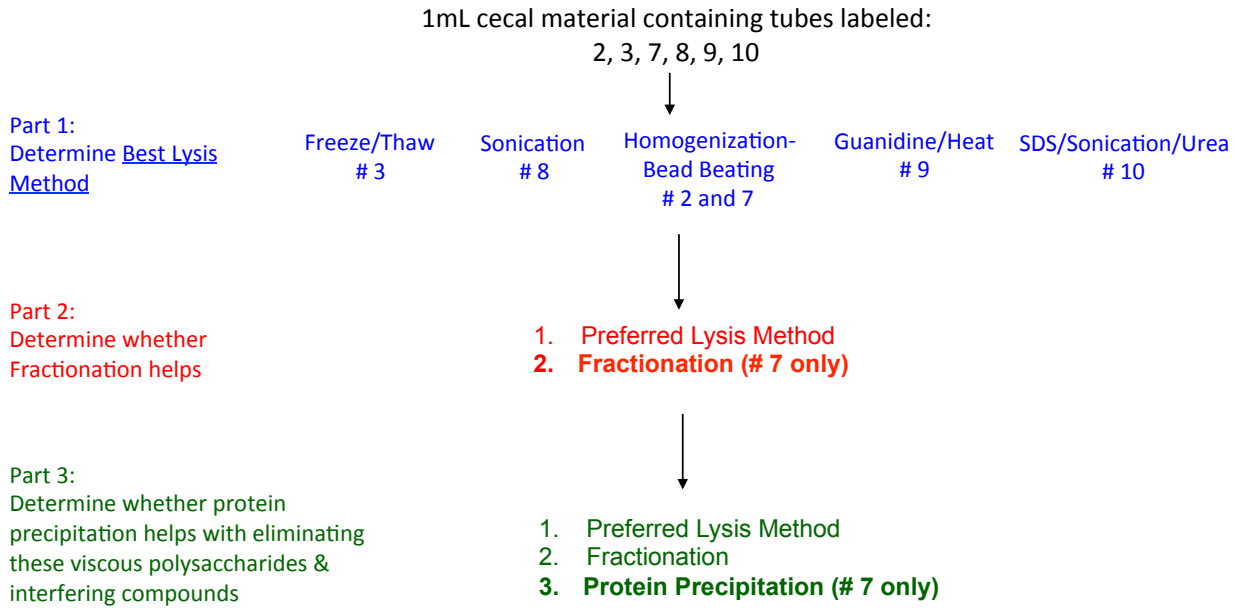


Figure 4.1. Experimental design for method optimization of microbial and host cell lysis and protein extraction of a model human-derived gut microbial community in gnotobiotic mice.

4.2: Experimental Methods

4.2.1: Sample collection

Dr. Jeffrey Gordon and Nate McNulty (Wash. Univ) provided a total of six ceca samples for the 7-member community proteomics experiments. The C57BL/6 adult male germ-free mouse cecum, labeled as either 2, 3, 7, 8, 9 and 10, were gavaged with an equal inoculum of the following species: *Bacteroides caccae*, *B. ovatus*, *B. uniformis*, *B. WH2*, *B. thetaiotaomicron*, *B. vulgatus*, and *Parabacteroides distasonis* and fed a standard BK diet *ab libitum*. The total microbial does was ~ 8.7 x 10^7 corresponding to 1.2-1.3x10^7 CFUs/microbe. The cecum was harvested at 14 days post-gavage, frozen in liquid nitrogen immediately, and shipped overnight on dry ice to ORNL.

4.2.2: Bacterial lysis and protein extraction

A total of five bacterial lysis methods were each performed on single mouse cecum (~1 mL). As described in **Figure 4.1**, the following widely-accepted bacterial lysis methods were applied and varied with respect to each of their individual protocols: sonication (8), freeze-thaw cycles (3), bead-beating (2 and 7), no physical disruption (9), and chemical disruption via sodium dodecyl sulfate (SDS) (10) with a brief description of the protocols to follow. Cecum 2 (bead-beating) was first solubilized with 1mL 6M guanidine in 0.1 mm zirconia/silica beads and beat using a RETSCH Mixer Mill MM 400 for a total of 2 minutes (30 second intervals with 2 minute break) at room temperature with a frequency of 20Hz. The homogenized cecum was centrifuged for 5 minutes at 3,000 rpm to remove excess debris and pellet all beads. The supernatant was removed and beads washed with 6M Guanidine.

Cecum 7 (bead-beating) was treated using the same steps described above for cecum 2; however, the cecum was solubilized in 50mM Tris/10mM $CaCl_2$ instead of 6M guanidine to allow for proper cell separation via ultracentrifugation. Following homogenization, the collected supernatant was transferred to a glass test tube and centrifuged at room temperature for 1 hour at 100,000x g using a Ti 40 fixed angle rotor ultracentrifuge (Beckman Coulter). The supernatant (soluble fraction) was extracted

away from the pellet (membrane fraction) and treated separately for protein denaturation and digestion.

Cecum 3 (freeze-thaw lysis method) was initially frozen in liquid N_2 for 1 minute followed by 60°C treatment in a water for 1 minute and repeated for a total of three cycles.

Cecum 8 (sonication only method) was initially solubilized in 6M Guanidine and exposed to sonication for a total 5 minutes at 20% amplitude on ice.

Cecum 9 (no physical or chemical disruption) wa processed via single tube cell lysis[86] and protein digestion. The cecum (~1mL) was suspended in 6M Guanidine/10mM DTT at 60°C for 1 hour to lyse cells and denature proteins.

Lastly, cecum 10 was solubilized in 1mL SDS lysis buffer (4% w/v SDS, 100mM Tris•HCl, pH 8.0, 10mM dithiothreitol (DTT)) and lysed mechanically by sonication followed by incubation for 5 minutes at 95°C. Cells were centrifuged at 21,000 x g. Following an overnight tricholoroacetic acid (TCA) precipitation, the TCA precipitates (protein mixtures) were resolubilized in 500uL of 8M urea, 100mM Tris•HCl, pH 8.0, and reduced by incubation at a final concentration of 10mM DTT for 1 hr at room temperature. Samples were sonicated and an aliquot taken to determine the protein concentration using a bicinchonic acid-(BCA) based protein assay kit (Pierce). Approximately 3mg of protein was extracted were diluted with 100mM Tris•HCl, 10mM CaCl2, pH 8.0 to a final urea concentration below 4M. Proteolytic digestions were initiated with sequencing grade trypsin (1/100, w/w; Promega) overnight at room temperature. A second aliquot of trypsin was added (1/100) and diluted with 100mM Tris•HCl, pH 8.0 to a final urea concentration below 2M. Following a 4 hr incubation at room temperature, samples were reduced to a final concentration of 10mM DTT. The peptides were acidified (protonated) in 200mM NaCl, 0.1% formic acid, filtered, and concentrated with a 10k molecular weight cutoff spin column (Sartorius). A total of ~100mg of peptides were used for each LC-MS/MS experiment.

For ceca 3, 8, and 9 the lysed cecum (~1mL) was centrifuged at room temperature for 10 minutes at 3,000 rpm to pellet all debris and any small contaminating molecules. An aliquot was taken from all six ceca to determine the protein concentration using a BCA (Pierce). A total concentration of 3mg protein was extracted for protein denaturation in 6M Guanidine and proteolytic digestions as follow. For ceca 2, 3, 7, 8, and 9 protein extractions, the guanidine concentration was diluted from 6M to 1M with 50mM Tris buffer/10mM $CaCl_2$ and proteolytic digestions initiated with sequencing grade trypsin (1/100, w/w; Promega) overnight at 37°C to digest proteins into peptides. A second aliquot of trypsin was added (1/100) and incubated for 4 hours at room temperature. Samples were reduced to a final concentration of 10mM DTT. The complex peptide solution was desalted via C_{18} solid phase extraction, concentrated and filtered (0.45um filter). For each LC-MS/MS analyses below, ~100mg of the total peptide sample was used for LC-MS/MS.

4.2.3: LC-MS/MS analysis

Peptides were loaded onto a two-dimensional (C_{18} and SCX) 15cm length column packed in-house and separated with a 12 step, multidimensional high-pressure liquid chromatographic elution method using an Ultimate HPLC system (Dionex, Sunnyvale, CA) consisting of eleven salt pulses followed by a 2 hr reverse-phase gradient from 100% solvent A (A: 95% H2O, 5% acetonitrile, 0.1% formic acid) to 50% solvent B (B: 30% H2O, 70% acetonitrile, 0.1% formic acid). The HPLC system was coupled on-line with an LTQ-Orbitrap XL (Thermo Fischer Scientific) via the Proxeon nanospray source. Full MS scans were acquired in the Orbitrap mass analyzer (from 400-1700 *m/z*) with resolution 30,000 followed by five data dependent tandem MS/MS scans in the LTQ with normalized collision energy of 35%. For all sequencing events, dynamic exclusion was enabled.

4.2.4: Data analyses and informatics

All MS/MS spectra were searched with the SEQUEST v 0.27 algorithm[66] and filtered with DTASelect/Contrast[98] at the peptide level [Xcorrs of at least 1.8 (+1), 2.5 (+2), 3.5 (+3)] with a deltCN 0.08. Only proteins identified with two fully tryptic peptides from

the 22 hr runs were considered for further biological inference. Tandem MS/MS spectra were searched against a protein sequence database (**Table 4.1**) containing the 7 relevant Bacteroides species (*B. WH2, B. ovatus, B. vulgatus, B. thetaiotaomicron, B. caccae, and Parabacteroides distasonis*) in addition to 8 distractor (non-relevant) species, the host (mouse) genome, diet components (rice and yeast), and common contaminants.

Table 4.1: Protein sequence database composition for the 7-member database searches. Bolded genome names are relevant microbes that were gavaged in the gnotobiotic mice.

Genome	Proteins	Size (MB)
Bacteroides caccae	3855	1.772
Bacteroides ovatus	5536	2.536
Bacteroides thetaiotaomicron	4778	2.376
Bacteroides uniformis	4663	1.952
Bacteroides vulgatus	4065	1.932
Bacteroides WH2	5244	2.656
Collinsella aerofaciens	2367	0.996
Clostridium scindens	3995	1.528
Clostridium spiroforme	2465	1.016
Dorea longicatena	2970	1.18
Eubacterium rectale	3631	1.42
*Faecalibacterium prausnitzii*_M212	3493	1.368
Parabacteroides distasonis	3850	1.86
Ruminococcus obeum	4175	1.536
Ruminococcus torques	2875	1.124
Mouse	34966	19.2
Yeast	6345	3.33
Rice	66710	36
Common contams	36	0.02

4.3: Results and Discussion

4.3.1: General MS-based proteome metrics

As first demonstrated in chapter 3, MS-based proteomics for the human gut microbiome in gnotobiotic mice has proven successful and applicable to *in vivo* gut-derived ecosystems[37]. However, with increasing complexity and diversity of the microbial community(s) membership, traditional up-front sample processing methods were evaluated and optimized for increased depth and coverage of proteomes derived *in situ* from cecum. All MS experiments were performed in duplicate for each method and demonstrated high technical reproducibility, with an $R^2 \geq 0.9184$ for all six ceca samples. Using general MS-based proteomics metrics (i.e., assigned spectra and protein identifications), a comparison of all six samples and methods suggest that fractionation via ultracentrifugation and protein precipitation (TCA) of the soluble fraction provided the greatest identification of non-redundant spectra, peptides and proteins (**Table 4.2**). Thus, the approach used for the 'soluble fraction' significantly improved the quantity of protein and peptide identifications with a 2.72X gain in protein identifications from, on average, 411 to 1,118 proteins and 2.925X gain in peptide identifications from, on average, 1,907 to 5,578 peptides per run. Therefore, from this point on, we will only focus on using the protein precipitated soluble fraction to represent the soluble fraction of sample and method #7.

Table 4.2: Proteome sample metrics of total non-redundant protein, peptide and spectra counts per method.

Sample	Run	Proteins	Peptides	Spectra	PPMs (%)	Method
7 (membrane)	1	3159	11675	23901	86.88	Bead-beating; Ultracentrifugation; Guanidine
	2	3149	11607	23338	86.99	
7 (soluble)	4	1188	5961	14789	88.91	Bead-beating; Ultracentrifugation; TCA; Guanidine
	5	1047	5194	13220	88.94	
	1	424	1935	10331	87.7	Bead-beating, Ultracentrifugation; Guanidine
	2	398	1879	10709	88.6	
10	1	3204	11162	20265	89.51	Boiling SDS; TCA; Urea
	2	2931	9986	21976	89.63	
8	1	2207	6874	16730	85.6	Sonication, Guanidine
	2	2079	6422	17107	85.79	
3	1	2062	6470	16641	85.45	Freeze-Thaw, Guanidine
	2	2178	6845	16467	86.29	
2	3	1826	5416	11951	86.55	Bead-beating, Guanidine
	4	1834	5448	11380	87.43	
9	2	1832	5908	13404	85.57	Guanidine only
	3	1728	5573	14175	86.88	

The results found in **Table 4.2** demonstrate that method #7, ultracentrifugation and protein precipitation performed the best based on overall MS-based metrics with a total sum of, on average, 4,272 proteins (3,154 membrane and 1,118 soluble proteins per run). However, these numbers can be misleading due to insufficient fraction purity and degree of protein overlap between the two fractions. As described further below (**Figure 4.2**), a total of 3,118 and 521 proteins are unique to the membrane and soluble (protein precipitation) fractions, respectively, with 1,248 shared proteins for a total of 4,887 non-redundant proteins. Although method #7 (ultracentrifugation) significantly out-performs method #2 (without ultracentrifugation) and all other methods, a gain of only 521 proteins were uniquely identified from the soluble TCA fraction. Following closely is the second best performing lysis method #10 with, on average, 3,068 proteins, 10,574 peptides, and 21,121 spectra per run. Of the physical disruption techniques (methods 2, 3 and 8), 8 and 3 (sonication and freeze-thaw, respectively) performed the best followed by #9 and #2 (guanidine only and bead-beating, respectively) which provided the least assigned spectra, peptide and protein identifications. Because an equal quantity of tandem MS/MS spectra were acquired for each method and MS run, the variation in assigned spectra is not a reflection of the

technical (mass spectrometer) variability, but rather a reflection of the lysis and protein extraction method.

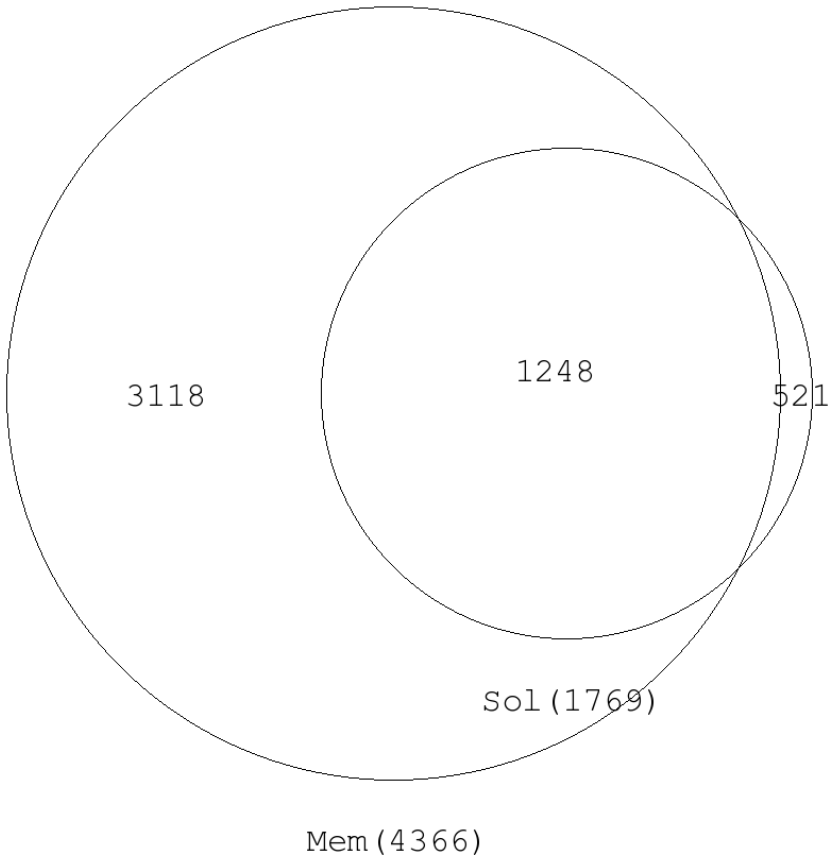


Figure 4.2: A comparison of identified proteins that are shared (1,248) between and unique to the membrane (left circle; 3,118 proteins) and soluble (TCA; right circle; 521 proteins) fractions for method 7 only. A total of 4,366 and 1,769 non-redundant proteins were indentified in the membrane and soluble (TCA) fraction, respectively.

4.3.2: Protein-level comparison

A four-way protein comparison of methods: 2 (bead-beating), 3 (freeze-thaw), 8 (sonication), and 9 (no physical disruption) identified 1,972 shared proteins (~61-73% of all protein identifications) across all four methods, with 6-9% of all identified proteins unique to only one method (**Figure 4.3**). A three-way protein comparison of methods containing physical disruption only (2, 3, and 8) identified 2,292 shared proteins (81%,

73%, and 71%, respectively, of all identified proteins) (**Figure 4.4**). These two comparisons suggest that the methods 2, 3, 8, and 9 do not vary significantly from one another, since the majority of identified proteins are similar in one or more methods (i.e., only 6-9% of all identified proteins are unique to a single method). On the other hand, these methods (with and without physical disruption) vary significantly from a method that uses a combination of chemical lysis (SDS) and physical disruption (sonication) simultaneously to lyse microbial and host cells, where only ≤ 48% of all identified proteins are shared with methods 2, 3, 8, and 9 and ≥ 30% are unique to method 10 only (data not shown).

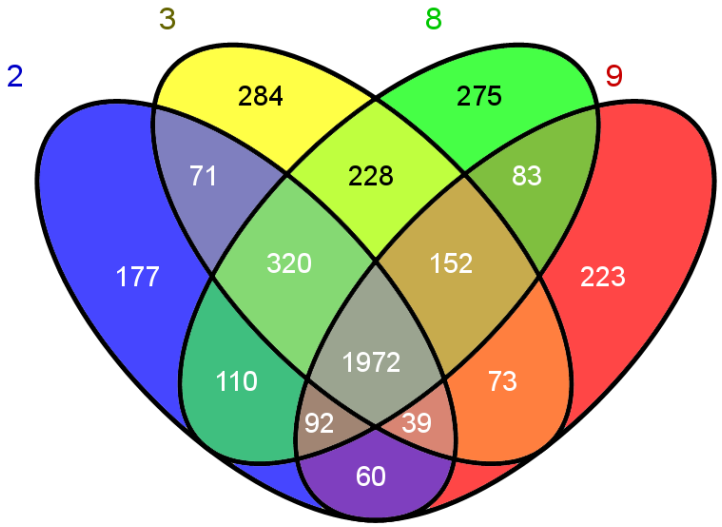


Figure 4.3: A four-way comparison of identified proteins from four different methods: 2 (bead-beating), 3 (freeze-thaw), 8 (sonication), and 9 (no physical disruption).

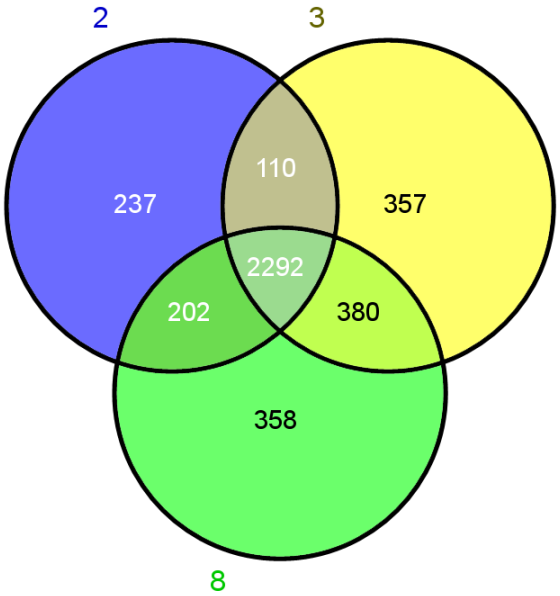


Figure 4.4: A three-way comparison of identified proteins from three methods containing physical disruption only (2, 3, and 8).

For method 7, on average, 3,154 proteins were identified in the membrane fraction per run (**Table 4.1**) with a total of 4,366 non-redundant proteins across both runs. On the other hand, an average of 1,118 and 411 proteins were identified per run for the soluble fraction with and without protein precipitation, respectively, for a total of 1,769 and 676 non-redundant identified proteins. As mentioned previously, a comparison of both fractions from method 7 indicated that 3,118 proteins were uniquely identified in the membrane fraction, as compared to 521 in the soluble fraction, with 1,248 proteins shared between the two fractions. Therefore, a total of 4,887 proteins were identified for method 7 (both membrane and soluble TCA fraction shared and unique proteins summed).

Based on total spectra, peptides, and proteins identified, method # 7 (with protein precipitation) is the best overall performing method, with a total of 4,887 non-redundant proteins, followed by method # 10 with 4,495 non-redundant proteins. Of all proteins identified, a two-way comparison of methods 7 and 10 indicate that 3,506 proteins are shared (72 and 78% of all identified proteins, respectively) with 1,381 and 989 proteins unique to methods 7 (28%) and 10 (22%), respectively (**Figure 4.5**). While a significant

portion of identified proteins were found using both methods, method # 7 had ~40% more unique proteins relative to method 10. Although method # 7 has out-performed the other four lysis methods, we wanted to determine whether any potential biases existed on a functional level as a result of each lysis and extraction method. Functional comparisons will be explored in more detail below.

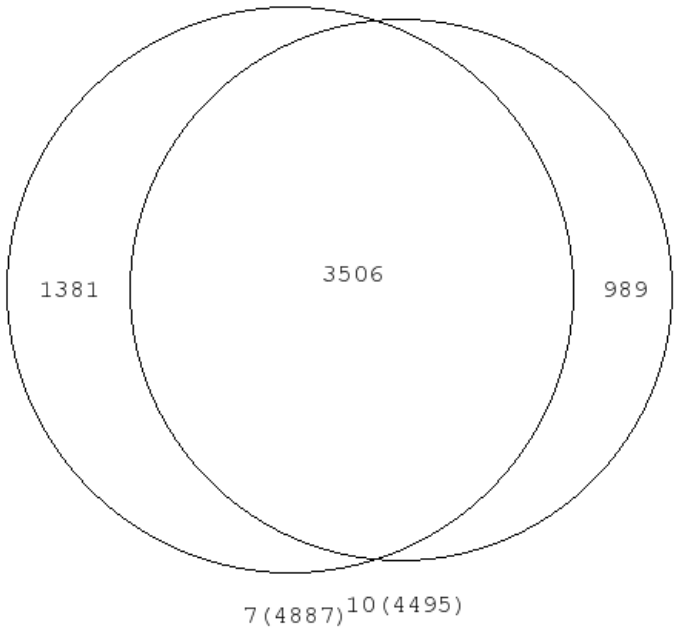


Figure 4.5: A comparison of identified proteins that are shared (3,506 proteins) between and unique to method 7 (membrane and soluble fractions combined; left circle) and 10 (right circle).

4.3.3: Comparison of methods based on protein-species assignments

Due to the differences in lysis methods (i.e., chemical or physical), we wanted to evaluate whether any of the methods were biased and preferentially lyse one or more specific bacterial species within the 7-member consortium. For example, does any method preferentially lyse one particular species better than another or enrich for proteins with specific functional roles, such as outer membrane receptor proteins? A comparison of identified protein counts per species (**Figure 4.6**) suggested that the majority of proteins were derived from *B. WH2* and *B. thetaiotaomicron* for all methods, with the exception of method 9 (guanidine only) and the soluble fractions (# 7).

Contrary to the majority of methods, *B. thetaiotaomicron* was one of the least abundant microbes in method # 9. These results may suggest that additional chemical (SDS) and/or physical (i.e., sonication) disruption is necessary for the complete lysis of all Bacteroides. Interestingly, *M. musculus* was the most abundant contributor based on protein counts followed by *B. WH2* in both soluble fractions contrary to the membrane fraction for method 7.

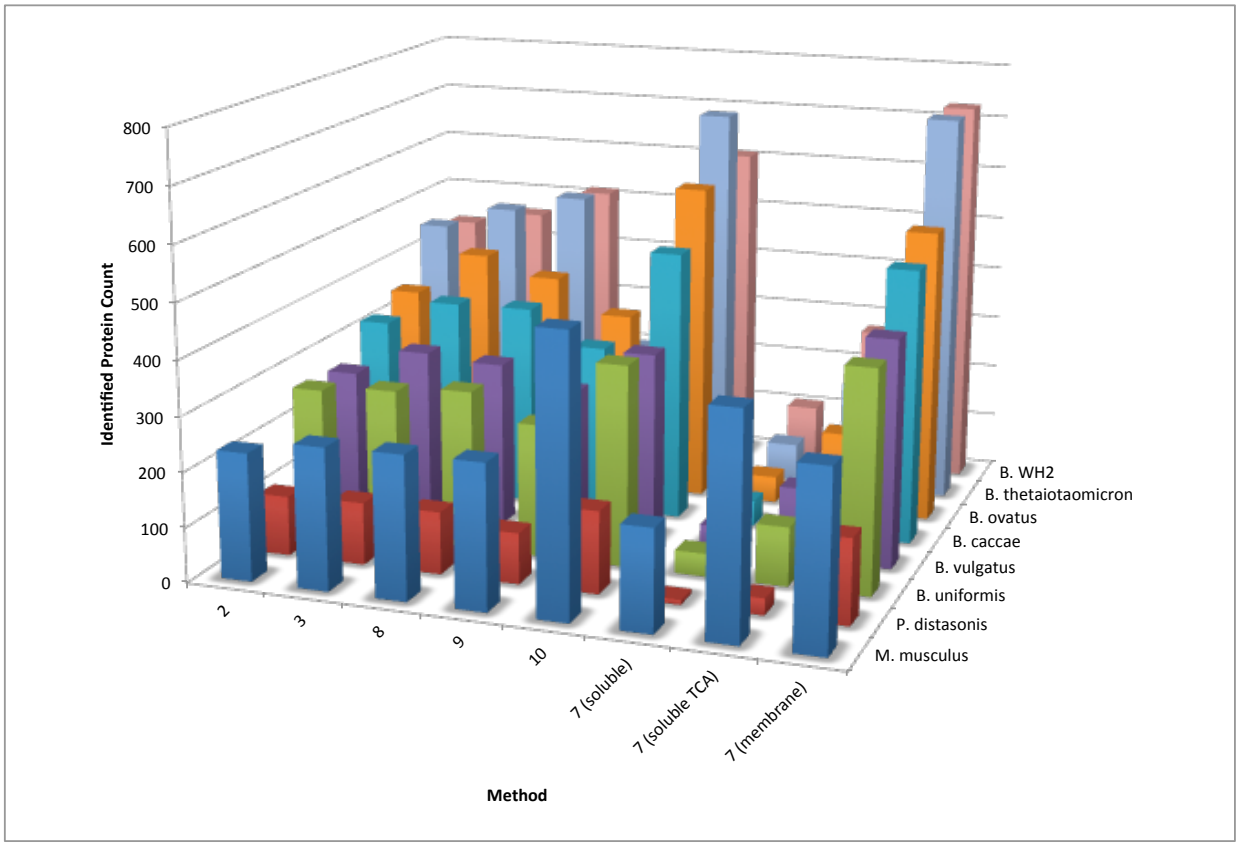


Figure 4.6: Distribution of identified protein counts for all relelvant database components (microbial and host) for all 7 cecal samples.

Due to protein sequence redundancy between the seven-closely related Bacteroides phylotypes, a similar comparison was performed with only unique peptides (**Figure 4.7**) and total spectra counts (**Figure 4.8**) to determine whether similar trends were observed. The distribution of unique peptides also suggest that *B. WH2* and *B. thetaiotaomicron* are the most abundant microbial species for all methods, including method 9 (guanidine only), with the exception of both soluble fractions (# 7). In

agreement with the protein count distribution, *M. musculus* was the most abundant based on unique peptide and spectra counts, followed by *B. WH2* for both soluble fractions. While the spectra counts distribution agrees with the protein and unique peptide distribution, slightly more spectra were assigned to *B. thetaiotaomicron* relative to *B. WH2* for the majority of methods (exception are the soluble fractions).

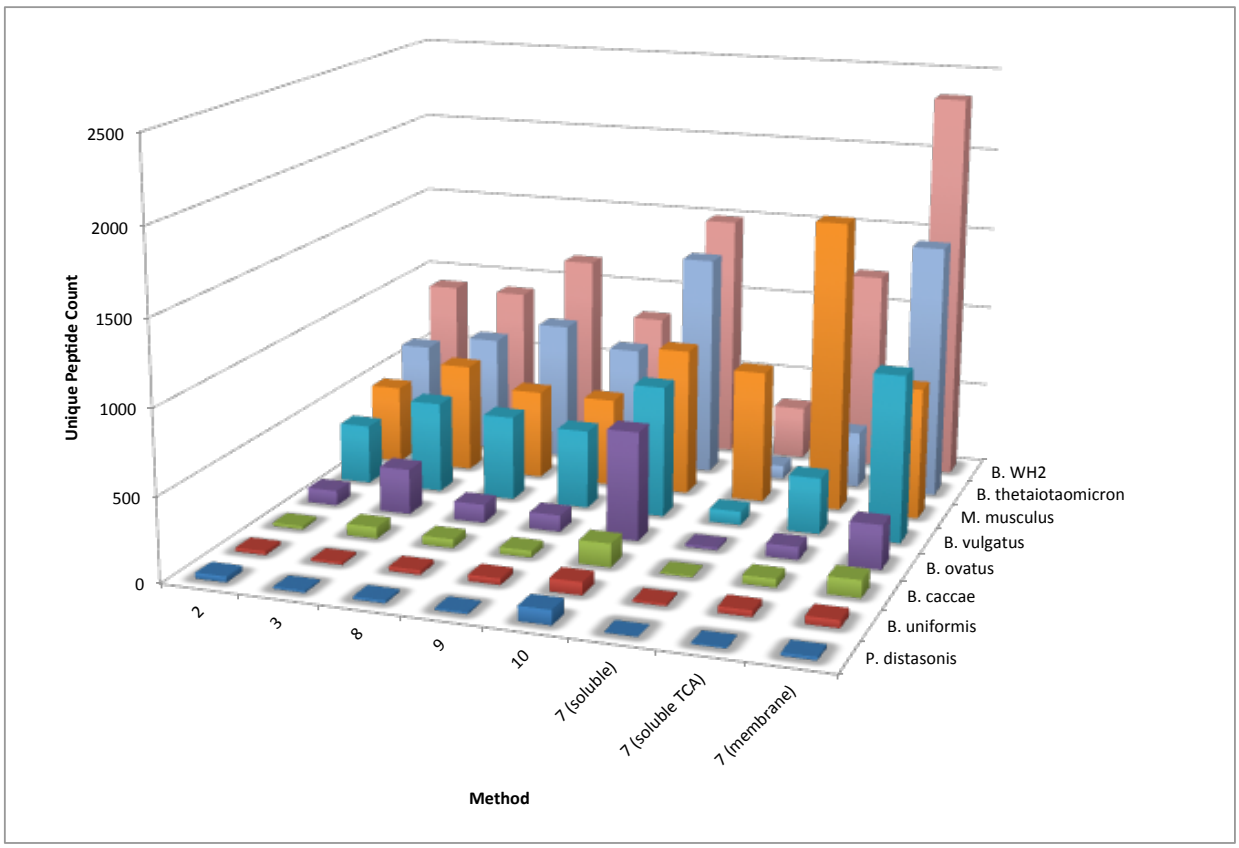


Figure 4.7: Distribution of unique peptide counts for all relevant database components (microbial and host) for all 7 cecal samples.

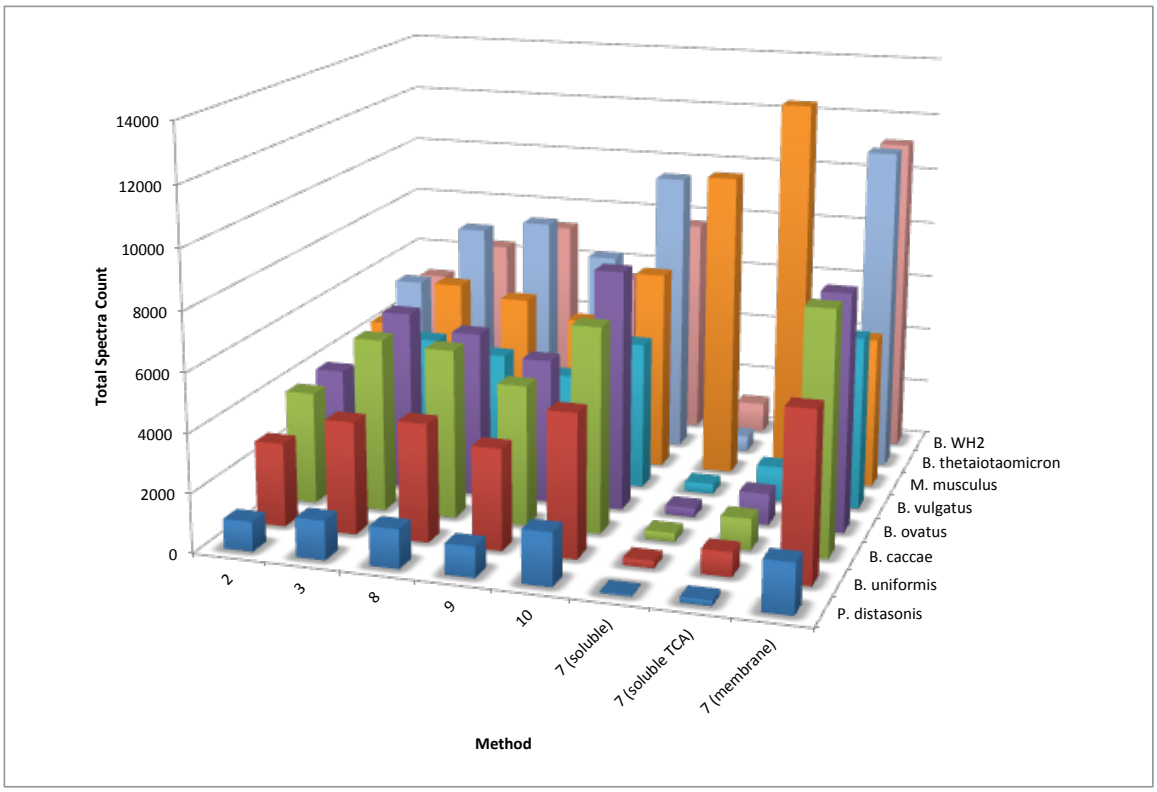


Figure 4.8: Distribution of total assigned spectra counts for all relevant database components (microbial and host) for all 7 cecal samples.

Why the soluble fractions contain significantly more host-related (*M. musculus*) proteins, peptides, and assigned spectra relative to the membrane fraction and other methods (**Figure 4.6-4.8**) is unclear. Although the unique peptide and protein counts agree with the majority of identifications belonging to *B. WH2* and *B. thetaiotaomicron* for most of the methods, we have yet to determine why these two microbes dominate the community proteome relative to the other five species belonging to Bacteroides. It may be that *B. WH2* and *B. thetaiotaomicron* contain significantly more 'unique' peptides in the genome (predicted proteome) relative to the other microbes; hence, more unique peptides were identified for these two species. Alternatively, there may be a biological (non-technical) difference between these two species relative to the other members of the consortium. To assess whether a subset of proteins related to a specific function are enriched in one method compared to the others, we classified the identified proteins in Clusters of Orthologous Groups (COGs). **Figure 4.9** indicated that

we are not enriching for specific functional subsets of proteins with any particular method. However, proteins classified with functions related to translation and metabolic pathways commonly involved in the gastrointestinal tract were highly abundant across all six methods. Therefore, while differences are evident between the types of COGs present in the proteomes, there are insignificant biases in the methods' enriching for specific functional groups of proteins. Although method # 7 identified more proteins relative to the other five methods, **Figure 4.10** (A and B) suggests that methods 7 and 10 provide very similar results in terms of identified protein COGs. In conclusion, while method # 7 has the overall best performance for proteome identification relative to the other methods, method 10 may be equally sufficient in the *in situ* lysis of microbial communities derived from the cecum of mice, with the advantage of less sample handling, fewer surface exposures, less time-consuming, and does not require additional expensive instrumentation (i.e., ultracentrifuge).

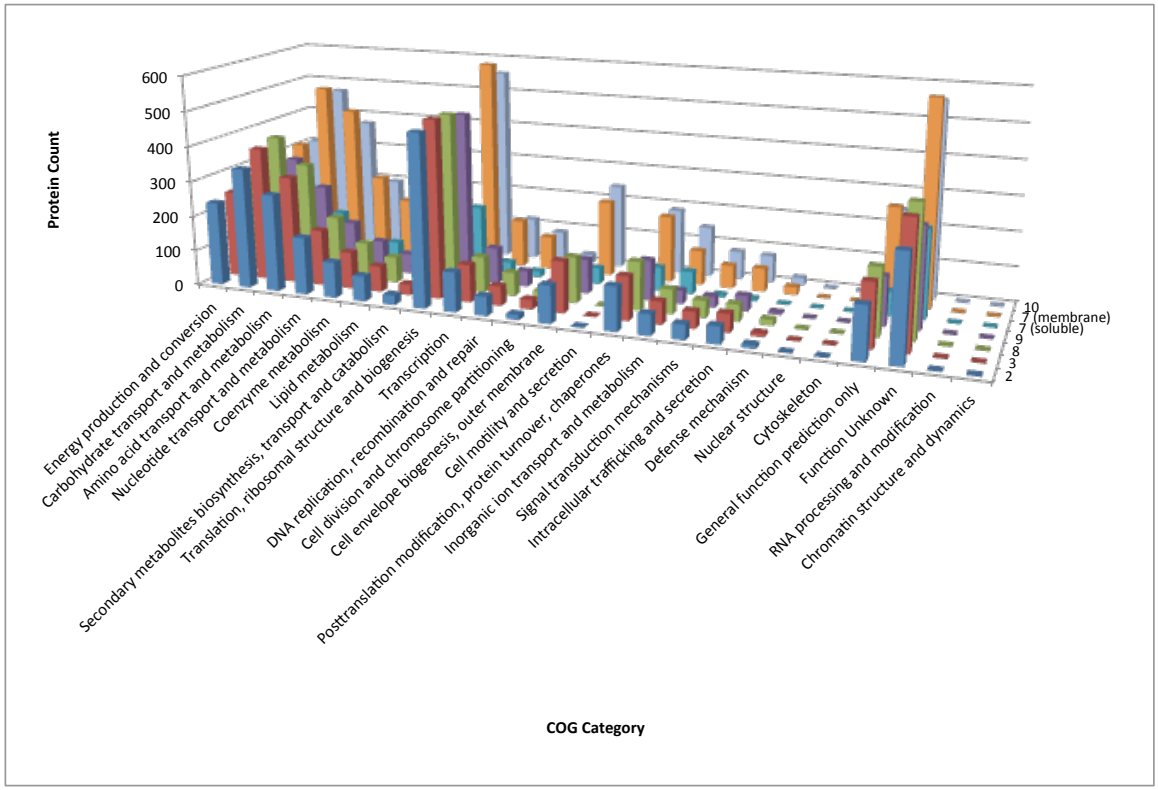


Figure 4.9: COG classification of all identified proteins for all cecal samples.

Figure 4.10: Distribution of identified proteins for (**A**) method 7 (membrane and soluble TCA fraction combined) and (**B**) compared to method 10 based on COG categories.

(A)

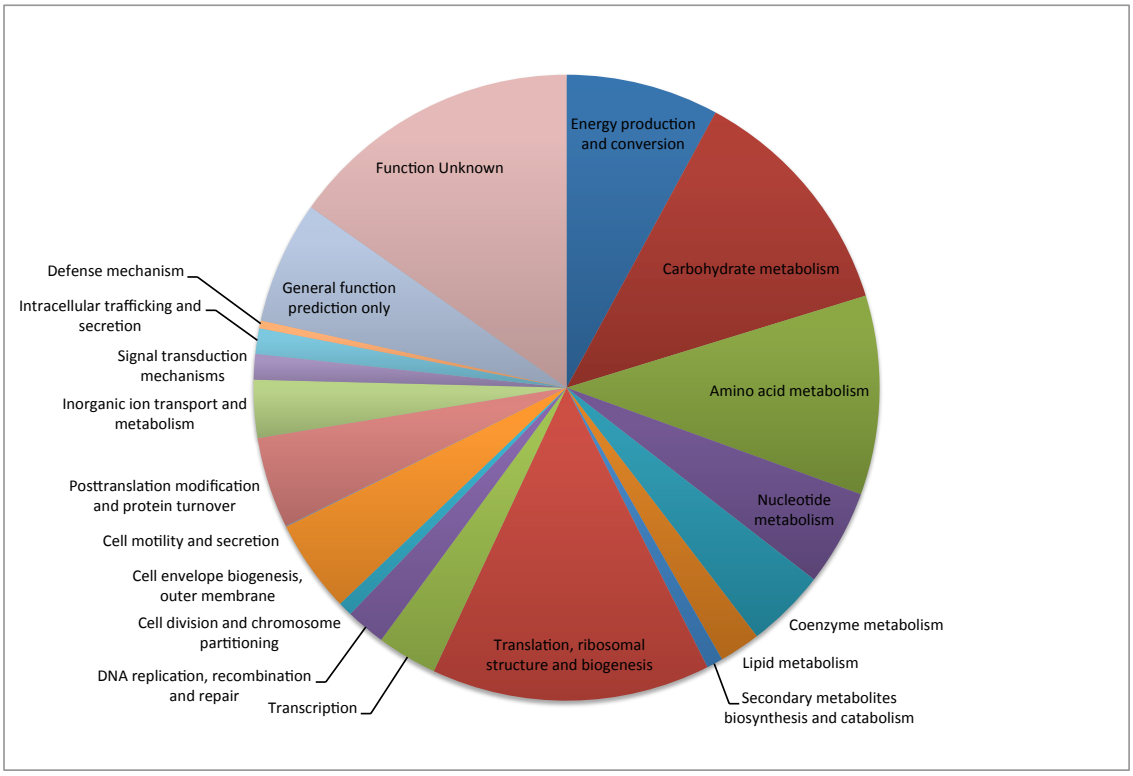
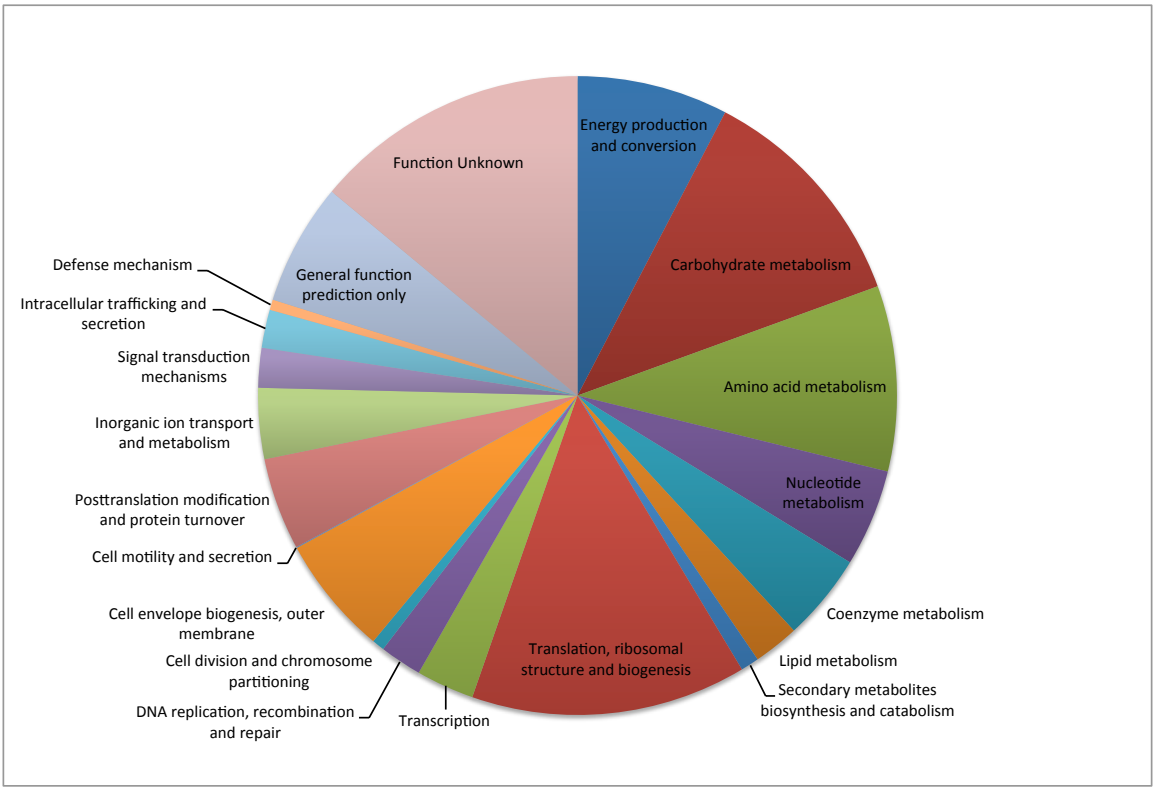


Figure 4.10 (B)



4.4: Conclusions

The method that demonstrated the greatest performance for the lysis and extraction of microbial and host (mouse) proteins from a model 7-member human microbial gut community is bead-beating homogenization, ultracentrifugation, and protein precipitation of the soluble fraction with a total of 4,887 non-redundant proteins. Although this method performed the best based on overall MS-based proteome metrics, the soluble fraction (TCA) only contributed 521 unique proteins. Method #10 (lysis via SDS solubuilization and sonication) followed closely with a total of 4,495 non-redundant proteins and may be more successful for some research groups whom do not have access to an ultracentrifuge.

