



12-2012

Functional Characterization of Microbial Symbiotic Associations by Metaproteomics

Jacque Caprio Young
jcaprio@utk.edu

Recommended Citation

Young, Jacque Caprio, "Functional Characterization of Microbial Symbiotic Associations by Metaproteomics." PhD diss., University of Tennessee, 2012.
https://trace.tennessee.edu/utk_graddiss/1595

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

To the Graduate Council:

I am submitting herewith a dissertation written by Jacque Caprio Young entitled "Functional Characterization of Microbial Symbiotic Associations by Metaproteomics." 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:

Steven Wilhelm, Kurt Lamour, Mircea Podar, Loren Hauser

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

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

**Functional Characterization of Microbial Symbiotic Associations by
Metaproteomics**

A Dissertation
Presented for the
Doctor of Philosophy
Degree

The University of Tennessee, Knoxville

Jacque Caprio Young
December 2012

Copyright © 2012 by Jacque Caprio Young
All rights reserved.

This dissertation is dedicated to my three wonderful children: Cody, Camryn, and Colin, who are my pride, joy, and inspiration for all I do.

Acknowledgments

First, I would like to thank my mentor, Dr. Bob Hettich, whose support and dedication have guided me through this challenging endeavor, and taught me not only about science, but about the kind of person and mentor I would like to one day become. I would also like to send my sincere gratitude to my dissertation committee members: Dr. Steve Wilhelm, Dr. Kurt Lamour, Dr. Mircea Podar, and Dr. Loren Hauser. Each has gone above and beyond to offer me support and valuable input throughout this journey. In addition, I would like to thank all of the GST students and ORNL staff in our group who have provided a team environment and support throughout the last few years. Namely, I would like to acknowledge Greg Hurst and Rich Giannone for their valuable input.

In addition, during my dissertation work, I had the privilege of working with and learning from fantastic collaborators. Specifically, I would like to acknowledge Dr. Rodolphe Barrangou, and Dr. Philippe Horvath from Danisco, Inc., Manuel Kleiner and Dr. Nicole Dubilier from the Max Plank Institute for Marine Microbiology, Dr. Mike Morowitz from the University of Pittsburg, and Dr. Jill Banfield and her group at University of California, Berkeley.

I would also like to thank the Genome Science and Technology Program at the University of Tennessee for both educational and financial support. Funding for the CRISPR proteomics project was through the US Department of Energy's Office of Science, Biological and Environmental Research Program. A portion of the gutless marine worm project was funded through Laboratory Directed Research and Development support at the Oak Ridge National Laboratory. And, the infant microbiome portion of this work was funded through Dr. Michael Morowitz through a grant from the March of Dimes. All research was performed at ORNL, which is managed by UT-Battelle, LLC for the U.S. Department of Energy.

Finally, I would like to thank my family and friends for their continued support and love. To my children, thank you for being patient and supportive through this process. Thank you to my parents for your unconditional love and support, and all your help with the kids. And, thanks to all my friends, including my softball, baseball, and football moms and dads who have taken my kids to games and practices, and provided valuable emotional support.

Abstract

Rarely are microbes found in isolation in the environment, but rather form *symbiotic associations* with other microbes or eukaryotic hosts. The advent of the systems biology era has allowed global characterization of these symbiotic associations at levels not previously possible. However, while metagenomic studies have revealed microbial membership and *potential* genomic information among members taking part in the symbiosis, there is still a significant lag in the functional characterization within these symbiotic associations. Thus, in this dissertation, we utilized a metaproteomic approach to study microbial symbiotic associations. We have developed and applied this robust platform to investigate various symbiotic associations ranging in complexity. Beginning with perhaps one of the simplest symbiotic systems, we investigated the proteomic response of infection of *S. thermophilus* with bacteriophage 2972, to reveal insights into the anti-viral CRISPR/Cas response. Then, transitioning to a more complex but tractable symbiotic interaction, we evaluated co-occurring proteobacterial endosymbionts of the marine worm *Olavius algarvensis* and uncovered novel pathways for carbon and energy use, in addition to unraveling abundant transposase protein expression. Finally, we progressed to a complex microbial community and its commensalistic association with its human host in the infant gut microbiome. Simultaneous measurements of microbial and human proteins over a time course during early infant development revealed functional adaptation of the host in response to the changing microbiome, resulting in a dynamic interplay between the host and its resident microbes. In each of these symbiotic systems, we found that a proteomics/metaproteomics approach was very powerful for the characterization of the functional signatures of all members of the symbiotic interaction, and yielded biological insights into each system that would have been unattainable by any other platform.

Table of Contents

<u>CHAPTER ONE: INTRODUCTION TO PROTEOMIC ELUCIDATION OF MICROBIAL SYMBIOTIC ASSOCIATIONS</u>	1
MICROBIAL SYMBIOTIC RELATIONSHIPS	1
SYMBIOSIS IN THE SYSTEMS BIOLOGY ERA	2
MASS SPECTROMETRY-BASED SHOTGUN PROTEOMICS AND METAPROTEOMICS	4
VIRUS: HOST SYMBIOTIC INTERACTIONS	6
SYMBIOTIC ASSOCIATIONS OF CHEMOSYNTHETIC BACTERIA WITH EUKARYOTIC HOSTS	8
COMMENSALISM IN THE HUMAN GUT MICROBIOME	10
SYMBIOSIS BEGINS: COLONIZATION IN THE NEWBORN INFANT	12
TRANSPOSABLE ELEMENTS: PARASITES OR NOT?	14
OBJECTIVE / SCOPE OF DISSERTATION	15
<u>CHAPTER TWO: EXPERIMENTAL DESIGN FOR PROTEOMIC ELUCIDATION OF MICROBIAL SYMBIOTIC INTERACTIONS</u>	17
SHOTGUN PROTEOMICS VIA NANO-2D-LC-MS/MS: EXPERIMENTAL OVERVIEW	17
SAMPLE PREPARATION FOR MASS SPECTROMETRY MEASUREMENTS	19
MASS SPECTROMETRY INSTRUMENTATION	26
MS-BASED INFORMATICS	29
<u>CHAPTER THREE: PHAGE-INDUCED EXPRESSION OF CRISPR-ASSOCIATED PROTEINS IS REVEALED BY SHOTGUN PROTEOMICS IN STREPTOCOCCUS THERMOPHILUS</u>	35
ABSTRACT	35
INTRODUCTION	36
MATERIALS AND METHODS	39
RESULTS	42
DISCUSSION	54
<u>CHAPTER FOUR: METAPROTEOMICS OF A GUTLESS MARINE WORM AND ITS SYMBIOTIC MICROBIAL COMMUNITY REVEAL UNUSUAL PATHWAYS FOR CARBON AND ENERGY USE</u>	60
ABSTRACT	60
INTRODUCTION	61
MATERIALS AND METHODS	63
RESULTS/DISCUSSION	66
CONCLUSIONS	87
<u>CHAPTER FIVE: ABUNDANT TRANSPOSASE EXPRESSION IN MUTUALISTIC ENDOSYMBIONTS IS REVEALED BY METAPROTEOMICS</u>	90
ABSTRACT	90
INTRODUCTION	91
MATERIALS AND METHODS	93
RESULTS AND DISCUSSION	94

CONCLUSIONS	100
<u>CHAPTER SIX: METAPROTEOMICS REVEALS TIME-DEPENDENT FUNCTIONAL SHIFTS IN MICROBIAL AND HUMAN PROTEINS IN THE PREMATURE INFANT GUT</u>	103
ABSTRACT	103
INTRODUCTION	104
MATERIALS AND METHODS	106
RESULTS	110
DISCUSSION	127
<u>CHAPTER SEVEN: APPLYING A METAPROTEOMICS APPROACH UNRAVELS INTRA- AND INTER-INDIVIDUAL VARIATION IN THE PRETERM INFANT GUT MICROBIOME</u>	133
USING METAPROTEOMICS TO MEASURE FECAL MICROBIOMES FROM MULTIPLE INFANTS	133
MICROBIAL SPECIES DISTRIBUTION IN THE CARROL BABY FECAL MICROBIOME	134
HUMAN PROTEINS IN THE CARROL BABY	139
VARIABILITY AMONG RATIOS OF HUMAN AND MICROBIAL PROTEINS ACROSS MULTIPLE INFANT FECAL MICROBIOME SAMPLES	142
INITIAL METHOD DEVELOPMENT: BUILDING THE OPTIMAL SEARCH DATABASE	145
IMPROVEMENT WITH NEXT-GENERATION MASS SPECTROMETERS	147
ATTEMPTS TO DEplete ABUNDANT HUMAN PROTEINS: MASS EXCLUSION LIST & DIFFERENTIAL CENTRIFUGATION	148
ADDITIONAL ATTEMPTS TO DEplete ABUNDANT HUMAN PROTEINS IN FECAL SAMPLES	150
<u>CHAPTER EIGHT: CONCLUSIONS AND INSIGHTS INTO MICROBIAL SYMBIOTIC INTERACTIONS GAINED THROUGH METAPROTEOMIC INVESTIGATIONS</u>	153
<u>LIST OF REFERENCES</u>	161
<u>VITA</u>	178