Based on the host and microbial protein, unique peptide, and spectra count distributions per species, the abundance-levels were not evenly distributed among all seven phylotypes of Bacteroides and the host, but rather skewed with a higher

abundance of *B. WH2* and *B. thetaiotaomicron* in all six methods with the exception of the soluble fraction in method 7. We believe that this is not a reflection of technical or instrumental variation, but is a result of the microbial community dynamics and functional signature differences because the abundance-level trends are very similar across all lysis and protein extraction methods. A classification of all microbial proteins by general function (e.g., COG) suggests that all six methods provide very similar functional trends with translation and metabolic pathways highly abundant across all methods. Therefore, the five different methods described in this study are not biased and do not preferentially lyse one species or functional group (COG) more efficiently relative to others within the 7-member consortium.

Chapter 5

Temporal profiling of a defined 12-member human gut microbial community in gnotobiotic mice in response to changing diets

Alison R. Erickson, Nathan P. McNulty, Nathan C. VerBerkmoes, Jeffrey I. Gordon, Robert L. Hettich

5.1: Introduction

The human gut microbiome is host to a large population of microbes that heavily rely upon the host (human) and diet for maintenance and growth. It is the microbial inhabitants that shape the establishment, diversity, and stability of the host-associated microbial community. While the phylogenetic composition and membership differs between human individuals[62], [112], gut microbiomes are functionally very similar[112]. Although host genetics accounts for a small fraction of the variation observed between human individuals[147] and mice[148], studies have indicated that environmental conditions (e.g., diet) and stochastic factors affect the relative abundances of microbes in the gut community of both humans and mice[149,150,151,152]. As our diets have evolved over time, from a starch-rich diet to high-starch plant and dairy foods)[153], the host and gut microbiota subsequently adapted by competing for dietary substrates. For example, the modern Western diet, low in complex carbohydrates but high in simple sugars and fat, has shown a relatively higher abundance of Firmicutes[131,154] relative to other phyla. As a result, the host selects for simple carbohydrates while the microbes consume and degrade complex polysaccharides (e.g., starch). To understand the interrelationship between the host-gut microbiota and diet, studies have focused on using gnotobiotic mice inoculated with a defined quantity of sequenced human gut bacteria and monitored their response to diet perturbations[155].

We have demonstrated effective proteomic methods for binary microbial communities and optimized the MS-based sample preparation methodology such that we are ready to scale up to more complex, interactive systems. Here, we used high

resolution MS-based proteomics to evaluate the functional differences and estimate the relative abundance of closely-related species within a defined microbial community in gnotobiotic mice in response to diet perturbations. An *in vivo* model microbial community was designed at Wash. Univ. to represent the diversity in the human gut and was comprised of 12 human gut-derived phylotypes belonging to the Firmicutes, Bacteroidetes, and Actinobacteria. The microbial consortium consisted of the same seven phylotypes used in the 7-member consortium (*Bacteroides caccae, B. ovatus, B. uniformis, B. WH2, B. thetaiotaomicron, B. vulgatus*, and *Parabacteroides distasonis*) in addition to four Firmicutes (*Dorea longicatenta, Ruminococcus obeum, Clostridium spiroforme*, and *C. scindens*) and one Actinobacteria (*Collinsella aerofaciens*). The gnotobioic mice and their 12-member consortium were exposed to two diet oscillation perturbations for which several mice initially consumed a high fat and simple sugar diet ('western' diet) and others consumed a standard BK diet. Expression profiling of the 12-member endpoint communities was performed using transcriptomics and MS-based proteomics to explore community robustness and system-wide microbial responses in the context of dietary disturbances. We wanted to gain insight into the communities' establishment, assembly, and adaption prior to and after a change in diet (**Figure 5.1**). Additionally, we wanted to investigate to what extent we can characterize the function of entire defined communities with significant dynamic range between species at the level of both transcription and protein expression.

There are several challenges associated with the 12-member consortium, (i) the Actinobacteria and Firmicutes are not as abundant as the Bacteroides in these samples (0.5-2.0% of the community), which complicates the ability to achieve complete proteome coverage of the lesser abundant species and (ii) an increased number of degenerate peptides (peptides that match to multiple proteins due to sequence overlap) and protein redundancy. The GeneChip expression data was also used to compare the protein abundance data to determine how well these methodologies agree and disagree in their revelation of the microbiotas' functional signature and differences in response to diet.

A challenge in mass spectrometry-based proteomics is the ability to accurately assign a PSM to the protein from which it originated. When studying microbial communities, this assignment is complicated by homology between proteins of different microbial strains/species. In several environmental communities, the majority of identified PSMs are non-unique, and spectral counting may not be sufficient to estimate a protein's abundance and concomitant phylogenetic origin. Thus, with increasing complexity and microbial composition, differentiation of closely-related species compared to species that are more evolutionarily divergent becomes challenging for mass spectrometry. Although proteomics is not the preferred method to evaluate the abundance of species or to differentiate between species, the use of either unique peptide counts or total spectra counts with the total number of identified proteins (per species) have been evaluated for their ability to differentiate and estimate the relative abundance of each species on a proteome level.

5.2: Experimental Methods

Adult germ-free mice were gavaged with: *Bacteroides thetaiotaomicron, B. caccae, B. ovatus, B. uniformis, B. vulgatus, B. WH2, Parabacteroides distasonis, Clostridium scindens, C. spiroforme, Ruminococcus obeum, Dorea longicatena*, and *Collinsella aerofaciens*. Proteins were extracted from the cecum of four gnotobiotic mice fed a diet rich in plant polysaccharides (BK diet) or a 'Western' diet high in fat and sugars.

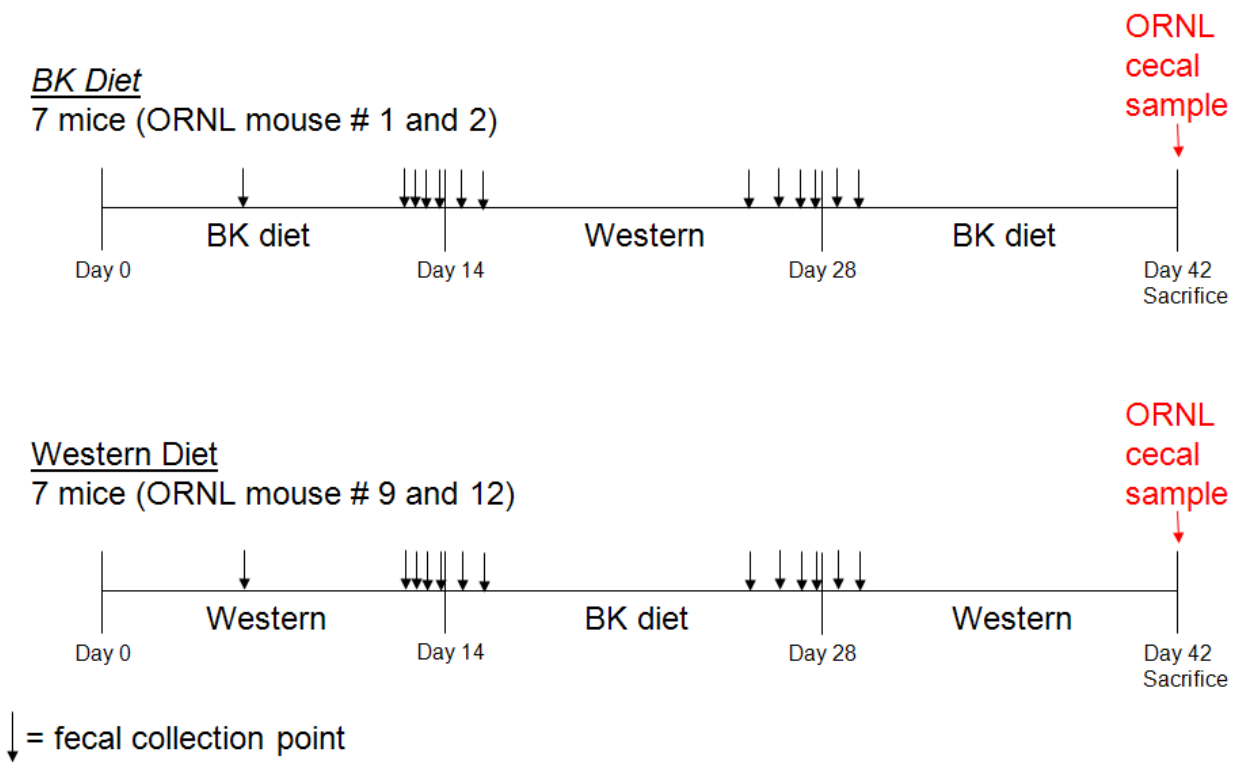


Figure 5.1: Experimental setup and example of sample collection timepoints for host-microbial community diet oscillations. Fecal and cecal collections were used for GeneChip expression experiments. Cecal collections were also used for MS experiments.

5.2.1: Sample preparation

Cecal contents were collected from four mice and analyzed via nano-2D (strong cation exchange – reverse phase)-LC-MS/MS on a hybrid LTQ – Orbitrap Velos mass spectrometer (Thermo Fisher Scientific). Cecal contents were solubilized in 1mL SDS lysis buffer (4% w/v SDS, 100mM Tris•HCl, pH 8.0, 10mM dithiothreitol (DTT)) and lysed mechanically by sonication followed by incubation for 5min at 95°C. Cells were centrifuged at 21,000xg. Following an overnight tricholoacetic acid (TCA) precipitation, the TCA precipitates (protein mixtures) were resolubilized in 500uL of 8M urea, 100mM Tris•HCl, pH 8.0, and reduced by incubation at a final concentration of 10mM DTT for 1 hr at room temperature. Samples were sonicated and an aliquot taken to determine the protein concentration using the widely available bicinchonic acid-(BCA) based protein

assay kit (Pierce). Samples were diluted with 100mM Tris•HCl, 10mM CaCl2, pH 8.0 to a final urea concentration below 4M. Proteolytic digestions were initiated with sequencing grade trypsin (1/100, w/w; Promega) overnight at room temperature. A second aliquot of trypsin was added (1/100) and diluted with 100mM Tris•HCl, pH 8.0 to a final urea concentration below 2M. Following a 4 hr incubation at room temperature, samples were reduced to a final concentration of 10mM DTT. The peptides were acidified (protonated) in 200mM NaCl, 0.1% formic acid, filtered, and concentrated with a 10k molecular weight cutoff spin column (Sartorius).

5.2.2: LC-MS/MS data collection and analyses

Peptide mixtures were desalted and separated utilizing a split phase 2D-LC column (SCX-C_{18}) over a 12-step gradient with 22 hr runs per sample. All MS analyses were performed in positive ion mode. One full MS scan was acquired in the Orbitrap Velos at 30K resolution followed by ten data-dependent MS/MS scans (*m/z* 400-1700) at 35% normalized collision energy with dynamic exclusion enabled at 1. The data was searched using SEQUEST against a database containing the following predicted proteomes to be encoded by the genomes of the 12-member community: *Bacteroides caccae, B. ovatus, B. thetaiotaomicron, B. uniformis, B. vulgatus, B. WH2, Clostridium scindens, C. spiroforme, Collinsella aerofaciens, Dorea longicatena, Parabacteroides distasonis*, and *Ruminococcus obeum* in addition to potential food components (e.g., rice and yeast) and common contaminants (e.g., keratins). Additionally, *Eubacterium rectale, Faecalibacterium prausnitzii* M212 and *R. torques* were included as distractors that were not expected to be present. The SEQUEST settings were the following: enzyme type, trypsin; parent mass tolerance, 3.0; fragment mass tolerance, 0.5; up to 4 missed cleavages, and fully tryptic peptides only. All datasets were filtered with DTASelect, parameters included: Xcorrs of 1.8, 2.5, and 3.5 for singly, doubly, and triply charged precursor ions, minimum deltCN of 0.08, and a minimum requirement of two fully tryptic peptides per protein. An *in silico* tryptic digested protein sequence database was used to generate a theoretical peptidome of unique peptides within a mass range of 600-4,890Da and ≤1 miscleavages.

Spectral normalization by community and individual species was generated with the help of Chongle Pan in which the p value is calculated using the Mann–Whitney U test (non-parametric version of t test). The spectral count difference is the difference between the median spectral counts of the two diets. A protein is labeled as "UP" regulated if the p value is less than 0.05 and the spectral count different is greater than 5 and equally for proteins are labeled as "DOWN."

5.3: Results and Discussion

5.3.1: Overview of proteome metrics

The proteomes displayed high technical reproducibility in terms of total number of assigned spectra per mouse with an $R^2 \geq 0.9883$ for all four mice. The BK-fed mice proteomes were highly correlated with an R^2 of 0.9059 for mouse 1 and 2, as compared to the Western-fed mice with R^2 0.7815 for mouse 9 and 12. This would suggest that the variation in the western fed mice is a reflection of biological variability and is not a result of technical deviation (i.e., mass spectrometer and HPLC effects). On average, a total of 4,827 and 3,251 proteins were identified for the BK diet and Western diet, respectively (**Table 5.1**). Of all proteins identified in the BK-fed mice, 4,220 proteins were found across both mice (1 and 2) and all four MS runs. Of all identified proteins in the Western-fed mice, ~2,658 proteins were identified across both mice (9 and 12) and all four MS runs. However, when both diets are directly compared, only ~1,824 microbial and host proteins were identified across both diets where 2,322 proteins were unique to the BK diet and 805 proteins unique to the western diet (**Figure 5.2**). Upon accumulation of all identified proteins from both diets and all MS runs, ~2-20% of each species' genome was identified (Table 5.2) with the majority of proteins identified from *B. WH2* (20%) and the least from *C. spiroforme* (2%). Although the phylotypes belonging to Bacteroides collectively had higher proteome coverage in terms of non-redundant protein identifications (>11%) compared to the Firmicutes and Actinobacteria in this study (<8%), the range of abundance varies between the two different diets.

Table 5.1: Overall MS metrics for the BK and western diet fed mice.

Diet	Sample	Run	Proteins	Peptides	Spectra	PPMs (%)
BK	1	3	4,778	22,270	47,508	86.92
		4	4,793	22,363	49,454	88.07
	2	3	4,828	21,908	43,413	89.31
		4	4,909	22,245	41,873	89.87
Western	9	4	3,420	17,102	42,979	87.98
		5	3,203	16,035	40,753	88.7
	12	4	3,093	14,703	37,131	88.74
		5	3,286	15,248	34,490	89.15

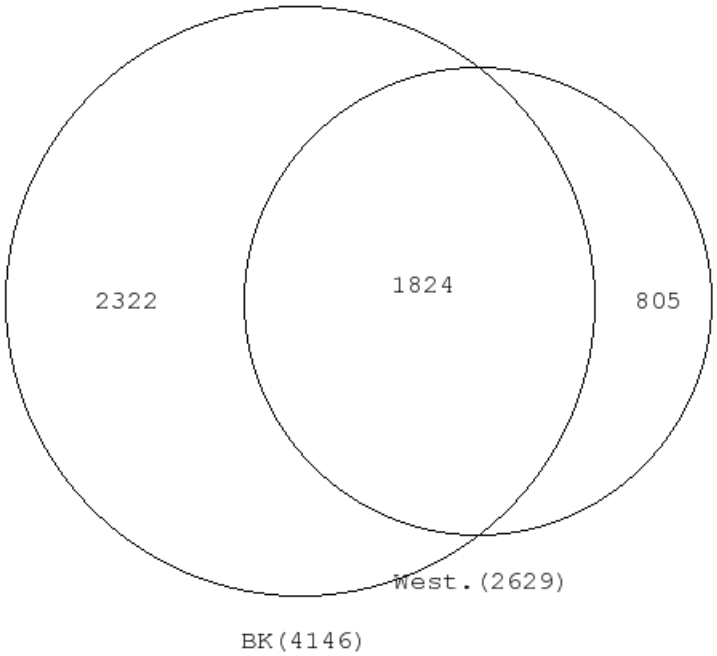


Figure 5.2: Comparison of shared and unique microbial and host identified proteins across both diets with proteins unique to the BK diet (left; 2,322 proteins), shared proteins (center; 1,824 proteins) and proteins unique to the Western diet (right; 805 proteins).

Table 5.2: Relevant database components and % of each species' genome identified via MS-based proteomics.

Relevant Database Components			Total Protein Identifcations				Community-wide analysis and significantly differential proteins with preference for one diet					
							BK fed mice			Western diet fed mice		
Genome	Acroynmn	Predicted Proteins	Identified Proteins	% of Predicted Proteome Identified	Proteins identified in BK diet	Proteins identified in Western diet	# of statistically differential proteins	% of Total Identified Proteome per Species	% of Total BK-Identified Proteome per Species	# of statistically differential proteins	% of Total Identified Proteome per Species	% of Total Western-Identified Proteome per Species
Bacteroides caccae	BACCAC	3855	669	17.35%	594	362	148	22%	25%	65	10%	18%
Bacteroides ovatus	BACOVA	5536	693	12.52%	667	277	176	25%	26%	16	2%	6%
Bacteroides thetaiotaomicron	BACTHE	4778	722	15.11%	660	343	164	23%	25%	52	7%	15%
Bacteroides uniformis	BACUNI	4663	563	12.07%	550	204	160	28%	29%	9	2%	4%
Bacteroides vulgatus	BACVUL	4065	783	19.26%	755	357	244	31%	32%	20	3%	6%
Bacteroides WH2	BACWH2	5244	1025	19.55%	996	353	339	33%	34%	24	2%	7%
Clostridium scindens	CLOSCI	3995	220	5.51%	105	203	1	0%	1%	82	37%	40%
Clostridium spiroforme	CLOSPI	2465	39	1.58%	18	32	0	0%	0%	4	10%	13%
Collinsella aerofaciens	COLAER	2367	154	6.51%	124	101	2	1%	2%	6	4%	6%
Dorea longicatena	DORLON	2970	114	3.84%	76	91	0	0%	0%	22	19%	24%
Parabacteroides distasonis	PARDIS	3850	414	10.75%	405	118	87	21%	21%	4	1%	3%
Ruminococcus obeum	RUMOBE	4175	325	7.78%	291	231	7	2%	2%	12	4%	5%
Mus musculus	Mus	34966	881	2.52%	575	698	41	5%	7%	203	23%	29%

5.3.2: Peptidome Comparisons

Several quantitative metrics were evaluated for their ability to differentiate and estimate the abundance of closely-related and divergent microbial species within a defined human gut consortium. Three quantitative methods that use spectra counts only, spectra and protein counts, or unique peptides to estimate species abundance were compared to GeneChip expression data from the same samples. All three quantitative methods suggest the Bacteroides phylotypes are significantly more abundant in mice consuming the plant polysaccharide-rich diet. More specifically, the unique peptide quantitative metric indicated that *B. caccae* was significantly more abundant in the western diet, in agreement with the GeneChip expression data. A phylotype's estimated abundance based on unique identified peptides, however, may be affected by the proportion of unique predicted peptides assignable to that phylotype in the database.

A method that estimates abundance via identified unique peptides may be biased and skewed by the degree of predicted unique peptides in the sequence database. For example, more unique peptides belonging to *Bacteroides WH2* may be present because there are more unique peptides in the database relative to the other species. To address this potential bias, a "theoretical peptidome" was created for each species (12 relevant species and 3 unrelated distractor species) in the sequence database with an *in silico* peptide digest that takes into account (i) tryptic miscleavages (0-4) and (ii) the standard peptide mass range that can be experimentally detected. A "theoretical peptidome" provides the ability to compare the *identified* peptidome (both unique and non-unique peptides per species) to the *predicted* (theoretical) peptidome to determine what percentage of each species is unique relative to all of the other species within the entire consortium. By comparing the percentage of each species' predicted unique peptidome to the identified peptidome, you could determine whether (i) the closely-related species (i.e., Bacteroides) are contributing any unique peptides to the entire predicted peptidome and (ii) whether the identified unique peptides are following the same trends as the predicted peptidome.

To generate a theoretical peptidome, several parameters (miscleavages and peptide mass range) were first compared for the identified peptide results prior to their application of the sequence database (15 microbial genomes, mouse, rice and yeast genomes, and common contaminants) used in this study. The majority of experimentally identified peptide sequences fell within 0-1 miscleavages (94%; **Table 5.3**) and a mass range of 600-4,890 Da. Based on this data, the theoretical peptidome was generated *in silico* with the same filters for the entire protein sequence database at ≤1 miscleaveage.

Table 5.3: Proportion of total non-redundant identified peptides for all mice with 0, 1, 2, 3 or 4 miscleavages.

# of Miscleavage(s)	# of Non-redundant Peptides	%
0	137,331	65.25%
1	60,235	28.62%
2	10,716	5.09%
3	1,208	0.57%
4	990	0.47%
Total Peptides:	210,480	

Although the mouse and diet genomes contributed the majority of unique predicted peptides (20.44%-34.57%), all twelve phylotypes (including closely-related species) contributed comparable percentages of unique peptides (~2-4%) to the entire unique peptidome (**Table 5.4**) with *B. WH2* containing the majority of unique predicted peptides of the human-derived bacteria. Within each individual species, ~61-89% of all predicted peptides are classified as “unique” (**Table 5.4**). This would suggest that although there is significant genome sequence overlap, this is does not equate on a peptide level since the majority of tryptic peptides are unique per species. Significantly, 89% of all peptides belonging to the single species (*C. aerofaciens*) of Actinobacteria are unique, whereas only ~61-79% of peptides belonging to the majority of Bacteroides are unique within this protein sequence database.

Table 5.4: Distribution of total predicted peptides (unique and non-unique) for the protein sequence database with ≤1 miscleavage.

Sequence Database: 1 Miscleavage; Peptide Mass Range: 600-4,890 Da										
Species	**Non-Unique**	**% Non-Unique**	**Unique**	**% Unique**	**Total**	**% Total of Total DB peptides**	**% Unique of Total DB peptides**	**% Unique of Total DB unique peptides**	**#Proteins**	**Size (MB)**
Rice	2,733,343	56.35%	2,117,393	43.65%	4,850,736	44.86%	19.58%	34.57%	66,710	36.00
Mouse	1,174,515	48.40%	1,252,078	51.60%	2,426,593	22.44%	11.58%	20.44%	34,966	19.20
Yeast	64,216	13.40%	414,911	86.60%	479,127	4.43%	3.84%	6.77%	6,345	3.33
B. WH2	71,238	20.96%	268,594	79.04%	339,832	3.14%	2.48%	4.39%	5,244	2.66
B. ovatus	105,322	33.19%	211,964	66.81%	317,286	2.93%	1.96%	3.46%	5,536	2.54
P. distasonis	29,795	12.83%	202,464	87.17%	232,259	2.15%	1.87%	3.31%	3,850	1.86
B. thetaiotaomicron	101,982	34.06%	197,434	65.94%	299,416	2.77%	1.83%	3.22%	4,778	2.38
B. vulgatus	58,625	23.66%	189,181	76.34%	247,806	2.29%	1.75%	3.09%	4,065	1.93
B. uniformis	70,619	30.25%	162,806	69.75%	233,425	2.16%	1.51%	2.66%	4,663	1.95
R. obeum	26,276	14.96%	149,371	85.04%	175,647	1.62%	1.38%	2.44%	4,175	1.54
C. scindens	32,174	18.16%	144,987	81.84%	177,161	1.64%	1.34%	2.37%	3,995	1.53
E. rectale	26,300	15.80%	140,179	84.20%	166,479	1.54%	1.30%	2.29%	3,631	1.42
B. caccae	87,509	39.39%	134,675	60.61%	222,184	2.05%	1.25%	2.20%	3,855	1.77
F. prausnitzii M212	20,667	14.55%	121,341	85.45%	142,008	1.31%	1.12%	1.98%	3,493	1.37
D. longicatena	29,210	20.72%	111,738	79.28%	140,948	1.30%	1.03%	1.82%	2,970	1.18
R. torques	25,607	18.95%	109,523	81.05%	135,130	1.25%	1.01%	1.79%	2,875	1.12
C. spiroforme	18,658	15.44%	102,145	84.56%	120,803	1.12%	0.94%	1.67%	2,465	1.02
C. aerofaciens	11,628	11.11%	93,021	88.89%	104,649	0.97%	0.86%	1.52%	2,367	1.00
Contaminants	1,026	43.24%	1,347	56.76%	2,373	0.02%	0.01%	0.02%	36	0.02

While *B. WH2, B. ovatus, P. distasonis*, and *B. thetaiotaomicron* are the most abundant phylotypes based on the representation of unique predicted peptides in the theoretical peptidome, experimental phylotype abundance estimated using unique identified peptides did not follow this same trend (**Figure 5.3**). Instead, *B. WH2, B. vulgatus, R. obeum, B. caccae*, and *C. scindens* are most abundant in the 'Western' diet. Although there is significant sequence overlap in the microbiota, less sequence overlap exists on a peptidome level, suggesting that a unique peptide quantitative metric, rather than spectra counts, can be used to quantitate the relative abundance of species and proteins in microbial communities with varying ranges of diversity. In addition, the expression data, provided by Jeff Gordon's group, indicates that *B. ovatus* and *B. WH2* have a strong preference for the BK diet while *B. caccae* prefers the western diet. Unlike the spectra and protein counts metric, the unique peptide counts' distribution strongly supports the genechip expression data.

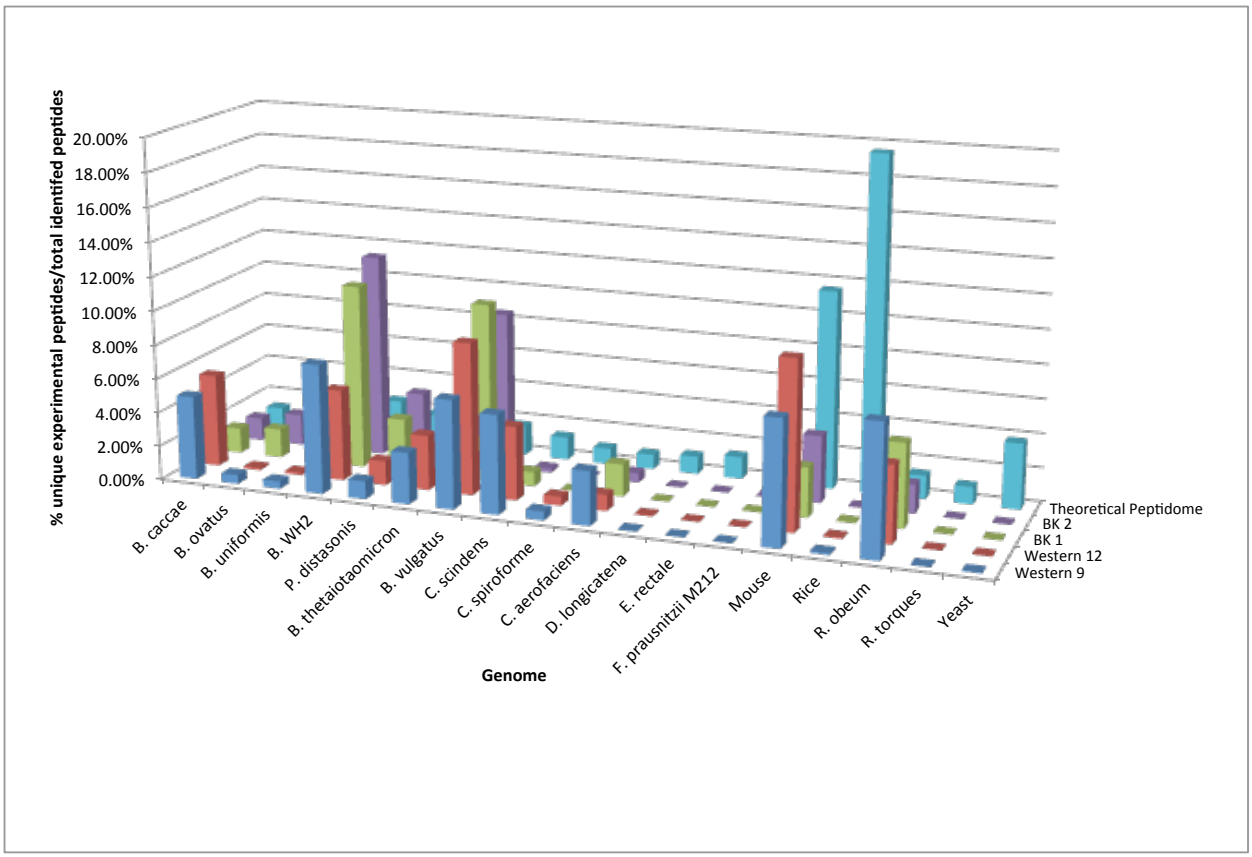


Figure 5.3: Theoretical and experimental unique peptidome comparison per database component (genome) for the BK and Western-fed mice. The % of unique peptides (predicted or identified) out of the total (unique and non-unique) peptides plotted for each database component.

Finally, the predicted unique peptidome was compared to the experimentally identified unique peptidome distribution to identify which microbes are hurt most by proteome overlap with the other community members. Although *B. WH2* is the most abundant species followed by *B. ovatus, B. thetaiotaomicron*, and *B. vulgatus* based on predicted unique peptides, the western diet fed mice did not follow this same trend. Comparing the ratio of identified unique peptides to total predicted unique peptides per species, *B. caccae* and *C. scindens* continue to be the most abundant microbes in the western diet fed mice and are more abundant than in the BK fed mice. Based on these results, our unique peptide data is not skewed by the sequence database and this

‘theoretical peptidome’ provides additional support for evaluating the relative distribution and abundance of species using MS proteomic data.

5.3.3: Community-wide functional comparisons

To understand to what extent we can monitor shifts in the community proteome in response to a diet change, we began by looking at the community as a one functional entity. Within the community, all of the Bacteroides phylotypes are more abundant in the BK diet compared to the Western diet, with the exception of *B. caccae,* which is approximately proportional across all samples and runs, using the sum of normalized spectra counts (**Figure 5.4**). *C. scindens*, on the other hand, is more abundant in the Western-fed mice. This would suggest that although some species, e.g., *C. scindens* are at very low abundance based on total proteome coverage, they still contribute significantly to the pool of proteins and shift in response to diet. In addition, several statistically differential abundant proteins were identified with higher abundance (preference) for one diet relative to the other. Out of all the proteins identified within the 12-member consortium, ~21-33% of the Bacteroides phylotypes’ identified proteins were identified with higher abundance in the BK diet relative to the western diet (p-value ≤0.03). Of the identified BK diet-only proteomes, *B. WH2* and *B. vulgatus* expressed the majority of differentially abundant proteins with a preference for the BK diet (34% and 32%, respectively). On the other hand, ~40% and 29% of all proteins identified as derived from *C. scindens* and *Mus musculus*, respectively, were significantly more abundant in the Western-fed mice (**Table 5.2**) compared to the BK diet (1% and 5%, respectively).

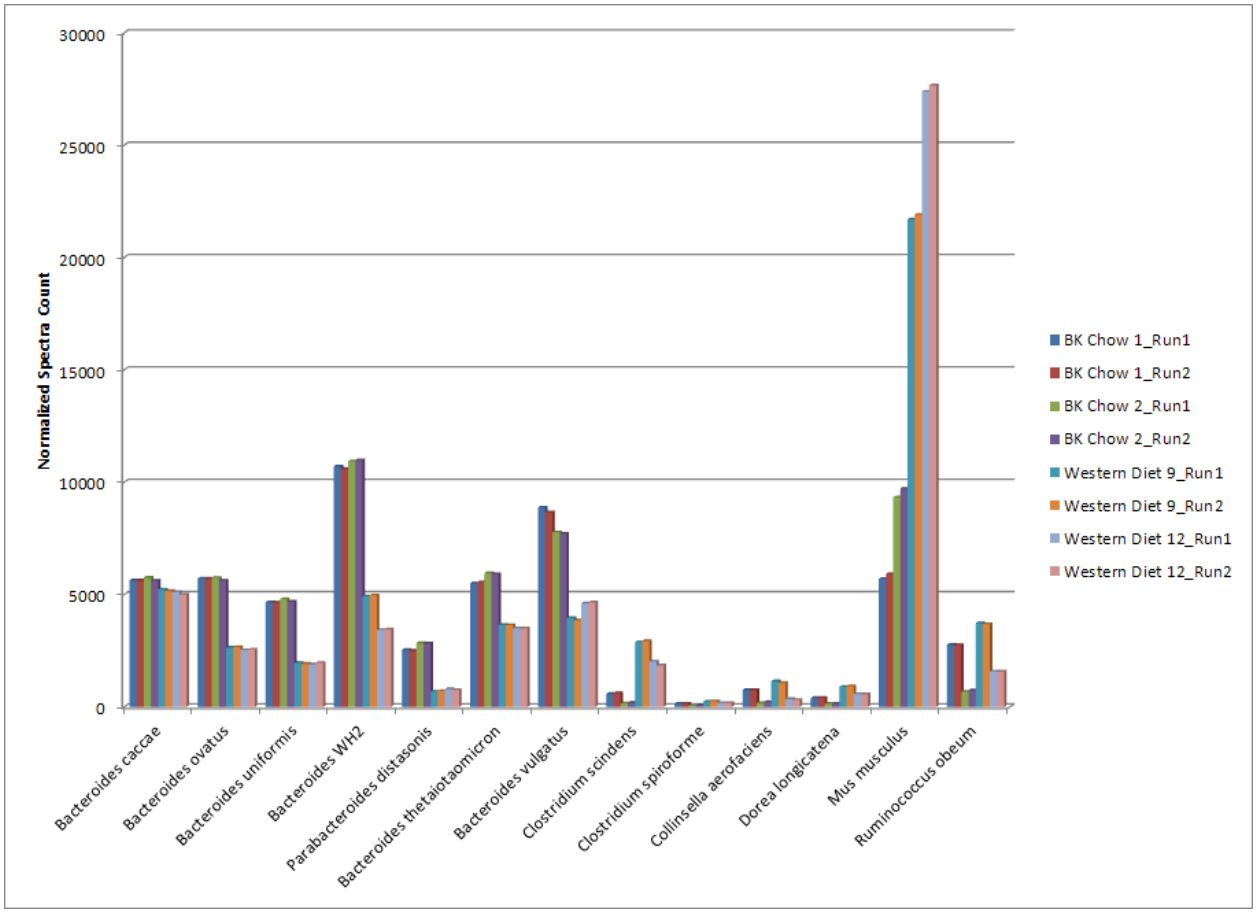


Figure 5.4: Community-wide normalized spectra counts for the host and relevant microbial proteins per MS run for both diets.

Classification of all identified microbial proteins by Clusters of Orthologous Groups (COG) for the two diets indicate that proteins involved in critical gut-associated functions, e.g., carbohydrate metabolism and energy production, are highly abundant across all mice regardless of diet based on normalized spectra counts (technical replicates average and biological replicates summed) (**Figure 5.5**). This would suggest that the microbial community members are actively working together to carryout vital metabolic functions necessary for host-microbiota gut homeostasis. Interesting, many poorly characterized proteins (proteins with unknown function) are highly abundant across both communities and diets, with ~800 and 400 hypothetical proteins identified in the BK and Western-fed mice, respectively, suggesting that the gut microbiota encode beneficial, yet many unknown functions across both communities. On the other hand,

translation was less represented in the Western-fed mice and protein functions related to cellular processes that include DNA replication, chromatin and nuclear structure, and RNA processing were not active in either diet.

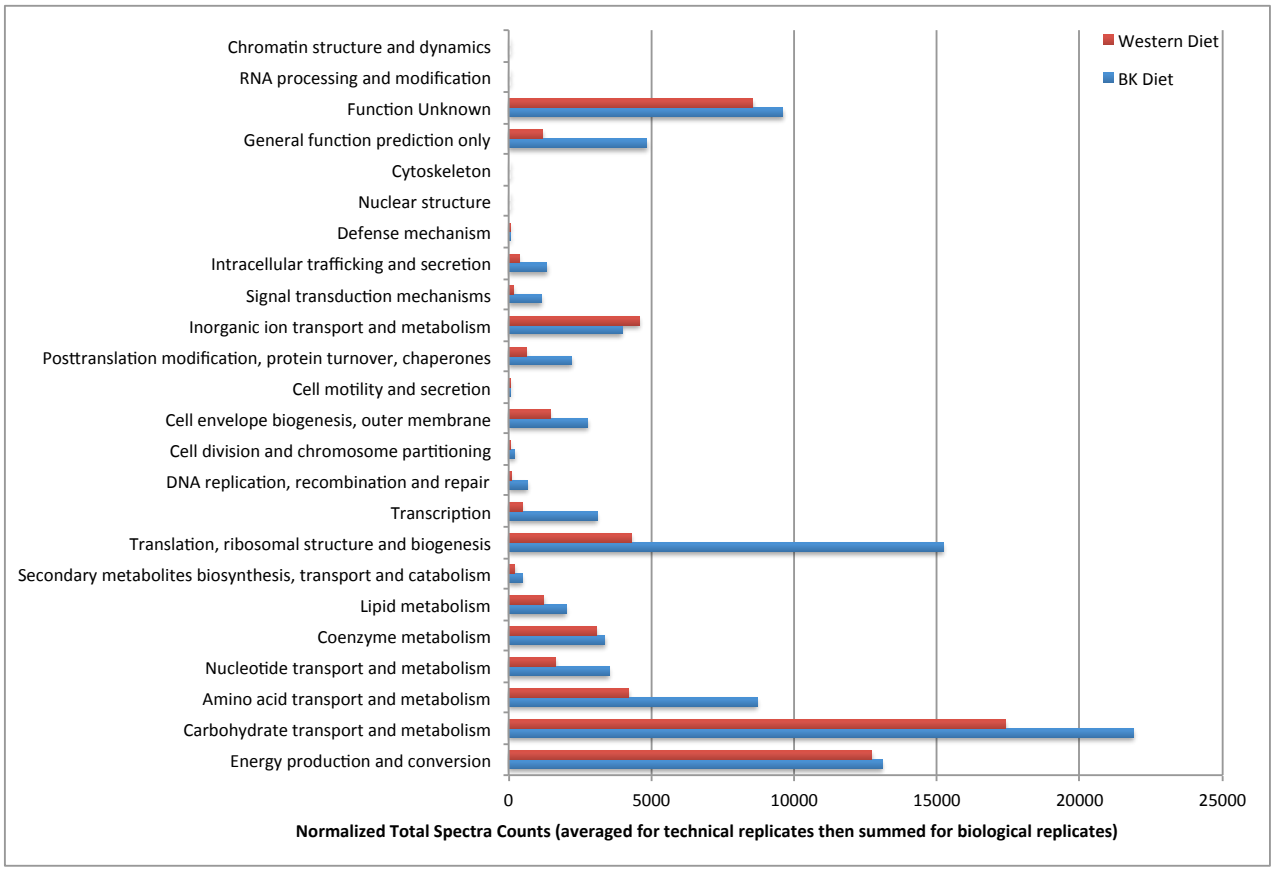


Figure 5.5: Classification of all differentially abundant proteins by COGs for both diets.

The classification of differentially abundant proteins by COGs suggests that many of the over-abundant proteins are also involved in similar key functions in the BK diet (carbohydrate and amino acid metabolism, and energy production (**Figure 5.6**)); however, translation-related proteins are statistically less abundant (down-regulated) in the Western-diet relative to the BK diet (**Figure 5.7**). While translation is present and active across all mice, many essential proteins associated with mRNA translation and its machinery: aminoacyl tRNA synthetases, ribosomes, and translation initiation and elongation factors are less abundant in the Western diet. Of the 13 over-abundant aminoacyl tRNA synthetases identified, all 13 are derived from phylotypes belonging to

Bacteroides. As a result of dietary protein deficiency or diet disturbances, protein synthesis by Bacteroides phylotypes may be suppressed in the western diet because they are growing slower since the diet does not contain the same level of nutrients that are needed by these species. On the contrary, of the 19 translation-associated proteins that are over-abundant in the Western diet, the majority (16) belong to *C. scindens*. Of the 332 microbial proteins identified with over-abundance in the Western diet, 16% (82 proteins) are derived from *C. scindens* and 39% from *M. musculus*. These findings suggest that the *C. scindens* and the host may be acting to provide one or several unique protein functions in the Western-fed mice that are not acquired or significantly active in the microbiota of the BK-fed mice. Further investigation of these individual microbes (*B. WH2* and *C. scindens*), their differences, and general role in the community is provided under 'species-level functional comparisons.'

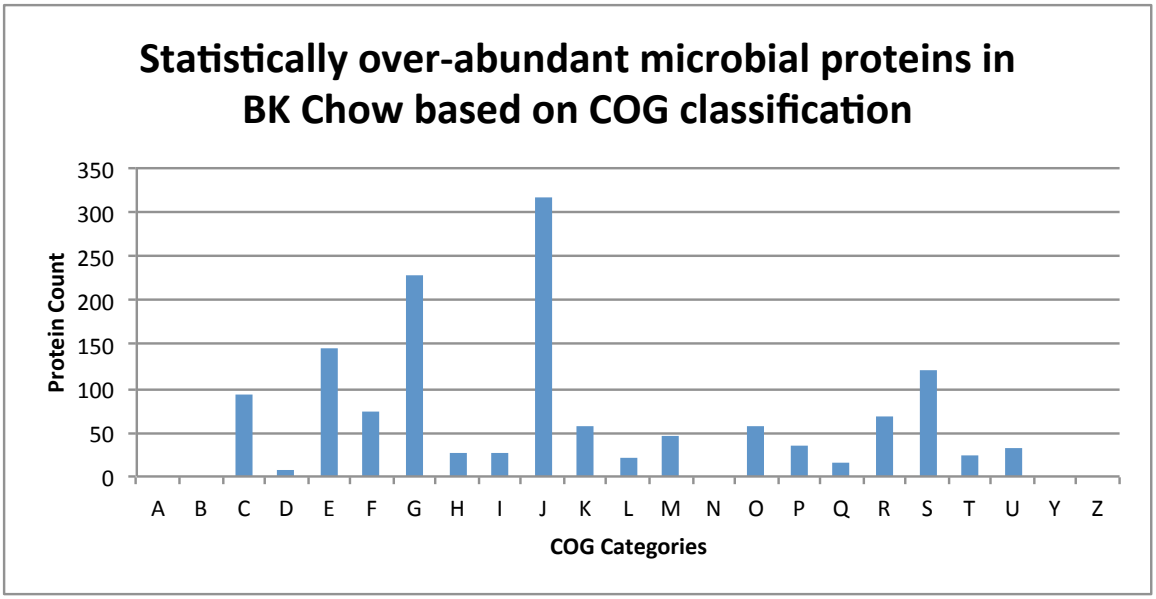


Figure 5.6: Statistically over-abundant proteins based on COG assignments in the BK diet relative to the Western diet.