List of Tables

TABLE 2.1: FIGURES OF MERIT COMPARISONS BETWEEN MASS SPECTROMETERS	27
TABLE 3.1: NUMBER OF PROTEINS, PEPTIDES, AND SPECTRA IDENTIFIED BY LC-MS/MS IN CELLULAR (MOI= 0.1 AND 1) OR PEG-ENRICHED VIRAL FRACTIONS (MOI=1) AT EACH TIME POINT OF INFECTION	43
TABLE 3.2: SEQUENCE COVERAGES OF PHAGE 2972 PROTEINS FROM VIRUS-ENRICHED AND CELLULAR FRACTIONS ACROSS INFECTION TIME POINTS.	46
TABLE 3.3: EXPRESSION OF CAS PROTEINS FROM <i>S. THERMOPHILUS</i> DGCC710 ACROSS TIME.	53
TABLE 5.1: OVERVIEW OF ALL EXPRESSED TRANSPOSASES GROUPED ACCORDING TO SHARED PEPTIDE MATCHES	96
TABLE 5.2: ORGANISM NORMALIZED NSAF VALUES (RELATIVE ABUNDANCE) FOR THE Γ 1-SYMBIONT TRANSPOSASES	97
TABLE 5.3: ORGANISM NORMALIZED NSAF VALUES FOR THE Δ 1-SYMBIONT TRANSPOSASES	98
TABLE 6.1: NUMBER OF PROTEINS, PEPTIDES, AND MS/MS SPECTRA IDENTIFIED ACROSS ALL TIME POINTS.	111
TABLE 6.2: TOPMOST ABUNDANT HUMAN FECAL PROTEINS WITH RELEVANCE TO HOST-MICROBE INTERACTIONS	122
TABLE 6.3: DETECTED PROTEINS INVOLVED IN EPITHELIAL BARRIER FUNCTIONS	124
TABLE 7.1: OVERVIEW OF PROTEOMIC RESULTS FROM MULTIPLE INFANT FECAL MICROBIOMES	134
TABLE 7.2 MICROBIAL PROTEINS DETECTED PER SPECIES FROM BABY CARROL (INFANT #74)	137
TABLE 7.3: ABUNDANT HUMAN PROTEINS FROM BABY CARROL (INFANT #74)	140
TABLE 7.4: COMPARISON OF ABUNDANT HUMAN PROTEINS IN THE UC1 BABY AND THE CARROL BABY AT SIMILAR TIME POINTS	141
TABLE 7.5: EXPERIMENTAL ATTEMPTS AT DEPLETING ABUNDANT HUMAN PROTEINS	149
TABLE 7.6: COMPARISON OF SDS-TCA AND FASP SAMPLE PREP METHODS	151

List of Figures

FIGURE 1.1: THE SYSTEMS BIOLOGY ERA	4
FIGURE 2.1: EXPERIMENTAL DESIGN: SHOTGUN PROTEOMICS	18
FIGURE 2.2: SCHEMATIC DIAGRAM OF A BIPHASIC COLUMN	22
FIGURE 2.3: EXAMPLE OF MUDPIT EXPERIMENTAL OUTPUT	24
FIGURE 2.4: SCHEMATIC OF PEPTIDE ION FRAGMENTATION PATTERNS	25
FIGURE 2.5: DERIVING AN AMINO ACID SEQUENCE FROM MS/MS SPECTRA	26
FIGURE 2.7: SCHEMATIC DIAGRAM OF AN LTQ-ORBITRAP	28
FIGURE 2.8: DTASELECT OUTPUT	30
FIGURE 3.1: PHAGE 2972 SPECTRAL ABUNDANCES	45
FIGURE 3.2: COG CLASSIFICATION OF <i>S. THERMOPHILUS</i> PROTEOMES ACROSS INFECTION TIME POINTS	49
FIGURE 3.3: VOLCANO PLOT OF PROTEIN ABUNDANCE CHANGES DURING PEAK INFECTION AT MOI=1	50
FIGURE 3.4: RESTRICTION MODIFICATION PROTEIN SUBUNITS INCREASED AT PEAK INFECTION TIMES	51
FIGURE 3.5: CAS PROTEINS CHANGING IN RESPONSE TO PHAGE 2972 INFECTION	52
FIGURE 4.1: POSITIVE EFFECT OF SYMBIONT ENRICHMENT USING DENSITY GRADIENT CENTRIFUGATION	68
FIGURE 4.2: OVERVIEW OF SYMBIOTIC METABOLISM BASED ON METAPROTEOMIC AND METABOLOMIC ANALYSES	70
FIGURE 4.3: MODIFIED VERSION OF THE 3-HYDROXYPROPIONATE BI-CYCLE	77
FIGURE 4.4: SUGGESTED ROLE OF THE PROTON-TRANSLOCATING PYROPHOSPHATASE	81
FIG. 4.5: COMPARISON OF THE "CLASSICAL" CALVIN CYCLE WITH A PROPOSED MORE ENERGY EFFICIENT VERSION	86
FIGURE 5.1: SEQUENCE COVERAGE OF A REPRESENTATIVE TRANSPOSASE	100
FIGURE 6.1: REPRODUCIBILITY BETWEEN TECHNICAL REPLICATES	112
FIGURE 6.2: DISTRIBUTION OF HUMAN AND MICROBIAL PROTEINS	114
FIGURE 6.3: TOTAL UNFILTERED MS2 SPECTRA COLLECTED FOR EACH RUN	114
FIGURE 6.5: MICROBIAL PROTEINS DETECTED ACROSS TIME	116
FIGURE 6.6: ANALYSIS OF MICROBIAL PROTEINS BY COG CATEGORY CLASSIFICATIONS	118
FIGURE 6.7: TOP CANONICAL PATHWAYS	120
FIGURE 6.8: TIGHT JUNCTION SIGNALING PATHWAY	123
FIGURE 6.9: HUMAN PROTEINS CHANGING ACROSS TIME	128
FIGURE 6.10: TOP CANONICAL PATHWAYS EXPRESSED AT DAY 20	130
FIGURE 7.1: MICROBIAL SPECIES DISTRIBUTION FROM INFANT #74 AS DETERMINED BY METAGENOMIC SEQUENCING	136
FIGURE 7.2: MICROBIAL SPECIES DISTRIBUTION FROM BABY CARROL (INFANT #74) AS DETERMINED BY METAPROTEOMICS	138
FIGURE 7.3: DISTRIBUTION OF MICROBIAL AND HUMAN PROTEINS	143
FIGURE 7.4: DISTRIBUTION OF MICROBIAL AND HUMAN SPECTRA	144
FIGURE 7.5: CHOOSING OPTIMAL DATABASE DESIGN	146
FIGURE 7.6: GREATER PROTEOME DEPTH ACHIEVED WITH NEXT-GENERATION VELOS INSTRUMENT	148
FIGURE 7.7: ATTEMPT TO DEplete ABUNDANT HUMAN PROTEINS USING A 50KDA MWCO FILTER	150

List of Symbols and Abbreviations

ABI	Abortive infection
ACN	Acetonitrile
ALPI	Intestinal-type alkaline phosphatase
AMD	Acid mine drainage
ANPEP	Aminopeptidase
BCA	Bicinchronic assay
BIMs	Bacteriophage insensitive mutants
BSL2	Biosafety level 2
C-18	Reverse phase
CARD-FISH	Catalyzed reporter deposition-fluorescence <i>in situ</i> hybridization
CAS	CRISPR-Associated
CID	Collision-induced dissociation
CLCA	Calcium-activated chloride-channel
CLDN	Claudin
CO	Carbon monoxide
COG	Clusters of orthologous groups
CRISPR	Clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
DE	Dynamic exclusion
DEFA	Defensins
DTT	Dithiothreitol
ELANE	Neutrophil elastase
ESI	Electrospray ionization
FA	Formic acid
FASP	Filter aided sample preparation
FCRBP	IgG Fc-receptor binding protein
FDR	False discovery rate
FT	Fourier transformed
FWHM	Full width at half maximum
GO	Gene ontology
HPI	Hours post infection
HPLC	High performance liquid chromatography
IAA	Iodoacetamide
IPA	Ingenuity pathway analysis

IS	Insertion sequence
ITLN	Intelectin
LC	Liquid chromatography
LCN	Lipocalin
LM17	M17 medium supplemented with 0.5% lactose
LTQ	Linear trapping quadrupole
LTF	Lactotransferrin/ Lactoferrin
LYZ	Lysozyme
MHC	Major histocompatibility complex
MOI	Multiplicity of infection
MPO	Myeloperoxidase
MS	Mass spectrometry
MS1	Full mass spectrum
MS/MS	Tandem mass spectra
MTASE	Methyltransferase
MUC	Mucin
MUDPIT	Multidimensional Protein Identification Technology
MWCO	Molecular weight cutoff
<i>M/Z</i>	Mass-to-charge
NANO-ESI	Nanoelectrospray ionization
NEC	Necrotizing enterocolitis
NSAF	Normalized spectral abundance factor
OCLN	Occludin
OD	Optical density
OLFM	Olfactomedin
ORF	Open reading frame
PCA	Principal component analysis
PHA	Polyhydroxyalkanoate
PIGR	Polymeric immunoglobulin receptor
PTM	Post-translational modification
PPI	Inorganic pyrophosphate
PPM	Parts per million
PSM	Peptide spectral matching
RF	Radio frequency
R-M	Restriction modification
RNA-SEQ	RNA sequencing
RP	Reversed-phase

SAP	Single amino acid polymorphism
SCX	Strong cation exchange
SDS	Sodium dodecyl sulfate
TE	Transposable element
TCA	Trichloroacetic acid
TJP	Tight junction protein
VLP	Virus like particles
1D-PAGE	One-dimensional polyacrylamide gel electrophoresis
2D-LC	Two-dimensional liquid chromatography

CHAPTER ONE

Introduction to Proteomic Elucidation of Microbial Symbiotic Associations

Microbial Symbiotic Relationships

Symbiosis, as translated from ancient Greek, means ‘living together’, and in the context of this work, will be focused on *long-term interactions between two or more species* (1). The nature of symbiotic relationships can vary depending on whether they are beneficial for both/all species, detrimental to one species, or one organism benefits without affecting the other(s). These are classified as **mutualistic**, **parasitic**, or **commensalistic**, respectively (1). Microbial symbiotic relationships can range in their complexity and nature, as well as in the number of organisms involved in the relationship. Within the environment, microbes are rarely found in isolation, but rather usually exist and function in some type of symbiotic interaction with other organisms. While some microbes exist in a **free-living** state, or are **facultatively** associated with their host, others are **obligately** host-associated and are incapable of reproducing outside their host. The nature of these states can vary, with organisms transitioning from one to the other. In addition, hosts can acquire symbiotic microbes either through **vertical** transmission (from parent to offspring during reproduction), or **horizontal** transmission (through the environment each host generation) (2). Throughout this dissertation, we will highlight and examine several types of symbiotic interactions, and specify how a proteomics/ metaproteomics approach can provide detailed insights. Specifically, we will begin with investigation of a simple dualistic symbiosis: a single bacterial isolate