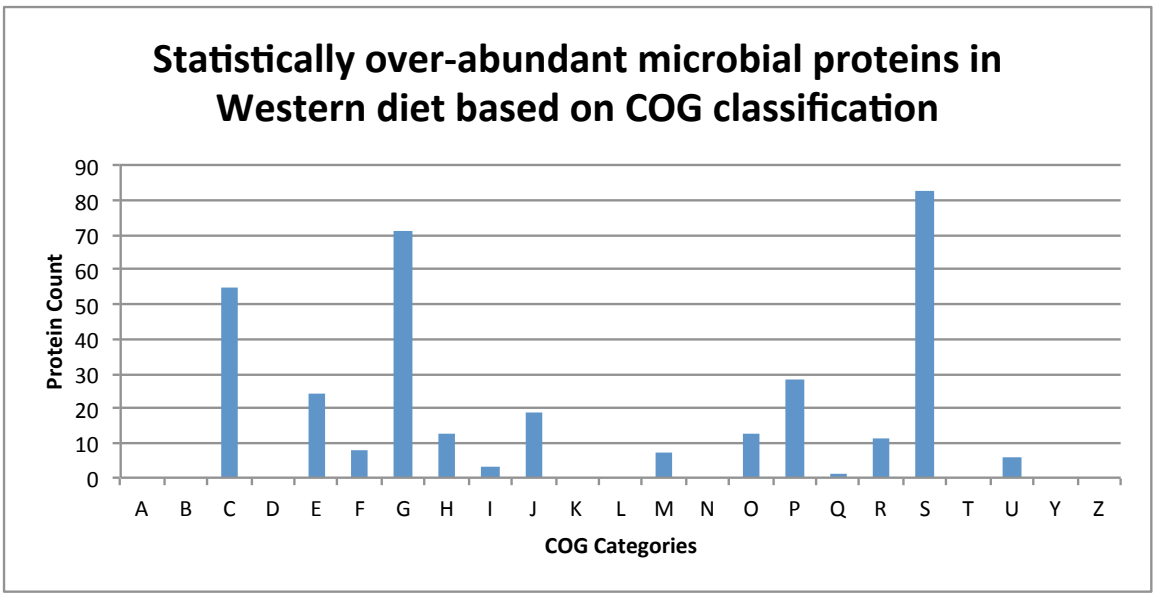


Figure 5.7: Statistically over-abundant proteins based on COG assignments in the Western diet relative to the BK diet.

While a similar number of proteins were identified and classified as 'inorganic ion metabolism' for both diets (225 and 144 proteins for BK and Western diet, respectively), the abundance of these proteins varies with higher abundance in the Western-fed mice. The majority of proteins classified as involved in inorganic ion metabolism belongs to phylotypes of Bacteroides and serve as outer membrane receptor proteins and superoxide dismutase. Of those with significantly higher abundance for the Western diet were outer membrane receptors of *B. caccae* with a preference for iron, ferrienterochelin and colicins. On the contrary, *B. WH2* identified outer membrane receptors associated with iron transport were significantly more abundant in the BK-fed mice. Although inorganic ion transport and metabolism is active and represented across both communities, these results may suggest that *B. caccae* and *B. WH2* may be overcompensating or primarily responsible for iron transport relative to the other Bacteroides phylotypes in the Western and BK-fed mice, respectively.

Due to protein sequence redundancy between closely-related species, statistically high abundance microbial proteins were clustered (UCLUST v2.1.) via sequence identity to group proteins with ≥97% sequence identity. Following clustering

analysis of the statistically differentiated microbial proteins, the total number of over-abundant proteins in the BK fed mice reduced from 1,330 proteins to 1,092 protein clusters and from 332 proteins to 303 clusters in the Western fed mice. Many of these protein clusters were related to similar functions described previously confirming a lack of translation-related protein abundance when exposed to a western fed diet relative to a high protein BK diet.

5.3.4: Species-level functional comparisons

While the community-wide analyses provide a broad functional understanding of the microbiota collectively, in order to directly compare the abundance levels of specific proteins of interest, it was necessary to evaluate the proteomes of each microbe within the 12-member community individually. In addition, by using a species-level approach, we can determine whether one species up- or down-regulates gene(s) in a specific pathway in response to the host's diet.

Based on the community-wide proteome analyses (**Figure 5.4** and **Figure 5.8**) and GeneChip expression data, *B. WH2* has a strong preference for the BK diet. However, why such significant abundance differences exist and the significant functional role of *B. WH2* in the BK-fed mice relative to the other microbial proteomes is unclear. Of the 1,025 *B. WH2* identified proteins, 269 were statistically differentiated, with 199 significantly more abundant in the BK-fed mice. The majority of these proteins (199) were classified with functions related to carbohydrate metabolism, amino acid metabolism, translation or hypothetical proteins. On the other hand, the *B. WH2* proteins that had a preference (up-regulated) for the Western diet were related to inorganic ion transport, coenzyme transport, carbohydrate metabolism, and hypothetical proteins. Coenzyme metabolism related proteins included an outer membrane cobalamin receptor, 7-keto-8-aminopelargonate synthetase, and phosphoserine aminotransferase that were significantly more abundant in the Western-fed mice (p value<0.03 and difference of 52-84 spectra). Inorganic ion metabolism related proteins include superoxide dismutase and outer membrane receptors for ferrienterochelin and colicins (p value<0.03 and difference of 17-280 spectra).

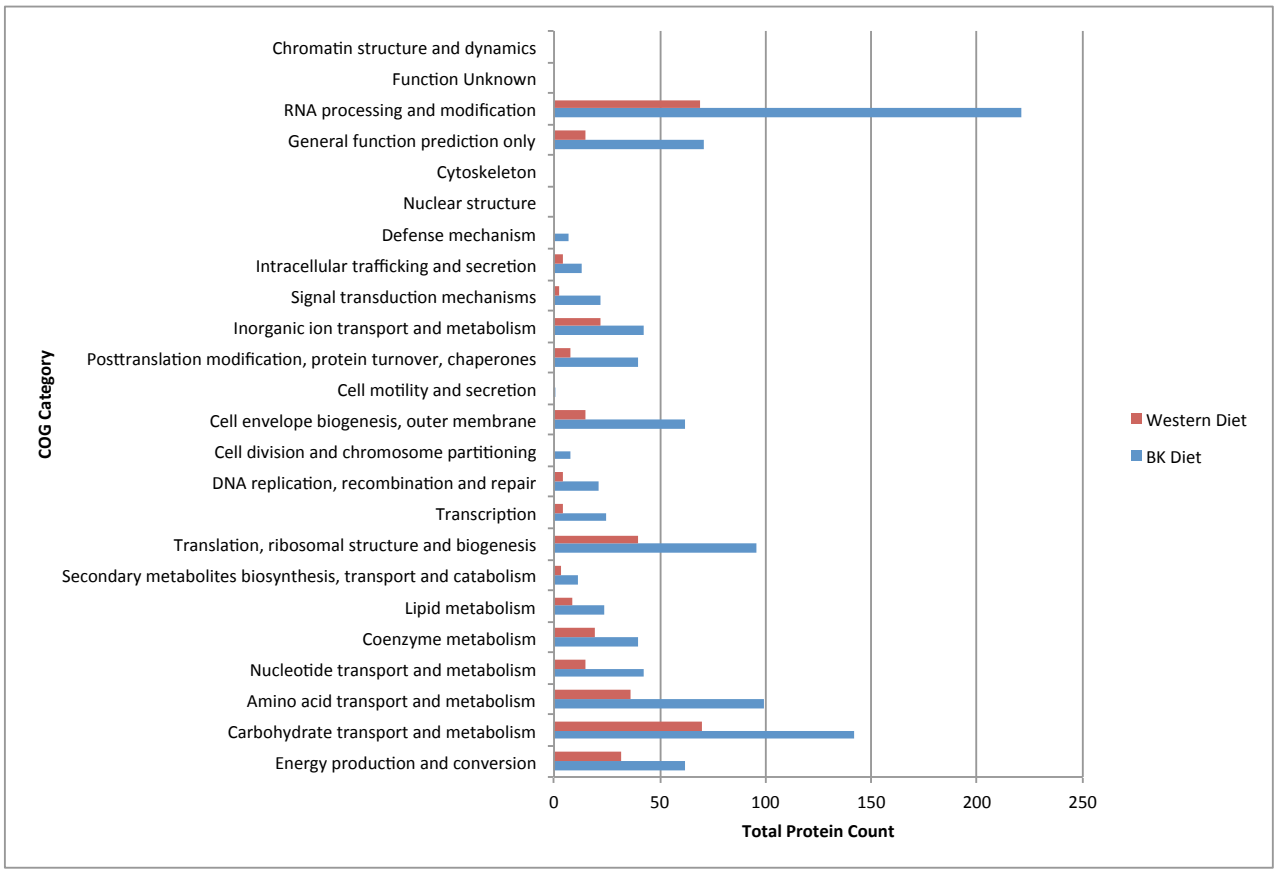


Figure 5.8: *B. WH2* (only) total identified proteins' distribution based on COG categories for both diets.

Due to the significant abundance of *B. WH2* relative to the other 12 phylotypes (**Figure 5.9** and **Figure 5.10**), is *B. WH2* performing differently from the other Bacteroides that might explain its dominance? After comparing each of the Bacteriodes phylotypes individually for each diet, carbohydrate metabolism, outer membrane-related proteins, and signal transduction were identified as being dominated by *B. WH2* proteins while, for example, translation is relatively proportional amongst all of the phylotypes (**Figure 5.10**). The proteomic data may suggest that *B. WH2* expresses a larger proportion of proteins that are involved in carbohydrate metabolism and signal transduction relative to the other Bacteroides phylotypes in the BK diet. In addition, a significant portion of B. WH2 identified proteins were annotated as 'unknown function.' In the Western-fed mice, *B. WH2* identified proteins are not as functionally different from the other Bacteroides phylotypes (**Figure 5.11** and **Figure 5.12**) compared to the BK-

fed mice. This may suggest that most members of the Bacteroides phylotypes are able to cooperate and function individually with respect to the host and Western diet whereas the BK-fed mice are heavily dependent upon *B. WH2* to process certain biological functions. Further investigation (ie, KEGG pathway analysis) would help reveal whether *B. WH2* is responsible for processing specific pathways within these broad functional terms relative to the other phylotypes.

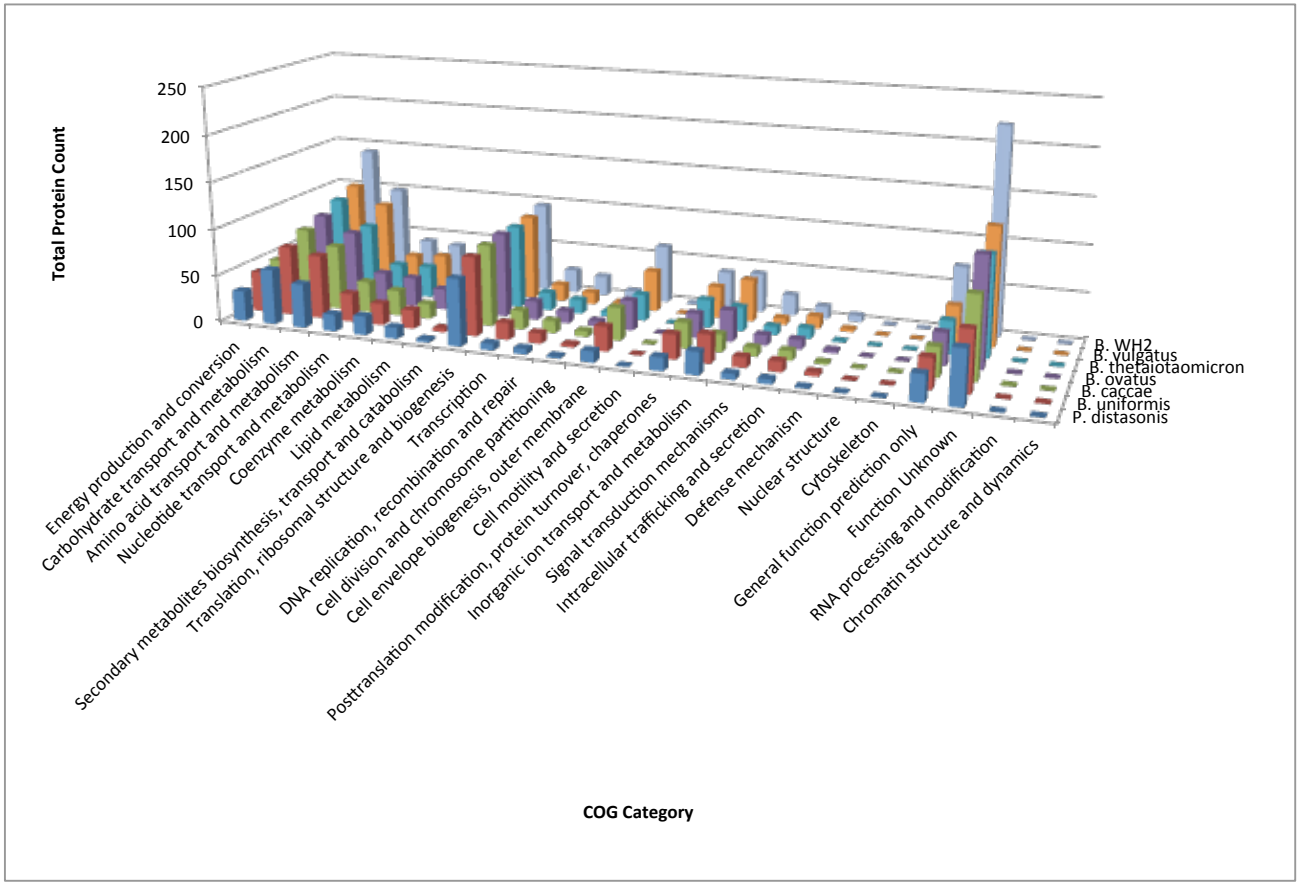


Figure 5.9: Distribution of protein counts per COG category for each phylotype of Bacteroides for the BK-fed mice.

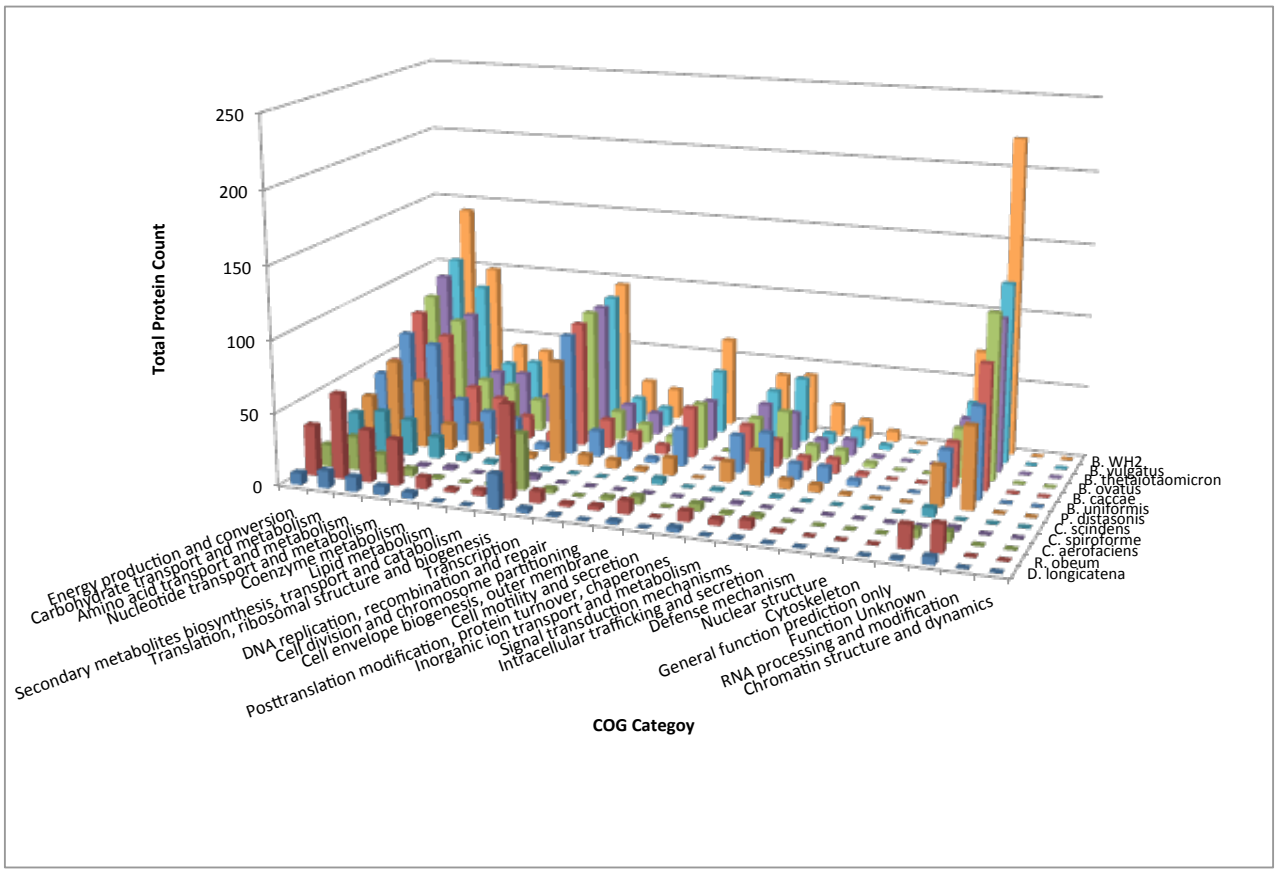


Figure 5.10: Distribution of identified protein counts per COG category for all 12 phylotypes in the BK-fed mice.

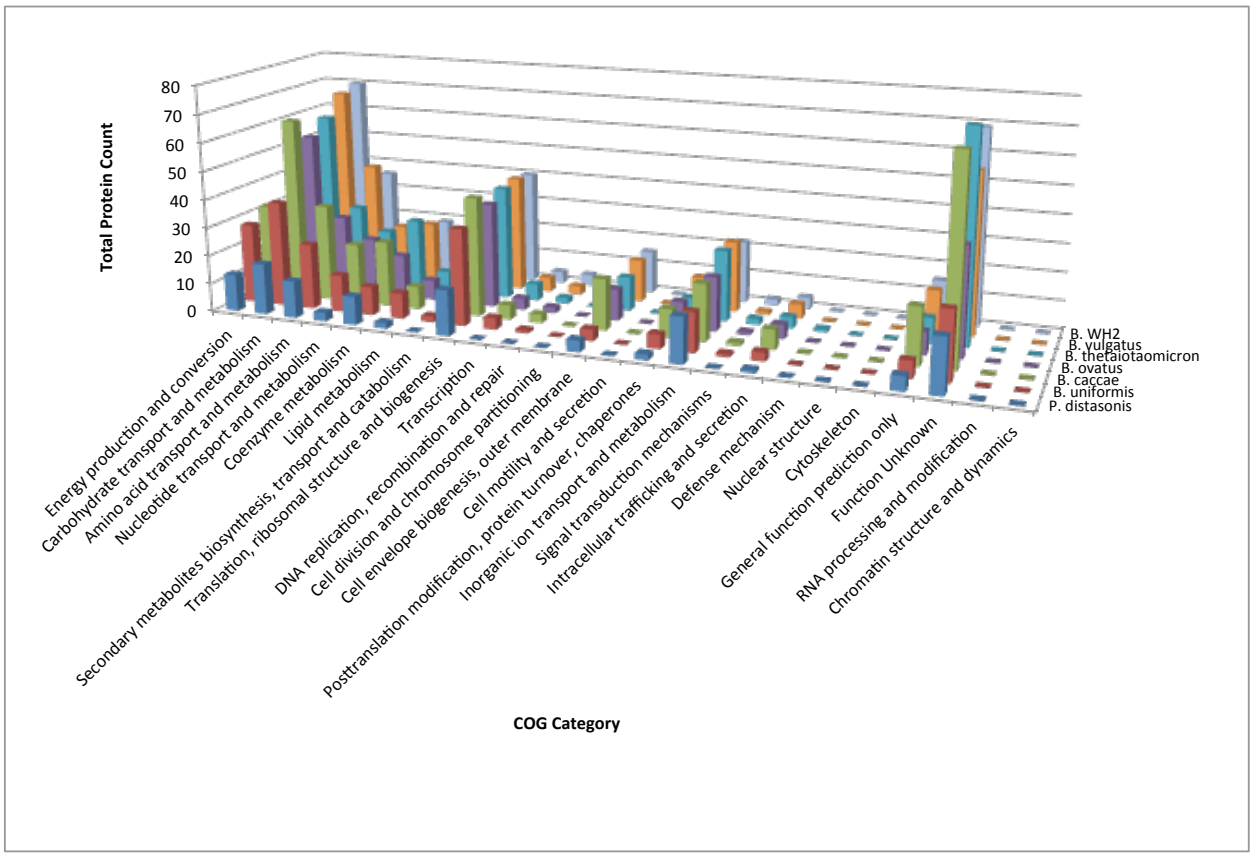


Figure 5.11: Distribution of protein counts per COG category for each phylotypes of Bacteroides for the Western-fed mice.

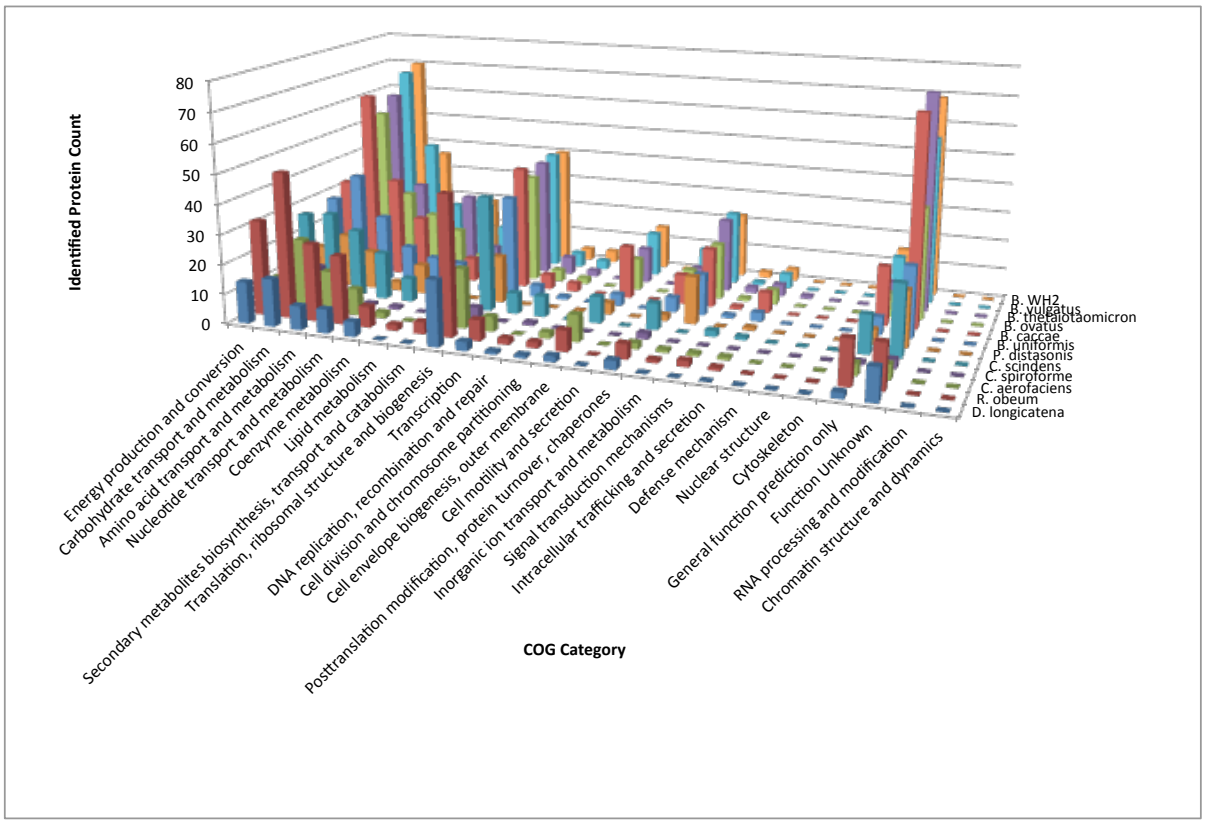


Figure 5.12: Distribution of protein counts per COG category for all 12 phylotypes for the Western-fed mice.

Compared to BK-fed mice, *C. scindens* was identified with higher abundance in the Western-fed mice (**Figure 5.4** and **Figure 5.13**). A significant portion of the Western-fed *C. scindens* proteome is involved in translation followed by carbohydrate metabolism. Of the 220 *C. scindens* identified proteins, 36 were statistically differentiated, with the majority (25) being significantly over-abundant in the Western-fed mice. These proteins were classified with functions in carbohydrate metabolism and energy production. For example, the *C. scindens* ABC-type sugar transport system was identified with over-abundance in the Western-fed mice compared to the BK-fed mice. Due to differences in the dietary components, the increased abundance of *C. scindens* may be a result of their ability to grow more efficiently on sugar compared to the other phylotypes in mice that are fed a diet rich in fat and sugar. The community-wide and individual species-level proteomic data both indicate that translation is significantly more abundant in the Bacteriodes phylotypes of BK-fed mice whereas particular proteins

involved in inorganic ion and carbohydrate metabolism and transport are more abundant in the Western-fed mice based on protein and/or normalized spectra counts.

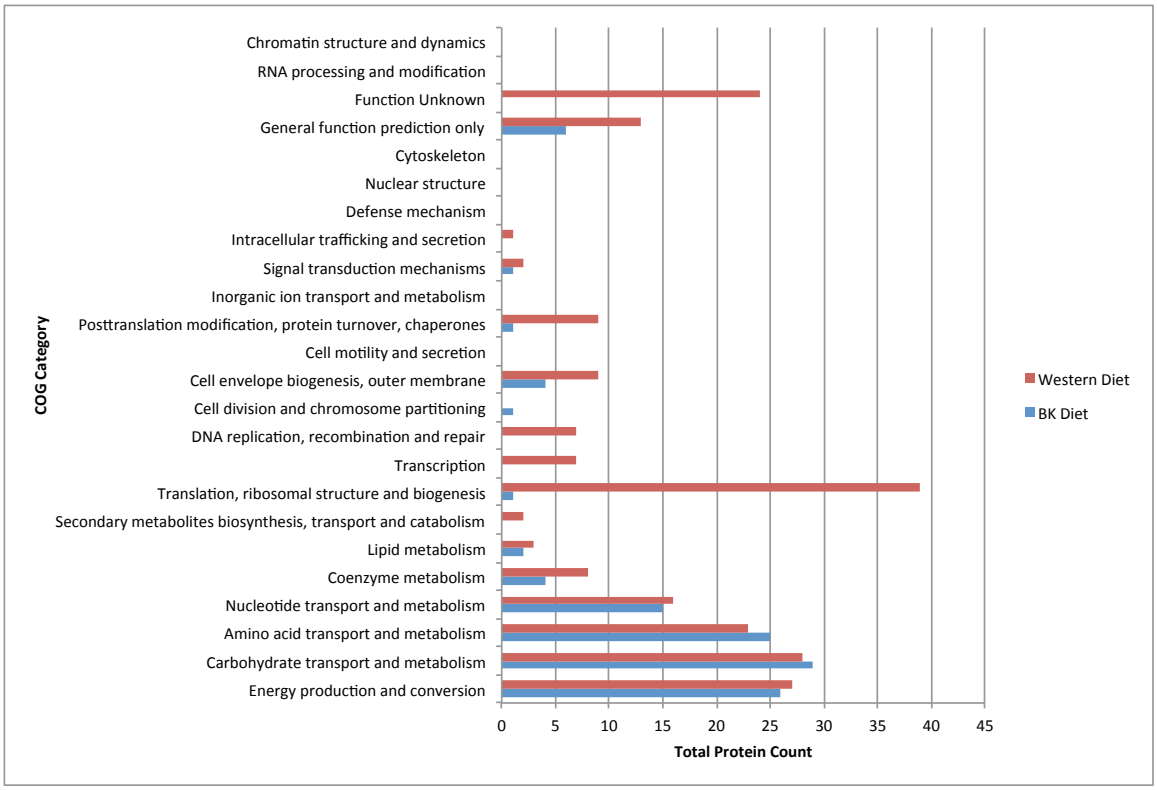


Figure 5.13: *C. scindens* (only) total identified proteins' distribution based on COG categories for both diets.

5.4: Conclusions

Based on all of the collective analyses, shotgun proteomics methods is quite effective for complex consortia comprising species that share a great deal of gene content and should be expandable to larger, more complex communities. In this study, we compared the functional differences and estimated the relative abundance of closely-related species within a 12-member model consortium of phylotypes belonging to Bacteroides, Firmicutes, and Actinobacteria. The microbial community proteomes were evaluated by (i) community-wide normalization (i.e., community-wide response) and (ii) normalization per individual species in response to diet perturbations and (iii) the prediction (theoretical peptidome) and identification of unique peptides as a method for

the relative estimation of species abundance with significant dynamic range. These analyses suggested that the community structure is dictated by the host's diet (i.e., diet is shaping overall community structure).

In conclusion, *B. WH2* is highly abundant across proteomes in both diets, but is a strong diet responder with a preference for the BK diet. *B. caccae* and *C. scindens*, on the other hand, are strong diet responders with a preference for the Western diet in agreement with the GeneChip expression data. We would hypothesize that either (i) *C. scindens* and *B. caccae* both may directly benefit from one or more compounds in Western diet (e.g., they share some common traits/preferences with respect to metabolic niches) or (ii) other species in the community that normally strongly compete with *C. scindens* and *B. caccae* are at a disadvantage in the Western diet where their loss is these two species' gain. To confirm either or additional hypotheses, further analysis and supporting genomic data is necessary.

Chapter 6

Shotgun metaproteomics of the human distal gut microbiota

The text is adapted from:

> *A. L. Russell, N. C. VerBerkmoes, M. Shah, A. Godzik, M. Rosenquist, J. Halfvarsson, M. G. Lefsrud, J. Apajalahti, C. Tysk, R. L. Hettich, and J. K. Jansson. "Shotgun metaproteomics of the human distal gut microbiota." ISME J., 2009, volume 3, pages 179-189.*

Alison R. Erickson's contributions include experimental preparation of microbial samples for proteomics, experimental LC-MS/MS measurements and analysis, and shared primary authorship with Nathan VerBerkmoes.

6.1: Introduction

The human gastrointestinal (GI) tract is host for myriads of microorganisms (approximately 10^{11}/gram feces) that carry out vital processes for normal digestive functions of the host and play an important, although not yet not fully understood, role in maturation of human immunity and defense against pathogens. Recent findings suggest that each human has a unique and relatively stable gut microbiota, unless disrupted by external factors such as antibiotic treatment[156]. Increasing evidence suggests that the composition of the GI microbiota is linked to inflammatory bowel diseases[157], such as Crohn's disease[158], and can even influence the propensity for obesity[131]. Current estimates based on sequencing of 16S rRNA genes in DNA extracted from feces, are that 800-1000 different microbial species and >7000 different strains inhabit the GI tract[159] and that the majority of these (> 80%) have not yet been isolated or characterized[62]. Therefore, there is a vast microbial diversity with largely unknown function that is waiting to be explored.

Recently, metagenomic sequencing has revealed information about the complement of genes in the gut microbiota of two healthy individuals[22]. Although this data set did not represent the entire GI microbiota, analysis of identified genes revealed that the GI microbiome has significantly enriched capacities for glycan, amino acid, and xenobiotic metabolism, methanogenesis, and synthesis of vitamins and isoprenoids.

This indirect evidence suggested that there are unique microbial functions carried out in the gut environment.

A major limitation of DNA based approaches is that they predict potential functions, but it is not known if the predicted genes are expressed at all or if so, under what conditions and to what extent. In addition, it is not possible to determine whether the DNA is from active viable cells, dormant inactive cells, or even dead cells. These limitations can be overcome by directly assessing proteins, because the genes must have been transcribed and translated to produce a protein product. However, to date only a couple of microbial proteins have been identified from the human gut and these were obtained by 2 dimensional polyacrylamide gel electrophoresis (2D PAGE)[47], followed by excision and *de novo* sequencing of targeted spots on the gel.

With an established and successful method to study the proteomes of lower-complexity microbiota in gnotobiotic mice, we expanded this methodology into higher complexity representative human gut microbiomes to evaluate how well this method would work in human feces. Here, our aim was to develop a novel high throughput, non-targeted mass spectrometry (MS) approach to determine the identities of thousands of microbial proteins in the most complex sample type to date (i.e. feces) and to test the feasibility of using a non-matched metagenome data set for protein identification. This MS-based shotgun proteomics approach relies on detection and identification of all proteins in a lysed cell mixture without the need for gel based separation or *de novo* sequencing. Instead, the resulting peptides from an enzymatic digest of the entire proteome are separated by liquid chromatography and infused directly into rapidly scanning tandem mass spectrometers (2D-LC-MS/MS) via electrospray ionization. The resulting peptide mass information and tandem mass spectra are used to search against protein databases generated from genome sequences. To date, the shotgun metaproteomics approach has only been demonstrated in a limited number of studies and only for microbial communities with low diversity, such as acid mine drainage systems[42,43], endosymbionts[160], and sewage sludge water[44]. It remains a technical challenge to apply this shotgun approach to more complex microbial communities, such as those inhabiting the human gut.

For this study, it was first necessary to develop the shotgun proteomics approach to work with fecal samples containing large amounts of particulate matter and undigested food and a large diversity of microbial cells. **Figure 6.1** provides an overview of the experimental approach developed. Fecal samples were chosen because sampling is non-invasive and feces have been shown to provide material that is representative of an individual's colonic microbiota[62]. Our goal was the qualitative identification of the range and types of proteins that can be confidently and reproducibly measured (i.e. with high specificity and low false positive rates; 1-5% maximum) from gut microorganisms by comparing to available metagenome databases[22] and available gut isolate genomes and to determine if unmatched data sets could suffice for accurate protein identifications. An additional goal was to apply a novel bioinformatics approach to assign putative functions to unknown proteins not covered by standard analysis of clusters of orthologous groups (COGs). Ultimately, our aim was to use the protein data to provide direct evidence of dominant and key microbial functions in the human gut for the first time, some of which could serve as indicators of a healthy or diseased state. In addition, this non-targeted approach enables identification of human proteins associated with the gut microbiota, thus illustrating potential interactions between the human microbiome and host.

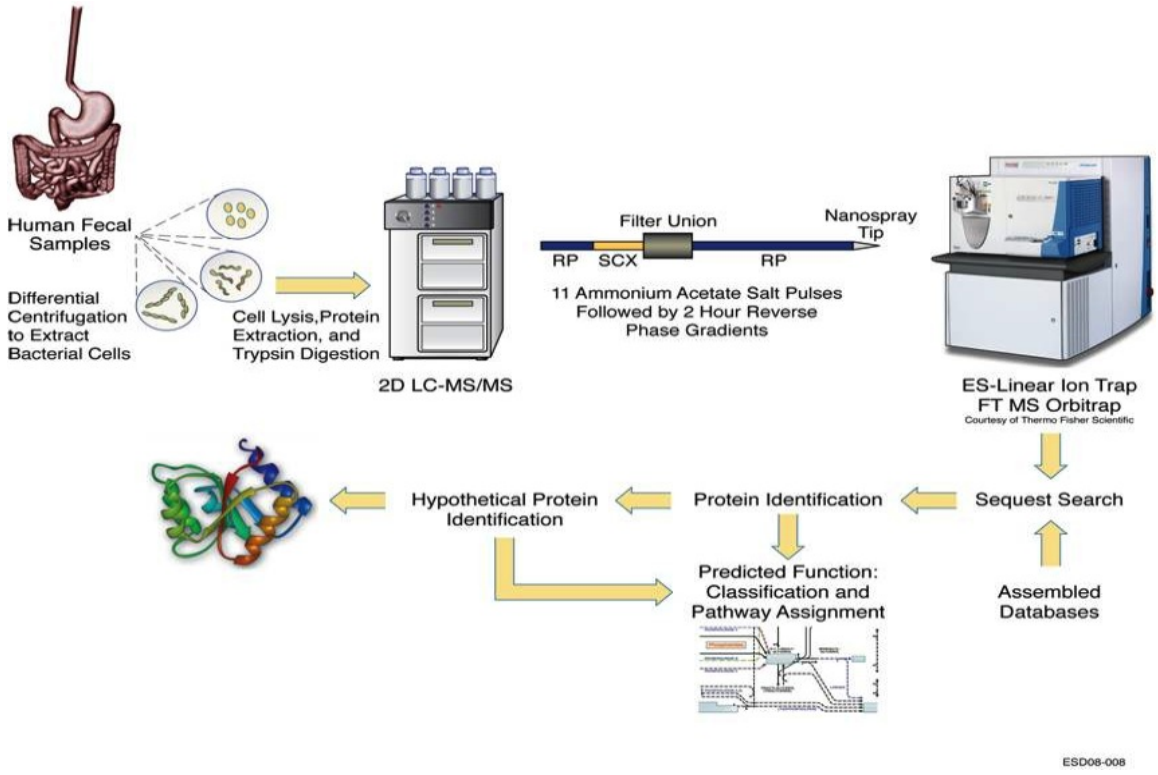


Figure 6.1: Shotgun metaproteomics approach used to identify thousands of microbial proteins in human fecal samples.

6.2: Experimental methods

6.2.1: Fecal sample collection

A female healthy monozygotic twin pair born in 1951 was invited to take part in a larger double blinded study, and details of these individuals with respect to diet, antibiotic usage, etc. are previously described: individuals numbered 6a and 6b[158], that provided Samples 7 and 8, respectively, thus were the focus of this study. The only differences between the individuals according to the submitted questionnaire data were that Individual 6a had gastroenteritis and Individual 6b had taken NSAIDs the last 12 months. Fecal samples were collected in 20 ml colonic tubes by the twins and immediately sent to Örebro University Hospital on the day of collection, where they were placed at –70°C and stored. The Uppsala County Ethics Committee and the ORNL human study review panel approved the study.

6.2.2: Microbial cell extraction from fecal samples

Fecal samples were thawed at +4°C and microbial cells were extracted from the bulk fecal material by differential centrifugation, as previously described[85]. This cell extraction method has previously been found to result in a highly enriched bacterial fraction from complex samples, such as soil and chicken feces, with negligible bacterial cell loss and a good representation of fecal microbiota[85]. The resulting bacterial cell pellets were immediately frozen at –70°C and stored until use.

6.2.3: Cell lysis and protein extraction from cell pellets

The microbial cell pellets (~100 mg) were processed via single tube cell lysis and protein digestion. Briefly, the cell pellet was resuspended in 6M Guanidine/10mM DTT to lyse cells and denature proteins. The guanidine concentration was diluted to 1M with 50 mM Tris buffer/10mM CaCl2 and sequencing grade trypsin (Promega, Madison, WI) was added to digest proteins to peptides. The complex peptide solution was desalted via C18 solid phase extraction, concentrated and filtered (0.45um filter). For each LC-MS/MS analyses below, ~1/4 of the total sample was used.

6.2.4: 2D-LC-MS/MS

Both samples were analyzed in technical duplicates via two-dimensional (2D) nano-LC MS/MS system with a split-phase column (RP-SCX-RP)[97] on a LTQ Orbitrap (Thermo Fisher Scientific) with 22 hr runs per sample (LC as previously described[42,43]. The Orbitrap settings were as follows: 30K resolution on full scans in Orbitrap, all data-dependent MS/MS in LTQ (top five), 2 microscans for both Full and MS/MS scans, centroid data for all scans and 2 microscans averaged for each spectra, dynamic exclusion set at 1.

6.2.5: Proteome informatics

All MS/MS spectra were searched with the SEQUEST algorithm[66] and filtered with DTASelect/Contrast[98] at the peptide level [Xcorrs of at least 1.8 (+1), 2.5 (+2), 3.5 (+3)]. Only proteins identified with two fully tryptic peptides from a 22 hr run were

considered for further biological study. Tandem MS/MS spectra were searched against four databases. The first database (db1) contained two human subject's metagenomes[22], a human database, and common contaminants such as trypsin, human keratins, etc. The existing metagenome databases[22] were deficient in *Bacteroides* sequences and as *Bacteroides* are known to be common and abundant in the human intestine[62], *Bacteroides* genome sequences were also included in a second database (metadb), plus other sequences from representatives of the normal gut microbiota deposited and available at the Joint Genome Institute (JGI) IMG database (http://img.jgi.doe.gov/cgi-bin/pub/main.cgi). In addition, we included distracters that one would not commonly expect in the healthy gut. The third and fourth database were made by reversing or randomizing the db1 and appending it on the end of db1; these databases were used primarily for determining false positive rates, as described earlier[43,96].

6.2.6: Hypothetical Protein Prediction

Hypothetical proteins were submitted to the distant homology recognition server FFAS03[161]. For 80% of the hypothetical proteins, a statistically significant match (Z-score below 9.5) to one of the proteins in the reference databases was obtained. Functions of the matching proteins were used to assign a provisional function for the hypothetical proteins identified in this study.

6.3: Results and Discussion

6.3.1: Metaproteomics of fecal samples

Our results present the first large-scale investigation of the human gut microbial metaproteome. The metaproteomes were obtained from two fecal samples (samples 7 and 8) collected from two healthy female identical twins (subjects 6a and 6b, respectively, see Dicksved et al. (2008) for a description of the individuals). The shotgun approach used enabled us to identify thousands of proteins by matching peptide mass data to available isolate genome and metagenome sequence databases. The total number of proteins identified from searching the first database (db1) that contained all

predicted human proteins and the gut metagenomes were 1822 redundant and 1534 non-redundant proteins, with approximately 600 to 900 proteins identified per sample and replicate (**Table 6.1**). From the entire non-redundant dataset, ~ 1/3 matched human proteins, ~ 2/3 matched predicted proteins from the microbial metagenome sequence data.

Table 6.1: Number of protein, peptide, and spectra identifications for Samples 7 and 8 (2 technical runs each) using the db1 and metadb databases (see supplementary material).

db1 database				
Sample ID	Protein identifications*	Peptide identifications	MS/MS Spectra	Peptides between 10 and -10 ppm**
Sample 7, Run 1	634	1886	4069	81.70
Sample 7, Run 2	722	2253	4440	80.42
Sample 8, Run 1	974	3021	5829	83.41
Sample 8, Run 2	983	2948	6131	81.47
metadb database				
Sample 7, Run 1	970	2441	4829	84.47
Sample 7, Run 2	1098	2977	5364	81.67
Sample 8, Run 1	1341	3586	6509	84.71
Sample 8, Run 2	1275	3374	6635	82.92

*Numbers given are non-redundant identifications
** Mass accuracy

The second database (metadb) contained all of the sequences in the db1 database above, in addition to sequences from representatives of the normal gut microbiota, including strains of *Bacteroides*, *Bifidobacteria*, *Clostridia*, and *Lactobacilli*, plus human pathogens and distracters that one would not commonly expect in the healthy gut, such as environmental isolates. The rice (*Oryza Sativa*) genome was included to help identify plant (food)-related proteins. From the metadb, the total number of proteins identified were 2911 redundant and 2214 non-redundant; between 970 and 1340 proteins were identified per sample and replicate (**Table 6.1**). The categorical breakdown of identified proteins from each major database type is shown **Table 6.2**. In three out of four runs, the highest percentage of protein identifications corresponded to

the bacterial genome sequences that were screened. In the fourth run (that is, run 2, sample 8), most protein identifications matched to one of the metagenomes. By contrast, 30-35% of spectra matched to the human protein database, most likely due to a few highly abundant human proteins in the samples with a large number of spectral counts. The proteins matching to both rice and environmental isolate distracters were low, between 2 and 9%, indicating that the majority of the sequences matched to bacterial types and human sequences that one would expect in the human gut environment. Among the microbial genomes screened, the highest protein matches were to expected sequences from gut isolates. Of the ~10,000-13,000 total spectra observed from each run, ~2,000 matched *Bacteriodes* or *Bifidobacterium* species, with the *Bacteriodes* species always having slightly more spectra, emphasizing the dominance of these groups and their functional significance in the human distal intestine. These data correlate well with our previously published microbial fingerprint data showing an abundance of Bacteroides spp. in both of the individuals studied here[158].

Table 6.2: Categorical breakdown of all identifications for each database component per MS run.