Streptococcus thermophilus and its phage 2972. Next, we will move on to a moderately complex system involving interaction of multiple co-symbionts with their invertebrate host: namely **proteobacterial endosymbionts of the gutless marine worm *Olavius algarvensis***. Then, we will progress to a complex microbial community and its commensalistic association with its human host in the **infant gut microbiome**. In addition, we will extend the discussion to the very basic genetic level while examining **transposable elements**, which are oftentimes considered to be parasitic. These various systems were chosen for their range in complexity and nature of their symbiotic associations. In addition, due to limited available studies which use “omics’ platforms in these systems, there is a significant knowledge gap which has led to incomplete characterization of these symbiotic interactions at the global level. However, with rapid advancements in the systems biology field, exciting new avenues have been opened up for symbiosis research. Therefore, in this dissertation work, we have utilized a common **proteomics/ metaproteomics approach** to study these various symbiotic associations, and unraveled novel biological insights unattainable by other available methods.

Symbiosis in the Systems Biology Era

The advent of high throughput DNA sequencing has revolutionized the field of biology such that entire complements of genes can be measured and characterized through a **genomics** approach (Figure 1.1). Likewise, the recent development of RNA sequencing (RNA-Seq) has enabled suites of RNA transcripts to be measured, escalating the field of **transcriptomics** (3). In turn, complete set of proteins can be characterized through **proteomics**, and sets of small molecules measured through **metabolomics**:

primarily due to recent rapid advances in mass spectrometry techniques (4, 5). Each of these –omics techniques provides unique, yet valuable information. While genomic data provides information about the encoded potential of a cell, and transcriptomics tells which genes are turned on, it is only through proteomics, a direct measure of actual proteins produced, that a viewpoint of functional signatures and metabolic activities at a particular time or under certain conditions can be obtained. Through a proteomic approach, proteins can be identified, quantified, localized, and post-translational modifications characterized (6).

While the majority of microbes in the environment still remain unidentified and uncultivable, the ability to “shotgun sequence” environmental samples *in situ* through **metagenomics** has offered the advantage of unraveling previously unattainable information about microbial community memberships and encoded metabolic potentials. Likewise, the rapidly developing field of **metaproteomics**, measuring the complement of proteins expressed by microbial communities in the environment, has offered the advantage of determining functional signatures and metabolic activities of a microbial community as a whole as well as those from individual members. Through the advent of these –omics technologies (Figure 1.1), a new era of *Systems Biology* has arisen in which microbes and microbial communities can be characterized at a level not previously possible, and in turn, novel insights into symbiotic relationships can be obtained.

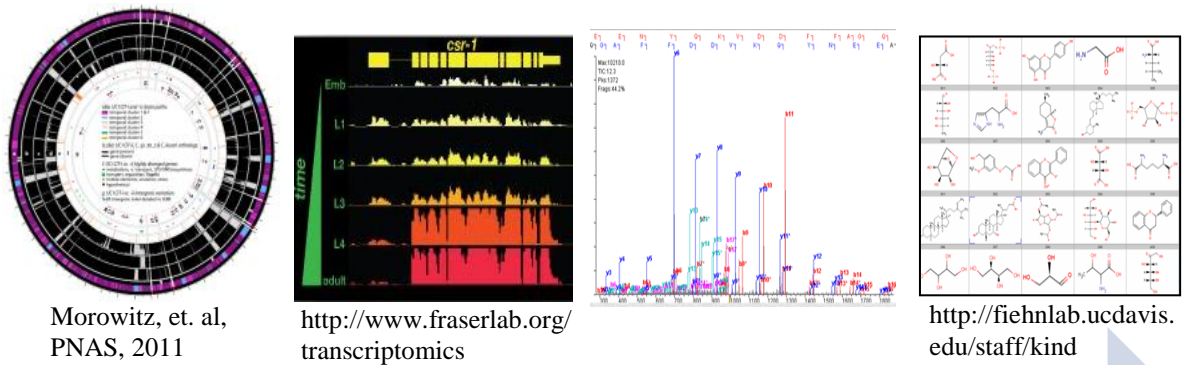


Figure 1.1: The Systems Biology Era. The application of –omics techniques allows characterization of microbial symbiotic relationships at a level not previously possible. Genomics allows characterization of complete set of genes within a cell, as is depicted above for the *Citrobacter* (7) genome. Transcriptomics measures suites of RNA transcripts expressed under certain cellular conditions/ times, for example via RNA-Seq as shown above. Proteomics characterizes complete sets of proteins expressed by a cell, typically via mass spectrometry and metabolomics measures suites of small molecules.

Mass Spectrometry-Based Shotgun Proteomics and Metaproteomics

Mass spectrometry has become an unparalleled platform for proteomic measurements due to very recent advances including: the development of electrospray ionization (ESI) (8), improvements in multi-dimensional chromatographic separations using high performance liquid chromatography (HPLC) (9, 10), rapid advances in mass spectrometry instrumentation (11-15), increased availability of genomic sequence information, and development of MS-based bioinformatic tools and algorithms (16, 17). The combination of these developments have allowed high throughput, sensitive, and accurate measurements which identify thousands of peptides and proteins from microbial isolates, as well as complex environmental samples.

Proteolytic digestion of proteins into peptides, followed by peptide mass and sequence analysis by tandem mass spectrometry, termed “shotgun” proteomics or “bottom-up” proteomics (9, 18) has become the preferred method for large-scale analyses of high-complexity samples. Through this method, microbial proteomes can be characterized with greater depth, accuracy, and levels of quantitation than ever observed before (9, 19-23). Specifically, it is possible to identify 1500-3500 proteins from a single growth state of a microbial isolate in 1-2 days, as well as make quantitative comparisons on hundreds to thousands of these proteins (22, 24, 25). Furthermore, by analyzing multiple growth states, it is possible to identify 50-90% of the predicted proteome from microbial isolates (26). In addition, complex microbial community proteomes can now be measured at depths in which representation of even the lower abundance organisms can be obtained.

It has just been within the last decade that landmark studies were carried out from microbial communities in Acid Mine Drainage (AMD) biofilms. These include the first metagenomic sampling and reconstruction of whole genomes directly from environmental samples (27) followed by subsequent metaproteomic characterization (28), and strain level resolution to infer sequence types via proteogenomics (29, 30). Studies from this system have revealed novel biological insights into ecological divergence, niche partitioning, and metabolic roles of different members within the community (30-32). These and other groundbreaking investigations, in combination with technological advancements, have paved the way for additional metaproteomics analyses in more complex microbial ecosystems including plants, soil, oceans, and the human body (25,

33-36), and enabled comprehensive characterization of symbiotic relationships discussed throughout this dissertation work.

Virus: Host Symbiotic Interactions

Viruses are typically classified as obligate intracellular parasites, relying on the host's cell machinery to replicate and reproduce. However, viruses can oftentimes affect their microbial hosts in a beneficial way by transmitting genes which add certain capabilities to the microbial host, such as toxins, virulence genes, or other functional genes (37). Bacteriophages, viruses that infect bacteria, are the most abundant and ubiquitous organisms on the planet, with an estimated 10^{31} particles found in seawater alone (38). Phages are found in virtually every natural environment, including seawater, soil, and the human body, and play a major role in the structure and function of microbial communities by altering host fitness, facilitating genetic exchange, and driving host evolution. In addition, phage are ubiquitous in many industrial settings, such as microbial fermentation processes used in making yogurt and cheeses, where they can prove detrimental in interrupting batch fermentation (39).

Despite the fact that the first completely sequenced genome was bacteriophage phiX174 (40) and to date, over 3,000 phage genomes have been sequenced (41), the majority of viruses are still uncultivated and unidentified, and thus one of the most unexplored life forms on Earth. The newly emerging field of viral metgenomics allows *in situ* high-throughput sequencing of viruses and virus-like particles from environmental samples; however, it is estimated that 65-68% of virus sequences show no similarity to any other sequences in the NCBI non-redundant database, and most viral open reading

frames (ORFs) are novel (42). In addition, there is no single gene common to all viruses, like the 16S rRNA gene in bacteria, which can aid in their identification. While viral metagenomics studies have provided biological insights into microbial ecosystem functions, for example, in the human gut microbiome (1, 41, 43), and acid mine drainage communities (44), the field of phage proteomics has lagged behind (45) consisting mostly of studies characterizing bacteriophage structural proteomes (proteins comprising the structure of the phage particles) (46-49), with only one study elucidating the proteomic response of the host upon phage infection (26) (discussed in chapter 4).

The parasitic nature of bacteriophage is to infect microbial hosts, using the host's machinery to replicate itself, and either lyse the host (in the case of lytic phages) or incorporate itself into the host's chromosome (lysogenic phages). However, bacteria have developed counter strategies to protect themselves against phage attack including anti-viral mechanisms such as the restriction-modification (R-M) system, abortive infection system (ABI), and the CRISPR-Cas system (Clustered Regularly Interspaced Short Palindromic Repeats) (discussed in chapter 3). This dynamic interplay between phage and bacteria is constantly evolving, redefining the nature of this symbiotic relationship.