Sample7_Run1

Database	Proteins	%	Spectra	Total %
Gut Isolate Genomes	547	38.17	2926	28.11
Contams	7	0.49	177	1.7
Human Proteins	166	11.58	3276	31.48
Gill Metagenome Set7	205	14.31	1304	12.53
Gill Metagenome Set8	328	22.89	1977	19
Rice	45	3.14	226	2.17
Isolate Distracters	135	9.42	522	5.02
Totals:	**1433**		**10408**	

Sample8_Run1

Database	Proteins	%	Spectra	Total %
Gut Isolate Genomes	604	32.26	3047	22.12
Contams	5	0.27	85	0.62
Human Proteins	232	12.39	4440	32.24
Gill Metagenome Set7	304	16.24	1835	13.32
Gill Metagenome Set8	568	30.34	3720	27.01
Rice	53	2.83	273	1.98
Isolate Distracters	106	5.66	373	2.71
Totals:	**1872**		**13773**	

Sample7_Run2

Database	Proteins	%	Spectra	Total %
Gut Isolate Genomes	600	38.73	3111	27.99
Contams	6	0.39	154	1.39
Human Proteins	187	12.07	3752	33.76
Gill Metagenome Set7	243	15.69	1477	13.29
Gill Metagenome Set8	365	23.56	2013	18.11
Rice	57	3.68	246	2.21
Isolate Distracters	91	5.87	360	3.24
Totals:	**1549**		**11113**	

Sample8_Run2

Database	Proteins	%	Spectra	Total %
Gut Isolate Genomes	515	30.13	2658	20.08
Contams	3	0.18	63	0.48
Human Proteins	214	12.52	4671	35.28
Gill Metagenome Set7	303	17.73	1810	13.67
Gill Metagenome Set8	556	32.53	3535	26.7
Rice	34	1.99	197	1.49
Isolate Distracters	84	4.92	305	2.3
Totals:	**1709**		**13239**	

By using established methods of reverse database searching[43,96]; we estimated a false-positive rate at the peptide level of 1-5% for all identified peptides depending on the method. If only those peptides with corresponding high mass accuracy measurements (<10 p.p.m.) were considered (80-85% of all identified peptides per run), then the rate dropped to 0.05-0.23%.

6.3.2: COG categories in the gut metaproteome

The proteins identified from the db1 search were classified into COG categories and when compared between the two samples and the two technical runs, the data were highly reproducible and consistent (**Figure 6.2**). By comparison to the average metagenomes previously published from other individuals[22], we found that several COG categories were more highly represented in the average microbial metaproteomes of the individuals in the present study (**Figure 6.3**). The metaproteomes were significantly skewed, with a more uneven distribution of COG categories than those represented in the average metagenomes. The majority of detected proteins were involved in translation, carbohydrate metabolism, or energy production; together representing more than 50% of the total proteins in the metaproteome. In addition, more proteins in the metaproteomes were representative of COG categories for post-translational modifications, protein folding, and turnover. By contrast, other COG categories were under represented in the metaproteomes when compared with the metagenomes, including proteins involved in inorganic ion metabolism, cell wall and membrane biogenesis, cell division and secondary metabolite biosynthesis.

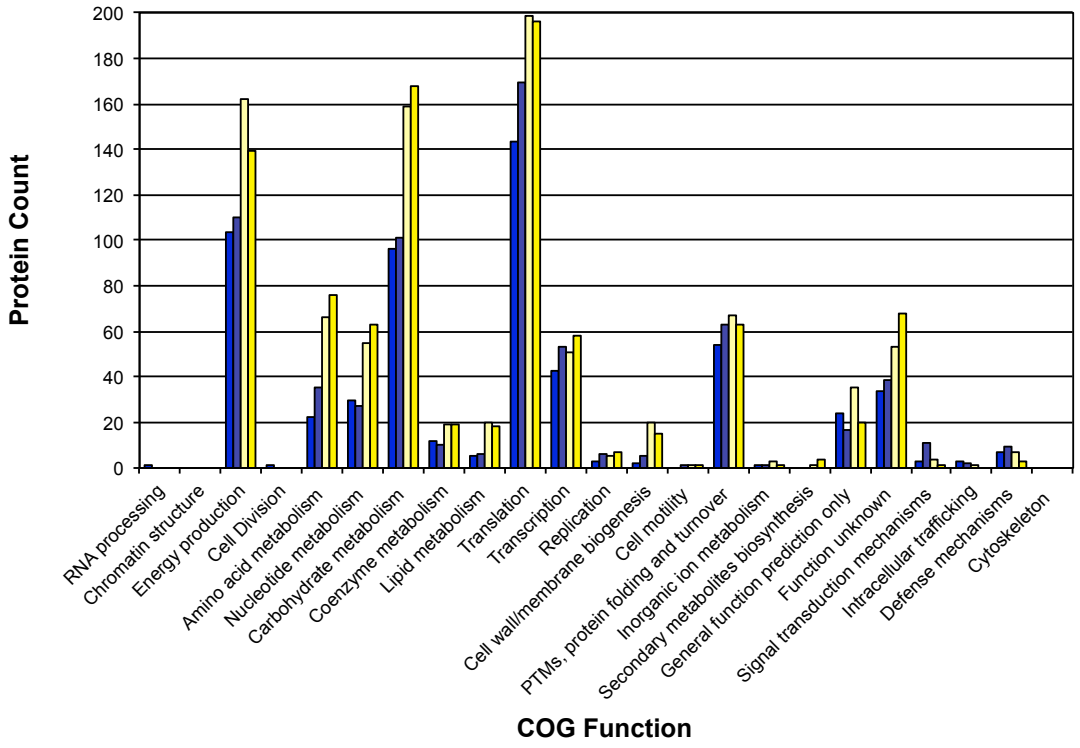


Figure 6.2: Microbial proteins identified from fecal samples 7 (blue bars) and 8 (yellow bars) according to clusters of orthologous group (COG) functions. Bars represent technical proteome runs 1 and 2.

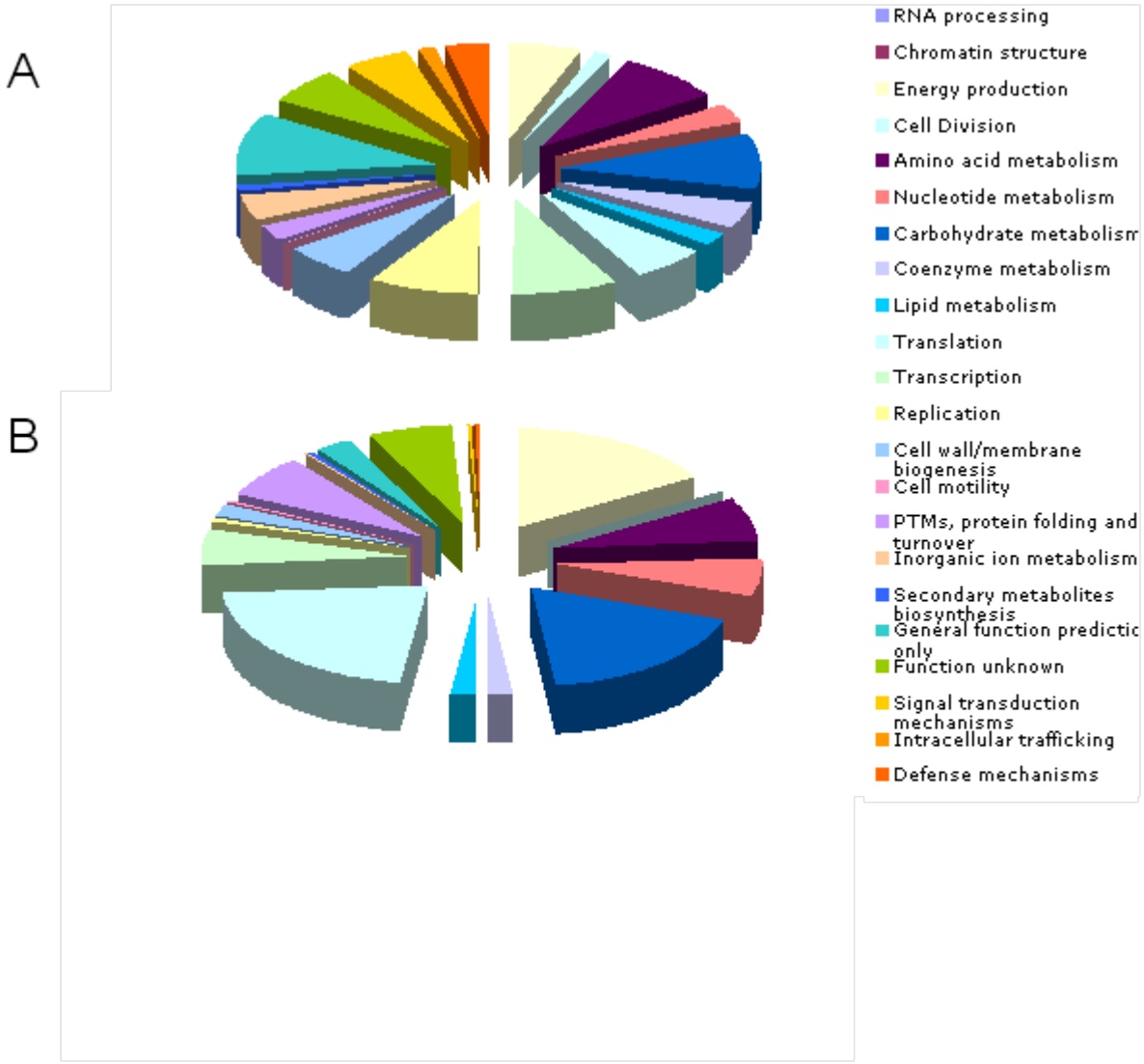


Figure 6.3: Comparison of average clusters of orthologous group (COG) categories for available human metagenomes and metaproteomes. (**A**) Average COG categories of the two *metagenomes* from the gut microbiota of two individuals from a previous study (Gill *et al*. 2006), (**B**) compared to average COG categories of the *metaproteomes* from the gut microbiota of two individuals in the present study.

6.3.3: Label-free estimation of relative protein abundance by normalized abundance factor

We estimated the relative abundances of the thousands of proteins that were detected in each sample by calculating normalized spectral abundance factors (NSAF)[127,162]. By comparing the NSAF data from each sample and technical run to each other, it was clear that the technical runs were highly reproducible for a given sample; R^2 values of 0.77 and 0.85 for samples 7 and 8, respectively (**Figures 6.4** and **6.5**).

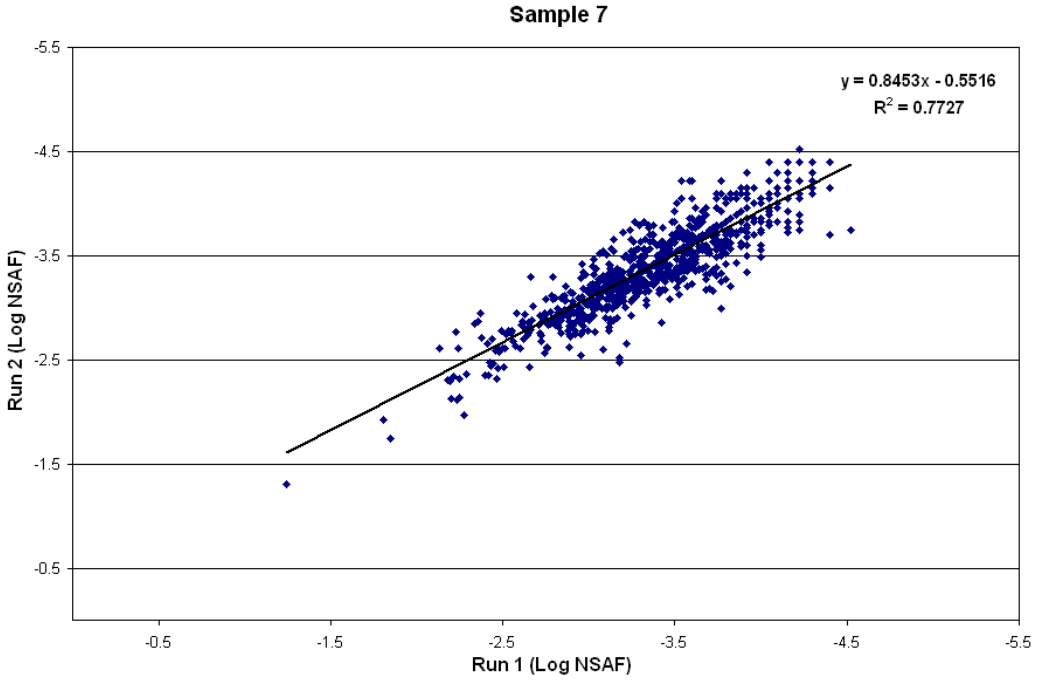


Figure 6.4: Comparison of NSAF values. Sample 7, run 1 and run 2 NSAF values are plotted on a log scale. The dark blue squares represent all of the proteins that were identified in both runs from metadb.

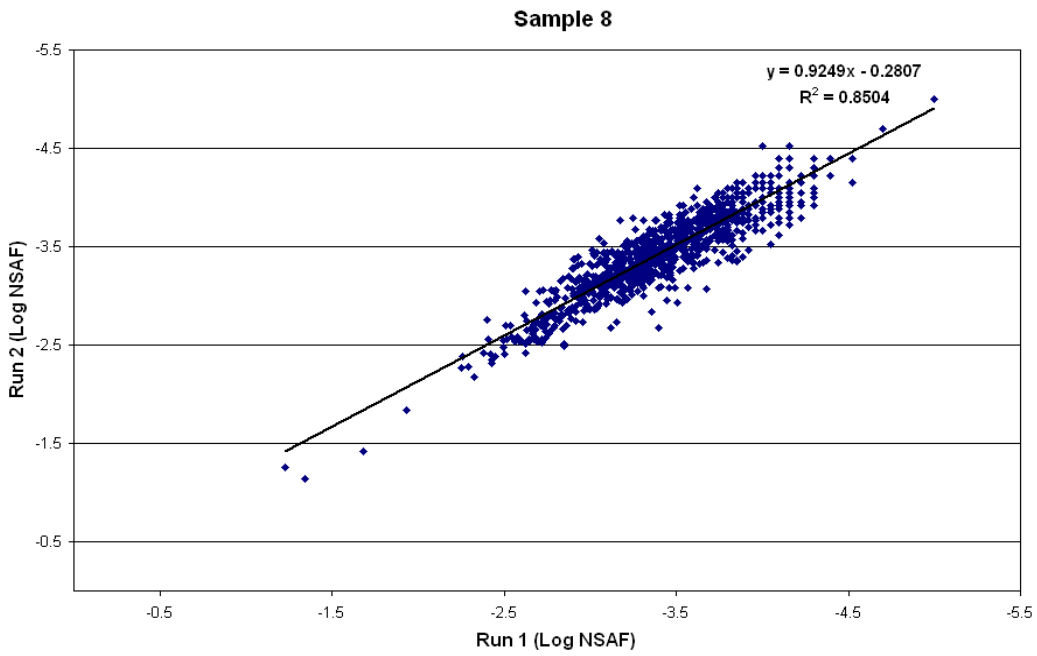


Figure 6.5: Comparison of NSAF values. Sample 8, run 1 and run 2 NSAF values are plotted on a log scale. The dark blue squares represent all of the proteins that were identified in both runs from metadb.

The most abundant proteins based on this prediction were common abundant human-derived digestive proteins such as elastase, chymotrypsin C, and salivary amylases. The most abundant microbial proteins included those for expected processes, such as enzymes involved in glycolysis (for example, glyceraldehyde-3-phosphate dehydrogenase). Ribosomal proteins (in particular for *Bifidobacterium*) were also relatively abundant, as were DNA binding proteins, electron transfer flavoproteins, and chaperonin GroEL/GroES (HP60 family).

The gut microbiomes previously published[22] were enriched for many COGs representing key genes in the methanogenic pathway, consistent with H_2 removal from the distal gut ecosystem through methanogenesis. By contrast, we found very few proteins represented by methanogens. One example is a hypothetical protein from *Methanobrevibacterium* found in sample 8. Instead, analysis of the list of proteins based on the NSAF ranking in our study revealed a high relative abundance of formyltetrahydrofolate synthetase, a key enzyme in the acetyl-CoA pathway of

acetogens[163]. Acetogenic bacteria utilize H_2 to reduce CO_2 and form acetate. Although methanogenesis is an important H_2 disposal route in about 30-50% of people in Western countries, in the remainder H_2 is consumed by sulfate reduction or reductive acetogenesis, and this seems to be the situation for the samples we have studied here.

Similar to the finding of COGs responsible for host-derived fucose utilization that were enriched in the human gut microbiome[22], we also found several proteins involved in fucose metabolism, including fucose isomerase and propanediol fermentation (later steps in the pathway). In particular, we detected proteins corresponding to polyhedral bodies that are assumed to protect the cell by sequestering the toxic propionaldehyde intermediate of this pathway[164].

Butyrate kinase was the most highly enriched COG in the previous metagenomic study by Gill et al. (2006). This enzyme is the final step in butyrate fermentation. Although we did not identify butyrate kinase, we did find that butyryl-CoA dehydrogenase had a relatively high abundance based on the NSAF analyses. This enzyme catalyzes one of the previous steps in the same pathway; interestingly, this protein was strongly expressed in sample 8 but was not detected in sample 7. Additional proteins of interest that were relatively abundant included NifU-like homologs and rubrerythrin. The role of NifU has been proposed as a scaffold protein for Fe-S cluster assembly[165]. Rubrerythrin is found in anaerobic sulfate-reducing bacteria and is a fusion protein containing an N-terminal iron-binding domain and a C-terminal domain homologous to rubredoxin. The physiological role of rubrerythrin has not been identified, but it has been shown to protect against oxidative stress in *Desulfovibrio vulgaris* and other anaerobic microorganisms[166].

Average NSAF values were compared to determine unique and shared proteins in samples 7 and 8 (**Figure 6.6**, metadb database; **Figure 6.7**, db1 database). The scatter plot reveals five distinct areas: proteins found in similar abundances in both samples along the diagonal, proteins found in only one sample on the respective axis, and two distinct lobes that are overexpressed in one sample or the other but present in both (**Figure 6.6**). We suggest that the group of approximately equally abundant

proteins (747 total) represent core gut populations and functions, supported by the finding that a high proportion of these proteins were from common gut bacteria (*Bacteroides*, *Bifidobacterium* and *Clostridium*) and represented housekeeping functions: translation (19%), energy production (14%), post-translational modification and protein turnover (12%) and carbohydrate metabolism (16%) (Supplementary Table S10, first tab). By contrast, the proteins found in only one sample contained proportionately fewer COG categories for housekeeping functions and from common gut species, but a higher proportion with unknown functions (28% compared to 11% found in both). These results suggest that the proteins present or over-represented in only one sample could represent bacterial populations and functions that change according to environmental influences, such as immediate diet. For example, 33% of the unique proteins only found in sample 7 are prolamin proteins, that is, plant storage proteins having a high proline content found in seeds of cereals, suggesting recent ingestion of cereal grains by that individual. Although these individuals did not specify any particular dietary habits in the questionnaire data that accompanied the samples[158], we do not have any detailed information about their specific dietary intake immediately prior to sampling that would enable us to verify this finding.

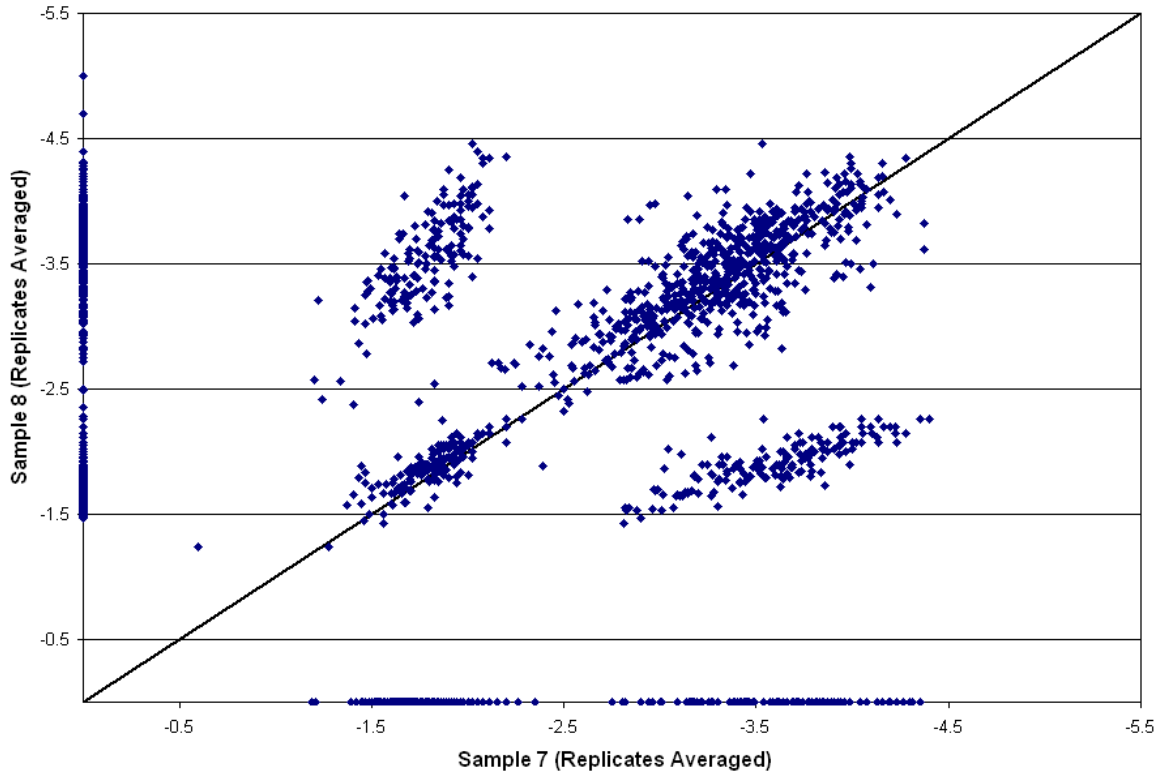


Figure 6.6: Comparison of relative abundances (NSAF values) of proteins detected in Samples 7 and 8. NSAF values for Samples 7 and 8 were averaged amongst their individual technical runs and plotted on a log scale. The dark blue squares represent all of the proteins identified in each sample from screening the metadb database. The straight diagonal line represents the location of all proteins that had approximately equal expression in both samples.

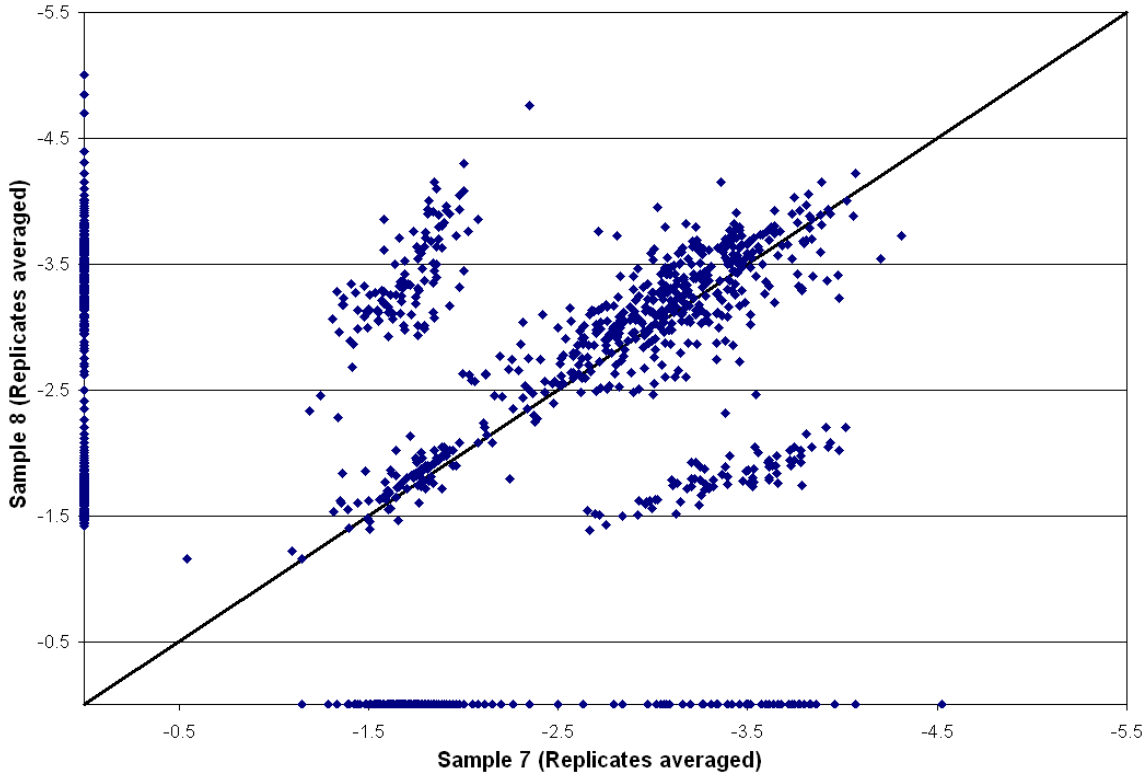


Figure 6.7: Comparison of NSAF values for Samples 7 and 8. NSAF values were averaged amongst two individual technical runs pre sample and plotted on a log scale. The dark blue squares represent all of the proteins identified in each sample from db1. The straight diagonal line is for visualizing the location of all proteins that had approximately equal expression in both samples.

6.3.4: Analysis of unknown hypothetical proteins

We performed detailed analyses of the unknown proteins (116 from the published metagenomes[22] and 89 from bacterial isolate genomes) that could not be classified into COG families. The majority belonged to novel protein families that are over-represented in genomes of gut microbes (**Figure 6.8a**). Five of the ten most abundant hypothetical proteins in the metaproteome belong to the novel protein family represented by hypothetical protein CAC2564, identified earlier in human metagenomes[22], whereas four out of the top ten belong to another novel protein family represented by a hypothetical protein BF3045 from *Bacteroides fragilis*. Members of both families are present in several *Bacteroides*, *Clostridium*, and *Vibrio*

species, where they are always associated with each other (see the red and green arrows in **Figure 6.8b**) and various metabolic enzymes and transport systems. The neighborhood of these two proteins resembles a typical amino acid metabolic pathway, and we hypothesize that they are involved in amino acid metabolism, most likely cysteine or methionine.

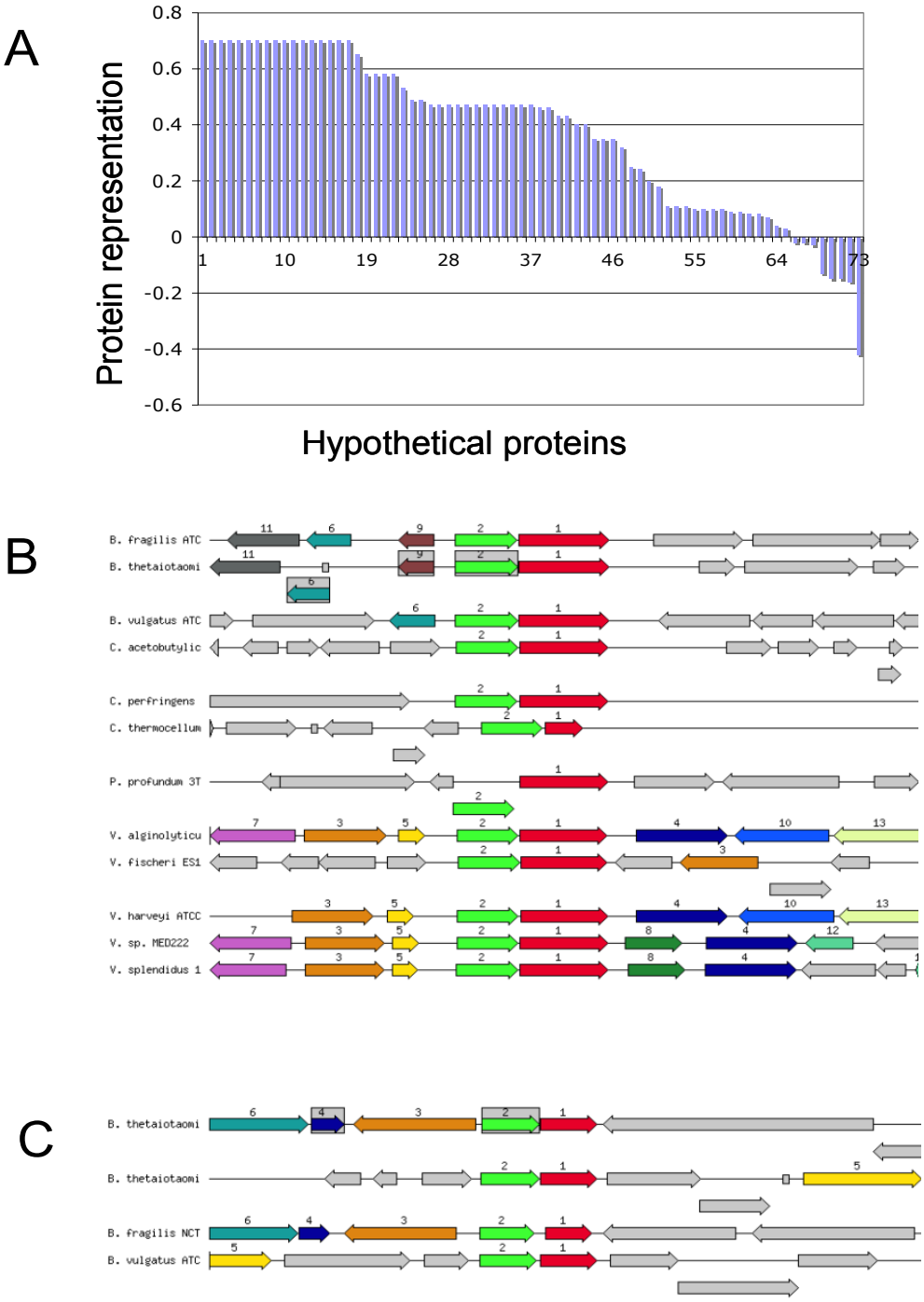


Figure 6.8: Detailed analysis of hypothetical proteins identified in human gut metaproteome. (**A**) Protein representation in the genomes of human gut associated microbes; scale changes from 1 (only found in human gut microbes) to -1 (never found there), 0 represents even distribution. Conserved genomic neighborhoods of the CAC2564 (**B**) and BT2437 (**C**) families.

Another interesting example is the CPE0573 family of hypothetical proteins, originally identified in the human gut metagenome[22]. A distant homolog from this family was recently shown to belong to a novel lacto/galacto-*N*-biose metabolic pathway, identified in *Bifidobacterium bifidum*[167] and *Bifidobacterium longum*[168]. Other proteins from this pathway were also found in the metaproteome samples, suggesting that it was active in our subjects who apparently ingested lactose in their diet. Additionally, an operon formed by a hypothetical protein BT2437 from *Bacteroides thetaiotaomicron* VPI-5482 was found which codes for a putative lipoprotein[169]. Proteins from this family are always associated with channel forming eight-stranded beta-barrel proteins from the OprF family[170] (**Figure 6.8c**).

6.3.5: Identification of human proteins

Almost 30% of all identified proteins were human. The two largest groups of human proteins identified in our study were digestive enzymes and structural cell adhesion and cell-cell interaction proteins. However, the third largest category was comprised of human innate immunity proteins, including antimicrobial peptides, scavenger receptor cysteine-rich (SRCR) proteins (represented by the DMBT1 (deleted in malignant brain tumors) protein), and many other proteins linked to innate immunity and inflammation response (intellectin, resistin, and others). Most of the abundant human proteins were similar in the two individuals, but some differences were found in less abundant proteins.

We were particularly interested in further investigation of DMBT1 (also called salivary agglutinin and glycoprotein-340) that is predominantly expressed in epithelial cells and secreted into the lumen. This protein has several proposed beneficial functions including tumor suppression, bacterial binding, and anti-inflammatory effects[171,172]. Detailed analysis of the distribution of DBMT1 peptides shows that they had fairly uniform distribution along the protein, including hits from all 17 domains present in the DBMT1 protein (**Figure 6.9**), suggesting that the DBMT1 protein was present in our samples as a complete, intact protein, that which we postulate is indicative of a healthy gut environment.

Figure 6.9: Positions of DMBT1 peptide fragments along the length of the DMBT1 protein are shown as blue boxes (figure is not to scale). DBMT1 has a length of 1785 amino acids. PFAM domain names: SRCR (Scavenger receptor cysteine-rich domain); CUB (from complement C1r/C1s, Uegf, Bmp1) is a domain found in many in extracellular and plasma membrane-associated proteins; Zona pellucida, a large, cysteine rich domain distantly related to integrins, found in a variety of mosaic eukaryotic glycoproteins, usually acting as receptors.

6.4: Conclusions

This is the first demonstration of an overall method for obtaining metaproteomics datasets from complex material, in this case human feces, and successful demonstration of the deepest coverage of a complex metaproteome to date. By comparison with earlier work on environmental samples with only a few dominant species[42,43,44], the gut microbiota represents a highly diverse community with thousands of species. Therefore, we are testing the technical limit of the use of the shotgun proteomics approach. We were encouraged that the sample extraction and preparation methods worked well for fecal samples. Although there remain experimental and computational challenges, this general approach should be applicable to other complex environments, such as marine and soil microbial communities.

We also successfully demonstrated that it was feasible to use an unmatched metagenome dataset to obtain valid protein identifications in fecal samples. It is currently more rapid and less expensive to obtain metaproteome data, as we have demonstrated here, than metagenome data. This finding is promising for future metaproteomics studies of other environments that do not have available matched metagenomics sequence data.

One particular challenge is to estimate protein abundances in complex samples. Here, we used label free methods based on spectral counting and NSAFs[127,162]. NSAF is based on spectral counts but also takes into account protein size and the total number of spectra from a run, thus normalizing the relative protein abundance between samples. Efforts are underway to develop better tools for label-free methods, such as the absolute protein expression (APEX) method recently developed by Lu et al.[173]. However, the APEX method was derived specifically for isolate data and is not applicable to complex microbial communities because it requires an estimate of the number of expressed proteins in the system and this is not known, for example, in our case.

Although our results present the largest coverage of the human gut microbial metaproteome to date, increasing the dynamic range beyond this initial study will be necessary in the future to more fully understand the function of the human gut microbiota and its interactions with the human host. Previous studies[42] and current work (NCV, unpublished results with artificial mixtures) suggest that proteins can be detected from populations that represent at least 1% of a mixed community. However, the number of proteins detected (dynamic range) dramatically decreases from thousands to hundreds of proteins for those populations that are present at lower abundances. One possibility to increase the dynamic range of detection would be to enhance the protein separation steps prior to analysis. The trade-off for increasing the number of separation steps would be the requirement for a greater amount of starting material and instrument time. Enrichment or depletion techniques could also be attempted to increase the coverage of community members present at low levels, but care must be taken to not affect the proteome during any manipulations. Increasing the dynamic range is a clear challenge for all proteomic applications and this will be a pressing area for research and method development in the future.

We made several comparisons of our metaproteome data to the existing metagenome data[22]. Some matches could be made between pathways predicted to be functioning based on abundant genes detected in the metagenome data to abundant proteins we found, such as those involved in fucose and butyrate fermentation. There

were also some interesting discrepancies, such as the implication of methanogenesis in the former study and the apparent lack of methanogenesis in the samples we analyzed. The few, low-level, nonunique peptide hits to methanogens that we found were not sufficient to indicate that these organisms were present or functioning. Instead, our data suggest that acetogenesis was occurring in our samples, implicating different hydrogen scavenging routes in the subjects in the two studies.

Although about the same percentage of proteins with 'unknown function' was found in both the metagenomes and the metaproteomes, the metaproteome data provide direct proof that such proteins are actually expressed. Overall, 67% of hypothetical proteins identified in this study could be recognized as distant homologs of already characterized families, allowing putative function assignments, with most of them further enriching the amino-acid and carbohydrate metabolism categories, but also including proteins involved in cell-cell signaling and active transport of nutrients across bacterial membranes. Also, fold recognition level structure predictions are possible for 55% of them, opening doors for modeling and more detailed function analysis.

There were additional discrepancies between some proteins predicted in the metagenomes that were not detected in the metaproteomes and reasons for this include all or some of the following: (1) the microbial community compositions and proteins produced were different in the different individuals, (2) the proteins were produced, but below the dynamic range of detection, (3) they might not have been expressed at significant levels at the time of sampling, or (4) the proteins may have mutated to a point that they are no longer detected by screening an unmatched metagenome[174]. Therefore, although we successfully identified thousands of proteins using an unmatched dataset, it would still be very valuable to have matching metagenome and metaproteome data from the same samples and this will certainly be achieved through ongoing and future initiatives, such as the NIH Human Microbiome Project (http://nihroadmap. nih.gov/hmp/) and the European Union Meta-HIT project (http://www.international.inra.fr/ press/metahit). Recently, 13 additional human metagenome sequences were published from Japan[24] and more representative genome sequences from commensal gut isolates are currently being sequenced [157].

Taken together, these represent valuable resources that should eventually aid in identification of more proteins from the human gut.

A large proportion of the proteins detected in the samples (approximately 30%) were human proteins. This finding can be explained by the differential centrifugation method that we used to obtain a bacterial cell fraction, which is not pure but highly enriched in bacterial cells when compared to human cells and particulate matter in the original fecal sample. Any human protein that adhered to the microbial cells would have been collected in the bacterial pellet. Also, there are many more proteins in human cells than in bacterial cells. Therefore, even a minor contamination of the bacterial fraction with human cells could represent a significant number of human proteins. In hindsight, this was advantageous because it enabled us to detect and identify human proteins, such as antimicrobial peptides, that reflect interaction between the host and the microbiota. Furthermore, this highlights the power of this technology to distinctly identify both microbial and human proteins in a combined mixture.

In summary, although it is evident that this massive dataset would require substantial effort to completely define and characterize, our goal was to develop an approach to obtain a first large-scale glimpse of the functional activities of the microbial community residing in the human gut. A wealth of information about functional pathways and microbial activities could be gleaned from this data, thereby providing one of the first views into the complex interplay of human and microbial species in the human gut microenvironment. It is clear that proteomics allows us to directly see potential host-commensal bacterial interactions. While the human immune response is usually described in terms of response to infection, it is clear that innate immunity proteins are part of a normal gut environment, shaping the gut microflora to the desired shape.

Chapter 7

Strategies for Metagenomic-Guided Whole Community Proteomics of Complex Microbial Environments

The text is adapted from:

> *Alison R. Erickson, Brandi L. Cantarel, Nathan C. Verberkmoes, Brian K. Erickson, Patricia A. Carey, Chongle Pan, Manesh Shah, Emmanuel F. Mongodin, Janet K. Jansson, Claire M. Fraser-Liggett, and Robert L. Hettich. "Strategies for Metagenomic-Guided Whole Community Proteomics of Complex Microbial Environments." Submitted and in review at PLoS One (2011).*

Alison R. Erickson's contributions include experimental preparation of microbial samples for proteomics, experimental LC-MS/MS measurements, integrated comparisons and analyses of all protein database search results, and shared primary authorship with Brandi Cantarel.

7.1: Introduction

Key questions in environmental microbiology include: (i) what microorganisms are present in a particular environment, (ii) how are they functioning, and (iii) how does community structure and function vary in response to environmental conditions/changes? Recent technological advances have provided powerful experimental approaches to address these questions, with 16S rRNA-based taxonomic profiling providing extensive information about microbial composition, and metagenomic whole-genome shotgun (WGS) sequencing/shotgun community proteomics, or "metaproteomics," providing insights into the composition and functional activities of microbial communities. In particular, metagenome sequencing with next-generation platforms has revolutionized the ability to measure and fully characterize the genomic repertoire in microbial communities.

In order to successfully identify peptide sequences using mass spectrometry (MS)-based proteomics methods, a relevant database of predicted genes derived from genome or metagenome sequences is necessary. Peptide identifications result from matching tandem mass spectra (MS/MS) against predicted fragmentation patterns of all

possible *in silico* digested peptides using well-established programs[66,67,68]. Therefore, successful MS/MS sequence-database searching is critically dependent on the quality and accuracy of the metagenomic predicted sequence database.

Although traditional MS-based proteomic analyses of single bacterial isolates are well established, applying these methods to complex microbial communities can be challenging for several reasons, including the lack of deep sequence coverage and difficulty in assembling metagenomes from 454-reads. Considerable improvements in mass spectrometers and chromatography have been made over the past decade; however, the development of tools for optimizing metagenome-metaproteome sequence matching has not kept pace, especially when using the shorter sequence reads associated with next generation sequencing platforms such as the 454 pyrosequencer[175] and Illumina GAII[176].

While an increasing number of studies have developed computation methods for proteogenomics[177,178] and begun to integrate metagenomic sequence data with proteome measurements[42,43,44], these studies have primarily focused on either single eukaryotic genomes or populations with low diversity, allowing for sufficient depth of sequence coverage of abundant community members that facilitate proteome identifications as compared to more complex microbial communities (e.g., human microbiome, ocean, and soil). In the human distal gut, there are approximately 1,000 estimated species which represent >7,000 prokaryotic strains; therefore, the complete metagenome is estimated to be >100 times the human genome [159]. Based on previous studies of these exact same samples, we would expect ~ 30% of the proteins identified by proteomics to be of human origin[26]. The challenges inherent in a metagenomic-metaproteomic characterization of complex environmental samples include (i) considerable sequence diversity among closely related strains/species, (ii) large number of organisms for which no reference genome sequence is available and (iii) low nucleotide sequence coverage for the microorganisms, especially low abundance members.

In the previous chapter, we demonstrated that whole community proteome measurements were possible in the human gut microbiome, but were concerned with the accuracy for how to handle extensive microbial protein redundancy in the metagenomes. Thus, we stepped back to a applying a systematic bioinformatics comparison and analysis of how to construct metagenomic sequence databases for optimum metaproteome measurements. Here we present a benchmarking of strategies for integration of metagenomic and metaproteomic data derived from the same human gut microbiome samples. Although the metagenomes were not sequenced to saturation, they were sufficient to enable us to evaluate how protein predictions based on metagenome data impact peptide-spectrum assignments in matched metaproteomic datasets (i.e., metagenome and metaproteome of the exact same sample). Using 454 pyrosequencing, 1,079 Mbp of DNA sequence was obtained from two fecal samples obtained from a pair of healthy twins[158]. Using these data, four protein sequence databases were created using several different assembly and gene finding strategies (**Figure 7.1**). The resulting databases were evaluated for their utility in MS sequence-database searching.

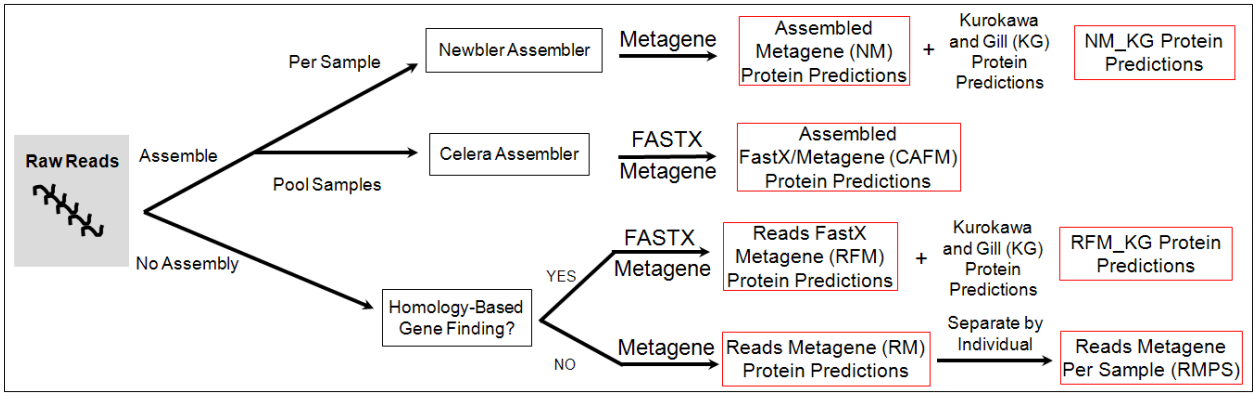


Figure 7.1: Creation of protein sequence databases. Protein sequence databases were created from metagenomic sequence reads using a variety of methods for assembly and gene finding.

Assembly of metagenomic reads can potentially generate errors by joining sequence reads that share sequence identity but are derived from different strains or species. This can be further complicated by sequencing errors, such as issues with

homopolymer tracts in 454 pyrosequencing datasets[179,180]. The metagenome assembly strategies examined in this study were (i) assembly by sample, exemplifying the traditional approach used for single isolate genomes, (ii) whole-dataset assembly, in order to increase sequence coverage, and (iii) no assembly, which will theoretically capture all sequence diversity present in a sample. Since sequencing errors can also introduce frameshifts and in-frame stop codons, resulting in fragmented gene predictions, we explored homology-based gene finding, as it allows the ability to 'gap' over sequencing errors, and *de novo* based gene finding which uses models of known gene structure for prediction.