Studying the interaction of a bacteriophage infecting its bacterial host, while dynamic in nature, is a relatively simple (two-party) interaction. This becomes more complicated when multiple phage and different microbial species are present in an ecosystem, which is usually the case in the environment, where the rise and fall of different microbial populations can be regulated by bacteriophage predator-prey interactions (50). Hence, the complexity of symbiotic associations increases dramatically when dealing with multiple microbial members. In the following section, we will begin to

look at increasingly complex symbiotic interactions while examining multiple co-occurring bacterial symbionts, and how they interact with each other and their eukaryotic host.

Symbiotic Associations of Chemosynthetic Bacteria with Eukaryotic Hosts

Many marine invertebrates form symbiotic associations with chemoautotrophic bacteria that prove beneficial to both parties by increasing their metabolic capabilities, and thus the number of ecological niches (51) . Specifically, in these chemosynthetic symbiotic associations, chemoautotrophic microbes provide the host with organic compounds through CO₂ fixation using reduced compounds, while the host provides access to substrates needed for the symbionts energy and biomass (51-53). The first demonstration of a chemoautotrophic bacteria-marine invertebrate association was discovered in *Riftia pachyptila*, a gutless tubeworm found in hydrothermal vents possessing a γ - Proteobacterial symbiont, which is solely responsible for providing nutrition to the worm by oxidizing sulfide and fixing carbon dioxide (51, 53). While this is an eloquent example of one symbiont providing for the needs of its host, many marine invertebrates house multiple co-occurring symbionts.

A notable example of a more complex chemosynthetic symbiotic association is the marine worm *Olavius algarvensis* and its co-occurring proteobacterial endosymbionts. *O. algarvensis* is a small worm living in shallow water sediments in the Mediterranean, and belongs to a group of oligochaetes which lack a mouth, gut, anus, and nephridia (54). To compensate for this, the worm relies solely on four proteobacterial endosymbionts to provide its nutrition. Specifically, it possesses two sulfur oxidizing Gammaproteobacteria

and two sulfate reducing Deltaproteobacteria working in a beneficial *syntrophic* relationship (nutritional relationship where two organisms combine metabolic capabilities to use a substrate neither could use alone) (53).

Studying chemosynthetic symbiotic relationships proves challenging due to many of the same abovementioned issues: the majorities of these microbes are uncultivable, and need to be studied in their natural environment. The first genomes sequenced from chemosynthetic endosymbionts were those inhabiting *Olavius algarvensis* (55) in a metagenomics analysis which revealed the metabolic potential of these symbionts to fix carbon, oxidize sulfur (gammas), reduce sulfate (deltas), take up and recycle worm waste products, and demonstrating capabilities of a versatile metabolism needed for optimal energy to shuttle between changing oxygen conditions in the worm's environment (55). However, it was only through metaproteomic and metabolomics analyses that direct evidence of novel pathways, which compensate for nutrient and energy limitations in this system, was achieved (specific results discussed further in chapters 4 and 5 (56)).

Metaproteomic analysis was also performed on the Gammaproteobacterial symbiont from *Riftia pachyptila*, revealing the symbiont uses the reductive TCA cycle and Calvin cycle for CO₂ fixation (57). These findings, which would not have been possible by any other methodology, highlight the importance of proteomics in elucidating functional capabilities in chemosynthetic symbiotic associations. Of note, this is the only metaproteomic study on chemosynthetic symbiotic associations to date, besides that from *O. algarvensis* endosymbionts reported in this dissertation work (chapters 4 & 5 (56)).

While chemosynthetic symbioses represent associations of relatively low complexity, being comprised of only a few microbial members and their eukaryotic hosts,

it is vital to evaluate the metaproteomic approach for a more complex ecosystem of microbes and their eukaryotic host, microbial communities in the human gut, where the number of organisms and dynamic nature of the symbiosis increases exponentially, as discussed below.

Commensalism in the Human Gut Microbiome

The human gut contains a consortium of trillions of bacteria that carry out numerous functions essential for human health, including: nutrient absorption, carbohydrate assimilation, informing the developing immune system, stimulating angiogenesis, regulating host fat storage, and providing protection from invasion of pathogenic bacteria (58-61). With this intricate symbiotic association, it is amazing that the microbial communities inhabiting our body provide us with traits we have not had to evolve on our own (62), while at the same time being tolerated as “self”, avoiding targeting by the innate and adaptive immune systems. Indeed, the human host has adapted intricate molecular mechanisms to tolerate its resident microbes (63). While this mutualistic association is generally beneficial to both the human host and microbiota, careful maintenance of homeostasis must be maintained so that any disruption, or dysbiosis, does not occur and lead to disease states. While the microbial compositions in the adult gastrointestinal tract is highly diverse among different individuals (64, 65), aberrant microbial compositions have been reported to be associated with diseases such as inflammatory bowel disease, Crohn’s disease, obesity, and metabolic syndrome (62, 66-68).

Homeostatic balance is kept intact structurally by the intestinal barrier, which is composed of the mucus layer and specialized epithelial cells: absorptive enterocytes, goblet cells, Paneth cells, M cells, and plasma cells (63). The thick mucus layer covers and protects intestinal epithelial cells and is composed of two layers. The outer mucus layer harbors commensal bacteria while the thicker, impenetrable inner layer offers protection by providing a physical barrier, as well as secreting antimicrobial compounds and secretory IgA (69). However, translocation of bacteria through the inner mucosal layer can occur, and oftentimes proves detrimental. Mucus is composed of mucins, which are glycoconjugates composed of a polypeptide core covered in O-linked carbohydrate side chains, and are secreted by goblet cells. The O-linked glycans provide an energy source for bacteria in the outer mucus layer (61).

While 16S rRNA and metagenomic surveys have provided important insights into the microbial community compositions, and potentially important genes involved in gut commensalism, there is an important need to characterize the functionality of the microbial members as well as the host response required to maintain the mutualistic association. This can be addressed using a metaproteomics approach. However, not surprisingly, only a few proteomic studies from the adult human gut microbiome have been reported to date (33, 70). This is likely due to the many challenges in acquiring metaproteomic data from adult fecal material including: a.) dealing with the complexity of fecal samples in terms of microbial community composition and raw material, b.) incompletely matched metagenomic information and c.) challenges in bioinformatics and data analysis. However deep proteomic measurements from complex human fecal

samples have been demonstrated (33), and options for data handling have been presented (70).

Symbiosis Begins: Colonization in the newborn infant

During the prenatal period, the human gut is sterile; colonization of microbial species begins just after birth. There is still much to be learned about the process of colonization, including how it works, what types of microbes are seen, how these patterns differ between and within individuals over time, what causes this variability, where the inoculating microbes come from, and how microbial colonization affects overall health of an individual (71). Several initial studies have begun to answer these questions, however there is still much to be unraveled. It has been shown that many factors influence initial colonization in the gut including: delivery mode (cesarean vs. vaginal delivery), the type of feeding (breast vs. bottle), the gestational age of the infant, as well as the geographic location where the infant is born (72, 73). This is complicated by the fact that there is a great deal of inter- and intra-individual diversity in the taxa of microbes that colonize the infant gut, and membership changes over time, primarily due to environmental factors such as dietary adjustments or antibiotic treatment.(7, 72, 74, 75). While colonization patterns vary initially between individuals, there appears to be some convergence around 2.5-3 years of age to an adult-like profile (72, 75). Even though investigations of microbial colonization in the infant gut have achieved rapid advances within the last few years, due to 16S rRNA gene-based surveys determining broad taxonomic memberships and metagenomics analyses, evaluating genetic and metabolic potentials in the infant gut, there is a growing need to understand this process at lower taxonomic levels. Only one

study to date has characterized the infant gut at the species and strain level using whole genome reconstruction of dominant community members from metagenomic sequence data (7). Metaproteomic analyses are currently underway on these same samples (discussed in chapter 5). Surprisingly, only one other metaproteomics study from the infant GI tract has been published to date: using two-dimensional gel electrophoresis combined with MALDI-TOF mass spectrometry, the authors concluded insufficient sequence information was available to identify the proteins, and only reported one peptide with high sequence similarity to transaldolase from a *Bifidobacterium* (76). Perhaps most importantly, this study did not monitor any human proteins. This highlights the important need for metaproteomics investigations which can simultaneously measure human and microbial proteins in order to elucidate the functional signatures playing important roles in the symbiotic interactions developing in the infant gut.

As previously mentioned, aberrant colonization or changes in microbial compositions disrupting homeostasis could be cause for development of certain diseases. It has been suggested that necrotizing enterocolitis (NEC), an inflammatory bowel disease, could be linked to improper colonization in the infant gut; however a single causative agent has not been identified as the pathogenic agent (77-79). Infants born prematurely have a higher susceptibility to developing NEC, possibly compounded by immature immune systems and under-developed epithelial barriers in these infants (80, 81) . Therefore, many studies are underway to try to determine the cause of NEC with the hopes of developing therapeutic interventions. However, much still needs to be learned about the process of colonization in healthy preterm infants before accurate comparisons can be made with sick infants. There have been no studies published to date that have

comprehensively characterized the microbial and human functions simultaneously in order to obtain an overall picture of how the symbiotic associations work, and how they may potentially be disturbed, however current work is underway to address this (presented in chapter five of this dissertation).

Transposable Elements: Parasites or Not?

We will conclude this examination of symbiotic associations by narrowing in to look at mobile genetic elements, specifically transposable elements, which are (controversially) considered parasitic or “selfish” genetic elements (82). Transposable Elements (TEs) are mobile genetic elements that can move within and between genomes. They are broadly classified into two types: DNA transposons, which contain *transposases*, enzymes that catalyze the movement of DNA, and retrotransposons, which go through an RNA intermediate and thus encode a reverse transcriptase gene. For the purpose of this discussion, we will be focusing on DNA transposons.