Proteomics approaches were also benchmarked to identify the parameters necessary to create accurate peptide-spectrum matches (PSMs; a match of a given MS/MS spectrum to a specific database peptide sequence) and increase protein discovery by *de novo* peptide sequencing. Several MS-related parameters (spectral quality, delta correlation (deltCN), and high mass accuracy (±10 ppm (parts-per-million)) were examined and proved to be helpful in providing more comprehensive, confident PSMs. Moreover, we investigated how much *de novo* peptide sequencing would increase peptide identification, since it provides novel sequences that were not originally present in the sequence database (e.g., polymorphisms). By utilizing the genomic and proteomic tools described in this study, we identified a strategy that increased the number of PSMs and protein identifications in a complex microbial community that can provide a more comprehensive and accurate characterization of the human gut microbiome.

7.2: Experimental Methods

7.2.1: Samples, DNA and protein extraction

Fecal samples from two healthy female human individuals (a concordant twin pair), numbered 6a and 6b, were collected under a separate study, as described and studied previously[26]. Both samples were used for DNA and protein extraction. An additional three twin pairs corresponding to six human fecal samples: numbered 15a and 15b (concordant pair with Crohn's disease), 16a and 16b (discordant pair, healthy 16a and

16b with Crohn's disease), and 18a and 18b (discordant pair, healthy 18a and 18b with Crohn's disease), were used for metagenomic sequencing only and were included in several of the sequence databases as described in "Protein Database Construction." Therefore, a total of two healthy samples (6a and 6b) were used for metaproteomics and eight (four healthy: 6a, 6b, 16b, and 18b and four diseased: 15a, 15b, 16a, 16b) samples were used for metagenomics. Throughout the manuscript, the diseased samples and individuals other than 6a and 6b are referred to as "unrelated" because we are only focusing on 6a and 6b samples' metaproteomes, thus, we have a "matched" or "related" metagenome-metaproteome. Since these fecal samples were collected under a separate research program and were supplied as de-identified information for this study, this work was approved in March 2010 by the Oak Ridge Site-wide Institutional Review Board (ORSIRB; Dr. Leigh Greeley, chair-person) as "human studies exemption 4", IRB REFERENCE #: ORNL EX(10)-3.

Total genomic DNA was extracted using the MoBio PowerSoil DNA Isolation kit (MoBio Laboratories, Carlsbad, CA) following the manufacturer's recommendations. Sample 6a was also extracted using the Zymo extraction protocol recently published by Ravel and colleagues[181]. Each sample was then sequenced using Roche 454 FLX-Titanium pyrosequencing according to manufacturer specifications. Raw sequence data were processed using the Roche/454 run processing software to filter short, mixed, and low quality reads. The sequencing generated 418K - 627M passed-filter reads and 170 – 381 Mbp per sample for the eight human fecal samples (15a, 15b, 16a, 16b, 18a, 18b, 6b, and 6a). Microbial cells (~100 mg cell pellet) and proteins were extracted and processed for 2D-LC-MS/MS. The protocol for cell lysis and protein extraction has been rigorously tested and developed by our laboratory[74,86] with specific details corresponding to these samples detailed in Verberkmoes *et al*.[26].

7.2.2: Protein database construction

Starting with 454 pyrosequencing reads, four metagenomic processing methods (NM, RM, RFM, and CAFM) were evaluated for the construction of predicted protein databases (**Figure 7.1** and **Table 7.1**). Sequences were first filtered for human

contamination by alignment of reads to the human genome (v 36) using NUCMER[182] using default parameters. The Newbler-Metagene (NM) protein sequence database was created using the single-genome strategy by generation of a *de novo* assembly followed by *de novo* gene finding. While there are a variety of gene prediction algorithms available, we chose to focus on MetaGene Annotator[183], a platform that we have extensive experience with for 454 sequencing datasets. Certainly, newer approaches, such as Orphelia[184], MetaGeneMark[185], and FragGeneScan[186] have appeared and shown promise for Illumina datasets; the accuracy of these algorithms do not differ greatly for 300-400 bp reads and MetaGene Annotator is well suited for assembled datasets.

Table 7.1: Performance and comparison of the metagenomic predicted protein sequence databases. The database composition and SEQUEST/DTASelect search results (compute time, identified non-redundant spectra and peptides) with a 2-peptide and deltCN of 0.08 filters are shown for samples 6a (Run 2 and 3) and 6b (Run 1 and 2).

Metagenomic Predicted Protein Sequence Database		Celera Assembler, Fastx, Metagene	Newbler, Metagene	Newbler, Metagene + Kurokawa/Gill	Raw Reads, Metagene	Raw Reads, FastX, Metagene	Raw Reads, FastX, Metagene + Kurokawa/Gill	Raw Reads, Metagene Paired Search
Database Acronym		CAFM	NM	NM_KG	RM	RFM	RFM_KG	RMPS
Number of Sequences (thousand)		1,844	190	540	1,903	1,520	1,907	2,146
Number of Amino Acids (million bp)		200	45	115	189	173	262	191
Compute Time Per Run (minutes)		670	80	320	750	1,060	1,030	435
Number of Non-redundant Spectra	6a Run 2	5,179	6,235	10,441	9,100	9,074	10,975	13,806
	6a Run 3	4,326	5,376	9,272	8,152	8,538	10,330	18,401
	6b Run 1	4,092	5,615	10,830	8,639	8,480	11,254	12,363
	6b Run 2	3,873	5,800	10,724	8,775	8,573	11,167	12,212
Total Spectra		***17,470***	***23,026***	***41,267***	***34,666***	***34,665***	***43,726***	***56,782***
Total number of PSMs within ±10ppm		***14,317***	***16,906***	***31,289***	***26,181***	***25,997***	***33,347***	***39,681***
Number of Non-redundant Peptides	6a Run 2	4,383	3,093	5,678	4,710	4,669	5,911	7,592
	6a Run 3	3,655	2,403	4,617	3,804	3,963	5,068	6,303
	6b Run 1	3,404	2,426	5,409	3,919	3,879	5,549	5,923
	6b Run 2	3,216	2,297	5,088	3,747	3,690	5,238	5,605
Total Peptides		***14,658***	***10,219***	***20,792***	***16,180***	***16,201***	***21,766***	***25,423***
Total NR Peptides		***8,632***	***5,994***	***12,406***	***9,618***	***9,608***	***13,111***	***16,055***

Shotgun sequences from each sample were assembled using the Newbler Assembler (v2.0.01.14), and genes were predicted on contigs greater than 500 bp using Metagene[187], resulting in a total of 153,586 predicted open reading frames (ORFs) larger than 50 nt across a total of the seven metagenome samples included in this study. The second database, Reads-Metagene (RM), was created by directly predicting ORFs from raw sequencing reads to prevent loss of sequence diversity when collapsing unrelated sequencing reads during genome assembly. ORFs were predicted using Metagene, yielding 1,866,893 predicted ORFs larger than 50 nt. Sequencing errors often seen in pyrosequencing datasets[179,180] can lead to artificially fragmented predicted ORFs. Because these errors cause frameshifts and in-frame stop codons in gene predictions, we used protein-to-DNA alignments, generated by sequence similarity searches against NCBI's NR using FASTX[188] with an expectation value threshold of $1e^{-6}$, to predict genes by homology. Homology-based gene finding was performed on raw 454 sequencing reads yielding 1,483,958 predicted ORFs larger than 50 nt, called Reads-FASTX-Metagene (RFM) protein database.

Additionally, three databases were created from assembled reads, with the intent of creating longer genes and fewer protein fragments. The combination of short sequencing reads, averaging 369 bp, and the high bacterial diversity found in the human gut, produced a dataset with many fragmented genes. Since assembled sequences were not much longer than raw sequencing reads, these genes were also fragmented, therefore, we were unable to validate proteins identified by multiple peptide matches. Thus, an assembly was created by combining the shotgun sequence data from these samples using the Celera Assembler (v5.4), called Celera Assembler-FASTX-Metagene (CAFM), yielding 1,807,963 predicted proteins on all contigs and singletons larger than 50 nt. Homology-based gene finding was also used for this CAFM database, using the same parameters as RFM. In addition to sequences generated in this study, we included the following published human gut metagenomic datasets: two metagenomes from Gill *et al*. [22] and thirteen metagenomes from Kurokawa *et al*.[24], that were concatenated with the NM (termed NM_KG) and RFM (termed RFM_KG) sequence databases to provide additional sequence variation and

increase proteome coverage. The metagenomes published from Gill *et al*.[22] (17,688 contigs; ORFs ≥ 20 amino acids; ~50,000 predicted proteins; available at the Joint Genome Institute (JGI) IMG database under NCBI project ID 16729) and Kurokawa *et al*.[24] (81,968 contigs; ORFs ≥ 50 amino acids; ~300,000 predicted proteins; available at CAMERA (2007)) studies were sequenced via Sanger-based methods. The amino acid sequence of the proteins belonging to the two samples' metagenomes used in this study (6a and 6b in addition to 15a, 15b, 16a, 16b, 18a, 18b) can be accessed through the NCBI Protein Database under NCBI project ID 46321.

For each of the protein sequence databases described above (NM, CAFM, RFM, NM_KG, and RFM_KG), we concatenated the metagenomic protein predictions from multiple individuals into a single database. For example, NM, RM, RFM, and CAFM each contain metagenomic sequences from seven individual human samples from this study (15a, 15b, 16a, 16b, 18a, 18b, and 6b), which include an unrelated healthy sample 16b (Figure 2 comparisons). The NM_KG and RFM_KG protein databases contain the same 7 metagenomic predicted protein sequences (15a, 15b, 16a, 16b, 18a, 18b, and 6b), but unlike NM and RFM, contain the published 13 Japanese metagenome sequences[24] and 2 American metagenome sequences[22] for a total of 22 concatenated metagenomes per protein sequence database.

Deeper whole genome shotgun sequencing was obtained from an extra run on 6b and an additional sample (6a), extracted using the Zymo and MioBio method, which resulted in a four-fold increase in sequence data for these two healthy samples (**Table 7.2**). Due to the limitations of analyzing this larger metagenomic sequence dataset, these sequences were processed similar to the RM strategy and compiled into 2-independent protein databases, termed RMPS, for 6a and 6b in this assessment. Each of these 8 protein databases (NM, NM_KG, CAFM, RFM, RFM_KG, RM, RMPS-6a and RMPS-6b) included human reference sequences (July 2007 release, NCBI; ~36,000 protein sequences) and common contaminants (i.e., trypsin and keratin; 36 protein sequences). Lastly, a 6-frame translation library was generated for sample 6a and searched against one MS experiment.

Table 7.2: Metagenomic sequencing metrics.

Metagenome	LC Prep	DNA Extraction	Total Reads	Total KiloBP	Avg Read Length
15a	1	MoBio	563,893	220,685	391
15b	1	MoBio	550,360	204,845	372
16a	1	MoBio	585,262	229,389	392
16b	1	MoBio	418,759	147,547	352
18a	1	MoBio	627,543	221,475	353
18b	1	MoBio	424,935	142,718	336
6a	1	MoBio	1,079,211	403,882	374
6b	1	MoBio	584,264	220,354	377
6b	2	MoBio	599,107	188,577	315

7.2.3: Spectral analysis

Microbial proteins were extracted and processed for 2D-LC-MS/MS as described[26] using an Ultimate HPLC system (Dionex, Sunnyvale, CA) coupled to a high resolution LTQ-Orbitrap (Thermo Fisher Scientific, San Jose, CA). Peptide mixtures from the two samples, 6a and 6b, were separated by a 12 step, multidimensional high-pressure liquid chromatographic elution consisting of eleven salt pulses followed by a 2 hr reverse-phase gradient from 100% solvent A (A: 95% H2O, 5% acetonitrile, 0.1% formic acid) to 50% solvent B (B: 30% H2O, 70% acetonitrile, 0.1% formic acid). Precursor full MS spectra (from 400-1700 m/z) were acquired in the Orbitrap with resolution = 30,000 followed by five data-dependent MS/MS scans at 35% normalized collision energy in the LTQ with dynamic exclusion enabled. All RAW files were converted to mzXMLs using ReAdW (v4.3.1; 2009) and mzXMLs subsequently converted to dta files using MzXML2Search (v4.3.1; 2009). All MS/MS were searched with SEQUEST (v.27)[66] for fully tryptic peptides (≤ 4 missed cleavages, 3 Da parent mass tolerance window, 0.5 Da fragment ion window) against each of the 8 custom-made FASTA formatted protein sequence databases described above. Since it is well established that trypsin cleaves primarily C-terminal to Arg and Lys[189], we have found in a variety of microbial communities[42,190,191] that using fully tryptic searches provides increased confidence in the peptide assignments while minimizing the potential for increased false positives due to incorrect candidate peptide sequences. All SEQUEST output files were assembled and filtered using DTASelect (v1.9)[98] at either a 2-peptide level for all seven: NM, NM_KG, RM, RFM_KG, RFM, CAFM, and RMPS databases and also 1-

peptide level for the RMPS database searches with the following widely accepted parameters: cross correlation scores (XCorr) of at least 1.8, 2.5, 3.5 for +1, +2, and +3 charge states[42,92,98], respectively and a minimum deltCN of either 0.08 (default) for all seven databases (NM, NM_KG, RM, RFM_KG, RFM, CAFM, and RMPS databases) and/or 0.0 for NM, NM_KG, RFM, RFM_KG, RMPS-6a and -6b, and target-decoy databases (described under “false discovery rates”). Post-translational modifications and other fixed modifications were not included in the search criteria.

We used the high mass accuracy capabilities of the Orbitrap with a wide mass tolerance to measure precursor ion (peptides) masses at low parts-per-million (ppm) and the ion trap to efficiently measure fragment ions at lower resolution. A “post-database search” filter with high precursor mass accuracy was used by comparing the theoretically derived peptide from the SEQUEST mass with what was observed in the Orbitrap in the full scan preceding the MS/MS scan. Recently, Hsieh et al. indicated that a wide precursor mass window in a database search[192] and a post-database high precursor mass accuracy filter is a more superior method to control false positives. Therefore, for post-filtering the database results by high mass accuracy, the mass deviation (in ppm) of a PSM was calculated using the measured monoisotopic mass and theoretical monoisotopic mass of the peptide. For all of the database searches (NM, NM_KG, RM, CAFM, RFM, RFM_KG, RMPS-6b and -6a, and target-decoy databases) and comparisons, DTASelect was run with a t0 option to report all MS/MS spectra, in which case two spectra per protein, rather than two peptides, are required for identification. We compared each of the database results in a relative fashion such that all comparisons (degenerate peptides) are consistent to one another. Every MS/MS spectrum that is assigned to a peptide (unique and non-unique peptides) was noted and handled by DTASelect as described[98]; therefore, we recognize peptides that are shared (non-unique) among multiple proteins. While we recognize that non-unique peptides are somewhat problematic for label-free quantification using spectral counts, this was not the focus of the current study.

Spectral quality assessment was accomplished utilizing an in-house developed script that parses the SEQUEST output and mzXML formatted spectral data. All

spectra collected during an analysis were categorized according to type: full scan (MS1) or tandem mass spectra (MS/MS). MS/MS spectra assigned to a peptide by SEQUEST were noted while the remaining unassigned MS/MS spectra were classified as high-quality or poor based on the following conditions: a. the charge state of the parent ion must be greater than 1, b. the minimum absolute intensity must be greater than 2500 counts, and c. greater than three fragment peaks within 20% of the based peak must be present (all other details in preparation to be submitted for publication). To quantify the peptide-spectrum success, MS/MS were categorized as (i) assigned or unassigned to a peptide and (ii) if unassigned, a score of high-quality or poor as reflected by four methods (NM, CAFM, RFM, and RMPS) and six databases (NM, CAFM, RFM, RFM_KG, and RMPS-6a and -6b).

7.2.4: False discovery rates

A target-decoy database[96,126] was generated for each of the five metagenomic processing methods (NM, CAFM, RM, RFM, RMPS), for a total of six forward-reverse databases (RM, RFM, CAFM, KG, NM_KG, and RMPS-6b) and searched against one of the two samples (6b) used in this study to estimate the peptide-level false discovery rate (FDR) with the new metagenomic processing methods. One sample and technical run (6b, Run1) was used to represent the entire sample set (2 samples; 4 runs) for each target-decoy database search in order to reduce the total number of target-decoy databases, search time, and complexity of comparisons. All target-decoy SEQUEST output files were assembled and filtered using DTASelect (v1.9)[98] with the same XCorr filters as described previously, and either a ≥1 peptide per protein with a deltCN filter of 0.0, or a ≥2 peptide per protein with a deltCN of 0.0 (RMPS-6b) or 0.08 (NM_KG, CAFM, RM, RFM, and KG), with an empirical FDR threshold of ≤ 2.0%. The initial, 1-peptide filter and deltCN 0.0, forward-reverse database searches provide FDRs for NM_KG, CAFM, RM, RFM, KG, and RMPS-6b (read-based) database analyses while the latter, 2-peptide and deltCN 0.08 filter, forward-reverse database searches contain the same filtering criteria as the original forward databases (NM, NM_KG, RM, RFM_KG, RFM, CAFM, and RMPS databases; Table 1 results) described earlier. Finally, a forward-reverse database was also created for the final paired metagenome

sequence strategy (RMPS) for 6b and searched against the spectra collected from 6b, Run 1 and Run 2 using a deltCN 0.0, 1-peptide minimum, and high mass accuracy filtering. The identified peptides (both forward and reverse) were then mapped back to the protein sequences derived from the assembled metagenomic sequences using a post-database 2-peptide filter by exact string comparisons. Although the peptides with corresponding high mass accuracy measurements (±10 ppm) were considered for all downstream analyses, the peptide-level FDRs were estimated for both, with (-10 ≤ ppm ≤ 10) and without (ppm < -10 and ppm > 10) high mass accuracy, for 6b, Run1 against six genomic processing methods (NM_KG, CAFM, RM, RFM, KG, and RMPS-6b). Each protein entry (sequence) was reversed, i.e., the original N-terminus became the C-terminus. The new reverse (false) sequences were then appended onto the backend of the original forward sequences where each set, forward and reverse, represents 50% of the entire database. A peptide-level FDR was calculated based on the calculation: $2[n_{rev}/(n_{rev} + n_{real})]*100$ where n_{rev} is the number of peptides identified from the reverse database and n_{real} is the number of peptides identified from the real (forward) database[96].

7.2.5: Sequences similarity searches

Peptides obtained from our SEQUEST/DTASelect searches were searched against the 6b and 16b protein databases using the FASTS algorithm and against raw sequencing reads using TFASTS[193], algorithm that compares peptides to DNA sequence, using an e-value cutoff of 10^{-5}.

7.2.6: *De novo* sequencing of peptides by MS

PepNovo+[105] and PEAKS[109] algorithms were used to *de novo* sequence MS/MS spectra collected from both samples, independent of all sequence databases. The PEAKS (v4.5 SP2) algorithm computes the best possible sequence among all probable amino acid combinations at a full peptide length confidence followed by individual amino acid confidence per residue in the predicted sequence for a MS/MS. PEAKS was run with default parameters with a parent mass error tolerance of 0.5 Da, fragment mass error tolerance of 0.5 Da, and trypsin digestion. First, a 90% confidence level was

required for the overall, full length prediction to be correct and second, an 80% confidence level was required for each residue within that sequence, which is consistent with Ma et al.[109]. PepNovo+ (v3.1) was executed using the following recommended parameters: -model CID_IT_TRYP -digest TRYPSIN -pm_tolerance 0.05 -num_solutions 5 -output_cum_probs. The top-scoring tags of all spectra were filtered using a cumulative probability cutoff of 0.5. In the sequence tags produced from both algorithms, the isobaric amino acid pair of Isoleucine (I) and Leucine (L) and the nearly isobaric pair of Lysine (K) and Glutamine (Q) are considered equivalent. L and I were both substituted with the letter, J, for convenience. Additionally, Q and K were substituted with the letter, U, since they are not easily resolvable (small mass difference of 0.036 Da) with ion trap MS/MS data. For all three algorithms, SEQUEST, PEAKS, and PepNovo+, a minimum of 3 residues has to be assigned to a spectrum for it to be considered for any additional analysis and comparison to other algorithms. For PEAKS, only the high confidence sequence tag was used for all analyses, not the predicted full-length peptide sequence. For the comparison of PSMs between all three algorithms, a "partial" consensus sequence was considered as a peptide sequence that has ≥3 amino acids that are exactly the same for the same mass spectrum between either SEQUEST peptide sequences, Peaks' high confidence sequence string, and/or Pepnovo+s' sequence tag. If a PSM has an "exact" consensus sequence with 100% sequence identity between any two or more algorithms, it would be considered a shared, exact consensus sequence. If a PSM does not have at least 3 residues within a peptide sequence string that match two or more algorithms, that spectrum would be considered unique to that algorithm. The identified SEQUEST/DTASelect PSMs for RMPS-6a and -6b sequence databases with a 1-peptide minimum and deltCN of 0.00 for 6a (Run 2 and 3) and 6b (Run 1 and 2) were compared to the PSMs from PEAKS and PepNovo+. The breakdown of partial and exact consensus sequences versus PSMs that are unique to a specific algorithm can be found in the Venn diagram. We did not take into account any single amino acid polymorphisms in the algorithms' consensus sequence comparisons. In this study, we controlled the false discovery rate by only using the high confidence consensus sequences tags found between the two *de novo* algorithms using their respective optimum parameters.

7.3: Results

7.3.1: Protein sequence database comparison

Four protein prediction strategies (**Figure 7.1**) were implemented for metagenomic DNA sequences obtained from two healthy human fecal samples (referred to as 6a and 6b), using a combination of assembly and gene prediction methods. Each protein sequence database has a defined acronym (2-4 letters), designating the strategy used (**Figure 7.1** and **Table 7.1**). Our goal was to increase peptide-spectrum matches (PSMs) using MS database searching for which the MS data was collected from the same samples as the DNA sequence data. The ability to accurately match peptides to tandem mass spectra (MS/MS) was assessed by comparing the number of PSMs and unique peptides identified for each database search with SEQUEST/DTASelect at a 2-peptide level, deltCN 0.08, and XCorr filtering against the same 2 samples, 6a (with spectra from runs 2 and 3) and 6b (with spectra from runs 1 and 2) (**Table 7.1**). These results illustrate how common metagenomic processing methods (assembly and ORF finding) affect peptide and spectra identification (**Table 7.1**). From these results, three major trends emerge: (A) Collapsing of the sequence data by assembly decreases the number of assigned spectra. There was a decrease of assigned spectra when all reads were assembled from all samples compared to assembly by individual sample (NM, 23,026 spectra vs. CAFM, 17,470 spectra). Additionally, if reads are annotated without assembly, PSMs increase (NM, 23,026 spectra vs. RM, 34,666 spectra). This can be largely attributed to the increased diversity of possible peptides, determined by *in silico* trypsin digestion, in the unassembled data, which is over 3 times what is found in assembled data (5,638,100 vs. 1,639,802). (B) An increase in spectrum assignment usually translates to an increase in unique peptide identifications. For example, the 11,640 gains in spectral assignment translate to a 3,624 gain in identification of unique peptide sequences for RM compared to NM (**Table 7.1**). However, this was not observed when comparing CAFM to NM, where the 5,556 gains in spectra assignment translated to a decrease of 2,638 unique peptides (**Table 7.1**). (C) *De novo* gene finding methods are sufficient for optimal spectrum assignment. The combined *de novo* and homology-based gene finding method did not increase PSMs as hypothesized

(RFM, 34,665 spectra vs. RM, 34,666 spectra) nor the number of identified unique peptides (RFM, 9,608 peptides vs. RM, 9,618 peptides; **Table 7.1**).

Because of the low relative sequence coverage of our metagenomic samples, we wanted to evaluate whether adding metagenomic sequences from 15 unrelated samples in two published studies would enhance our spectrum assignment. Therefore, to protein databases NM and RFM, we added the proteins sequences from predicted ORFs from two published human gut metagenomic studies, referred to as "KG" for Kurokawa et al. and Gill et al.[22,24], which are referred to as NM_KG and RFM_KG respectively. The KG database contains 13 metagenomes from a Japanese cohort[24] and 2 metagenomes from an American cohort[22], both geographically distinct from samples in this study. When compared to the metagenomic sequences in this study, only 9% of sequences align in KG at 99% identity or greater; thus, they provide over 2 million additional unique peptides for MS/MS assignment, that are not identified in any of the matched metagenomes. Because the assemblies from these studies are on average longer (average contig length of 2,300 nt for Kurokawa et al. compared to an average contig length of 1,128 nt in this study), the predicted proteins are more likely to be full-length compared to ORFs in this study (average protein length of 194.5 aa for Kurokawa et al. metagenomes; average protein length of 225 aa for Gill et al. metagenomes compared to an average protein length of 168.5 aa in this study). By including metagenomic sequence from additional sources[22,24], the number of identified spectra increased (NM versus NM_KG (23,026 versus 41,267 spectra) and RFM versus RFM_KG (34,665 versus 43,726 spectra)) for 6a (Run 2 and 3) and 6b (Run 1 and 2) in total (**Table 7.1**). However, the additional KG sequence data came at the cost of increased peptide degeneracy and subsequent protein redundancy (i.e., peptides mapping to multiple proteins or to the same protein in multiple metagenomes within the sequence database). Although the level of redundancy ranges with the sequence diversity of a sample and has no effect on the actual database search algorithms, this complicates protein inference and assigning its' corresponding phylogenetic origin in a complex environmental community.

While the four metagenomic processing methods were compared based on their ability to comprehensively assign all collected MS/MS spectra to peptides, the percentage of assigned and high-quality unassigned MS/MS is equally important to establish the utility of each sequence database. For the following spectral analyses, the collected and assigned spectra from sample 6a (Run 2 and 3) and 6b (Run 1 and 2) were assessed and categorized after applying the same filters described above (2-peptide level and deltCN 0.08 filter) with the following databases. Of the total MS/MS collected during one MS experiment (70,000-81,000), on average 6,600 spectra were assigned to a peptide sequence in the NM database (~8% of total collected MS/MS spectra for a single run; **Table 7.3**). In contrast, the processing strategy used to create RFM resulted in the assignment of an additional 1,800 MS/MS from the same sample, for a total of 8,430 peptide-spectrum matches on average (11% of total collected MS/MS). Furthermore, the addition of unrelated KG sequences to RFM (a 25% increase in sequence data) resulted in an increase of the number of assigned spectra by only 2-3%. Finally, the strategy used to create RMPS resulted in an additional 4,000 MS/MS spectra assigned, for a total of 12,461 peptide-spectrum matches on average per sample (16% of total collected MS/MS spectra). Although the total number of assigned MS/MS increased from NM < RFM < RFM_KG < RMPS, the number of unassigned, high-quality spectra decreased with database quality (NM > RFM > RFM_KG > RMPS).

Table 7.3: Database dependent distribution of acquired full MS and MS/MS and assigned MS/MS for samples 6a and 6b. Unassigned MS/MS were parsed into either quality or poor spectra.

Sample	Run #	Database	Total Spectra Collected	# MS1 Collected	# MS/MS Collected	# Assigned MS/MS	% Assigned MS/MS	# Unassigned MS/MS	# of Quality MS/MS (ID'd + Qual Unass.)	# Poor Unass. MS/MS	# Quality Unass. MS/MS
6a	Run 2	NM	94,379	15,998	78,381	6,163	7.86	72,218	13,375	65,006	7,212
		CAFM				8,872	11.32	69,509	15,681	62,703	6,809
		RFM				8,854	11.30	69,527	15,622	62,759	6,768
		RFM_KG				10,576	13.49	67,805	16,983	61,398	6,407
		RMPS				13,426	17.13	64,955	19,139	59,242	5,713
	Run 3	NM	85,191	14,685	70,506	5,301	7.52	65,205	6,752	63,754	1,451
		CAFM				7,918	11.23	62,588	9,215	61,291	1,297
		RFM				8,305	11.78	62,201	9,522	60,984	1,217
		RFM_KG				9,911	14.06	60,595	11,041	59,465	1,130
		RMPS				12,413	17.60	58,096	13,390	57,119	977
6b	Run 1	NM	97,907	16,705	81,202	7,434	9.15	73,768	16,380	70,222	8,946
		CAFM				8,391	10.33	72,811	16,921	64,281	8,530
		RFM				8,234	10.14	72,968	16,802	64,400	8,568
		RFM_KG				10,778	13.27	70,424	18,698	62,504	7,920
		RMPS				12,077	14.87	69,125	19,543	61,659	7,466
	Run 2	NM	97,301	16,504	80,797	7,498	9.28	73,299	16,893	63,904	9,395
		CAFM				8,517	10.54	72,280	17,594	63,203	9,077
		RFM				8,327	10.31	72,470	17,447	63,350	9,120
		RFM_KG				10,715	13.26	70,082	19,182	61,615	8,467
		RMPS				11,927	14.76	68,870	20,124	60,673	8,197

The effects of two common filtering parameters (deltCN and high mass accuracy) on MS/MS peptide assignment were examined by determining the quantity of MS/MS spectra not assigned to the same peptide in multiple database searches (Supporting Text). These results (**Figure 7.2**) suggest that filtering on high mass accuracy rather than deltCN can decrease ambiguous peptide-spectrum matches and provide more consistent and reproducible MS/MS identifications. In order to maintain high specificity and accuracy with increasing metagenomic sequence data, a false discovery rate (FDR) was estimated at the peptide level using an established method of reverse database searching[96,126] for each metagenomic processing method for a total of 6 target-decoy databases (RM, RFM, CAFM, KG, NM_KG, RMPS-6b). Because we are using methods that directly measure peptides, not proteins, the FDR was estimated at the peptide level. In addition, we are primarily comparing the performance of all databases by peptide-spectrum matches, not proteins, given the nature of the metagenomic processing methods and their corresponding databases (i.e., not all databases contain assembled contigs, but only reads). It has previously been noted[75] that false discovery rates can be difficult to accurately determine with metaproteome datasets due to problems associated with massive peptide degeneracy. In this study, for example, of all the identified peptides for 6a (Run 2), only 7-30% were unique peptides from each database. Consequently, if only unique peptides are used, the false discovery rate

would be overestimated; on the contrary, if all peptides are used the false discovery rate could be underestimated[75]. Therefore, to set a static FDR threshold and filter multiple databases (6 sequence databases in this study) of different sizes and internal levels of peptide redundancy to that threshold (i.e., 1%) becomes a challenge, in this case, for comparing and identifying the best metagenomic processing method for MS/MS database searching and peptide-spectrum matching. As the level of redundancy affects the FDR, we have chosen a set of fixed scoring filters in order to accurately compare database assignments. Thus, the same filter criteria to all database searches (i.e., Xcorr and ppm filtering) was applied to all database searches with a requirement that the FDR be less than or equal to, i.e., 2.0%. The FDRs for the 1-peptide level, deltCN 0.0, with and without HM filtering were 1.17%-2.03% and 16.09-31.47%, respectively for 6b, Run 1 (**Table 7.4**). The 2-peptide level and deltCN 0.08 filtered reverse database searches serve to represent the FDR of peptide identifications found in Table 1. The FDRs for these PSMs, with and without HM filtering were within 0.09%-0.38% and 2.17-4.15%, respectively for 6b, Run 1 (**Table 7.5**). Following the application of a post-database high precursor mass accuracy filter (± 10 ppm) to both, the 1- and 2-peptide filtered forward-reverse datasets, the number of identified reverse peptides decreased by, on average, 93% for each database which resulted in a reduction of the FDR to 0.09%-0.38%.

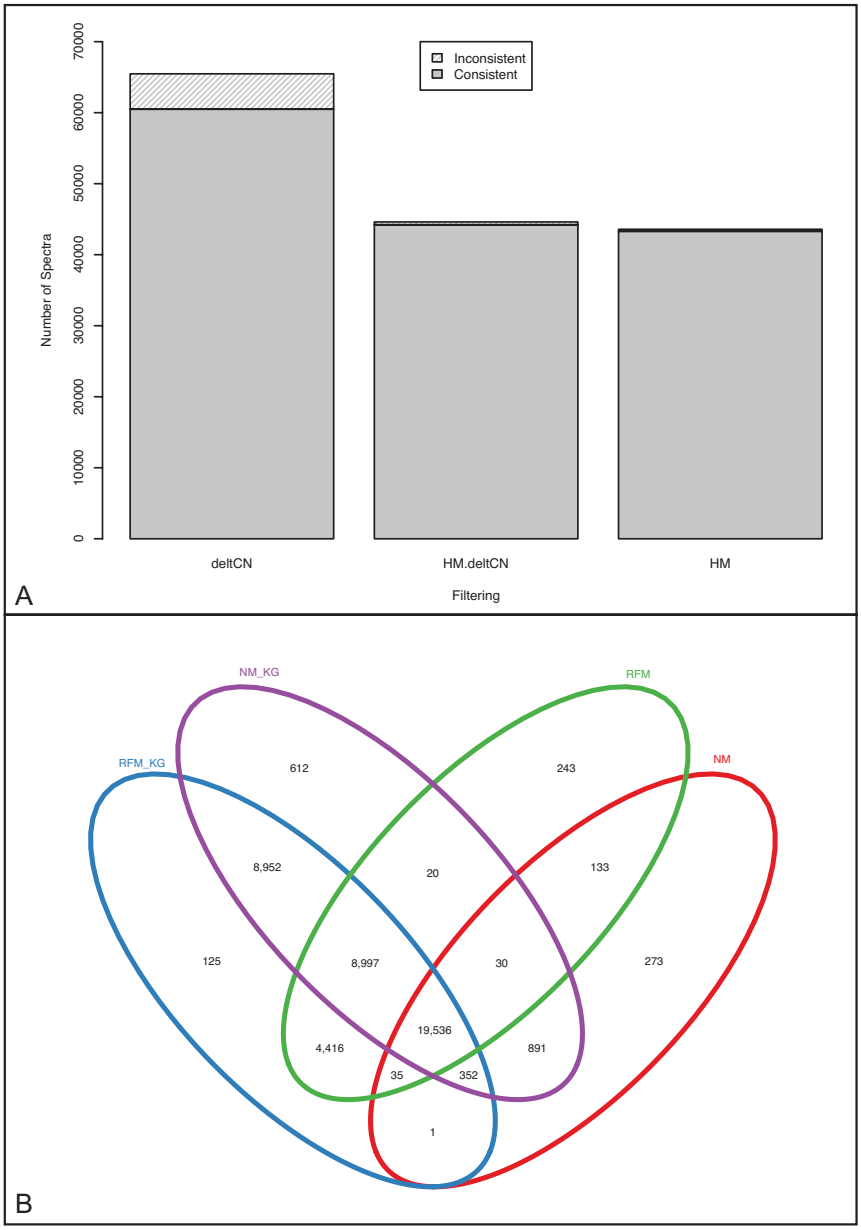


Figure 7.2: Accuracy Assessment by DTASelect Filtering. (**A**) For each DTASelect peptide prediction search, the number of identified spectra was calculated and compared using three different parameter combinations, deltCN filtered results at a deltCN of 0.08 only, both deltCN of 0.08 and HM (±10 ppm), and HM (±10 ppm) only, where identified peptide sequences were designated either 'Consistent' (solid gray) or 'Inconsistent' (diagonal stripes). (**B**) A VENN diagram with assignable spectra for RFM, RFM_KG, NM, and NM_KG databases, filtered by high mass accuracy, for both samples combined.

Table 7.4: False discovery rates for sample 6b (Run 1) against six different metagenomic-predicted sequence databases. The database results were filtered at a 1-peptide level with and without high mass accuracy.

Sample	Database (Forward/Reverse)	Total Identified Forward		% of Forward Identified		Total Identified Reverse		False Discovery Rate	
		<±10 ppm	>±10 ppm	<±10 ppm	>±10 ppm	<±10 ppm	>±10 ppm	Total FP	FP <±10 ppm
6b, Run1	RM	19,589	8,747	69.13%	30.87%	321	4,551	29.34%	1.93%
	RFM	18,229	8,375	68.52%	31.48%	320	4,648	31.47%	2.03%
	CAFM	17,443	7,676	69.44%	30.56%	277	4,094	29.64%	1.88%
	KG	20,059	7,672	72.33%	27.67%	256	3,419	23.40%	1.63%
	NM+KG	23,881	8,709	73.28%	26.72%	256	3,603	21.17%	1.40%
	RMPS	27,218	8,722	75.73%	24.27%	228	2,917	16.09%	1.17%

Table 7.5: False discovery rates for sample 6b (Run 1) against six different metagenomic-predicted sequence databases. The database results were filtered at a 2-peptide level with and without high mass accuracy.

Sample	Database (Forward/Reverse)	Total Identified Forward		% of Forward Identified		Total Identified Reverse		False Discovery Rate	
		<±10 ppm	>±10 ppm	<±10 ppm	>±10 ppm	<±10 ppm	>±10 ppm	Total FP	FP <±10 ppm
6b, Run1	RM	15,363	3,523	81.35%	18.65%	27	310	3.51%	0.28%
	RFM	13,971	3,057	82.05%	17.95%	33	328	4.15%	0.38%
	CAFM	13,783	2,902	82.61%	17.39%	16	291	3.61%	0.19%
	KG	16,840	3,695	82.01%	17.99%	15	248	2.53%	0.14%
	NM+KG	19,898	4,320	82.16%	17.84%	11	255	2.17%	0.09%

7.3.2: Tracking Missing Peptides

By adding the unrelated KG metagenomic sequences to the RFM protein database, the number of additional predicted unique peptide sequences increased by 40%. Therefore, we wanted to determine how many additional peptide-spectrum matches were gained by adding these KG proteins sequences to the database. The RFM_KG assigned MS/MS were distributed into three different categories: RFM only, KG only, and RFM plus KG (shared) for each sample (**Table 7.6**). The majority of RFM_KG assigned spectra were “shared” between both RFM and KG protein sequences. About 26% of the total spectrum assignments were unique to the RFM protein sequences (zero overlap with KG sequences) and only ~8% of the spectra were unique to the KG protein sequences (no overlap with the RFM sequences) (**Table 7.6**).

Table 7.6: Distribution of RFM_KG assigned PSMs for 6a (Run 2 and 3) and 6b (Run1 and 2). The assigned PSMs were distributed into three different categories: RFM only, KG only, and RFM plus KG based on their sequence uniqueness to each set of sequences. If a PSM was unique to protein sequences in RFM, but was not present in KG, the PSM was classified and categorized as RFM only and vice versa. If a PSM was found to match a protein in both, RFM and KG, the PSM was categorized as a shared spectrum.

RFM_KG Database										
Sample	RFM Only			Kurokawa & Gill (KG) Only			RFM and KG			Total Assigned Spectra
	Unique Spectra	% of RFM & KG Spectra	% of Total Spectra	Unique Spectra	% of RFM & KG Spectra	% of Total Spectra	Shared Spectra	% of RFM & KG Spectra	% of Total Spectra	
6a, Run 2	8,012	34.33%	30.35%	1,454	6.23%	5.51%	13,869	59.43%	52.53%	26,403
6a, Run 3	6,995	31.20%	27.73%	1,341	5.98%	5.32%	14,085	62.82%	55.84%	25,223
6b, Run 1	5,651	27.09%	23.64%	2,431	11.65%	10.17%	12,779	61.26%	53.46%	23,902
6b, Run 2	5,595	28.29%	23.93%	2,149	10.86%	9.19%	12,036	60.85%	51.48%	23,378
Runs Averaged	6,563	30.23%	26.41%	1,844	8.68%	7.55%	13,192	61.09%	53.33%	24,727

There are two possible hypotheses for why the metagenomes from these samples (i.e., RFM) cannot be used to assign peptides to spectra which are assignable by the unrelated protein database KG: (1) because of low sequencing depth, peptides are not assigned because our protein database is incomplete or (2) because of a sequencing error or limitation for predicting ORFs, we are unable to predict the proteins that are present. Therefore, we have aligned the RFM_KG (2-peptide, deltCN 0.08, HM filtered) identified peptides (**Figure 7.3**, y-axis) from 6a (left panel) or 6b (right panel) to predicted raw reads from the related/same sample (6b) and an unrelated sample (16b) (Fig. 7.3, x-axis) using TFASTS (Fig. 7.3, white, fine striped bars). Those results were compared to alignments of the same identified peptides to the predicted protein database from the related/same sample (6b) and the unrelated sample (16b) using FASTS[193] (Fig. 7.3, gray, solid bars). As expected, more peptides mapped to the related/same (matched metagenome-metaproteome) sample (15% for 6a: left panel, Fig. 7.3 and 6b: right panel, Fig. 7.3) than to the unrelated, 16b, predicted protein sequences (8% for 6a and 10% for 6b). When these same peptides were compared using TFASTS (algorithm that compares peptides to DNA sequence) to the raw sequencing reads (Fig. 7.3, white, fine striped bars), the number of peptides matching to reads increased by two-fold for both 6a and 6b.

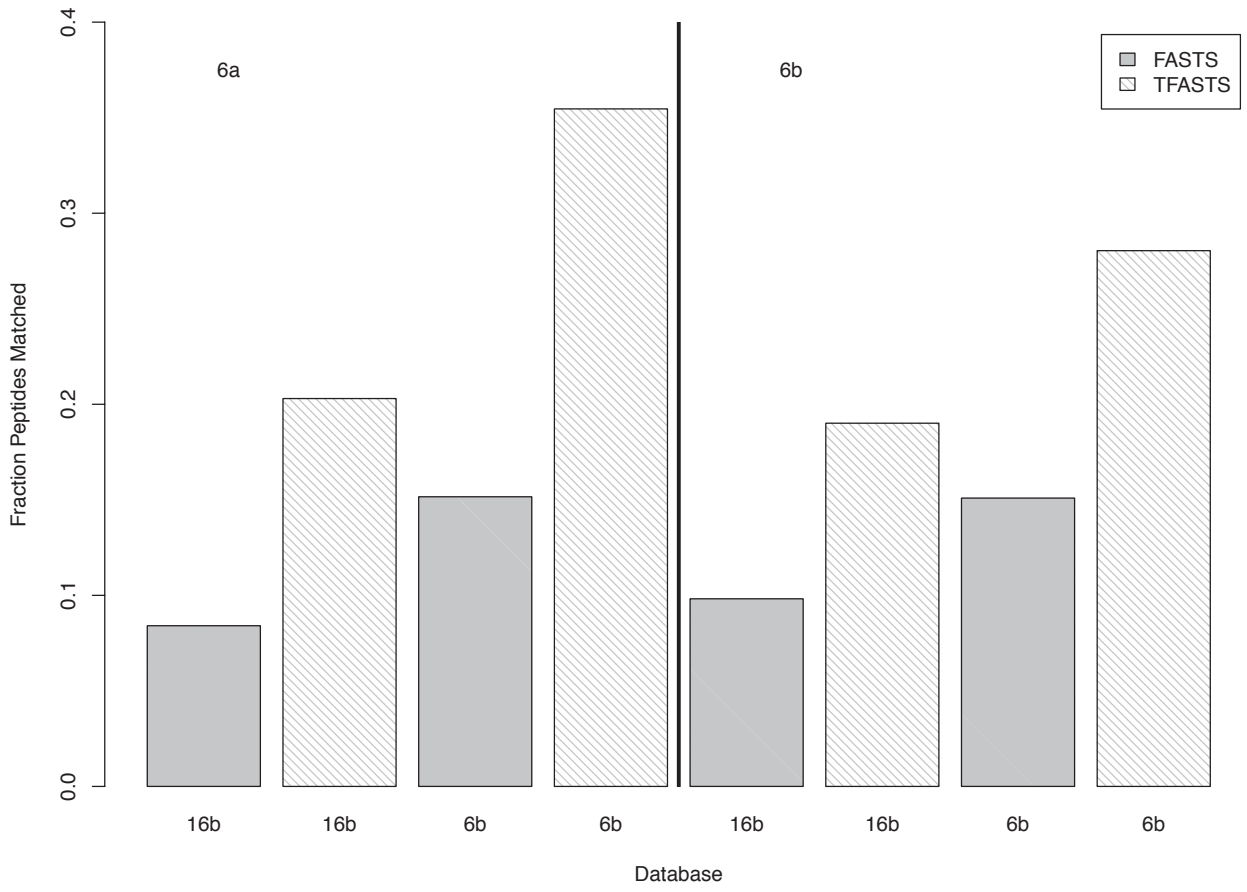


Figure 7.3: Comparison of identified peptides using sequence similarity techniques. Percentage of matches found when comparing identified peptides from sample 6a (left panel) or 6b (right panel) to predicted proteins using FASTS (gray bars) and raw sequencing reads using TFASTS (white striped bars).