Transposable elements are considered by some as “selfish” or “parasitic”, with the sole purpose of enhancing their own transmission, causing a neutral or detrimental effect on the organism as a whole (82, 83). However, a counterview to that argument is the TEs play important regulatory roles in cells and their evolution. A ‘selfish genetic elements’ model has been proposed portraying an “ecological” view of the genome as whole with mutualistic, commensalistic, and selfish interactions (83). The model suggests that genetic conflict is created among components of a genome with different transmission patterns, whereby transmission of an element is increased, even if it is detrimental to an organism/genome, and transmission of other genetic elements are decreased: creating an

evolutionary “arms race”(83). This genetic conflict is thought to promote evolutionary change, but has not been experimentally proven in any system (83).

Transposable elements are present in nearly all microbial genomes and are particularly abundant in endosymbiotic bacteria which have recently transitioned to an obligate host associated lifestyle (2, 38). However, factors which cause transposable elements to increasingly proliferate throughout genomes are still undefined. Typically, it is thought that proliferation is tightly regulated in order to avoid loss of essential genes (84). While recent genomic studies have demonstrated that selfish genetic elements, such as transposable elements, help shape the structure and function of symbiont genomes (2), little is known about the activation of transposase proteins in this process. In fact, few investigations reporting transposase proteins have been done at the proteomic level (28, 85, 86), and none of these were in symbiotic systems. In chapter four we will discuss abundant transposase protein expression in symbionts of the gutless worm *Olavius algarvensis*, which we think are contributing to transposable element expansion in the symbiont genomes.

Objective / Scope of dissertation

The scope of this dissertation work aims to elucidate symbiotic interactions in three different microbial ecosystems via mass spectrometry-based proteomics/metaproteomics. We will begin with an in-depth description of the proteomic/metaproteomic experimental design in chapter two. Chapter three will discuss the proteomic response of the lactic acid bacteria *Streptococcus thermophilus* upon infection with bacteriophage virus 2972, and its subsequent defense mechanism via acquired

immunity, the CRISPR/Cas response. Chapter four reveals novel metabolic pathways carried out by four microbial endosymbionts of the gutless marine worm, *Olavius algarvensis* discovered using metaproteomics. Chapter five reports the abundant transposase expression within these same endosymbionts, also determined by metaproteomics, revealing novel hypotheses pertaining to transposable element expansion in mutualistic endosymbionts. Chapters six and seven discuss a more complex ecosystem and symbiotic association: the infant gut microbiome, with the former evaluating the symbiotic association of host and microbial proteins contributing to colonization and establishment of homeostasis in one preterm infant, while the later chapter focuses on methodological application to multiple infants and commonalities of the human protein complements.

While the emerging field of systems biology has achieved rapid advances in characterizing genomes and proteomes of microbes, there is still a significant knowledge gap in the characterization of symbiotic associations at the global level where microbial functions are adapted in context of each member's role in the symbiotic interaction. Importantly, very few proteomic studies have been carried, leaving important information yet undiscovered regarding metabolic functions and interactions. Alternatively, many studies have offered an incomplete/one-sided view characterizing microbial symbionts and not their host (in the case of the human gut microbiome). Through this dissertation work, we have unraveled functional aspects of dynamic symbiotic interactions, ranging in complexity and gained novel biological insights via a proteomics/metaproteomics approach.

CHAPTER TWO

Experimental Design for Proteomic Elucidation of Microbial Symbiotic Interactions

Shotgun Proteomics via nano-2D-LC-MS/MS: Experimental Overview

The overall experimental design used throughout this dissertation work employs shotgun proteomics (18) via nanospray-2 dimensional liquid chromatography coupled with tandem mass spectrometry (nano-2D-LC-MS/MS) (Figure 2.1). It begins with sample collection from either microbial isolates grown in the laboratory, or from more complex microbial community samples from the environment, such as human microbiome fecal samples. Proteins are extracted, denatured, reduced, and enzymatically digested into peptides. These complex mixtures of peptides are then separated across time using multidimensional liquid chromatography. The HPLC is coupled directly to the mass spectrometer in this gel-free approach, so that the peptides are ionized and electrosprayed into the mass spectrometer and their mass-to-charge (m/z) ratios measured generating a full mass spectrum. The most abundant peptide ions are selected for fragmentation, by collision-induced dissociation (CID) and the resulting fragment ions measured, generating tandem (MS/MS) spectra. Peptides are identified by computationally matching experimental MS/MS spectra to theoretical spectra generated from *in silico* tryptic digestion of the predicted protein sequences. Proteins are inferred by computationally matching peptides to their corresponding protein sequences and spectral counts used as a quantifiable measure of relative protein abundances. Thus, this technology provides the capability of not only identifying, but also quantifying thousands of proteins in one MS run. Specific details and considerations of each step are discussed below.

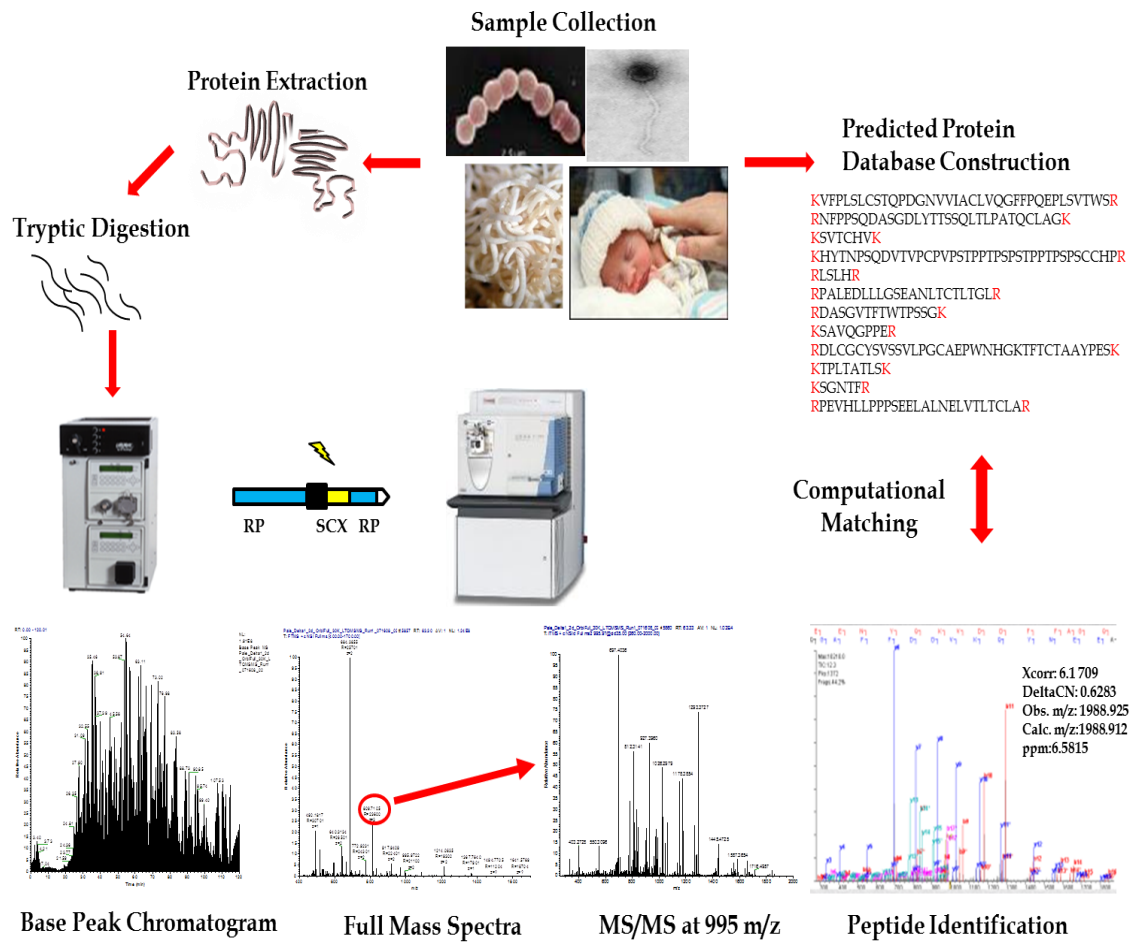


Figure 2.1: Experimental Design: Shotgun Proteomics via Nano-2D-LC-MS/MS. In general, samples are collected, proteins extracted, and enzymatically digested with trypsin. Peptides mixtures are separated by multidimensional liquid chromatography, electrosprayed into the mass spectrometer where full MS scans and tandem MS/MS scans are collected. Computational algorithms match the experimental spectra with theoretical spectra generated from a predicted protein database.

Sample Preparation for Mass Spectrometry Measurements

Sample collection: Shotgun proteomics is widely applicable to variety of sample types ranging from microbial isolates to environmental samples. In this dissertation, three different types of samples were analyzed: 1.) A bacterial isolate, *Streptococcus thermophilus*, infected with bacteriophage 2972 (chapter 3), 2.) Four proteobacterial symbionts found within the gutless worm *Olavius algarvensis* (chapters 4 and 5), and 3.) Fecal microbiome samples from preterm infants (chapters 6 & 7). Thus, each project required different methods for sample collection; specific details of which are discussed in the corresponding chapters. In general, around 1-10 mg of wet biomass from an environmental sample, or cell pellet is sufficient to obtain enough protein for shotgun proteomic measurements. However, when working with raw fecal material, around 250 mg is needed due to the complex matrix (which contains fiber, fat, inorganic matter, etc.).

Biosafety issues: Human fecal material is considered to be biohazardous and therefore, special precautions were taken during sample preparation. Specifically, personal protective gear such as gloves and lab coats were worn and protein extraction performed in a biosafety level 2 (BSL2) hood. In addition, all material was disposed of in the biohazard waste, which was subsequently autoclaved. Since fecal samples potentially contain human pathogens, all personnel working with fecal material completed bloodborne pathogen training.