7.3.3: Targeting Peptide Discovery

Throughout the course of our study, we were able to accumulate more metagenomic sequence data for the two healthy samples, 6a and 6b, by ~5 fold (**Table 7.2**, italicized text). Although this increase in predicted ORFs resulted in an increase in the number of assigned MS/MS spectra, it can reduce the throughput of MS/MS sequence-database searching. Therefore, we investigated the impact of searching a metagenomic-based protein database derived from the exact same single sample to that of a concatenated sequence library of all available metagenomic data from this study. The additional metagenomic sequences were used to construct a sequence database similar to that of

RM (non-assembled reads with 5.6 million predicted unique peptides), called RMPS (**Figure 7.1**) which has ~ 1.3 million predicted unique peptides, on average, per healthy sample 6a and 6b. Searching the RMPS sequence databases with SEQUEST using standard 2-peptide, deltCN 0.08, and high mass accuracy filtering decreased the compute time to ~ 300-500 minutes per MS raw file. By increasing the amount of metagenomic sequence data for a single sample, the total number of assigned spectra increased by 63% (from 34,666 to 56,782) and the number of total identified non-redundant (NR) peptides increased by 67% (from 9,618 to 16,055) (**Table 7.1**, RM versus RMPS), resulting in a 54% increase in protein identifications (3,394 to 5,233) when mapping these peptides to a protein dataset generated from assembled reads for the exact same metagenomic sample.

Other than limitations associated with computational resources, there was also a concern that real peptides predicted from 454-reads would be filtered out given a 2-peptide per protein minimum filter (**Table 7.7**, top panel). Therefore, the filtering parameters were readjusted with a deltCN 0.0, 1-peptide minimum, and a high mass accuracy filter (±10 ppm) for the SEQUEST RMPS database searches for both 6a (Run 2 and 3) and 6b (Run 1 and 2). The identified peptides were then mapped back to the predicted protein sequences derived from the assembled metagenomic sequences with a 2-peptide filter, resulting in an increase of protein identifications, from 5,233 to 6,186 (Table 2, RMPS top panel versus bottom panel). The filtering parameters were also readjusted with a deltCN 0.0 and a high mass accuracy filter (±10 ppm) for the SEQUEST-RFM database searches for both 6a (Run 2 and 3) and 6b (Run 1 and 2). The protein identifications also increased, from 3,431 to 3,706 (**Table 7.7**, RFM top versus bottom panel). While this increase might seem minimal, there is significantly less redundancy, less false positives, and no computational cost added to these filtering parameters. The false discovery rate, using the same filtering parameters (deltCN 0.0, 1-peptide minimum and HM) for the RMPS database was 1.17% for 6b (Table 7.4), however, these identified peptides (≥ 1 peptide/read) were mapped back to the predicted protein sequences derived from the assembled metagenomic sequences

using a post-database ≥2-peptide/protein filter. Following application of this 2-peptide/protein filter, the FDRs dropped to 0.1%-0.2% for 6b, Run 1 and 2 (**Table 7.8**).

Table 7.7: Comparison of RFM and RMPS database results with different filtering metrics and a post-database mapping strategy. Comparison of SEQUEST/DTASelect database search results, non-redundant spectra and protein counts with different filtering parameters and HM, post-database mapping of identified peptides to a protein dataset generated from assembled reads for the same metagenomic sample.

Protein Database	**RFM**		**RMPS**	
2-peptide, deltCN 0.08, HM Filter				
	Spectra	Protein	Spectra	Protein
6a Run 2	3,246	1,154	6,542	1,761
6a Run 3	3,091	1,010	6,237	1,544
6b Run 1	2,639	637	5,212	973
6b Run 2	2,552	630	4,870	955
Total	11,528	3,431	22,861	5,233
1- or 2-peptide, deltCN 0.0, HM Filter				
	Spectra	Protein	Spectra	Protein
Peptide Criteria	≥ 2 peptide		≥ 1 peptide	
6a Run 2	3,541	1,252	7,497	2,069
6a Run 3	3,346	1,088	7,048	1,808
6b Run 1	2,879	686	5,881	1,182
6b Run 2	2,786	680	5,502	1,127
Total	12,552	3,706	25,928	6,186

Table 7.8: False discovery rates for sample 6b (Run 1 and 2) against the RMPS database. An initial ≥1-peptide, deltCN 0.0, and high mass accuracy (±10ppm) filter were applied to the read-based identifications followed by a ≥2-peptide/protein post-database mapping filter.

Sample	Total Peptides	Non-redundant Peptides	Total Identified Forward Peptides	Non-redundant Forward Peptides	Total Identified Reverse Peptides	Non-redundant Reverse Peptides	Total FDR (%)	Non-redundant FDR (%)
6b, Run 1	5,500	3,765	5,498	3,763	2	2	0.07	0.11
6b, Run 2	5,325	3,538	5,316	3,535	9	3	0.34	0.17

7.3.4: *De novo* Peptide Sequencing

Two popular algorithms, PepNovo+[105] and PEAKS[109], were used to identify peptide sequences *de novo* from MS/MS spectra collected from both samples, independent of all protein sequence databases. Initially, the two algorithms were run independently on the same raw MS data and samples as described. The identified, high confidence consensus sequence tags (≥3 residues) were acquired from each *de novo* algorithm. The *de novo* consensus sequence tags (Supporting Text) for PEAKS and Pepnovo+ were compared for every MS/MS to identify the partial (≥3 residues) and exact consensus sequence tags that would represent the most confident PSMs identified by the two different *de novo* algorithms. In this study, it was not our goal to compare the performance of the two programs; instead, we want to combine the best results from the two programs using their respective optimum parameters. The final, representative *de novo* consensus tags were compared to the previously mentioned SEQUEST results from the RMPS sequence database searches that were filtered at a ≥1 peptide/read, deltCN 0.0, and high mass accuracy with a post-database ≥2 peptide/protein filters. On average, ~593-724 MS/MS spectra were assigned with a high confidence consensus peptide sequence between the two *de novo* algorithms, but were not assigned with the SEQUEST–RMPS database search (**Figure 7.4**). These *de novo* peptide sequences were mapped to protein sequences predicted from assembled contigs with a 2-peptide minimum per protein and compared to the peptides that were identified from the SEQUEST-RMPS database searches. A total of 421 new, non-redundant proteins were identified with the *de novo* sequenced peptides for metagenome 6b, and 333 non-redundant proteins for metagenome 6a; these proteins were not identified using SEQUEST. Approximately 450 *de novo* sequenced peptides (non-redundant) per sample could not be mapped to the matched metagenomic sequence data.

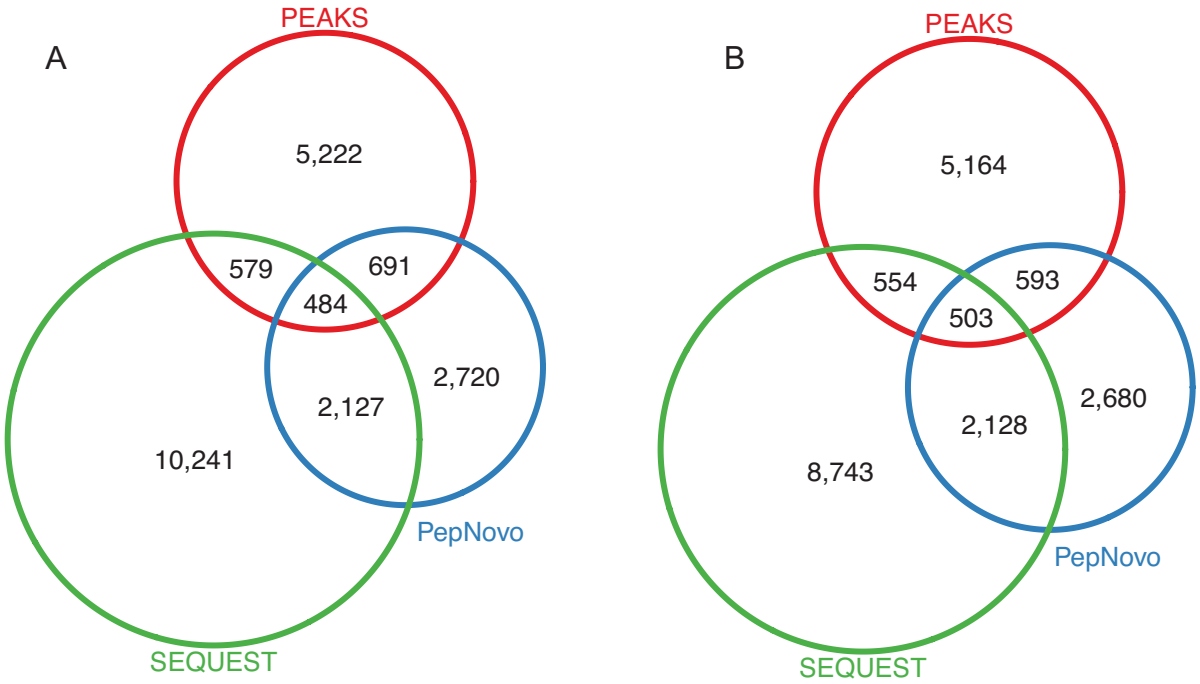


Figure 7.4: Performance and comparison of de novo peptide sequencing results. Distribution of assigned spectra per *de novo* algorithm with a predicted consensus sequence (partial and/or exact sequence match) among all three algorithms, PEAKS, PepNovo+, and SEQUEST. Identified peptides from SEQUEST and RMPS sequence database were compared to the *de novo* predicted peptides for (A) 6a and (B) 6b.

7.4: Discussion

One of the major goals of MS-based proteomics is to comprehensively identify the protein complement of a given sample (isolate, mixture, or community). The proteome(s) of microbial communities are highly complex and pose numerous challenges for MS experimentation and analysis. These challenges include the dynamic range of peptide abundances and a number of informatics hurdles, such as differentiation between closely related species, identification of sequence polymorphisms, and global identification of post-translational modifications. Many of the algorithms used in MS/MS database searching are based on the assumption that a protein is derived from a single organism with little sequence diversity. However, these assumptions are no longer valid in the case of complex microbial communities. This

study presents several strategies for improving metagenomic guided MS-based metaproteomic peptide-spectrum matching in complex samples.

It has become very clear that the quality of metagenomic sequence data and resulting protein sequence database has a significant impact on community MS-based proteomics and the ability to achieve deep proteome coverage. This study initially explored how assembly and gene finding methods for metagenomic sequences affects peptide-spectrum matching. Our findings suggest that predicting ORFs from an *ab-initio* gene finder on metagenomic reads provides the best database for maximal MS/MS assignment. While assembly of metagenomic data can greatly reduce the necessary compute time for gene finding and database searching, it essentially collapses sequence diversity; thus, it is sub-optimal for maximal spectral assignment. Yet, introducing a homology-based gene finding method (RFM) does not increase the number of assigned spectra. Lastly, with an increase in sequence coverage for a biological sample, our results suggest that predicted protein sequence databases derived from matched metagenomic sequenced reads (RMPS), increases the number of MS/MS spectra, peptides, and protein identifications. In conclusion, expanding the metagenomic sequence library for matched or related samples improved peptide-spectrum matching. However, improvements in gene finding are equally important to maximize protein identification and coverage.

As the matched metagenomic predicted protein sequence database (RMPS) more accurately reflected the “true proteome”, previously unassigned high-quality spectra are now being identified and provided greater proteomic depth. When these results were compared to a standard bacterial isolate (e.g., *E. coli*) with a well-curated genome, ~ 41,000 MS/MS spectra were assigned to peptides (37% of total collected MS/MS) (data not shown) using the same database searching filters (≥2 peptide and deltCN 0.08). This would suggest that underlying challenges are still inhibiting the identification of a majority of spectra collected from the community samples compared to that of a standard bacterial isolate. The classification of acquired and assigned MS/MS spectra and quantification of total identified peptides suggested that the RMPS processing method provided the most comprehensive assignment of MS/MS spectra.

When we examine why some peptides are assigned from the read-based ORFs (e.g., RMPS processing method) and not assigned from the contig-based ORFs (e.g., NM processing method), we find that these “lost peptides” fall into three categories: (i) some reads are not assembled and therefore their protein predictions are not in the contig-based ORF predictions, (ii) because of SNPs and frameshifts, the peptides are 100% similar to a predicted contig-based ORF, but are not 100% identical, and (iii) some peptides were very different (<50% identical) or missing from the contig-predicted ORF. A 6-frame translation protein database was generated for sample 6a to capture all possible candidate peptide sequences and searched against one MS experiment (Run 2). However, routine use of this sequence database is impractical due to the increased quantity of sequences which directly correlates with an increased quantity of candidate peptides, therefore, more scoring and prohibitively large search times (~134 hrs per MS experiment) (data not shown). As sequencing data generation increases, even a read-based strategy could become unsustainable, which will only worsen as new larger ‘omic’ datasets become available. The testing and comparison of new search algorithms that are faster, accurate, and developed for omic’ datasets may help researchers overcome these challenges.

Identifying the most reliable set of peptides from a MS-based metaproteomic experiment can be complicated, as we have shown that MS/MS assignments can vary and be assigned to different peptide sequences with different protein databases. While filtering on deltCN is a common practice for reducing false positives, this type of filtering may (i) continue to include many ambiguous peptides based on the different database predictions and (ii) remove many legitimate peptides as a result of a highly redundant database. Although filtering on deltCN and peptide-protein matches has proven effective for single genome searching, these filters decrease both precision and sensitivity in metagenomic predicted sequence databases. As common filtering strategies have proven to be less effective and practical for large-scale proteomics studies (e.g., post-translational studies), these and other challenges will surface as the MS field moves towards sampling more environmental communities. Alternatively, we propose that when high mass accuracy is used in conjunction with other filtering

metrics, such as, cross correlation (XCorr) and enzyme cleavage specificity, one can confidently identify the most comprehensive and reproducible set of PSMs and control false positives adequately in a complex environmental community sample. As shown, this strategy greatly reduces the rate of ambiguous peptide predictions thereby giving higher confidence to our final peptide-protein identifications. Once peptides are identified and mapped to metagenomic sequences, which have been assembled, the subsequent use of a 2-peptide filter greatly reduces the number of false positives in protein discovery for complex microbial environments.

Finally, *de novo* peptide sequencing can complement MS/MS database searching to identify peptides absent in the protein sequence database due to the limitations of the gene finding algorithms or low metagenomic sequence coverage. We believe that novel peptides were identified with high confidence in this study, because these peptides were independently identified by two *de novo* sequencing algorithms. However, there is no widely accepted method for us to use for rigorously evaluating the FDRs of novel peptides identified from our microbial community samples. Thus, *de novo* sequencing results should be used with the caveat of uncertain FDRs as supplement to database searching results[194].

7.5: Conclusions

By using a variety of MS filtering metrics, we were able to assess the quality and accuracy of MS/MS peptide sequencing for each MS experiment against four predicted protein sequence databases derived from whole genome shotgun sequences. Our findings suggest that: (i) proteomic data is twice as likely to match metagenomic data derived from the same sample, (ii) although unrelated metagenomic data may capture more sequence diversity, large protein databases can create unreasonable sequence redundancy, thereby hampering the ability to differentiate real peptide-protein identifications, (iii) the percentage of unassigned, high-quality MS/MS spectra decreases with increased quality of metagenomic sequences, (iv) metagenomic data processing, such as assembly and gene finding, affects the ability to assign peptides to spectra, (v) MS filtering metrics can affect the accuracy of peptide-spectrum matching,

(vi) deeper metagenomic sequencing coverage results in deeper coverage of matched metaproteomes and (vii) *de novo* peptide sequencing can overcome potential sequencing errors and provide evidence for novel sequences not yet sequenced or not identified by database searching methods. The high-quality unassigned MS/MS from sequence-database searching would be ideal target spectra to submit for *de novo* peptide sequencing whereby these sequences could be mapped back to help refine the metagenome and identify potential sequencing errors. Finally, this study illustrates how common metagenomic processing methods (assembly and ORF finding) and database construction can affect metaproteomics search results.

Chapter 8

Meta-omics reveals human host-microbiota signatures of Crohn's disease

The text is adapted from:

> *Alison R. Erickson, Brandi L. Cantarel, Regina Lamendella, Youssef Darzi, Emmanuel F. Mongodin, Chongle Pan, Manesh Shah, Jonas Halfvarson, Curt Tysk, Bernard Henrissat, Jeroen Raes, Nathan C. Verberkmoes, Claire M. Fraser-Liggett, Robert L. Hettich, and Janet K. Jansson. "Meta-omics reveals human host-microbiota signatures of Crohn's disease." Draft will be submitted to the journal, Molecular Systems Biology (2012).*

Alison R. Erickson's contributions include experimental preparation of all microbial samples for proteomics, experimental LC-MS/MS measurements, integrated matched MG-MP comparisons and analyses, biological inference of human proteins, and shared primary authorship with Brandi Cantarel and Regina Lamendella.

8.1: Introduction

Humans live in close association with communities of microorganisms (the human microbiota) that inhabit every exposed surface and cavity in the body[195]. The collective genetic information of the human microbiota represents a second genome, the human microbiome, currently the focus of intense international sequencing and research efforts[14],[196],[25]. Although most human host-microbe associations are beneficial, changes in the composition and function of the human microbiota are associated with a growing list of diseases, including inflammatory bowel disease (IBD)[197]. Several studies using both culture-dependent and molecular approaches have suggested that there is a dysbiosis in the gut microbiota of patients with Crohn's disease (CD) compared to healthy subjects[158],[198],[199].

Recent advances in DNA sequencing and proteomics technologies have opened the door to explore the structure and function of the gut microbiota without the necessity for cultivation. However, there have been very few reports to date that have used a multi-"omics" approach to study the complex ecosystem in the human gut. The ability to combine information about the identities of microbial community members (obtained from 16S rRNA gene-based measurements), metabolic potential (obtained from

metagenome sequence data) and expressed proteins (obtained from metaproteome data) enables explorations of the gut microbiota at multiple molecular levels simultaneously[200].

In the previous chapter, we compared metagenomic data processing methods, such as assembly and gene finding, and their affects on the ability to assign peptides to MS/MS spectra. Using the best performing informatics workflow, predicted protein sequence databases derived from matched metagenomic sequenced reads (RMPS), we can now apply this workflow to focus on biological inference and human disease as opposed to chapter 6, which focuses only on measurements of a healthy twin pair. This study was focused on a subset of fecal samples collected from a large Swedish twin cohort with inflammatory bowel disease (IBD) that was previously characterized, with respect to their bacterial community composition by deep 16S rRNA pyrotag sequencing[158,198,201] and metabolite profiling[202]. Previous data indicated that healthy twin pairs had a similar gut microbiota, even when they had been living separately for decades. By contrast, twin pairs in which one or both subjects had CD harbored very dissimilar gut microbial compositions[158]. This disparity of the gut microbiota was particularly striking for subjects with inflammation in the ileum (ileal CD, ICD) compared to healthy subjects[158],[202],[201] and was primarily characterized by the reduced abundance of several key beneficial members of the community, such as *Faecalibacterium prausnitzii*.

Here our aim was to further explore a subset of the same Swedish twin cohort that had demonstrated microbial dysbiosis in fecal samples according to CD phenotype, for functions that were correlated to CD by applying non-targeted metagenomics12 and metaproteomics[26]. This eco-systems biology approach[200] allowed us to detect and directly correlate genes and expressed proteins for the first time in the same samples. It was particularly valuable to include discordant twin pairs in the sample set, where one twin was diseased and one was healthy, to mitigate the influence of host genetics on the microbiome[158],[112],[150]. The sample cohort included one healthy twin pair with existing metaproteome data[26], one colonic Crohn's (CCD) twin pair, two ICD concordant twin pairs and two ICD discordant twin pairs (**Table 2.1**). All samples were

collected when the Crohn's subjects were in endoscopic remission, or had minor inflammatory activity in the neo-terminal ileum only.

8.2 Experimental Methods

8.2.1: Sample collection

Fecal samples were collected from 6 monozygotic twin pairs: 1 healthy twin pair with existing metaproteome data[26], 1 concordant pair with CCD, 2 concordant twins with ICD and 2 ICD discordant twin pairs (**Table 2.1**). DNA was extracted by the MoBio[158] and IGS-Zymo[203] protocols producing 3 – 5ug of purified metagenomic DNA from each sample. Proteins were extracted and processed for 2D-LC-MS/MS as previously described[26].

8.2.2: Metagenomics

Sequences were generated and processed using the 454 Titanium Roche platform and assembled using Newbler (v2.0.01.14). Genes were predicted on contigs greater than 500 bp using METAGENE[187] and for those less than 500 bp using a combination of METAGENE and FASTX[188] from alignments to homologous sequences in reference genomes. Proteins were clustered using BLASTP[204], using e-value cutoffs of 10^{-5}, and MCL[205], with an inflation value of 1.5. ORFs were searched against the eggnog[206], CAZY[129] and KEGG genes[207] using NCBI-BLAST[204] using e-value cutoff of 10^{-6} and bits per position cutoff of 1.

8.2.3: Metaproteomics

MS-based shotgun proteomics was performed as described[26] and acquired MS/MS searched against two databases: 1) matched metagenomic-predicted protein database (MM) and 2) human microbial reference genome database (HMRG); both including the human genome. The spectral count for a microbial protein cluster was calculated as the number of unique peptide identifications that can be attributed to proteins from that cluster only and not any other cluster. Spectral counts for human proteins were

calculated from both unique and non-unique peptide identifications. All spectral counts were normalized by the total numbers of MS/MS spectra per run.

8.2.4: Statistics

Non-metric multidimensional scaling was performed using normalized spectral abundances of identified proteins. Protein lists were generated for proteins that correlated with a particular phenotype using Indicator Species Analysis and Wilcoxon's rank sum tests.

8.3: Results

We generated whole genome shotgun metagenomic (**Table 8.1**) and shotgun mass spectrometry (MS)-based metaproteomic (**Tables 8.2** and **8.3**) datasets from the same samples for direct comparisons. Metagenomic data were used to assess community gene content and predicted functional capability, while metaproteomics was used to identify the most abundant expressed microbial and human proteins. The number of genes identified in each sample was two orders of magnitude greater than that of proteins (since proteins cannot be amplified like DNA), although both represent only the more abundant fraction of the total gene and protein reservoir in the human gut microbiome.

Table 8.1: Metagenomic sequence data and statistics.

Subject	Mega base pairs (Mbp)	Number of Reads (thousands)	Number of Singletons (thousands)	Number of Contigs	Mbp in Contigs	Average Contig Length	Contigs > 500 bp	Fraction Singleton	Fraction Contigs > 500
6a	684	1,901	299	55,650	66	1195	39,318	0.157	0.7065
6b	409	1,183	372	55,463	43	776	33,375	0.3147	0.6018
9a	245	782	155	19,893	21	1053	13,439	0.1988	0.6756
9b	290	845	197	35,571	27	754	20,011	0.2328	0.5626
10a	502	1,211	82	36,328	49	1337	24,726	0.0681	0.6806
10b	541	1,327	140	30,642	39	1259	20,219	0.1053	0.6598
15a	504	1,358	256	38,046	42	1109	25,909	0.1887	0.681
15b	484	1,270	272	36,871	36	981	21,555	0.2146	0.5846
16a	229	585	207	22,764	18	798	15,545	0.3544	0.6829
16b	319	1,126	511	46,845	28	591	23,857	0.4541	0.5093
18a	425	1,341	215	26,803	27	1014	17,971	0.1604	0.6705
18b	258	955	372	28,631	19	671	13,862	0.3898	0.4842

Table 8.2: Normalized total spectra counts across all subjects and 24 MS runs for the matched metagenome (MM) database searches.

Phenotype	Sample	Run	Proteins	Peptides	Total Identified MS/MS	Total Collected MS/MS	#DB Entries
healthy	6a	2	2,315	9,110	15,724	78,381	1,356,947
		3	2,009	8,005	15,452	70,492	
	6b	2	1,385	5,548	11,300	80,797	862,006
		1	1,413	5,796	11,258	81,202	
colonic CD	9a	1	871	3,738	7,628	92,865	527,904
		2	829	3,381	7,879	91,745	
	9b	1	723	2,551	5,153	82,546	594,986
		2	728	2,532	5,074	84,089	
ileal CD	10a	3	1,089	4,474	8,891	90,775	1,428,694
		1	1,049	4,286	8,244	91,145	
	10b	1	1,118	4,078	8,057	75,873	1,618,290
		2	1,141	3,984	8,546	74,574	
ileal CD	15a	1	946	3,913	9,276	83,254	1,005,286
		2	1,183	4,862	10,616	77,906	
	15b	2	787	3,890	8,647	81,970	982,092
		1	769	3,421	8,619	80,718	
ileal CD	16a	2	369	1,716	4,079	78,811	1,119,221
		1	407	1,844	4,012	73,878	
healthy	16b	2	1,248	4,581	11,659	92,460	843,556
		3	1,256	4,416	11,914	91,865	
ileal CD	18a	2	687	3,233	7,951	76,489	781,500
		1	654	2,733	7,281	78,661	
healthy	18b	1	794	3,018	7,640	92,795	628,384
		2	829	3,245	7,442	93,485	

Table 8.3: Normalized total spectra counts across all subjects and 24 MS runs for the human microbial isolate reference genome database (HMRG) searches.

Phenotype	Sample	Run	Proteins	Peptides	Total Identified MS/MS	Total Collected MS/MS
healthy	6a	2	3,138	8,679	13,254	78,381
		3	2,618	6,838	11,351	70,492
	6b	2	2,716	6,653	11,022	80,797
		1	2,612	6,308	10,919	81,202
colonic CD	9a	1	2,477	5,910	10,089	92,865
		2	2,337	5,416	10,388	91,745
	9b	1	2,089	5,231	8,485	82,546
		2	2,062	5,161	8,354	84,089
ileal CD	10a	3	2,172	5,997	10,199	90,775
		1	1,985	5,635	9,291	91,145
	10b	1	1,859	4,483	7,964	75,873
		2	1,878	4,071	7,919	74,574
ileal CD	15a	1	1,874	5,494	10,291	83,254
		2	2,122	6,783	11,698	77,906
	15b	2	2,146	6,235	9,816	81,970
		1	1,916	5,430	9,695	80,718
ileal CD	16a	2	1,733	4,738	7,728	78,811
		1	1,840	4,956	7,830	73,878
healthy	16b	2	3,049	6,926	12,610	92,460
		3	3,108	7,070	13,281	91,865
ileal CD	18a	2	1,964	5,931	11,161	76,489
		1	1,648	5,108	10,350	78,661
healthy	18b	1	2,855	6,830	12,669	92,795
		2	3,139	7,626	12,642	93,485

8.3.1: Taxonomic Community structural differences

Taxonomic profiles of the metagenomic data were determined using nucleotide alignments and compared based on disease status (healthy, CCD, ICD). Greater than 60% of the metagenomic sequence reads in the healthy samples could not be assigned at the phylum, family or genus level, as compared with ~40% of the reads in ICD or CCD subjects, potentially resulting from reduced diversity in CD. Of the metagenomic reads for which a taxonomic assignment could be made, 396 genera were represented in all of the samples, and nine of those were present at > 5%. Eleven genera of the Firmicutes phylum *(Faecalibacterium*, *Geobacillus*, *Desulfotomaculum*, *Desulfitobacterium*, *Holdemania*, *Thermoanaerobacter*, *Thermosinus*, *Carboxydothermus*, *Enterococcus*, *Alkaliphilus*, *Subdoligranulum* and *Anaerotruncus*)

and one genera of the Proteobacteria phylum (*Pelobacter)* were less abundant in the ICD compared to healthy subjects (**Table 8.4**). The depletion of *Faecalibacterium* and *Subdoligranulum* was previously reported using 16S rRNA gene sequencing in these same samples[201]. Bacteria that were more abundant in the microbiota of ICD compared to healthy subjects included genera of *Pasteurella*. These data are consistent with other reports of dysbiosis of the microbiota in subjects with CD[198], however, the list of differentially abundant genera are not entirely consistent between studies, which most likely reflects the different cohorts that have been studied to date.

Table 8.4: Relative abundances of differentially abundant species.

Taxonomy	CCD 16S	CCD WGS	ICD 16S	ICD WGS	H 16S	H WGS
Desulfitobacterium	0.00E+00	0.00E+00	0.00E+00	1.67E-06	0.00E+00	1.50E-05
Desulfotomaculum	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	5.00E-06
Geobacillus	0.00E+00	1.00E-05	0.00E+00	0.00E+00	0.00E+00	7.50E-06
Pasteurella	0.00E+00	2.00E-05	0.00E+00	8.33E-06	0.00E+00	2.50E-06
Thermoanaerobacter	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.50E-05
Enterocytozoon	0.00E+00	5.00E-06	0.00E+00	0.00E+00	0.00E+00	7.50E-06
Carboxydothermus	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	5.00E-06
Thermosinus	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.50E-06
Pelobacter	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.50E-06
Holdemania	0.00E+00	1.65E-04	6.67E-05	1.25E-04	0.00E+00	5.08E-04
Anaerotruncus	1.20E-03	1.85E-04	2.83E-04	2.00E-04	6.25E-04	1.18E-03
Enterococcus	2.45E-03	1.77E-02	0.00E+00	1.19E-03	0.00E+00	7.87E-04
Faecalibacterium	1.53E-02	2.69E-03	1.55E-03	1.65E-03	1.50E-02	1.25E-02
Neorickettsia	0.00E+00	1.50E-05	0.00E+00	0.00E+00	0.00E+00	0.00E+00
Fibrobacter	0.00E+00	5.00E-06	0.00E+00	0.00E+00	0.00E+00	7.50E-06
Alkaliphilus	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.75E-05
Subdoligranulum	8.50E-03	2.58E-03	7.75E-03	4.97E-04	3.25E-03	2.99E-03

8.3.2: Peptide-Spectrum Matching (PSM) and broad functional comparisons

In addition to taxonomic information, metagenomic data provides information on gene content of the gut microbiome; however, it does not reveal the identities and relative abundances of expressed gene products (proteins) under the conditions studied. Therefore, to directly address gene function and expression, we performed database searches with tandem mass spectra (MS/MS) of peptides from the same samples. These extensive MS/MS datasets were searched both against their corresponding matched microbial metagenome (MM) (**Table 8.1**) or a representative set of 51 sequenced human microbiome isolate reference genomes (HMRGs) (**Table 8.5**), each concatenated with the human genome. The HMRGs provide nearly complete DNA sequence coverage of a bacterial species, and the predicted genes are often full length, as compared to the MMs that are not sequenced to sufficient depth to contain complete genomes, and the predicted genes are often fragmented. However, many sequences captured from a clinical sample do not map to HMRGs. One challenge in using a HMRG database is gene redundancy between strains/species belonging to the same genera, which can make it difficult to uniquely assign a peptide to a mass spectrum. Therefore, we developed a novel method for clustering functionally similar proteins from the MMs to provide a more robust method of assigning peptide-spectrum counts for relative quantification. This approach enabled us to take advantage of both MMs and HMRGs to identify both ‘core’ proteins and disease-specific proteins associated with the human gut microbiota, including those with unknown function.

Table 8.5: Human microbial isolate reference genome database (HMRG) database components. 51 bacterial isolates were downloaded from the JGI IMG human microbiome project (IMG-HMP) into a single FASTA-formatted protein sequence database.

Database Identifier	Genome	GID	Total # of Predicted Proteins
Aero_hydr_hydr_ATCC_7966	Aeromonas hydrophila hydrophila ATCC 7966	639633004.faa	4,129
Cate_mits_DSM_15897	Catenibacterium mitsuokai DSM 15897	643886110.faa	2,977
Dore_form_ATCC_27755	Dorea formicigenerans ATCC 27755	641736133.faa	3,277
Ente_faec_PC4.1	Enterococcus faecalis PC4.1	647000238.faa	2,695
Shig_sp._D9	Shigella sp. D9	645058835.faa	4,463
Akke_muci_ATCC_BAA-835	Akkermansia muciniphila ATCC BAA-835	642555104.faa	2,176
Alis_putr_DSM_17216	Alistipes putredinis DSM 17216	641736205.faa	2,742
Anae_ster_DSM_17244	Anaerofustis stercorihominis DSM 17244	641736193.faa	2,331
Bact_cacc_ATCC_43185	Bacteroides caccae ATCC 43185	640963023.faa	3,855
Bact_dore_DSM_17855	Bacteroides dorei DSM 17855	642979370.faa	4,966
Bact_frag_NCTC_9343	Bacteroides fragilis NCTC 9343	637000024.faa	4,299
Bact_ovat_ATCC_8483	Bacteroides ovatus ATCC 8483	641380449.faa	5,536
Bact_pect_ATCC_43243	Bacteroides pectinophilus ATCC 43243	642979337.faa	3,246
Bact_sp._2_2_4	Bacteroides sp. 2_2_4	646206266.faa	5,959
Bact_sp._3_1_33FAA	Bacteroides sp. 3_1_33FAA	647533113.faa	4,666
Bact_sp._3_2_5	Bacteroides sp. 3_2_5	646206273.faa	4,505
Bact_sp._4_3_47FAA	Bacteroides sp. 4_3_47FAA	646206274.faa	4,613
Bact_sp._9_1_42FAA	Bacteroides sp. 9_1_42FAA	646206263.faa	4,871
Bact_sp._D4	Bacteroides sp. D4	646206258.faa	4,431
Bact_thet_VPI-5482	Bacteroides thetaiotaomicron VPI-5482	637000026.faa	4,816
Bact_unif_ATCC_8492	Bacteroides uniformis ATCC 8492	641380447.faa	4,663
Bact_vulg_ATCC_8482	Bacteroides vulgatus ATCC 8482	640753008.faa	4,076
Bifi_adol_L2-32	Bifidobacterium adolescentis L2-32	640963015.faa	2,428
Bifi_long_infa_ATCC_15697	Bifidobacterium longum infantis ATCC 15697	643348516.faa	2,486
Blau_hans_DSM_20583	Blautia hansenii VPI C7-24 DSM 20583	643886146.faa	3,218
Blau_hydr_S5a33_DSM_10507	Blautia hydrogenotrophicus S5a33 DSM 10507	643886199.faa	3,869
Citr_kose_ATCC_BAA-895	Citrobacter koseri ATCC BAA-895	640753015.faa	5,031
Clos_bart_DSM_16795	Clostridium bartlettii DSM 16795	641736113.faa	2,787
Clos_bolt_ATCC_BAA-613	Clostridium bolteae ATCC BAA-613	641380428.faa	7,284
Clos_lept_DSM_753	Clostridium leptum DSM 753	641380427.faa	3,923
Clos_nexi_DSM_1787	Clostridium nexile DSM 1787	642979369.faa	4,239
Clos_sp_M62-1	Clostridium sp M62-1	643886005.faa	4,266
Clos_sp._SS2-1	Clostridium sp. SS2-1	641736270.faa	3,167
Coll_aero_ATCC_25986	Collinsella aerofaciens ATCC 25986	640612206.faa	2,367
Coll_inte_DSM_13280	Collinsella intestinalis DSM 13280	642979320.faa	1,786
Copr_come	Coprococcus comes	643886116.faa	3,913
Dial_invi_DSM_15470	Dialister invisus DSM 15470	645951833.faa	1,954
Esch_coli_K-12_MG1655	Escherichia coli str. K-12 substr. MG1655	646311926.faa	4,148
Euba_rect_ATCC_33656	Eubacterium rectale ATCC 33656	644736367.faa	3,621
Faec_prau_A2-165	Faecalibacterium prausnitzii A2-165	645951831.faa	3,475
Faec_prau_M21-2	Faecalibacterium prausnitzii M21-2	641380420.faa	3,493
Lact_reut_CF48-3A	Lactobacillus reuteri CF48-3A	643886138.faa	2,164
Meth_smit_F1_DSM_2374	Methanobrevibacter smithii F1 DSM 2374	643886215.faa	1,710
Myco_tube_CDC1551	Mycobacterium tuberculosis CDC1551	637000172.faa	4,235
Para_dent_F0305	Parascardovia denticolens F0305	647533193.faa	1,481
Prev_copr_CB7_DSM_18205	Prevotella copri CB7 DSM 18205	643886200.faa	3,293
Rose_inte_L1-82	Roseburia intestinalis L1-82	642979356.faa	4,817
Rumi_gnav_ATCC_29149	Ruminococcus gnavus ATCC 29149	640963057.faa	3,913
Rumi_lact_ATCC_29176	Ruminococcus lactaris ATCC 29176	642791604.faa	2,750
Rumi_obeu_ATCC_29174	Ruminococcus obeum ATCC 29174	640963024.faa	4,175
Rumi_sp._5_1_39BFAA	Ruminococcus sp. 5_1_39BFAA	646206280.faa	3,525

On average, a total of 1,250 (healthy), 850 (ICD), and 788 (CCD) proteins were identified with MM database searches and 2,904 (healthy), 1,928 (ICD), and 2,241 (CCD) proteins using the HMRG database, thus represents the most extensive metaproteome characterization of the human gut to date (**Tables 8.2** and **8.3**). Due to the redundancy of homologous proteins, microbial proteins with >80% sequence identity were clustered to generate a total of 5,692 and 3,101 orthologous clusters (OC) from the HMRGs and MMs, respectively, across all 24 MS runs. Of the OCs that were identified using the MM searches, 52 were identified across all subjects (core; **Table 8.6**) and included primarily general housekeeping functions (such as ribosomal proteins); whereas 151, 3, and 88 OCs were unique to either the healthy, ICD, or CCD core metaproteomes, respectively (**Figure 8.1**). Post-cluster analysis revealed that 1,017 proteins from the MM database searches were unique (i.e., they did not fall into a protein cluster), in contrast, all identified proteins from the HMRGs did cluster, suggesting that there is considerable diversity of genes within the human gut microbiota that is not captured in reference genome sequences.

Table 8.6: Common core microbial proteins identified in the metaproteomes of all subjects included in the study (healthy, ileal CD and colonic CD).

Cluster	Protein
CLST000006	Ribosomal protein L14
CLST000011	Ribosomal protein S11
CLST000208	Annotation not available
CLST000254	Phosphoenolpyruvate carboxykinase (ATP)
CLST000335	NifU homolog involved in Fe-S cluster formation
CLST000603	Glyceraldehyde-3-phosphate dehydrogenase/erythrose-4-phosphate dehydrogenase
CLST000797	GTPases - translation elongation factors
CLST000998	Ribosomal protein L11
CLST001935	Rubrerythrin
CLST005531	Fructose/tagatose bisphosphate aldolase
CLST005618	Ketol-acid reductoisomerase
CLST005666	Chaperonin GroEL (HSP60 family)
CLST005804	Ribosomal protein L5
CLST005825	Ribosomal protein L7/L12
CLST005842	DNA-directed RNA polymerase, beta subunit/140 kD subunit
CLST005865	Ribosomal protein S7
CLST005872	Ribosomal protein S2
CLST005915	Molecular chaperone
CLST005929	Ribosomal protein S19
CLST005988	Ribosomal protein L20
CLST005996	Ribosomal protein S8
CLST006170	Ribosomal protein S9
CLST006191	Translation elongation factors (GTPases)
CLST006298	Ribosomal protein L1
CLST006300	Phosphoenolpyruvate synthase/pyruvate phosphate dikinase
CLST006373	ABC-type sugar transport systems, ATPase components
CLST006584	IMP dehydrogenase/GMP reductase
CLST006673	Ribosomal protein S10
CLST006805	Ribosomal protein L29
CLST006834	3-phosphoglycerate kinase
CLST006844	Ribosomal protein L13
CLST006883	F0F1-type ATP synthase, beta subunit
CLST006904	Ribosomal protein L23
CLST006921	Protein involved in phosphoenolpyruvate-dependent sugar phosphotransferase system
CLST007033	Pyruvate-formate lyase
CLST007119	DNA-directed RNA polymerase, beta subunit/160 kD subunit
CLST007120	Transaldolase
CLST007226	Ribosomal protein S4 and related proteins
CLST007262	Ribosomal protein S3
CLST007269	Ribosomal protein S5
CLST007338	Pyruvate/oxaloacetate carboxyltransferase
CLST007389	Ribosomal protein L6P/L9E
CLST007461	Co-chaperonin GroES (HSP10)
CLST007642	Ribosomal protein L17
CLST007797	Ribosomal protein S15P/S13E
CLST008351	Pyruvate:ferredox in oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, beta subunit
CLST008679	Triosephosphate isomerase
CLST014282	Penicillin tolerance protein
CLST017476	Acetyl-CoA acetyltransferase
CLST018911	Formyltetrahydrofolate synthetase
CLST020880	Carbon dioxide concentrating mechanism/carboxysome shell protein
CLST022770	Acyl-CoA dehydrogenases

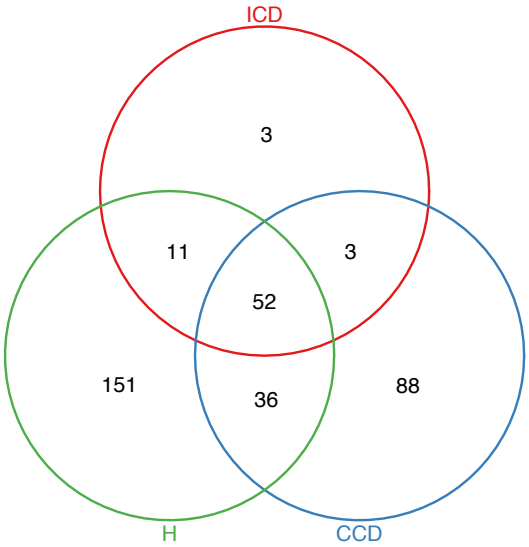


Figure 8.1: Venn diagram showing the number of protein clusters common or unique to each disease category: H (healthy), ICD (ileal Crohn's disease) and CCD (colonic Crohn's disease).

By broad comparison of the metagenomes and metaproteomes, two trends emerged: (a) CD samples clustered separately from healthy (**Figures 8.2a** and **8.3**) and (b) the percent of expressed genes compared to the total gene repertoire is lower in CD patients as compared to healthy subjects (**Figure 8.2b**), reflected by a significant decrease in protein family richness in ICD and CCD that was particularly pronounced for ICD (**Figure 8.2c**). The metaproteomes significantly differentiated by disease phenotype ($p<0.004$) based on spectra matching to HMRGs (**Figure 8.3a**). This was also shown with i) clustering based on function to MMs (**Figure 8.3b**) and ii) functional assignments of genes from MMs by KEGG (**Figure 8.3c**), suggesting that disease phenotype was a stronger discriminator than zygosity, similar to our previous analyses of the same samples[201], [202]. Although healthy and CCD metaproteomes could be distinguished from another, they clustered more closely together compared to the ICD metaproteomes that were clearly distinct (**Figure 8.2a**). Therefore, we primarily focused on functions that differentiated ICD from healthy, but included comparisons to the CCD twin pair when relevant.

Figure 8.2: (**A**) Non-metric multidimensional scaling (nMDS) of fecal metaproteomes. A matrix of normalized spectral counts per protein from each duplicate gut metaproteome was imported into PCORD v5 software. nMDS was performed using the Bray-Curtis distance measure A three-dimensional solution was found after 119 iterations. The final stress for the nMDS was 6.47458. (**B**) Fraction of proteins expressed as measured by comparison to the metagenome (**C**) Functional richness as measured by the Chao1 richness estimate of KEGG orthologous groups (KOs).

Figure 8.2

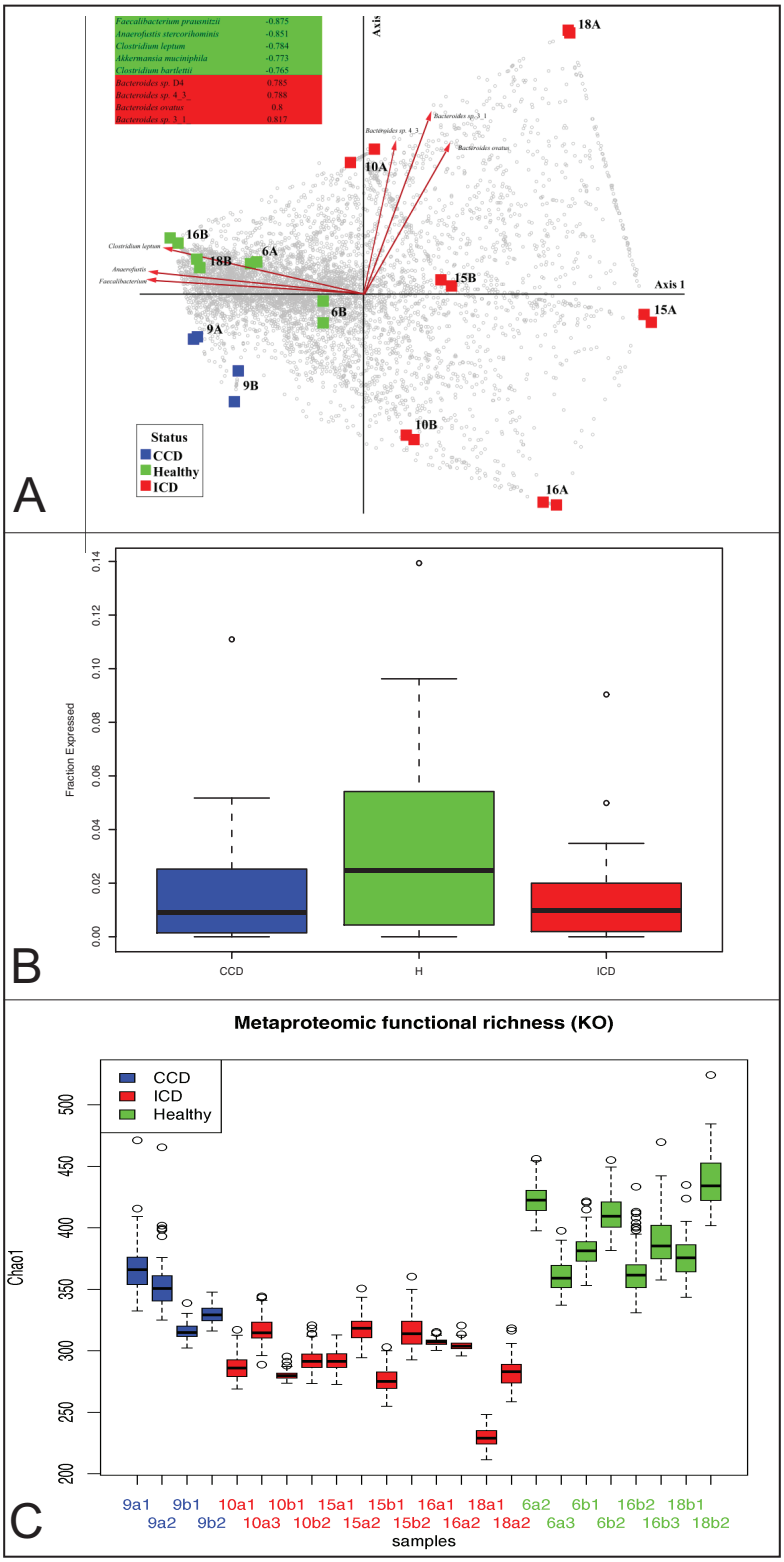
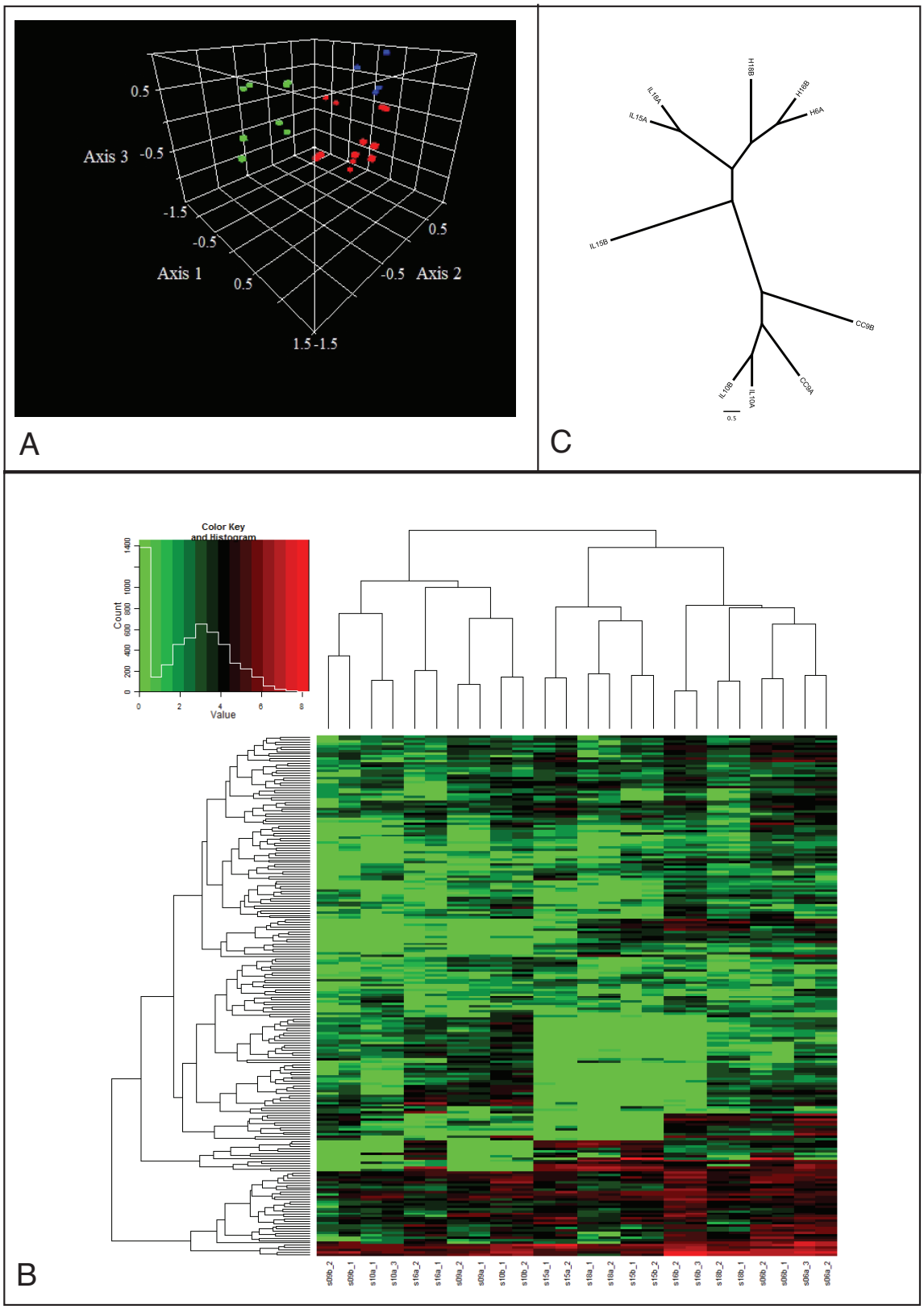

A
Faecalibacterium prausnitzii -0.875
Anaerofustis stercorihominis -0.851
Clostridium leptum -0.784
Akkermansia muciniphila -0.773
Clostridium bartlettii -0.765
Bacteroides sp. D4 0.785
Bacteroides sp. 4_3_ 0.788
Bacteroides ovatus 0.8
Bacteroides sp. 3_1 0.817
Axis 2
Axis 1
18A
10A
16B
18B
6A
15B
15A
6B
9A
9B
10B
16A
Status
CCD
Healthy
ICD
B
Fraction Expressed
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
CCD
H
ICD
C
Metaproteomic functional richness (KO)
CCD
ICD
Healthy
Chao1
200
250
300
350
400
450
500
9a1 9b1 10a1 10b1 15a1 15b1 16a1 18a1 6a2 6b1 16b2 18b1
9a2 9b2 10a3 10b2 15a2 15b2 16a2 18a2 6a3 6b2 16b3 18b2
samples

Figure 8.3: Clustering by Phenotype. (**A**) Non-metric multidimensional scaling (nMDS) of fecal metaproteomes. A matrix of normalized spectral counts per protein from each duplicate gut metaproteome was imported into PCORD v5 software. nMDS was performed using the Bray-Curtis distance measure A three-dimensional solution was found after 119 iterations. The final stress for the nMDS was 6.47458. (**B**) Heatmap of Metaproteomes prediced from matched metagenomes by protein clusters. (**C**) Hieractical Clustering of Metagenomes by KEGG KO relative abundances using Manhattan distance calculation and the 'average' clustering method with an arcsin square root transformation.

Figure 8.3


Axis 3
0.5
-0.5
-1.5
Axis 1
-0.5
0.5
1.5
-1.5
-0.5
0.5
Axis 2
A
H18B
H16B
H6A
IL19A
IL15A
IL15B
CC9B
CC8A
IL10B
IL10A
0.5
C
Color Key
and Histogram
Count
Value
B

While there were core microbial functions that were identified across all samples in the metagenome (**Figure 8.4a**) and metaproteomes (**Figure 8.4b**), proteins involved in translation, defense, organic metabolism, post-translational modification and signaling, and genes involved in intracellular trafficking, translation and defense differed in abundance between healthy and ICD subjects. To assess pathway abundance, KEGG module analysis was performed on metagenome and metaproteome datasets. Glycolysis, reductive pentose phosphate cycle and butyrate production were found to be under-represented in ICD compared with healthy microbiota, in both the metagenomic and metaproteomic datasets (data not shown). In the metagenomic analysis, conjugated bile acid biosynthesis, urea cycle, phosphonate transport system and type IV secretion system were found to be over-represented in ICD compared with healthy microbiota.

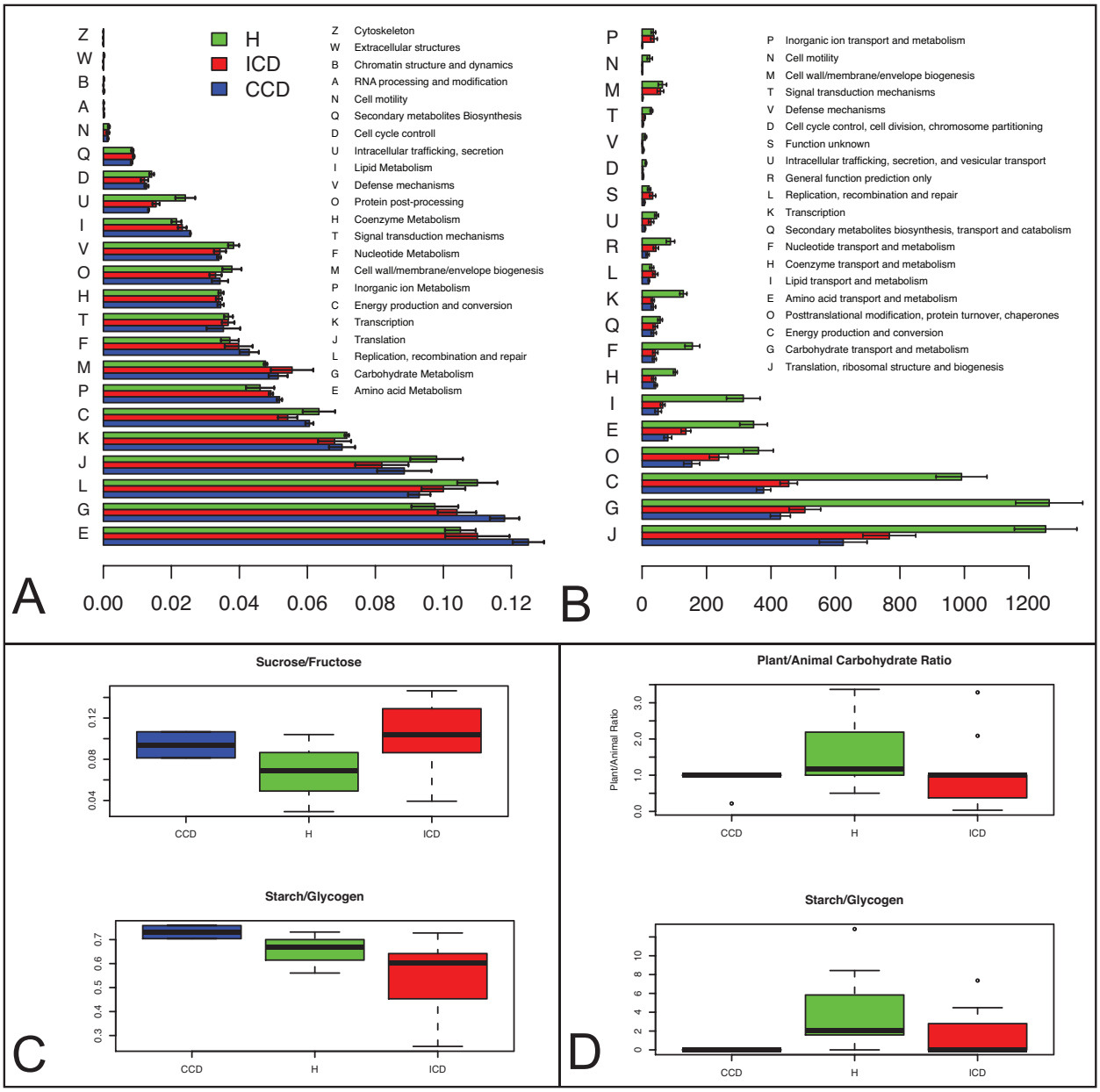


Figure 8.4: (**A**) Relative abundance of metagenomic reads assigned to COG categories. (**B**) Relative abundance of metaproteomic spectra assigned to COG categories. (**C, D**) Sugar utilization in the metagenome (C) and metaproteome (D) by comparison to the CAZy database.

Each dataset contained a subset of genes and proteins of unknown function. For example, ~17% of predicted ORFs were conserved with no known function or were not homologous to any proteins. Approximately 31% of identified HMRG proteins and 29%

of identified microbial OCs (including proteins that did not cluster in MMs) had no known function. Interestingly, one OC comprising 11 unknown proteins was significantly correlated with ICD, where five OCs (10-100s of unknown proteins) were significantly correlated with healthy. These findings support the need for better coupling of phenotypic assays with -omics strategies to aid in the characterization of important functional but unknown genes and proteins.

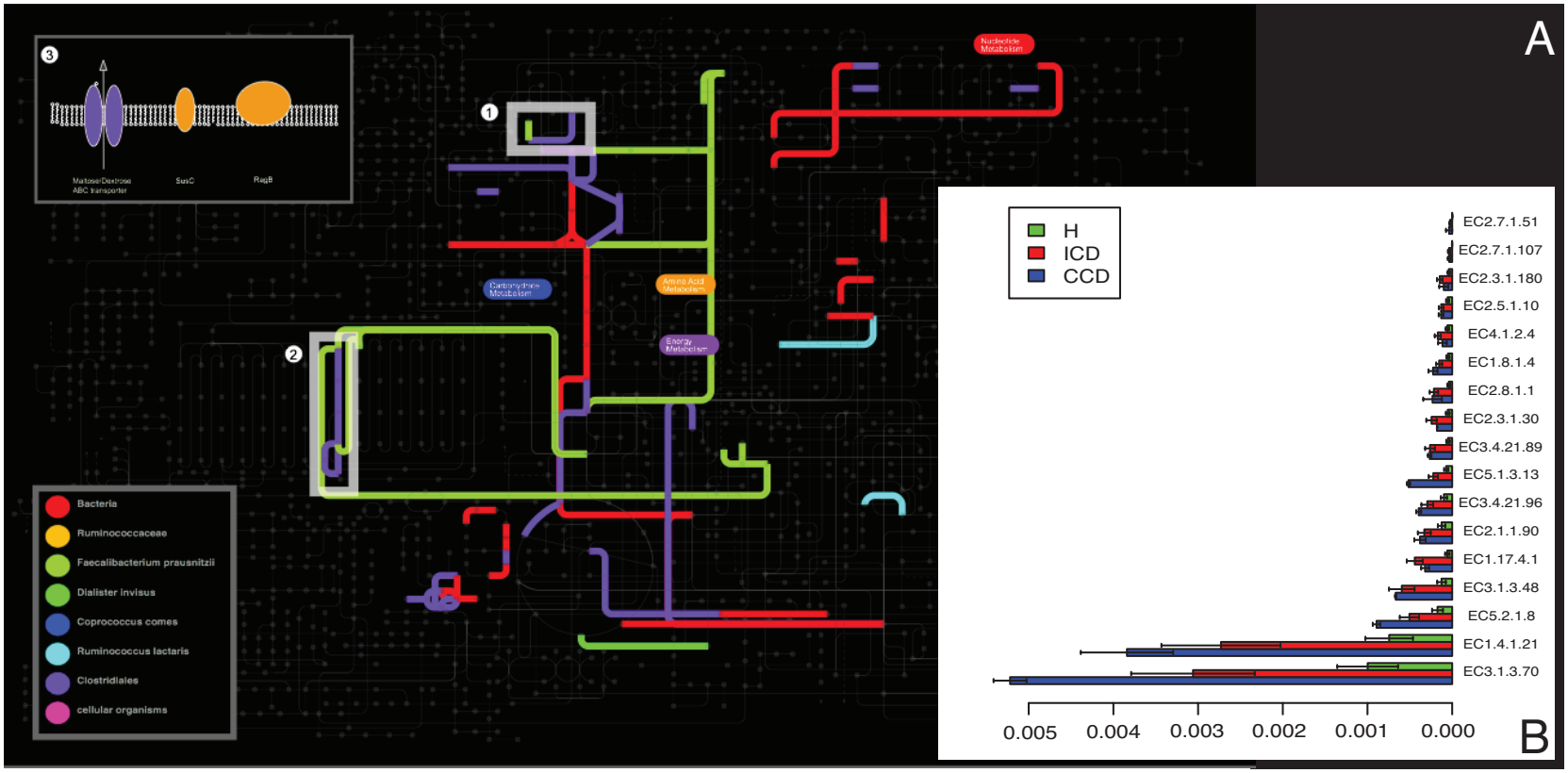
8.3.3: Metabolic pathways differentiate CD and healthy phenotypes

We identified several examples in both the metagenome and metaproteome datasets which suggested that functions related to carbohydrate transport and metabolism and energy production are depleted in the ICD microbiota (**Figure 8.4b**). In addition to the differential pathways identified by KEGG analysis, the abundance of genes for sucrose and fructose degradation is higher in ICD, while genes and proteins involved in starch, glycogen, and complex carbohydrate degradation are lower in abundance (**Figure 8.4c and d**). These results, along with pathway analysis, suggest that the microbiota of ICD subjects have a reduced capability to uptake complex carbohydrates and breakdown nutrients.

Many proteins that were less abundant in ICD reflected a decreased abundance of bacteria that contain metabolic pathways with relevance to the physiology of the human gut (**Figure 8.2a**). Butyrate, a major energy source for colonocytes, is involved in the maintenance of colonic mucosal health and can elicit anti-inflammatory effects[208], thus its depletion could be one reason for the inflammation in CD. *Faecalibacterium prauznitzii* is a major butyrate producer in the gut and the low abundance of this species (as revealed by 16S rRNA and metagenomic analyses) and proteins involved in the butyrate pathway (**Figure 8.5**) could contribute to the inflammation associated with ICD. Reduced butyrate production correlates to the depletion of known butyrate producers (e.g., *Roseburia*, *Faecalibacterium* and *Subdoligranulum*) in our CD subjects (**Figures 8.2a and 8.6**). The increased abundance of *F. prausnitzii* revealed abundant proteins central to butyrate production and other short-chain fatty acid production (e.g., acetate and proprionate) exclusively in

the healthy and CCD subjects but not in ICD (**Figure 8.7a**). Several other genes associated with anti-inflammatory responses and properties, such as lactocepin (EC3.4.21.96) and aspartate dehydrogenase (EC3.4.21.96), were significantly more abundant in CD relative to healthy (**Figure 8.5b**).

Figure 8.5: (**A**) Metabolic pathways differentiating by disease phenotype, as resulting from the metabolic module analysis ($p<0.05$; 5% FDR). Highlighted areas discussed in the main text: (1) butyrate production; (2) membrane proteins (**B**) Enzymes that were significantly different across Healthy, ICD, and CCD fecal metagenomes.



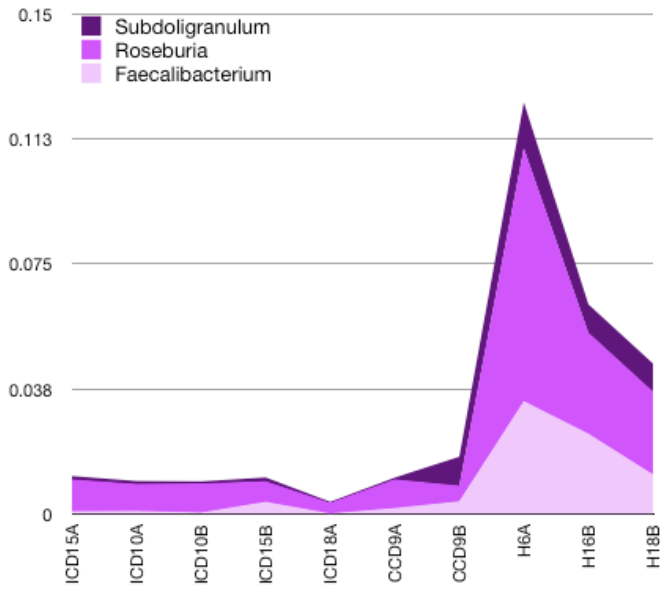


Figure 8.6: Differences in Butyrate Production in ICD compared to Healthy. Cumulative plots of fraction of total reads assigned to Faecalibacterium, Rosburia and Subdoligranulum genera per sample.

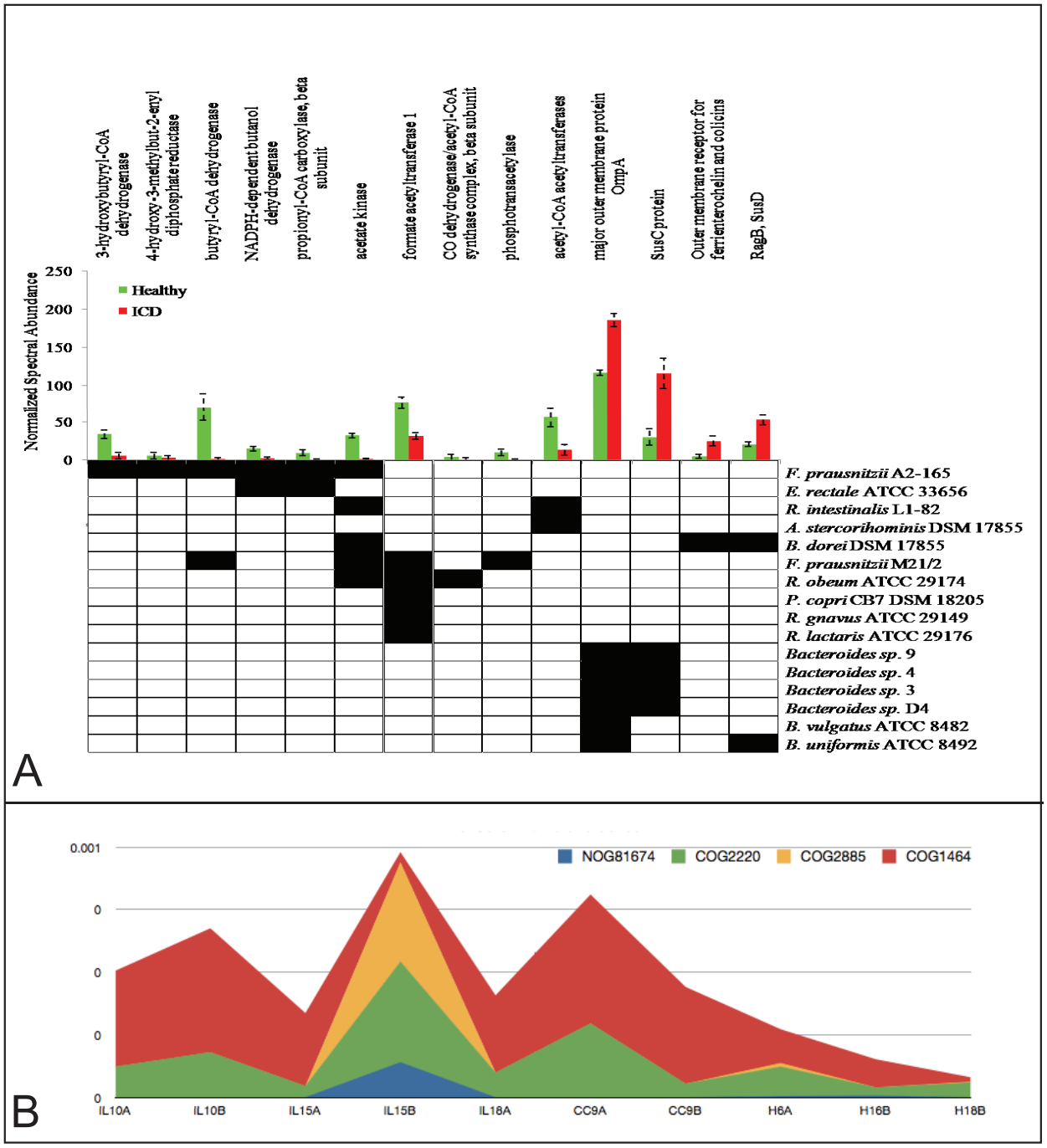


Figure 8.7: (**A**) Most significantly differential proteins from Healthy and CD subjects. Presence-absence heatmap shows which of the 51 bacterial strains the proteins matched to. (**B**) Enzymes that were significantly different across Healthy, ICD, and CCD fecal metagenomes.

Protein abundance measurements, based on MM searches, indicate an over-representation of bacterial TonB-dependent cell surface receptors, which are multi-functional but are involved in inorganic ion transport and metabolism, in the ICD microbiota. These data are also consistent with metagenomic analysis which reveals a greater abundance of genes involved in inorganic ion metabolism in the CD microbiota (**Figure 8.4a**). If the gut ecosystem is deficient in inorganic ions in CD, the gut microbiota may compensate by up-regulation of genes and proteins that are involved in ion acquisition and transport.

8.3.4: Bacterial-host interactions and defense

Several proteins involved in bacterial-host interactions and defense were more abundant in the ICD microbiota and included several bacterial outer membrane proteins (e.g., OmpA, RagB, and SusC/D) that were differentially present in both the metagenomes and metaproteomes (**Figure 8.7a** and **b**), supporting the current hypothesis that CD is manifested by an aberrant mucosal response to otherwise harmless bacterial antigens in genetically susceptible subjects[209,210,211]. OmpA, a pore-forming protein in the outer membrane of many Gram-negative bacteria, harbors diverse functions including maintenance of cell structure, binding various substances, adhesion, and resistance to antimicrobials[212], and is suggested to be involved in gut mucosal association[213]. One hypothesis is that because OmpA is highly represented and highly conserved in many enteric bacteria, the immune system has acquired the ability to recognize and to be activated by this class of protein[214]. Because these proteins are more abundant in ICD, this suggests that the immune system is functioning abnormally with respect to reduced levels of the corresponding bacteria expressing this protein, and supports the current hypothesis that CD is manifested by an aberrant mucosal response to otherwise harmless bacterial antigens in genetically susceptible individuals[209,210,211].

Our study provides the first evidence of elevated abundance of other major OMPs, such as RagB, SusC/D associated with CD (**Figure 8.5a**). An elevated IgG response to RagB was previously reported in subgingival samples of patients with

periodontitis[215] and virulence of the rag locus was demonstrated in *Porphyromonas gingivalis* strains[216]. While the role of RagB/Sus in the etiology of CD warrants further study, our data suggest that there is a shift from a healthy microbiota towards a microbial consortium that can elicit an inflammatory immune response.

In addition, an integration host factor (IHF) protein, which is linked to virulence gene regulation[217,218], was identified as being statistically more abundant in ICD metaproteomes using MMs, but not HMRGs. This finding highlights the importance of MMs to identify proteins that originate from bacteria not yet sequenced, or cultivated.

8.3.5: Broad Functional Comparisons of the Human Proteome

Because we are able to measure both bacterial and human proteins using metaproteomics, a total of 1,646 human proteins were experimentally identified. Gene ontology (GO) analysis revealed that human proteins found in all 3 subject groups (core) are enriched in functions associated with the structural integrity of the mucosal epithelium. Proteolysis, digestion, and carbohydrate catabolism were also among the most abundant 'core' functions, as would be expected in the GI-tract (**Figure 8.8a**). For human proteins that varied in healthy compared to CD, the majority were involved in epithelial integrity and function, as detailed below. To our knowledge, this is the first use of non-targeted shotgun proteomics to simultaneously assess both human and microbial proteins from the same fecal samples to assess a disease state.

Figure 8.8: Human proteins identified in the metaproteome data. (**A**) Human proteins' "core" Gene Ontology terms across all subjects (healthy (H), ileal Crohn's Disease (ICD), and colonic Crohn's Disease (CCD)). (**B**) Human proteins Gene Ontology terms that are enriched according to disease.

Figure 8.8 (A)

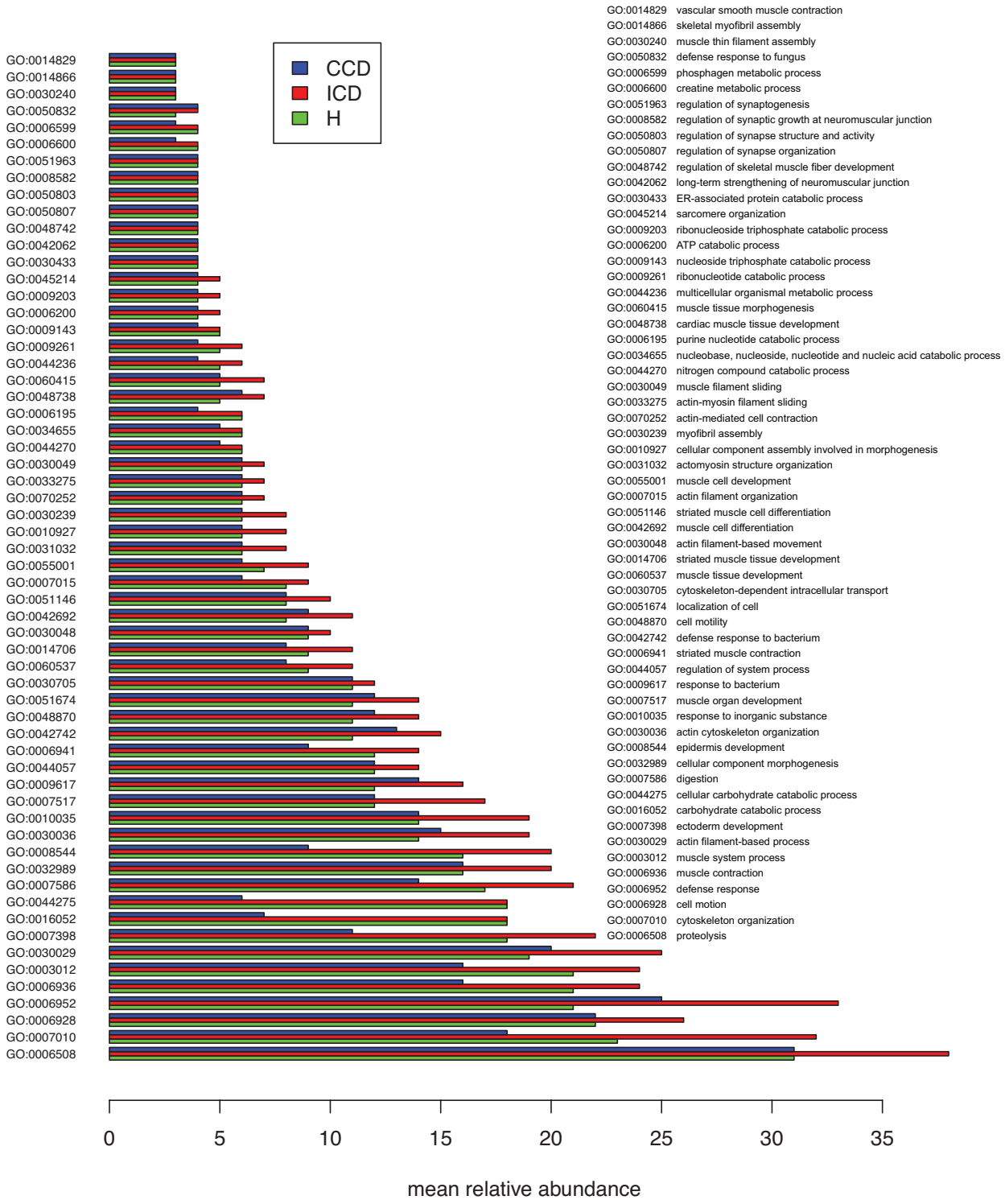
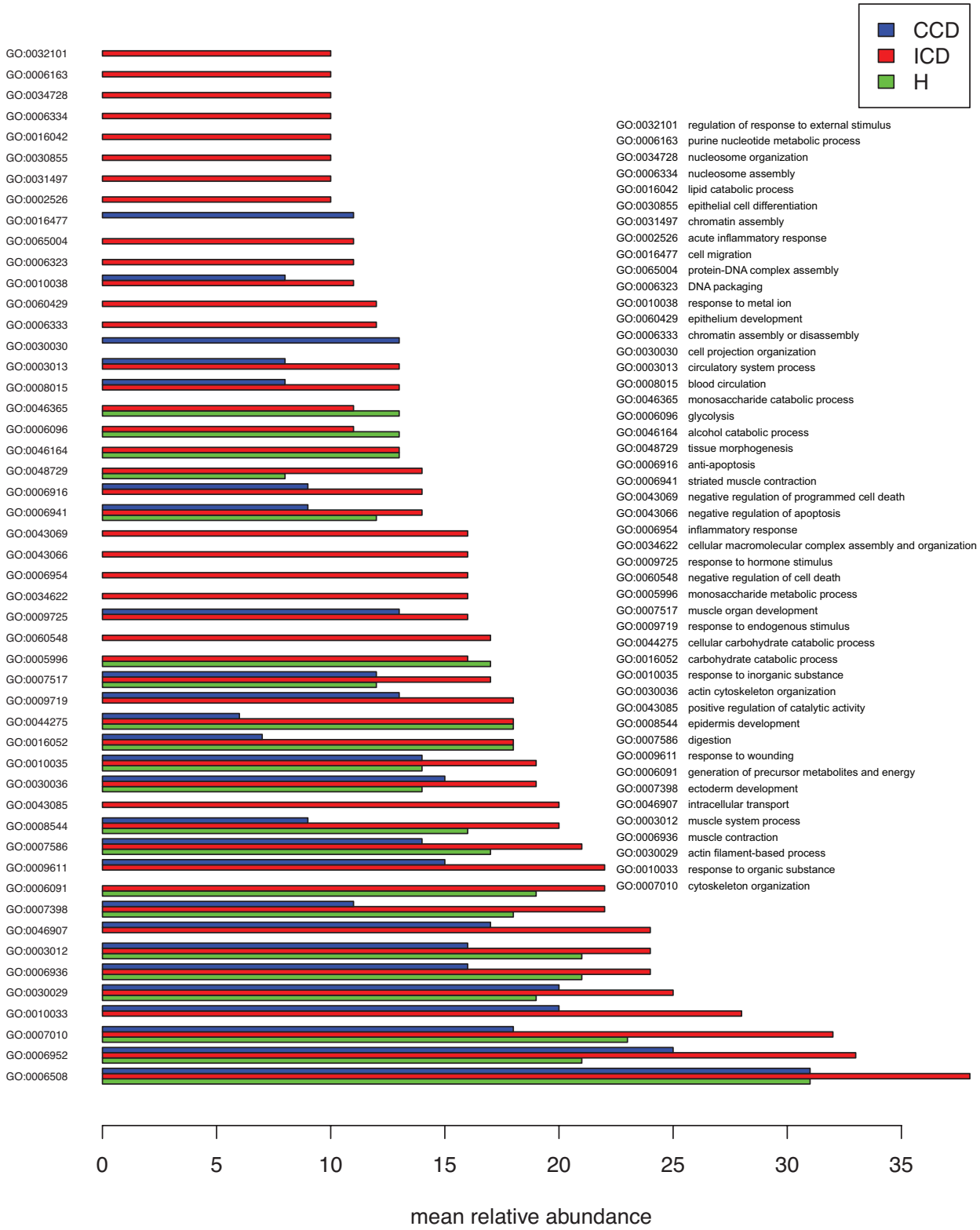

CCD
ICD
H
GO:0014829 vascular smooth muscle contraction
GO:0014866 skeletal myofibril assembly
GO:0030240 muscle thin filament assembly
GO:0050832 defense response to fungus
GO:0006599 phosphagen metabolic process
GO:0006600 creatine metabolic process
GO:0051963 regulation of synaptogenesis
GO:0008582 regulation of synaptic growth at neuromuscular junction
GO:0050803 regulation of synapse structure and activity
GO:0050807 regulation of synapse organization
GO:0048742 regulation of skeletal muscle fiber development
GO:0042062 long-term strengthening of neuromuscular junction
GO:0030433 ER-associated protein catabolic process
GO:0045214 sarcomere organization
GO:0009203 ribonucleoside triphosphate catabolic process
GO:0006200 ATP catabolic process
GO:0009143 nucleoside triphosphate catabolic process
GO:0009261 ribonucleotide catabolic process
GO:0044236 multicellular organismal metabolic process
GO:0060415 muscle tissue morphogenesis
GO:0048738 cardiac muscle tissue development
GO:0006195 purine nucleotide catabolic process
GO:0034655 nucleobase, nucleoside, nucleotide and nucleic acid catabolic process
GO:0044270 nitrogen compound catabolic process
GO:0030049 muscle filament sliding
GO:0033275 actin-myosin filament sliding
GO:0070252 actin-mediated cell contraction
GO:0030239 myofibril assembly
GO:0010927 cellular component assembly involved in morphogenesis
GO:0031032 actomyosin structure organization
GO:0055001 muscle cell development
GO:0007015 actin filament organization
GO:0051146 striated muscle cell differentiation
GO:0042692 muscle cell differentiation
GO:0030048 actin filament-based movement
GO:0014706 striated muscle tissue development
GO:0060537 muscle tissue development
GO:0030705 cytoskeleton-dependent intracellular transport
GO:0051674 localization of cell
GO:0048870 cell motility
GO:0042742 defense response to bacterium
GO:0006941 striated muscle contraction
GO:0044057 regulation of system process
GO:0009617 response to bacterium
GO:0007517 muscle organ development
GO:0010035 response to inorganic substance
GO:0030036 actin cytoskeleton organization
GO:0008544 epidermis development
GO:0032989 cellular component morphogenesis
GO:0007586 digestion
GO:0044275 cellular carbohydrate catabolic process
GO:0016052 carbohydrate catabolic process
GO:0007398 ectoderm development
GO:0030029 actin filament-based process
GO:0003012 muscle system process
GO:0006936 muscle contraction
GO:0006952 defense response
GO:0006928 cell motion
GO:0007010 cytoskeleton organization
GO:0006508 proteolysis
0
5
10
15
20
25
30
35
mean relative abundance

Figure 8.8 (B)


CCD
ICD
H
GO:0032101 regulation of response to external stimulus
GO:0006163 purine nucleotide metabolic process
GO:0034728 nucleosome organization
GO:0006334 nucleosome assembly
GO:0016042 lipid catabolic process
GO:0030855 epithelial cell differentiation
GO:0031497 chromatin assembly
GO:0002526 acute inflammatory response
GO:0016477 cell migration
GO:0065004 protein-DNA complex assembly
GO:0006323 DNA packaging
GO:0010038 response to metal ion
GO:0060429 epithelium development
GO:0006333 chromatin assembly or disassembly
GO:0030030 cell projection organization
GO:0003013 circulatory system process
GO:0008015 blood circulation
GO:0046365 monosaccharide catabolic process
GO:0006096 glycolysis
GO:0046164 alcohol catabolic process
GO:0048729 tissue morphogenesis
GO:0006916 anti-apoptosis
GO:0006941 striated muscle contraction
GO:0043069 negative regulation of programmed cell death
GO:0043066 negative regulation of apoptosis
GO:0006954 inflammatory response
GO:0034622 cellular macromolecular complex assembly and organization
GO:0009725 response to hormone stimulus
GO:0060548 negative regulation of cell death
GO:0005996 monosaccharide metabolic process
GO:0007517 muscle organ development
GO:0009719 response to endogenous stimulus
GO:0044275 cellular carbohydrate catabolic process
GO:0016052 carbohydrate catabolic process
GO:0010035 response to inorganic substance
GO:0030036 actin cytoskeleton organization
GO:0043085 positive regulation of catalytic activity
GO:0008544 epidermis development
GO:0007586 digestion
GO:0009611 response to wounding
GO:0006091 generation of precursor metabolites and energy
GO:0007398 ectoderm development
GO:0046907 intracellular transport
GO:0003012 muscle system process
GO:0006936 muscle contraction
GO:0030029 actin filament-based process
GO:0010033 response to organic substance
GO:0007010 cytoskeleton organization
GO:0006508
0
5
10
15
20
25
30
35
mean relative abundance

8.3.6: Impaired epithelial integrity in ICD

The human proteins detected primarily in CD subjects support the hypothesis that subjects with ICD, even in remission, have a defective epithelial barrier. A higher abundance of proteins in GO categories for inflammatory and host defense, wounding response, intracellular transport, and epithelial development and differentiation were enriched in ICD subjects (**Figure 8.8b**). For example, mucin 2 (MUC-2), the most prominent mucin secreted by intestinal epithelial cells, was also more abundant in ICD subjects. Similarly, thiosulfate sulfurtransferase, important in sulfate reduction and linked to mucin fermentation (PMID 3214155), was elevated in CD ($p<0.001$) based on metagenome analysis (**Figure 8.5b**).

Other proteins that function in maintaining mucosal integrity were identified as being statistically under-represented in ICD, including protocadherin LKC, a calcium dependent mediator of cell-cell adhesion that associates with the mucosal actin cytoskeleton[219] and type 1 collagen (alpha-2), the major collagen in the intestinal extracellular matrix[220]. A depletion of these proteins might compromise host defense at the mucosal interface.

A defective epithelial barrier is thought to result in an aberrant host response to luminal antigens leading to an exaggerated adaptive immune response and chronic inflammation[211]. Alpha defensin 5, a protein implicated in regulation of bacterial concentrations in the ileal intestinal crypt[221,222,223] was statistically more abundant in ICD, suggesting that the host may increase expression of defensins in response to aberrant microbiota in these subjects, or that the products are leaking from the intestinal site of action to feces.

8.3.7: Impaired intestinal absorption in ICD

The primary function of the small intestine is absorption and this appears to be impaired in subjects with ICD. For example, several pancreatic enzymes: chymotrypsinogen B1 and B2, pancreatic carboxypeptidase A1 and B1 and pancreatic lipase were identified with higher abundance in ICD. These enzymes are synthesized in the pancreas as

inactive precursors that are activated in the intestine where they aid in digestion. Relatively high amounts of pancreatic enzymes in feces may be indicative of pancreatitis, which has been linked to CD[224], but remains to be confirmed since the subjects in this study have not had active pancreatitis.

Several bile salts were previously found to be elevated in ICD fecal samples[202], supporting the hypothesis that there is malabsorption of secreted enzymes and metabolites by the gut epithelium in ICD. The reduced uptake of bile salts and pancreatic enzymes could also be due to surgery since all ICD patients had undergone resections of the ileum. Since uptake of bile salts occurs within the terminal part of the ileum, patients that have undergone resections, leaving them with a shorter ileum, might have a reduced uptake of bile salts. Bile salt malabsorption with secondary diarrhea is a common clinical feature in patients undergoing extensive ileal resections.

8.4: Conclusions

Here we have used a combination of extensive and complementary "-omics" datasets to provide a more comprehensive view of the role of the gut microbiota in CD than has been previously possible. The value of this approach comes from the ability not only to examine the structure and function of the microbiota from multiple perspectives, but also from the ability to integrate data from the gut microbiota and the host. The validity of our methods is supported by data at the species, gene, and protein levels that confirm previous reports that ICD is associated with a loss of *F. prauznitzii*. New findings from this study suggest several other malfunctions in CD, both with respect to the intestinal microbiota and the host. Dysbiosis of the bacterial community in ICD results in a higher abundance of bacterial surface proteins, many of which are antigenic and could contribute to an exaggerated immune response, and that could cause or aggravate inflammation associated with CD. This imbalance comes at the expense of loss of many beneficial members of the microbiota, including those that produce butyrate. At the same time, there are several indications that the host epithelial barrier is impaired, both with respect to structural integrity of the mucosal boundary and with respect to its ability to absorb secreted enzymes and metabolites. These functional changes may define the

CD phenotype, even when patients are in remission. It will be of great value to extend these studies to larger cohorts of CD patients and to also carry out longitudinal studies to assess how the structure and function of the gut microbiota changes in a given patient over time. We have also uncovered some interesting examples of where the meta-omics data does not completely overlap, indicating the need to further explore the fundamental differences and significance of genomic potential versus proteome abundances. Together, these data point towards several new targets for further investigation in the hunt for diagnostic targets and therapeutic treatments for CD.