Cell lysis, protein denaturation and reduction: Once samples are collected for MS analysis, cells are lysed and proteins extracted. Typically, this is done using either

guanidine HCl, a chaotropic denaturant, or SDS detergent, with the aid of heat. Dithiothreitol (DTT) is also included in order to reduce disulfide bonds. Physical disruption such as bead beating or sonication is oftentimes used to aid in cell lysis especially for microbial isolates, which are more difficult to lyse (i.e. Gram positives), or environmental samples in complex matrices. Three types of sample prep methods were used in this study, all of which are widely accepted methods for metaproteomics: 1.) a small-scale microbial biomass experimental approach (87), 2.) a thermally assisted SDS-TCA detergent based method (88, 89), and 3.) FASP: filter aided proteome preparation (90). The small-scale sample prep method has traditionally been used with success for microbial isolates (14, 26, 87) (chapters 3, 4, and 5). However, the more recently developed SDS-TCA method, provides improved protein purification from more complex matrices such as soils and feces (35). The inclusion of a protein precipitation step via trichloroacetic acid (TCA) in this method has been shown to help reduce humic acids and other interfering molecules in soils (35). Thus, this method has recently been successfully applied to soils, plants, and fecal material (25, 35) (Young et al. *in preparation*) (chapters 6 and 7). Another sample prep method, FASP has been used with some success, especially for small sample amounts. However, this method was applied to fecal material in our study and resulted in minimal improvement in overall results (discussed in chapter 7). Thus, sample preparation for MS analyses has required optimization for different sample types.

Tryptic digestion: Following protein extraction (by any method), proteins are enzymatically digested into peptides using sequencing-grade trypsin (Promega, Madison,

WI), which cleaves at the N-terminus of lysine and arginine residues. When looking for a particular protein of interest, consideration of the protein size and number of basic residues must be taken, since proteins which contain too many or too few lysines and/or arginines will generate tryptic peptides which are too big or too small that are not compatible with MS.

Prior to digestion, proteins are quantified using the Bicinchronic assay (BCA), a colorimetric assay, which uses a protein standard curve to extrapolate the measured values and obtain protein concentrations (91) optimal amount of protein is digested (between 1-3 mg). In addition, BCA assays are performed at the peptide level when using the SDS-TCA method (not feasible with small-scale sample prep method), and roughly equivalent amounts of peptides loaded onto the columns for MS measurements. Since the downstream analyses involves ESI-MS, which is not compatible with detergents, salts, etc., samples must be 'cleaned up' by desalting by C-18 solid-phase extraction (SepPak, Waters, Milford, MA) in the case of the small-scale sample prep method.

Multidimensional protein identification technology for shotgun proteomics

(MudPit):

Mass spectrometry measurements throughout this dissertation work applied a multidimensional protein identification technology for shotgun proteomics (MudPit) (18), which combines multidimensional liquid chromatography with electrospray ionization (8) and tandem mass spectrometry. The initial HPLC separation serves to simplify complex mixtures of tens of thousands of peptides by eluting them off as a few (or few hundred) at one time. Reducing the sample complexity is important in order to sufficiently resolve the

chromatographic peaks. Indeed, *peak capacity*, the number of peaks able to be separated at a specific retention window at a certain resolution (92), can be a limiting factor when dealing with complex mixtures, therefore multidimensional chromatographic separations are crucial to improve the separation power and thus reduce sample complexity. In our experimental design, a two-dimensional separation is applied using a biphasic column containing strong cation-exchange (SCX) and reversed-phase (C-18) resins, which elute peptides based on charge and hydrophobicity, respectively (18).

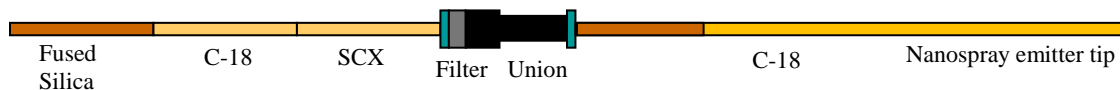


Figure 2.2: Schematic diagram of a biphasic column used in MudPit experiments. A fused silica back column is loaded with reverse phase (C-18) and strong cation exchange (SCX) material, then attached via filter and union to a nanospray emitter tip containing reverse phase material.

A fused silica back column (150 μm inner diameter) is loaded with C-18 (3-5 cm) and SCX (3-5 cm) resins via a pressure cell. Then peptides (50-150 μg) are loaded, binding to the reverse phase material. The back column is directly coupled (via filter and union) to a nanospray emitter tip front column (15 \pm 1 μm tip, 100 ID, from New Objective, Woburn, MA) containing 13-15 cm of reverse phase (C-18) material (Figure 2.2). During the chromatographic separation, sequential plugs of peptides are pushed onto the front column during the salt pulses, which are then separated over an organic gradient over time.

Specifically, in our experimental setup, peptides are eluted from the SCX resin in twelve steps consisting of increasing ammonium acetate salt pulses followed by reverse

phase resolution over two hour organic gradients (28, 29, 33). Typically, the first step involves a gradient from low to high organic: 100% Solvent A (95% H₂O, 5% acetonitrile (ACN), 0.1% formic acid (F.A.)) to 50% Solvent B (30% H₂O, 70% ACN, 0.1% F.A.), in order for peptides bound to the C-18 back column to move onto the SCX material. Then, in steps 2-12, using increasing ammonium acetate (500mM) salt pulses (typically 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 60%), peptides are sequentially eluted off of the SCX, bind to the C-18 front column, and are separated by the reverse phase gradient (increasing from 0-50% Solvent B over two hours). In addition, samples prepared using the SDS-TCA method require an off-line wash prior to starting the MS run. This entails fifteen minutes of Solvent A followed by five, two minute gradients of 100% Solvent A to 100% Solvent B, in order to wash any excess detergent or salt from the column, which could detrimentally affect the mass spectrometry runs.

In conjunction with sequential elution, peptides are analyzed by the mass spectrometer, which is comprised of three basic components: the ion source, where gas phase ions are produced, the mass analyzer which separates ions based on their mass to charge (m/z) ratios, and the detector, which measures the m/z values and abundances of the different ions. Throughout this work nanoelectrospray ionization (nano-ESI) was used as the ionization source. In electrospray ionization (ESI) (8) peptides in a liquid solution are pushed through a capillary and ejected through a fine-point needle ($15 \pm 1 \mu\text{m}$), which is held at high electrical potential with respect to the inlet of the MS, thereby causing desolvation and ionization of peptides into the gas phase. This 'soft' ionization permits measurement of fragile and large polar molecules, such as those of biological

interest like peptides, proteins, and nucleotides, without fragmentation. In nano-electrospray ionization (93) a low solvent flow rate (300 nl/min) is used to provide the advantages of low sample consumption, and enhanced sensitivity, in addition to direct coupling on-line with the HPLC.

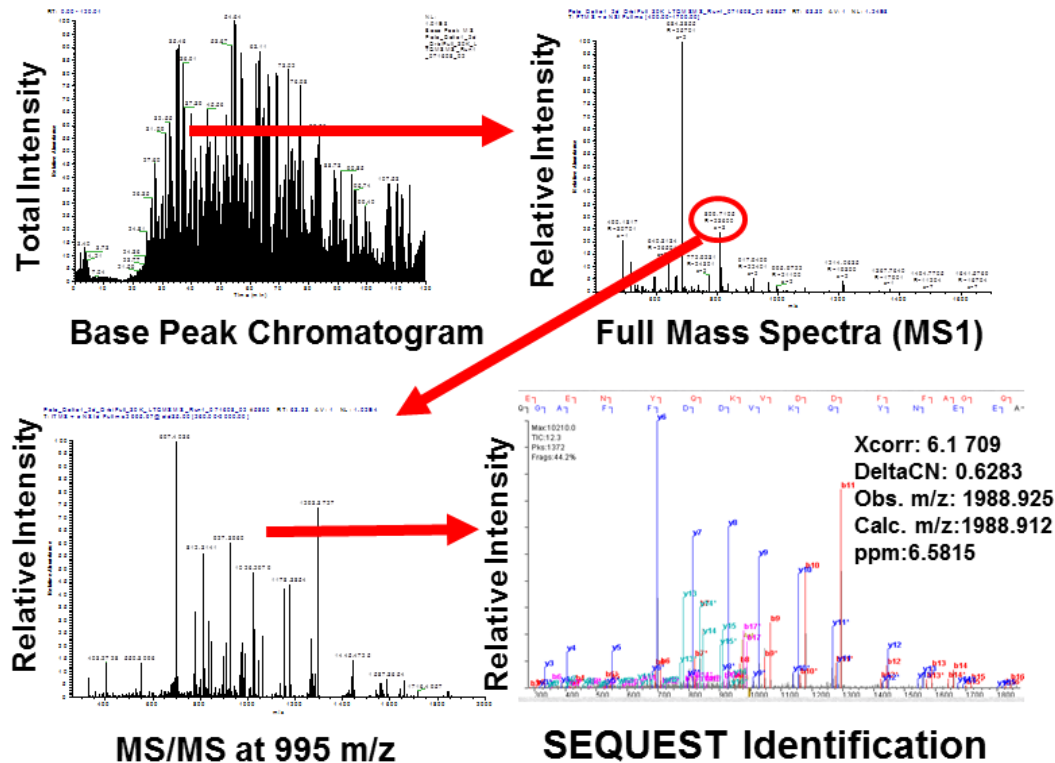
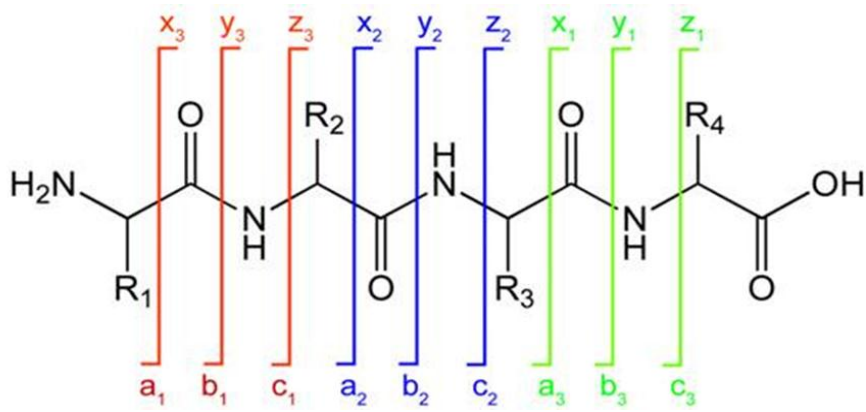


Figure 2.3: Example of MudPit experimental output: From top right to bottom left: A.) Base peak chromatogram of peptides separated over time via HPLC with the x-axis= time, and the y-axis=intensity, each peak represents ions eluted off at one particular time B.) Full mass spectra (MS1) collected of m/z ratios of parent ions, C.) MS/MS spectra of fragment ions, D.) b- and y-type ions used for peptide identification.