Chapter 9

Conclusions from the Metaproteomics Characterization of the Human Gut-Associated Microbiome

9.1: Conclusions

The human microbiome, the collective set of microbes inhabiting the human body, is a complex ecosystem that is poorly understood in both human health and disease. Although the HMP has focused tremendous efforts and funds to understand the human microbiota by sequencing the microbes present in and on the skin, oral and nasal cavity, vagina and gastrointestinal tract by metagenomics, this approach will only reveal the composition and 'potential' function. While genomics and metagenomics have laid the groundwork for many microbial communities including the human microbiome, proteomics and metaproteomics have evolved to provide an additional level of information, 'actual' protein abundance that is not possible with metagenomics.

The research presented in this dissertation represents a detailed characterization of the human gastrointestinal (gut) microbiome. Although not completely comprehensive, a fairly deep level of detail regarding the identity and functional signature of the host and gut microbial metaproteomes has been revealed through an integrative approach consisting of both community genomics and proteomics. The technology that enables high-throughput, unbiased, and highly reproducible community proteomics is high throughput, high performance mass spectrometry. MS-based proteomics can identify hundreds to thousands of proteins from a microbial community sample. As discussed previously, genomics and metagenomics (predicted protein database) is the foundation for MS-based proteomics. Therefore, the quality of DNA sequencing (i.e., depth of coverage and sequencing errors), assembly, and gene-finding has a tremendous effect on the ability of MS-based proteomics to assign all tandem mass spectra using protein database searching. For example, if the final genomes or metagenomes are not representative of the exact same samples used for proteomic measurements or are not sequenced to a sufficient depth, fewer quality MS/MS will be

assigned resulting in fewer peptide and protein identifications. Due to the interdependence of both technologies and future of systems biology, it is the development and advancement of genomic and proteomic technologies that will enable and improve biological inference of the complex human microbiome.

The objective of this dissertation research is a detailed and mechanistic understanding of the host and microbial functional signature in the human gut microbiome. Initially, we used a less complex and defined human-derived microbial community in gnotobiotic mice as a model system to study the human gut microbiome. This model system is advantageous for several reasons, including the ability to control the microbial membership in present in the gut. A defined human microbiota enables the functional study of each of microbial member, their interactions, cooperation, competition and adaptation in the gut. From the lower complexity binary and 12-member consortia, we progressed to a representative and higher complex human gut microbiome in human individuals. A non-targeted MS-based approach is ideal for studying complex communities based on its ability to directly measure expressed proteins from complex environmental matrices. This approach was applied to elucidate the functional 'core' and differences in the commensal microbiota of human twins with and without Crohn's disease. Although challenges are present in both approaches, both have provided different information that has contributed to a larger understanding of how the human gut, health and disease, functions with our microbial counterparts.

9.2: Experimental optimization and biological inference in the human gut microbiome

The experimental methodology and analytical technology originally developed on single bacterial isolates has been extended to low- (AMD) and high-complexity (soil, ocean, and the human microbiome) microbial communities, all of which with range in microbial composition and diversity. Microbial communities have many challenges not characteristic of single bacterial isolates including: environmental sample biomass quantity, interfering matrices, species and protein dynamic range, and microbial sequence redundancy. As revealed throughout this dissertation, experimental

challenges associated with human gut microbiome related samples include efficient, non-biased lysis and extraction of proteins from bacteria and host cells in complex sample matrices (feces and cece) and dynamic range (detection of lower abundant proteins and microbes). Informatics challenges are based on traditional database filtering metrics and the ability to uniquely assign MS/MS to a protein and its' corresponding microbial species within a large collection of closely related and diverse microbes. Although these challenges initially had a significant impact on the ability to perform deep proteome characterizations, we have identified new strategies that have and will enhance community MS-based proteomic studies of the human microbiome. For the two approaches described in this dissertation, liquid chromatography coupled with tandem mass spectrometry has been successful to characterize the gut microbial community proteomes of gnotobiotic mice and human twins.

9.2.1: Gnotobiotic mice

A defined human representative consortium of microbes has provided insight into how i) distinct members of a larger consortium of microbes initially establish themselves through cooperation and competition, and subsequently ii) compose the collective functional community. Chapters 3-6 outlined experimental and computational procedures used for proteomic assays of a model gut microbiota, and also illustrated some of the benefits in obtaining this type of information. Experimental methods that used a combination of pre-fractionation via ultracentrifugation and chemical solubilization and physical homogenization have significantly improved peptide-spectrum matching and protein identification of *in situ* extracted proteomes. Computational methods that compare and use unique peptide (theoretical peptidome), spectra, and protein counts enable the differentiation and assignment of proteins with high sequence similarity to a distinct phylotype. The binary community proteomic results revealed that the majority of identified proteins belonging to *B. thetaiotaomicron* and *E. rectale* are true unique identifications, and that these species can be easily differentiated by proteomics. Although this was a simplified two component human gut microbiota of two evolutionary divergent species, the 12-member proteomic results revealed similar conclusions with unique peptides as a preferred method for the relative

estimation of species abundance with significant dynamic range. These analyses suggested that the community structure is dictated by the host's diet (i.e., diet is shaping overall community structure), with many conserved hypothetical and pure hypothetical proteins identified whose presence had not been predicted in the initial annotation of the finished genome.

9.2.2: Human gut twin cohort

With a successful method to study the proteomes of lower-complexity microbiota in gnotobiotic mice, we extended this methodology into higher complexity representative human gut microbiomes in human feces. We have developed a novel non-targeted MS approach to measure the identities of thousands of microbial and host proteins in human feces using non-matched and/or matched metagenomes in addition to human-derived reference genomes for protein identification. Using this approach, we established the role of an integrated platform using MS-based proteomics and metagenomics in the human microbiome. Although these results presented the largest coverage of the human gut metaproteome, to fully understand the functional role of the gut microbiota and its interaction with the human host would require extensive efforts to comprehensively define and characterize each microbial member in addition to the community as a single collective entity in health and disease.

We have successfully demonstrated that whole community metaproteome measurements were achievable in the human gut microbiome and provided the first large-scale glimpse into the functional activities of the microbial community inhabiting a healthy gut. These results also provided key insight into the challenges that we and future studies will encounter as the omics' field progresses and accumulates thousands of metagenomic sequences, including extensive microbial sequence redundancy. In order to advance and apply this methodology to higher complexity microbiota and human subjects with disease, the field has to establish methods for how to tackle microbial protein sequence redundancy in environmental samples. Although non-matched metagenomic data may capture more sequence diversity, large protein databases can create unreasonable sequence redundancy. Therefore, we applied a

bioinformatics comparison and analysis of how to construct metagenomic sequence databases for optimum metaproteome measurements. These results suggested that proteomic data is twice as likely to match metagenomic data derived from the same sample and protein databases derived from matched metagenomic sequenced reads (RMPS), increased the number of MS/MS spectra, peptides, and protein identifications. Using this novel approach, we were able to increase PSM and coverage of metaproteomes collected from both healthy and individuals with Crohn's disease and revealed examples of where reference genomes and meta-omics data does not correlate, indicating the need for future studies to explore the differences between genomic potential versus proteome abundances.

9.3: Future directions

As this dissertation has demonstrated, biological advancements go hand in hand with technological developments. We cannot improve our understanding of the human microbiome unless the experimental and analytical tools are available and adapted for complex environmental samples with higher complexity in microbial composition and diversity (i.e., thousands of bacterial species with a wide range of abundances). These analyses have and will advance the field of metaproteomics in the human gut microbiome by providing novel experimental and bioinformatic strategies to identify and characterize the metaproteomes of complex microbiomes extracted from feces and ceca. Experimental comparisons and developments that lead to enhanced lysis and protein extraction methods will enable future studies to build upon these methods to increase protein identification and coverage of large-scale metaproteomes.

The future of metaproteomics in the human microbiome will likely focus on several of the challenges discussed within this dissertation. Because we are only sampling the surface with the identification of ~1-10% of the community proteome, technological advancements will enable deeper measurements and wider coverage of the entire community, but more importantly the lesser abundant microbes. Improvements in chromatographic peptide separation and/or fractionation and mass spectrometric measurements will provide better peptide separation and detection,

through-put, and higher resolution and mass accuracies to resolve single amino acid polymorphisms and post-translational modifications in the human microbiome which is currently not possible.

The current bottleneck in microbial community metaproteomics hinges on the available informatics algorithms and filtering metrics that were designed for single microbial isolates and mixtures of proteins. Contrary to single bacterial genomes, when faced with thousands of publically available metagenomes and reference genomes, computational resources will be stretched to their practical limits, and traditional database search algorithms will be ineffective and obsolete. New cost-effective computational resources (i.e., to store and create substantial omic' databases) and informatics algorithms that are designed for microbial communities will lead the future and enable the comprehensive and accurate assignment of all tandem mass spectra within microbial communities for which a large portion of the are closely-related microbes with high sequence similarity.

The future of the human microbiome, both in metagenomics and metaproteomics, includes the development of tools to characterize large numbers of proteins with unknown function. As evident throughout this dissertation in gnotobiotic mice and human individuals, a large percentage of the collective microbial community consists of proteins with unknown function that are not revealed by metagenomics. It is obvious that these proteins are critical for microbial survival and carry out important functions in the human gut. New experimental and biochemical assays focused on the profiling and characterization of proteins with unknown functions will likely unravel new microbial phylotypes and functions yet to be seen by traditional sequencing technologies.

It is the field of systems biology and the combination of omic approaches, with advancement in all areas of MS-based proteomics including technology and informatics workflows that will serve as the future revolutionary tool to fully characterize microbial community metaproteomes in the human microbiome.

9.4: Perspective

Over the last five years, the research presented here has helped establish the field of metaproteomics and its successfulness in the human gut microbiome, even though the majority of current research efforts and funds are focused on metagenomics. We have identified an experimental and analytical platform that supports an unbiased and deep identification of the human gut metaproteome. This platform can be effectively scaled from a less complex, controlled model microbiota to a highly complex gut microbiome derived from human subjects. We have developed a novel bioinformatics workflow for integrated omic studies that incorporates metagenomic sequence data and MS to provide optimum identification and characterization of human host-gut metaproteomes. In addition to designing and developing experimental, analytical, and informatics workflows, we have provided a glimpse into whether a healthy 'core' gut metaproteome exists and the metabolic functional differences between individual microbial species (e.g., *B. thetaiotaomicron* and *E. rectale*) and communities as a whole (e.g., healthy versus disease). These experiments and results represent substantial progress towards the ultimate goal of a complete identification of the human gut microbiome.

The next 5 years will undoubtedly focus on implementing and continually developing the platforms described herein to characterize human microbiomes collected from higher complexity model communities (e.g., a 100-member microbial community in gnotobioic mice) and other human body sites (e.g., oral cavity and vagina). The establishment of metaproteomics in the human microbiome should drive an increase in the funding and more extensive studies that focus on characterizing the actual functional metaproteome. As a result, metagenomic-related research groups will engage metaproteomics to not only enhance our understanding of the microbiome, but also improve metagenomic sequencing with respect to its impact on metaproteomics. With regards to mass spectrometry, new informatics workflows that combine traditional protein database searching with novel *de novo* sequencing algorithms will be developed and benefit integrated omic' studies with the identification of unknown proteins that are not sequenced and/or are missed in the assembly or gene-finding algorithms unique to metagenomics. Finally, MS will be challenged to another level where efforts will likely

begin to focus on designing new techniques and technologies in all omic' fields that permit the study of the 'web of events' rather than a static snapshot of the functional activities in a microbial community. It is not apparent that there are any fundamental inpenetratable roadblocks to this progress towards a comprehensive systems-biology characterization of the human microbiome, but rather only experimental and informatics hurdles that need to continue to be navigated. Since the ultimate goal is a gain in biological insight, a focus on mining biological inferences from integrated metagenomic-metaproteomic datasets will advance mass spectrometry in the human microbiome.

References

1. Brock TD. The study of microorganisms in situ: progress and problems; 1987. Cambridge University Press. pp. 1-17.
2. Bakken LR (1985) Separation and purification of bacteria from soil. Appl Environ Microbiol 49: 1482-1487.
3. Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59: 143-169.
4. Schloss PD, Handelsman J (2003) Biotechnological prospects from metagenomics. Curr Opin Biotechnol 14: 303-310.
5. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37-43.
6. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66-74.
7. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311: 496-503.
8. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5: e77.
9. Baker BJ, Banfield JF (2003) Microbial communities in acid mine drainage. FEMS Microbiol Ecol 44: 139-152.
10. Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, et al. (2007) Microbial population structures in the deep marine biosphere. Science 318: 97-100.
11. DeLong EF (1997) Marine microbial diversity: the tip of the iceberg. Trends Biotechnol 15: 203-207.
12. DeLong EF (2005) Microbial community genomics in the ocean. Nat Rev Microbiol 3: 459-469.
13. Kent AD, Triplett EW (2002) Microbial communities and their interactions in soil and rhizosphere ecosystems. Annu Rev Microbiol 56: 211-236.
14. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, et al. (2007) The human microbiome project. Nature 449: 804-810.
15. Pei Z, Bini EJ, Yang L, Zhou M, Francois F, et al. (2004) Bacterial biota in the human distal esophagus. Proc Natl Acad Sci U S A 101: 4250-4255.
16. Verhelst R, Verstraelen H, Claeys G, Verschraegen G, Delanghe J, et al. (2004) Cloning of 16S rRNA genes amplified from normal and disturbed vaginal microflora suggests a strong association between Atopobium vaginae, Gardnerella vaginalis and bacterial vaginosis. BMC Microbiol 4: 16.
17. Zhou X, Bent SJ, Schneider MG, Davis CC, Islam MR, et al. (2004) Characterization of vaginal microbial communities in adult healthy women using cultivation-independent methods. Microbiology 150: 2565-2573.
18. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE (2005) Defining the normal bacterial flora of the oral cavity. J Clin Microbiol 43: 5721-5732.

19. Bik EM, Eckburg PB, Gill SR, Nelson KE, Purdom EA, et al. (2006) Molecular analysis of the bacterial microbiota in the human stomach. Proc Natl Acad Sci U S A 103: 732-737.
20. Paster BJ, Boches SK, Galvin JL, Ericson RE, Lau CN, et al. (2001) Bacterial diversity in human subgingival plaque. J Bacteriol 183: 3770-3783.
21. Xie G, Chain PS, Lo CC, Liu KL, Gans J, et al. (2010) Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. Mol Oral Microbiol 25: 391-405.
22. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355-1359.
23. Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simon-Soro A, et al. (2011) The oral metagenome in health and disease. ISME J.
24. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, et al. (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res 14: 169-181.
25. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59-65.
26. Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, et al. (2009) Shotgun metaproteomics of the human distal gut microbiota. ISME J 3: 179-189.
27. Li M, Wang IX, Li Y, Bruzel A, Richards AL, et al. (2011) Widespread RNA and DNA sequence differences in the human transcriptome. Science 333: 53-58.
28. Torsvik V, Goksoyr J, Daae FL (1990) High diversity in DNA of soil bacteria. Appl Environ Microbiol 56: 782-787.
29. Alberts B (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 92: 291-294.
30. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141-147.
31. Bailly J, Fraissinet-Tachet L, Verner MC, Debaud JC, Lemaire M, et al. (2007) Soil eukaryotic functional diversity, a metatranscriptomic approach. ISME J 1: 632-642.
32. Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, et al. (2008) Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A 105: 3805-3810.
33. Gilbert JA, Field D, Huang Y, Edwards R, Li W, et al. (2008) Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS One 3: e3042.
34. Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP, et al. (2009) Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. Environ Microbiol 11: 1358-1375.
35. Urich T, Lanzen A, Qi J, Huson DH, Schleper C, et al. (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS One 3: e2527.
36. Shi Y, Tyson GW, DeLong EF (2009) Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. Nature 459: 266-269.

37. Mahowald MA, Rey FE, Seedorf H, Turnbaugh PJ, Fulton RS, et al. (2009) Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. Proc Natl Acad Sci U S A 106: 5859-5864.
38. Klaassens ES, Boesten RJ, Haarman M, Knol J, Schuren FH, et al. (2009) Mixed-species genomic microarray analysis of fecal samples reveals differential transcriptional responses of bifidobacteria in breast- and formula-fed infants. Appl Environ Microbiol 75: 2668-2676.
39. Turnbaugh PJ, Quince C, Faith JJ, McHardy AC, Yatsunenko T, et al. (2010) Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. Proc Natl Acad Sci U S A 107: 7503-7508.
40. Booijink CC, Boekhorst J, Zoetendal EG, Smidt H, Kleerebezem M, et al. (2010) Metatranscriptome analysis of the human fecal microbiota reveals subject-specific expression profiles, with genes encoding proteins involved in carbohydrate metabolism being dominantly expressed. Appl Environ Microbiol 76: 5533-5540.
41. Gosalbes MJ, Durban A, Pignatelli M, Abellan JJ, Jimenez-Hernandez N, et al. (2011) Metatranscriptomic approach to analyze the functional human gut microbiota. PLoS One 6: e17447.
42. Ram RJ, Verberkmoes NC, Thelen MP, Tyson GW, Baker BJ, et al. (2005) Community proteomics of a natural microbial biofilm. Science 308: 1915-1920.
43. Lo I, Denef VJ, Verberkmoes NC, Shah MB, Goltsman D, et al. (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. Nature 446: 537-541.
44. Wilmes P, Wexler M, Bond PL (2008) Metaproteomics provides functional insight into activated sludge wastewater treatment. PLoS One 3: e1778.
45. Giovannoni SJ, Bibbs L, Cho JC, Stapels MD, Desiderio R, et al. (2005) Proteorhodopsin in the ubiquitous marine bacterium SAR11. Nature 438: 82-85.
46. Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, et al. (2009) Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. ISME J 3: 93-105.
47. Klaassens ES, de Vos WM, Vaughan EE (2007) Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. Appl Environ Microbiol 73: 1388-1392.
48. Rudney JD, Xie H, Rhodus NL, Ondrey FG, Griffin TJ (2010) A metaproteomic analysis of the human salivary microbiota by three-dimensional peptide fractionation and tandem mass spectrometry. Mol Oral Microbiol 25: 38-49.
49. Grant MM, Creese AJ, Barr G, Ling MR, Scott AE, et al. (2010) Proteomic analysis of a noninvasive human model of acute inflammation and its resolution: the twenty-one day gingivitis model. J Proteome Res 9: 4732-4744.
50. Deutscher MP (2003) Degradation of stable RNA in bacteria. J Biol Chem 278: 45041-45044.
51. Condon C (2007) Maturation and degradation of RNA in bacteria. Curr Opin Microbiol 10: 271-278.
52. Yoder-Himes DR, Chain PS, Zhu Y, Wurtzel O, Rubin EM, et al. (2009) Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing. Proc Natl Acad Sci U S A 106: 3976-3981.

53. Liu JM, Livny J, Lawrence MS, Kimball MD, Waldor MK, et al. (2009) Experimental discovery of sRNAs in Vibrio cholerae by direct cloning, 5S/tRNA depletion and parallel sequencing. Nucleic Acids Res 37: e46.
54. Passalacqua KD, Varadarajan A, Ondov BD, Okou DT, Zwick ME, et al. (2009) Structure and complexity of a bacterial transcriptome. J Bacteriol 191: 3203-3211.
55. Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, et al. (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. PLoS Genet 5: e1000569.
56. Pradet-Balade B, Boulme F, Beug H, Mullner EW, Garcia-Sanz JA (2001) Translation control: bridging the gap between genomics and proteomics? Trends Biochem Sci 26: 225-229.
57. Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19: 1720-1730.
58. Anderson NL, Anderson NG (1998) Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis 19: 1853-1861.
59. Tew KD, Monks A, Barone L, Rosser D, Akerman G, et al. (1996) Glutathione-associated enzymes in the human cell lines of the National Cancer Institute Drug Screening Program. Mol Pharmacol 50: 149-159.
60. Wilmes P, Bond PL (2004) The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. Environ Microbiol 6: 911-920.
61. Gans J, Wolinsky M, Dunbar J (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science 309: 1387-1390.
62. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. Science 308: 1635-1638.
63. Kroes I, Lepp PW, Relman DA (1999) Bacterial diversity within the human subgingival crevice. Proc Natl Acad Sci U S A 96: 14547-14552.
64. Kazor CE, Mitchell PM, Lee AM, Stokes LN, Loesche WJ, et al. (2003) Diversity of bacterial populations on the tongue dorsa of patients with halitosis and healthy patients. J Clin Microbiol 41: 558-563.
65. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, et al. (2010) The human oral microbiome. J Bacteriol 192: 5002-5017.
66. Eng JK, McCormack, A.L., and Yates, J.R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry 5: 976-989.
67. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20: 3551-3567.
68. Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20: 1466-1467.
69. Peterson J, Garges S, Giovanni M, McInnes P, Wang L, et al. (2009) The NIH Human Microbiome Project. Genome Res 19: 2317-2323.
70. Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, et al. (2010) A catalog of reference genomes from the human microbiome. Science 328: 994-999.

71. Erickson AE, Cantarel, B.L., Lamendella, R., Darzi, Y., Mongodin, E.F., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., Raes, J., Verberkmoes, N.C., Fraser-Liggett, C.M., Hettich, R.L., and Jansson, J.K. (2011) Meta-omics reveals human host-microbiota signatures of Crohn's disease.
72. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, et al. (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci U S A 104: 13780-13785.
73. (!!! INVALID CITATION !!!).
74. Chourey K, Jansson J, VerBerkmoes N, Shah M, Chavarria KL, et al. (2010) Direct cellular lysis/protein extraction protocol for soil metaproteomics. J Proteome Res 9: 6615-6622.
75. Wilkins MJ, Verberkmoes NC, Williams KH, Callister SJ, Mouser PJ, et al. (2009) Proteogenomic monitoring of Geobacter physiology during stimulated uranium bioremediation. Appl Environ Microbiol 75: 6591-6599.
76. Cantarel BL, Erickson, A.R., Verberkmoes, N.C., Erickson, B.K., Carey, P.A., Pan, C., Shah, M., Mongodin, E.M., Jansson, J.K., Fraser-Liggett, C.M., Hettich, R.L. (2011) Strategies for Metagenomic-Guided Whole-Community Proteomics of Complex Microbial Environments. Plos One.
77. VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF (2009) Systems biology: Functional analysis of natural microbial consortia using community proteomics. Nat Rev Microbiol 7: 196-205.
78. Second TP, Blethrow JD, Schwartz JC, Merrihew GE, MacCoss MJ, et al. (2009) Dual-pressure linear ion trap mass spectrometer improving the analysis of complex protein mixtures. Anal Chem 81: 7757-7765.
79. Yates JR, Ruse CI, Nakorchevsky A (2009) Proteomics by mass spectrometry: approaches, advances, and applications. Annu Rev Biomed Eng 11: 49-79.
80. Karas M, Bachmann D, Bahr U, Hillenkamp F (1987) Matrix-assisted ultraviolet laser desorption of non-volatile compounds. International Journal of Mass Spectrometry and Ion Processes 78: 53-68.
81. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. Science 246: 64-71.
82. Olsen JV, de Godoy LM, Li G, Macek B, Mortensen P, et al. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. Mol Cell Proteomics 4: 2010-2021.
83. Rooijers K, Kolmeder C, Juste C, Dore J, de Been M, et al. (2011) An iterative workflow for mining the human intestinal metaproteome. BMC Genomics 12: 6.
84. Xie H, Onsongo G, Popko J, de Jong EP, Cao J, et al. (2008) Proteomics analysis of cells in whole saliva from oral cancer patients via value-added three-dimensional peptide fractionation and tandem mass spectrometry. Mol Cell Proteomics 7: 486-498.
85. Apajalahti JH, Sarkilahti LK, Maki BR, Heikkinen JP, Nurminen PH, et al. (1998) Effective recovery of bacterial DNA and percent-guanine-plus-cytosine-based analysis of community structure in the gastrointestinal tract of broiler chickens. Appl Environ Microbiol 64: 4084-4088.

86. Thompson MR, Chourey K, Froelich JM, Erickson BK, VerBerkmoes NC, et al. (2008) Experimental approach for deep proteome measurements from small-scale microbial biomass samples. Anal Chem 80: 9517-9525.
87. Abram F, Gunnigle E, O'Flaherty V (2009) Optimisation of protein extraction and 2-DE for metaproteomics of microbial communities from anaerobic wastewater treatment biofilms. Electrophoresis 30: 4149-4151.
88. Pierre-Alain M, Christophe M, Severine S, Houria A, Philippe L, et al. (2007) Protein extraction and fingerprinting optimization of bacterial communities in natural environment. Microb Ecol 53: 426-434.
89. Ogunseitan OA (1993) Direct extraction of proteins from environmental samples. Journal of Microbiological Methods 17: 273-281.
90. Sowell SM, Abraham PE, Shah M, Verberkmoes NC, Smith DP, et al. (2011) Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. ISME J 5: 856-865.
91. Benndorf D, Balcke GU, Harms H, von Bergen M (2007) Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. ISME J 1: 224-234.
92. VerBerkmoes NC, Shah MB, Lankford PK, Pelletier DA, Strader MB, et al. (2006) Determination and comparison of the baseline proteomes of the versatile microbe Rhodopseudomonas palustris under its major metabolic states. J Proteome Res 5: 287-298.
93. Gygi PM, Licklider, L.J., Peng, J., Gygi, S.P. (2002) In protein analysis: A laboratory manual; Simpson R, editor. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.
94. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, et al. (1999) Direct analysis of protein complexes using mass spectrometry. Nat Biotechnol 17: 676-682.
95. Washburn MP, Wolters D, Yates JR, 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19: 242-247.
96. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res 2: 43-50.
97. McDonald WH, Ohi, R., Mlyamoto, D.T., Mitchison, T.J., Yates, J.R. III (2002) Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. International J Mass Spectrom 219: 245-254.
98. Tabb DL, McDonald WH, Yates JR, 3rd (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J Proteome Res 1: 21-26.
99. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75: 4646-4658.
100. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74: 5383-5392.

101. Craig R, Beavis RC (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom 17: 2310-2316.
102. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, et al. (2004) Open mass spectrometry search algorithm. J Proteome Res 3: 958-964.
103. Tabb DL, Fernando CG, Chambers MC (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res 6: 654-661.
104. Chen T, Kao MY, Tepel M, Rush J, Church GM (2001) A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. J Comput Biol 8: 325-337.
105. Frank A, Pevzner P (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal Chem 77: 964-973.
106. Taylor JA, Johnson RS (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 11: 1067-1075.
107. Frank AM, Savitski MM, Nielsen ML, Zubarev RA, Pevzner PA (2007) De novo peptide sequencing and identification with precision mass spectrometry. J Proteome Res 6: 114-123.
108. Tabb DL, Ma ZQ, Martin DB, Ham AJ, Chambers MC (2008) DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. J Proteome Res 7: 3838-3846.
109. Ma B, Zhang K, Hendrie C, Liang C, Li M, et al. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 17: 2337-2342.
110. Mo L, Dutta D, Wan Y, Chen T (2007) MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. Anal Chem 79: 4870-4878.
111. Pan C, Park BH, McDonald WH, Carey PA, Banfield JF, et al. (2010) A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. BMC Bioinformatics 11: 118.
112. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457: 480-484.
113. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, et al. (2008) Evolution of mammals and their gut microbes. Science 320: 1647-1651.
114. Shipman JA, Berleman JE, Salyers AA (2000) Characterization of four outer membrane proteins involved in binding starch to the cell surface of Bacteroides thetaiotaomicron. J Bacteriol 182: 5365-5372.
115. Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, et al. (2007) Evolution of Symbiotic Bacteria in the Distal Human Intestine. PLoS Biol 5: e156.
116. Bjursell MK, Martens EC, Gordon JI (2006) Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, Bacteroides thetaiotaomicron, to the suckling period. J Biol Chem 281: 36269-36279.
117. Sonnenburg JL, Xu J, Leip DD, Chen CH, Westover BP, et al. (2005) Glycan foraging in vivo by an intestine-adapted bacterial symbiont. Science 307: 1955-1959.

118. Martens EC, Chiang HC, Gordon JI (2008) Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. Cell Host Microbe 4: 447-457.
119. Sonnenburg JL, Chen CT, Gordon JI (2006) Genomic and metabolic studies of the impact of probiotics on a model gut symbiont and host. PLoS Biol 4: e413.
120. Lozupone CA, Hamady M, Cantarel BL, Coutinho PM, Henrissat B, et al. (2008) The convergence of carbohydrate active gene repertoires in human gut microbes. Proc Natl Acad Sci U S A 105: 15076-15081.
121. Ley RE, Peterson DA, Gordon JI (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell 124: 837-848.
122. McHardy AC, Goesmann A, Puhler A, Meyer F (2004) Development of joint application strategies for two microbial gene finders. Bioinformatics 20: 1622-1631.
123. Samuel BS, Gordon JI (2006) A humanized gnotobiotic mouse model of host-archaeal-bacterial mutualism. Proc Natl Acad Sci U S A 103: 10011-10016.
124. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 57: 289-300.
125. Thompson MR, VerBerkmoes NC, Chourey K, Shah M, Thompson DK, et al. (2007) Dosage-dependent proteome response of Shewanella oneidensis MR-1 to acute chromate challenge. J Proteome Res 6: 1745-1757.
126. Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4: 207-214.
127. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, et al. (2006) Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. J Proteome Res 5: 2339-2347.
128. Liu H, Sadygov RG, Yates JR, 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76: 4193-4201.
129. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37: D233-238.
130. Brigham CJ, Malamy MH (2005) Characterization of the RokA and HexA broad-substrate-specificity hexokinases from Bacteroides fragilis and their role in hexose and N-acetylglucosamine utilization. J Bacteriol 187: 890-901.
131. Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. Nature 444: 1022-1023.
132. Barcenilla A, Pryde SE, Martin JC, Duncan SH, Stewart CS, et al. (2000) Phylogenetic relationships of butyrate-producing bacteria from the human gut. Appl Environ Microbiol 66: 1654-1661.
133. Duncan SH, Lobley GE, Holtrop G, Ince J, Johnstone AM, et al. (2008) Human colonic microbiota associated with diet, obesity and weight loss. Int J Obes (Lond) 32: 1720-1724.

134. Koropatkin NM, Martens EC, Gordon JI, Smith TJ (2008) Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. Structure 16: 1105-1115.
135. Cuff MA, Lambert DW, Shirazi-Beechey SP (2002) Substrate-induced regulation of the human colonic monocarboxylate transporter, MCT1. J Physiol 539: 361-371.
136. Li F, Hinderberger J, Seedorf H, Zhang J, Buckel W, et al. (2008) Coupled ferredoxin and crotonyl coenzyme A (CoA) reduction with NADH catalyzed by the butyryl-CoA dehydrogenase/Etf complex from Clostridium kluyveri. J Bacteriol 190: 843-850.
137. Lecona E, Barrasa JI, Olmo N, Llorente B, Turnay J, et al. (2008) Upregulation of annexin A1 expression by butyrate in human colon adenocarcinoma cells: role of p53, NF-Y, and p38 mitogen-activated protein kinase. Mol Cell Biol 28: 4665-4674.
138. Candido EP, Reeves R, Davie JR (1978) Sodium butyrate inhibits histone deacetylation in cultured cells. Cell 14: 105-113.
139. Tabuchi Y, Takasaki I, Doi T, Ishii Y, Sakai H, et al. (2006) Genetic networks responsive to sodium butyrate in colonic epithelial cells. FEBS Lett 580: 3035-3041.
140. Joseph J, Mudduluru G, Antony S, Vashistha S, Ajitkumar P, et al. (2004) Expression profiling of sodium butyrate (NaB)-treated cells: identification of regulation of genes related to cytokine signaling and cancer metastasis by NaB. Oncogene 23: 6304-6315.
141. Roediger WE (1982) Utilization of nutrients by isolated epithelial cells of the rat colon. Gastroenterology 83: 424-429.
142. Daly K, Shirazi-Beechey SP (2006) Microarray analysis of butyrate regulated genes in colonic epithelial cells. DNA Cell Biol 25: 49-62.
143. Comalada M, Bailon E, de Haro O, Lara-Villoslada F, Xaus J, et al. (2006) The effects of short-chain fatty acids on colon epithelial proliferation and survival depend on the cellular phenotype. J Cancer Res Clin Oncol 132: 487-497.
144. Hooper LV, Xu J, Falk PG, Midtvedt T, Gordon JI (1999) A molecular sensor that allows a gut commensal to control its nutrient foundation in a competitive ecosystem. Proc Natl Acad Sci U S A 96: 9833-9838.
145. Duncan SH, Belenguer A, Holtrop G, Johnstone AM, Flint HJ, et al. (2007) Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. Appl Environ Microbiol 73: 1073-1078.
146. Hubbell SP (2006) Neutral theory and the evolution of ecological equivalence. Ecology 87: 1387-1398.
147. Dethlefsen L, Eckburg PB, Bik EM, Relman DA (2006) Assembly of the human intestinal microbiota. Trends Ecol Evol 21: 517-523.
148. Benson AK, Kelly SA, Legge R, Ma F, Low SJ, et al. (2010) Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proc Natl Acad Sci U S A 107: 18933-18938.
149. Walter J, Ley RE (2010) The Human Gut Microbiome: Ecology and Recent Evolutionary Changes. Annu Rev Microbiol.

150. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, et al. (2009) The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Sci Transl Med 1: 6ra14.
151. Muegge BD, Kuczynski J, Knights D, Clemente JC, Gonzalez A, et al. (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science 332: 970-974.
152. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, et al. (2011) Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science.
153. Diamond J (2002) Evolution, consequences and future of plant and animal domestication. Nature 418: 700-707.
154. Turnbaugh PJ, Backhed F, Fulton L, Gordon JI (2008) Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. Cell Host Microbe 3: 213-223.
155. Faith JJ, McNulty NP, Rey FE, Gordon JI (2011) Predicting a human gut microbiota's response to diet in gnotobiotic mice. Science 333: 101-104.
156. Jernberg C, Lofmark S, Edlund C, Jansson JK (2007) Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. ISME J 1: 56-66.
157. Peterson DA, Frank DN, Pace NR, Gordon JI (2008) Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases. Cell Host Microbe 3: 417-427.
158. Dicksved J, Halfvarson J, Rosenquist M, Jarnerot G, Tysk C, et al. (2008) Molecular analysis of the gut microbiota of identical twins with Crohn's disease. ISME J 2: 716-727.
159. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI (2005) Host-bacterial mutualism in the human intestine. Science 307: 1915-1920.
160. Markert S, Arndt C, Felbeck H, Becher D, Sievert SM, et al. (2007) Physiological proteomics of the uncultured endosymbiont of Riftia pachyptila. Science 315: 247-250.
161. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A (2005) FFAS03: a server for profile--profile sequence alignments. Nucleic Acids Res 33: W284-288.
162. Florens L, Carozza MJ, Swanson SK, Fournier M, Coleman MK, et al. (2006) Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. Methods 40: 303-311.
163. Drake HL, Gossner AS, Daniel SL (2008) Old acetogens, new light. Ann N Y Acad Sci 1125: 100-128.
164. Havemann GD, Bobik TA (2003) Protein content of polyhedral organelles involved in coenzyme B12-dependent degradation of 1,2-propanediol in Salmonella enterica serovar Typhimurium LT2. J Bacteriol 185: 5086-5095.
165. Ayala-Castro C, Saini A, Outten FW (2008) Fe-S cluster assembly pathways in bacteria. Microbiol Mol Biol Rev 72: 110-125, table of contents.
166. Mukhopadhyay A, Redding AM, Joachimiak MP, Arkin AP, Borglin SE, et al. (2007) Cell-wide responses to low-oxygen exposure in Desulfovibrio vulgaris Hildenborough. J Bacteriol 189: 5996-6010.
167. Derensy-Dron D, Krzewinski F, Brassart C, Bouquelet S (1999) Beta-1,3-galactosyl-N-acetylhexosamine phosphorylase from Bifidobacterium bifidum

DSM 20082: characterization, partial purification and relation to mucin degradation. Biotechnol Appl Biochem 29 (Pt 1): 3-10.
168. Nishimoto M, Kitaoka M (2007) Identification of N-acetylhexosamine 1-kinase in the complete lacto-N-biose I/galacto-N-biose metabolic pathway in Bifidobacterium longum. Appl Environ Microbiol 73: 6444-6449.
169. Chang HJ, Sheu SY, Lo SJ (1999) Expression of foreign antigens on the surface of Escherichia coli by fusion to the outer membrane protein traT. J Biomed Sci 6: 64-70.
170. Saint N, El Hamel C, De E, Molle G (2000) Ion channel formation by N-terminal domain: a common feature of OprFs of Pseudomonas and OmpA of Escherichia coli. FEMS Microbiol Lett 190: 261-265.
171. Rosenstiel P, Sina C, End C, Renner M, Lyer S, et al. (2007) Regulation of DMBT1 via NOD2 and TLR4 in intestinal epithelial cells modulates bacterial recognition and invasion. J Immunol 178: 8203-8211.
172. Ligtenberg AJ, Veerman EC, Nieuw Amerongen AV, Mollenhauer J (2007) Salivary agglutinin/glycoprotein-340/DMBT1: a single molecule with variable composition and with different functions in infection, inflammation and cancer. Biol Chem 388: 1275-1289.
173. Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol 25: 117-124.
174. Denef VJ, Shah MB, Verberkmoes NC, Hettich RL, Banfield JF (2007) Implications of strain- and species-level sequence divergence for community and isolate shotgun proteomic analysis. J Proteome Res 6: 3152-3161.
175. Ronaghi M, Uhlen M, Nyren P (1998) A sequencing method based on real-time pyrophosphate. Science 281: 363-+.
176. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53-59.
177. Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, et al. (2008) Discovery and revision of Arabidopsis genes by proteogenomics. Proc Natl Acad Sci U S A 105: 21034-21038.
178. Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, et al. (2006) Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biol 7: R35.
179. Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, et al. (2006) 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7: 275.
180. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, et al. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nature Methods 6: 639-641.
181. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, et al. (2010) Microbes and Health Sackler Colloquium: Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A.
182. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biol 5: R12.

183. Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 15: 387-396.
184. Hoff KJ, Lingner T, Meinicke P, Tech M (2009) Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res 37: W101-105.
185. Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 38: e132.
186. Rho M, Tang H, Ye Y (2010) FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res 38: e191.
187. Noguchi H, Park J, Takagi T (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res 34: 5623-5630.
188. Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132: 185-219.
189. Olsen JV, Ong SE, Mann M (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. Molecular & Cellular Proteomics 3: 608-614.
190. Callister SJ, Wilkins MJ, Nicora CD, Williams KH, Banfield JF, et al. (2010) Analysis of biostimulated microbial communities from two field experiments reveals temporal and spatial differences in proteome profiles. Environ Sci Technol 44: 8897-8903.
191. Sowell SM, Abraham PE, Shah M, Verberkmoes NC, Smith DP, et al. (2010) Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. ISME J.
192. Hsieh EJ, Hoopmann MR, MacLean B, MacCoss MJ (2010) Comparison of database search strategies for high precursor mass accuracy MS/MS data. Journal of Proteome Research 9: 1138-1143.
193. Mackey AJ, Haystead TA, Pearson WR (2002) Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. Molecular & Cellular Proteomics 1: 139-147.
194. Kim S, Gupta N, Bandeira N, Pevzner PA (2009) Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. Molecular & Cellular Proteomics 8: 53-69.
195. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, et al. (2009) Bacterial community variation in human body habitats across space and time. Science 326: 1694-1697.
196. Erlich Y, Chang K, Gordon A, Ronen R, Navon O, et al. (2009) DNA Sudoku--harnessing high-throughput sequencing for multiplexed specimen analysis. Genome Res 19: 1243-1253.
197. Tamboli CP, Neut C, Desreumaux P, Colombel JF (2004) Dysbiosis as a prerequisite for IBD. Gut 53: 1057.
198. Willing B, Halfvarson J, Dicksved J, Rosenquist M, Jarnerot G, et al. (2009) Twin studies reveal specific imbalances in the mucosa-associated microbiota of patients with ileal Crohn's disease. Inflamm Bowel Dis 15: 653-660.
199. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, et al. (2006) Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut 55: 205-211.

200. Raes J, Bork P (2008) Molecular eco-systems biology: towards an understanding of community function. Nat Rev Microbiol 6: 693-699.
201. Willing BP, Dicksved J, Halfvarson J, Andersson AF, Lucio M, et al. (2010) A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. Gastroenterology 139: 1844-1854 e1841.
202. Jansson J, Willing B, Lucio M, Fekete A, Dicksved J, et al. (2009) Metabolomics reveals metabolic biomarkers of Crohn's disease. PLoS One 4: e6386.
203. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, et al. (2011) Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A 108 Suppl 1: 4680-4687.
204. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
205. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575-1584.
206. Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, et al. (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res 38: D190-195.
207. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32: D277-280.
208. Canani RB, Costanzo MD, Leone L, Pedata M, Meli R, et al. (2011) Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J Gastroenterol 17: 1519-1528.
209. Sartor RB (2000) New therapeutic approaches to Crohn's disease. N Engl J Med 342: 1664-1666.
210. Fiocchi C (1998) Inflammatory bowel disease: etiology and pathogenesis. Gastroenterology 115: 182-205.
211. Perera LaM, L. (2005) Immunologic defects underlying the IBD. Gastroenterology & Hepatology 1: 108-116.
212. Wexler HM (2002) Outer-membrane pore-forming proteins in gram-negative anaerobic bacteria. Clin Infect Dis 35: S65-71.
213. Sato K, Kumita W, Ode T, Ichinose S, Ando A, et al. (2010) OmpA variants affecting the adherence of ulcerative colitis-derived Bacteroides vulgatus. J Med Dent Sci 57: 55-64.
214. Soulas C, Baussant T, Aubry JP, Delneste Y, Barillat N, et al. (2000) Outer membrane protein A (OmpA) binds to and activates human macrophages. J Immunol 165: 2335-2340.
215. Curtis MA, Hanley SA, Aduse-Opoku J (1999) The rag locus of Porphyromonas gingivalis: a novel pathogenicity island. J Periodontal Res 34: 400-405.
216. Shi X, Hanley SA, Faray-Kele MC, Fawell SC, Aduse-Opoku J, et al. (2007) The rag locus of Porphyromonas gingivalis contributes to virulence in a murine model of soft tissue destruction. Infect Immun 75: 2071-2074.

217. Sieira R, Comerci DJ, Pietrasanta LI, Ugalde RA (2004) Integration host factor is involved in transcriptional regulation of the Brucella abortus virB operon. Mol Microbiol 54: 808-822.
218. Stonehouse E, Kovacikova G, Taylor RK, Skorupski K (2008) Integration host factor positively regulates virulence gene expression in Vibrio cholerae. J Bacteriol 190: 4736-4748.
219. Okazaki N, Takahashi N, Kojima S, Masuho Y, Koga H (2002) Protocadherin LKC, a new candidate for a tumor suppressor of colon and liver cancers, its association with contact inhibition of cell proliferation. Carcinogenesis 23: 1139-1148.
220. Graham CA, McLean WH, Hughes AE, Nevin NC (1988) Characterization of human skin fibroblast extracellular proteins by two-dimensional polyacrylamide gel electrophoresis. Electrophoresis 9: 343-351.
221. Schroder JM, Harder J (1999) Human beta-defensin-2. Int J Biochem Cell Biol 31: 645-651.
222. Ayabe T, Satchell DP, Wilson CL, Parks WC, Selsted ME, et al. (2000) Secretion of microbicidal alpha-defensins by intestinal Paneth cells in response to bacteria. Nat Immunol 1: 113-118.
223. Wilson CL, Ouellette AJ, Satchell DP, Ayabe T, Lopez-Boado YS, et al. (1999) Regulation of intestinal alpha-defensin activation by the metalloproteinase matrilysin in innate host defense. Science 286: 113-117.
224. Ravi K, Chari ST, Vege SS, Sandborn WJ, Smyrk TC, et al. (2009) Inflammatory bowel disease in the setting of autoimmune pancreatitis. Inflamm Bowel Dis 15: 1326-1330.

Vita

Alison Russell Erickson earned her Bachelor of Science degree in Forensic Science from Baylor University in 2003. She then earned her Master of Science degree in Biochemistry at Texas State University in 2006. She expects to complete her Ph.D. dissertation work in the Genome Science and Technology Graduate School at the University of Tennessee – Knoxville in October 2011. She lives in Knoxville, TN with her husband, Brian, dogs, Abby and Dexter, and bilingual parrot, Chester.