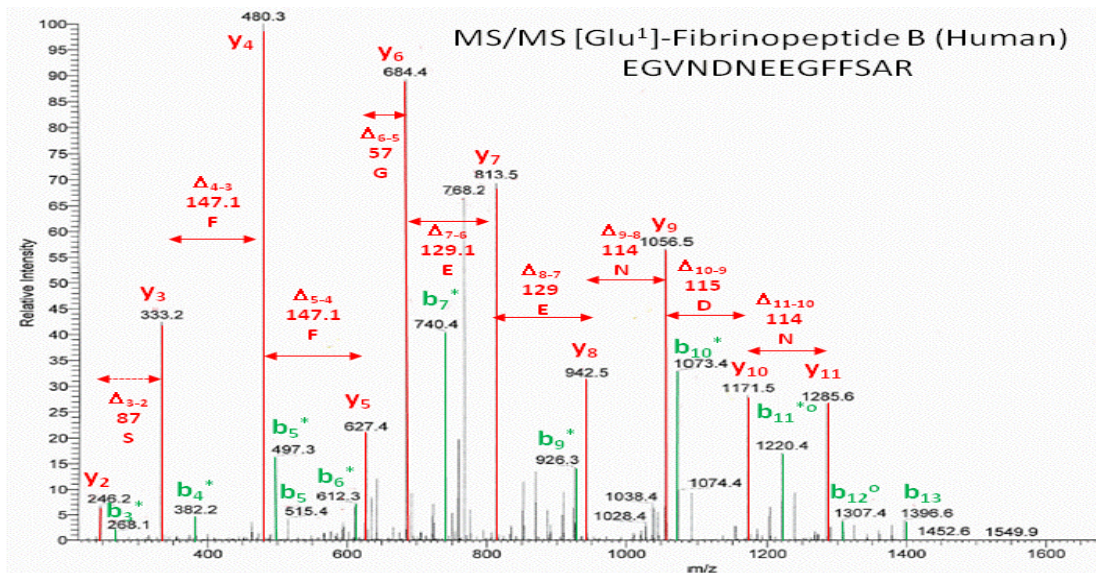
Once peptides are eluted, ionized, and transferred to the mass spectrometer, the m/z ratios of the peptide ions are measured in a full mass spectrum (MS1). (Most tryptic peptides carry a +2 charge). Since MS1s are still very complex, with hundreds of ion

masses measured, there is a need for further fragmentation. So, using a data dependent mode, the topmost abundant ions (10-20), are selected, undergo fragmentation, and their m/z ratios displayed in a tandem (MS/MS) spectra (Figure 2.3). The instrument oscillates between full scan and MS/MS scans, recording m/z ratios of parent and fragment ions. Fragmentation occurs through a process called collisional induced dissociation (CID) in which peptides ions within a narrow m/z window are isolated in the gas phase, collided with a target gas (helium) to break the peptide and create fragment ions. Fragment ions are named according to which end the charge is retained and numbered according to which amino acid position the break occur: a-type, b-type, c-type: N-terminus, x-, y-, z-: C-terminus (Figure 2.4). CID typically causes breakage at the peptide amide bond and thus predominantly produces b- and y- ions (Figure 2.5). The patterns of fragment ions are used to identify the peptides based on the genomic information using computational algorithms (discussed below).



http://www.mbc.manchester.ac.uk/images/clip_image002_0001.jpg

Figure 2.4: Schematic of peptide ion fragmentation patterns. Displayed is a representative peptide, with colored lines and numbers showing the bond where fragmentation occurs, and the letters indicating which end the charge is retained: a-, b-, and c-type ions on the N-terminus, and x-, y-, and z-type ions on the C-terminus.



<http://employees.csbsju.edu/hjakubowski/classes/ch331/protstructure/olcompseqconform.html>

Figure 2.5: Deriving an amino acid sequence from MS/MS spectra/fragment ions: Displayed is a representative MS/MS spectra with b- and y-type ions labeled. The mass differences between the sequential ions (i.e. y^6 and y^7) reveal the amino acid sequence.

Mass Spectrometry Instrumentation

Several different mass spectrometers were used for this dissertation work, with each chosen for a particular project based on various instrument capabilities and performance figures of merit (Table 2.1). Instruments employed included: a basic linear ion trap mass spectrometer, LTQ-XL (94), and hybrid instruments (composed of more than one mass analyzer): LTQ-Orbitrap-XL (13), LTQ-Orbitrap Velos (95), LTQ-Orbitrap Elite (96). The basic composition of the LTQ-XL linear ion trap consists of four parallel metal rods, in which a combination of dc and ac radio frequency (rf) voltages are applied. The combination of fixed and alternating electric fields electrostatically confines ions, acting as a mass filter and ion storage device. The ion storage capacity, fast scan times, affordability, and simplicity of construction of the LTQ-XL are optimal for protein

identification and label-free quantification of microbial isolates, providing sufficient *dynamic range* (the difference between the most and least abundant components of the samples which can be measured) for these relatively lower complexity samples (i.e. *S. thermophilus* measurements discussed in chapter 3) (Table 2.1).

Table 2.1: Figures of Merit Comparisons Between Mass Spectrometers

	LTQ-XL	LTQ-Orbitrap	LTQ-Orbitrap Velos	LTQ-Orbitrap Elite
Mass Accuracy	0.1 Da	< 3 ppm with external calibration < 1 ppm using internal calibration	< 3 ppm with external calibration < 1 ppm using internal calibration	<3ppm RMS with external calibration <1ppm RMS using internal calibration
Resolution	0.05 FWHM 1000-2000	7,500- >100,000 at m/z 400	7,500 ->100,000 at m/z 400	15,000->240,000 at m/z 400
Mass Range	m/z 15-200 m/z 50-2,000 m/z 200-4,000	m/z 50-2,000 m/z 200-4,000	m/z 50-2,000, m/z 200-4,000	m/z 50-2,000 m/z 200-4,000
Dynamic Range		> 4,000 within a single scan guaranteeing specified mass accuracy	>5,000 within a single scan guaranteeing specified mass accuracy	>5,000 within a single scan guaranteeing specified mass accuracy
MS/MS Sensitivity	25:1 signal-to-noise ratio	100:1 signal-to-noise ratio	100:1 signal-to-noise ratio	100:1 signal-to-noise ratio

For more complex samples (microbial symbionts of gutless worms discussed in chapters 4 & 5), a hybrid mass spectrometer, the LTQ-Orbitrap (13), was used. The basic makeup of this instrument is similar to the LTQ-XL in containing a linear trapping quadrupole, where MS/MS scans are performed, with the addition of a second mass analyzer, the Orbitrap, where *high resolution* full scans are performed (Figure 2.7). The Orbitrap is designed to radially traps ions around a central spindle electrode, measuring m/z values from the ion oscillation frequencies which are subsequently Fourier

transformed (FT) to generate mass spectra. The higher *mass resolving power* of the Orbitrap allows this mass analyzer to distinguish/pull apart adjacent peaks (calculated by measuring the full width at half of the maximum peak height (FWHM)) (Table 2.1). In addition, the LTQ-Orbitrap can perform *high mass accuracy* measurements, ensuring a calculated mass matches a measured mass ($\Delta m_{\text{accuracy}} = m_{\text{true}} - m_{\text{measured}}$) within an error rate of parts per million ($\text{ppm} = 10^6 \Delta m_{\text{accuracy}} / m_{\text{measured}}$) (Table 2.1). Since the purpose of the mass analyzer is to determine the mass-to-charge ratio of gas-phase ions, the ability of the instrument to accurately measure a particular m/z with little deviation from the known m/z is a crucial factor especially with complex samples.

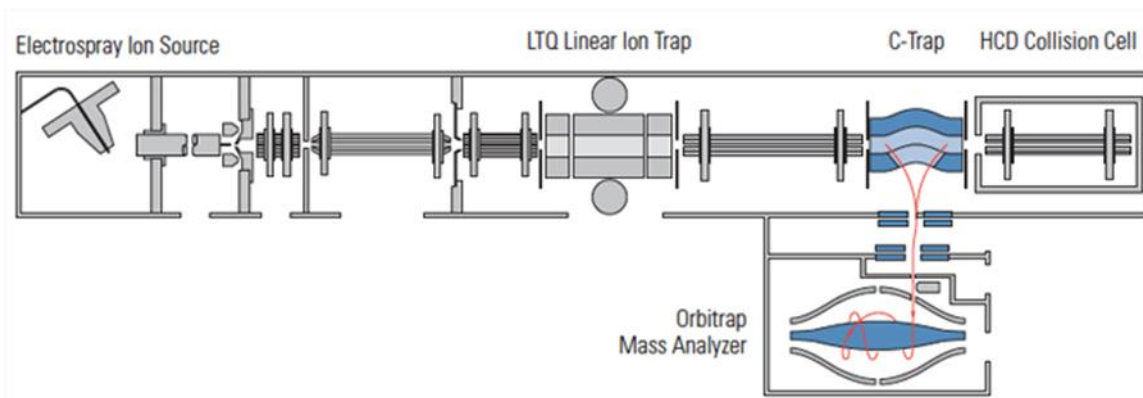


Figure 2.7: Schematic diagram of an LTQ-Orbitrap mass spectrometer. Ions are generated by the electrospray ionization source, transferred through and trapped in the LTQ-XL linear ion trap, axially ejected, collected in the C-trap, and then passed to the Orbitrap (<http://www.thermoscientific.com/>).

The newer generation of hybrid instruments, the LTQ-Orbitrap-Velos (95) utilizes the *ultra-high resolution* and *accurate mass* of the LTQ-Orbitrap, but goes one step beyond, providing improved *robustness* (due to an improved API source with generation I, stacked “S-lens” ion optics technology and neutral beam blocker) and *faster scanning times*, so that more measurements are taken across a chromatographic peak, yielding

better quantification (Table 2.1). Another key improvement to this instrument is the dual LTQ trap design, which contains a higher pressure trap for CAD and a lower pressure trap for ion measurement. In addition, this newer-generation instrument has the ability to measure a greater *dynamic range* than the LTQ-Orbitrap (Table 2.1), allowing deeper coverage of lower abundant proteins from complex samples, such as fecal microbiome samples (discussed in chapters 6 and 7).

In addition, the most recent advancement in mass analyzers to date, the Orbitrap Elite incorporates all of the features of the LTQ-Orbitrap-Velos but is equipped with a high-field mass analyzer with improved orbital trapping providing faster *scan speed* and higher *resolution* (Table 2.1). In addition, construction of this instrument provides increased *sensitivity*, while being more *robust* (due to generation II ion optics with neutral beam blocker). This instrument was recently purchased by our laboratory and was used in this dissertation work to analyze fecal microbiome samples (discussed in chapter 7).

MS-Based Informatics

Search algorithms: Once MS/MS spectra are collected, peptides are identified using specialized search algorithms (Sequest (97), Myrimatch (12), DBDigger (11)) which match the experimental spectra with theoretical spectra generated from a predicted protein database: a process called peptide spectral matching (PSM). The predicted protein database comes from *in silico* tryptic digestion of the translated genome sequences. Then, the Sequest algorithm (97) used throughout this dissertation work, queries the search database to find a linear combination of amino acid sequences that match the precursor m/z of the peptide within a certain mass tolerance ($\pm 1-3u$). From this list of ‘candidate’

Locus Key:

Validation Status	Locus	Sequence Count	Spectrum Count	Sequence Coverage	Length	MolWt	pI	Descriptive Name
-----------------------------------	-----------------------	--------------------------------	--------------------------------	-----------------------------------	------------------------	-----------------------	--------------------	----------------------------------

Similarity Key:

Locus	# of identical peptides	# of differing peptides
-----------------------	---	---

U	HOM_gi_4557577_ref_NP_001434.1	14	128	89.0%	127	14208	7.2	fatty acid binding protein 1, liver [Homo sapiens] # pt:6.60 MW:14209		
	Filename	XCorr	DeltCN	ObsM+H+	CalcM+H+	SpR	SpScore	Ion%	#	Sequence
*	20110227_Morowitz_6421_150ug_OrbiV_30K_01.12361.12361.2	73.2611	0.530481	1762.8	1763.97	14	202.0	87.0%	26	K.YQLSQENFEAFMK.A
*	20110227_Morowitz_6421_150ug_OrbiV_30K_01.10898.10898.2	63.607	0.447598	1210.8	1211.44	14	156.0	94.4%	2	K.AIGLPEELIQK.G
*	20110227_Morowitz_6421_150ug_OrbiV_30K_05.11403.11403.2	44.5427	0.410366	1752.0	1753.09	14	33.0	64.0%	1	K.AIGLPEELIQKGDIK.G
*	20110227_Morowitz_6421_150ug_OrbiV_30K_01.5565.5565.1	34.1154	0.297233	1030.5	1031.15	14	91.0	85.7%	1	K.GVSEIVQNGK.H
*	20110227_Morowitz_6421_150ug_OrbiV_30K_01.5570.5570.2	55.7073	0.453759	1030.6	1031.15	14	91.0	87.5%	1	K.GVSEIVQNGK.H
*	20110227_Morowitz_6421_150ug_OrbiV_30K_11.7367.7367.2	46.1484	0.293258	1444.0	1443.65	14	4.0	75.0%	2	K.GVSEIVQNGKHF.F
*	20110227_Morowitz_6421_150ug_OrbiV_30K_02.5930.5930.1	25.3975	0.208691	824.4	824.953	14	139.0	80.0%	1	K.FITAGSK.V
*	20110227_Morowitz_6421_150ug_OrbiV_30K_03.13885.13885.2	61.787	0.551872	2444.0	2444.7	14	85.0	74.1%	6	K.VIQNEFTVGEEC*ELETMTGEK.V
*	20110227_Morowitz_6421_150ug_OrbiV_30K_05.6432.6432.2	32.4363	0.310352	1331.6	1330.53	14	32.0	60.0%	2	K.VKTVVQLGDNK.L
*	20110227_Morowitz_6421_150ug_OrbiV_30K_02.10449.10449.2	34.0901	0.34779	1793.0	1793.07	14	75.0	44.4%	2	K.TVVQLGDNKLVITFK.N
*	20110227_Morowitz_6421_150ug_OrbiV_30K_06.20845.20845.2	80.2045	0.621132	2382.2	2382.74	14	154.0	81.2%	56	K.SVTELNQDIIITNMTLGDIVFK.R
*	20110227_Morowitz_6421_150ug_OrbiV_30K_10.20042.20042.3	67.7962	0.494242	2383.3	2382.74	14	233.0	54.3%	10	K.SVTELNQDIIITNMTLGDIVFK.R
*	20110227_Morowitz_6421_150ug_OrbiV_30K_08.19106.19106.2	80.5143	0.556387	2537.4	2538.93	14	110.0	77.4%	17	K.SVTELNQDIIITNMTLGDIVFKR.I
*	20110227_Morowitz_6421_150ug_OrbiV_30K_08.19076.19076.3	49.7405	0.266578	2538.4	2538.93	14	180.0	47.1%	1	K.SVTELNQDIIITNMTLGDIVFKR.I

U	HOM_gi_113414893_ref_XP_001127175.1	78	494	82.4%	512	56162	6.8	PREDICTED: similar to lactotransferrin [Homo sapiens] # pt:6.43 MW:56162		
	Filename	XCorr	DeltCN	ObsM+H+	CalcM+H+	SpR	SpScore	Ion%	#	Sequence
	20110227_Morowitz_6421_150ug_OrbiV_30K_01.19192.19192.3	91.9596	0.521451	3133.3	3134.27	14	238.0	73.1%	7	R.ESTVFEDLSDEAERDEYELLC*PDI
	20110227_Morowitz_6421_150ug_OrbiV_30K_01.19149.19149.2	42.2686	0.287375	3134.4	3134.27	14	4.0	66.7%	2	R.ESTVFEDLSDEAERDEYELLC*PDI
	20110227_Morowitz_6421_150ug_OrbiV_30K_11.4230.4230.2	40.482	0.134742	861.6	862.06	14	162.0	83.3%	3	R.KFVDFEK.D
	20110227_Morowitz_6421_150ug_OrbiV_30K_11.5371.5371.1	27.7741	0.253556	1046.5	1047.23	14	146.0	81.8%	1	K.FKDC*HLAR.V
	20110227_Morowitz_6421_150ug_OrbiV_30K_11.5372.5372.1	27.6362	0.181133	771.4	771.875	14	156.0	87.5%	1	K.DC*HLAR.V
	20110227_Morowitz_6421_150ug_OrbiV_30K_11.5186.5186.1	35.204	0.339993	935.5	936.103	14	124.0	76.9%	2	R.VPSSHAVVAR.S
	20110227_Morowitz_6421_150ug_OrbiV_30K_11.5096.5096.2	59.8638	0.593728	938.6	936.103	14	24.0	100.0%	12	R.VPSSHAVVAR.S
	20110227_Morowitz_6421_150ug_OrbiV_30K_04.14208.14208.1	38.6528	0.497657	1614.8	1615.83	14	140.0	76.2%	1	R.SVNGKEDAIWNLRL.Q
	20110227_Morowitz_6421_150ug_OrbiV_30K_03.13828.13828.2	64.1378	0.552479	1614.8	1615.83	14	178.0	86.4%	10	R.SVNGKEDAIWNLRL.Q
	20110227_Morowitz_6421_150ug_OrbiV_30K_03.13737.13737.3	58.316	0.164178	1618.0	1615.83	14	253.0	58.7%	6	R.SVNGKEDAIWNLRL.Q
	20110227_Morowitz_6421_150ug_OrbiV_30K_01.15334.15334.2	46.5594	0.275604	1129.6	1130.29	14	163.0	86.7%	2	K.EDAIWNLRL.Q
	20110227_Morowitz_6421_150ug_OrbiV_30K_11.3178.3178.1	31.5448	0.144604	802.5	802.958	14	106.0	80.0%	1	K.SVNGKEDAIWNLRL.Q

Figure 2.8: DTASelect Output. Depicted above is an example of a DTASelect .html output file displaying the protein id (red), sequence counts per protein, spectral counts per protein, % sequence coverage, length, molecular weight, pI, and protein description/annotation. In addition, for each peptide, the spectra files, Sequest scores, # spectral counts per peptide, and sequence are displayed.

peptides, theoretical fragment ions are calculated and theoretical spectra generated from the m/z ratios. Sequest compares the two and gives a correlation/ probability score, Xcorr, based on how similar they are. In this work, standard Xcorr cutoffs of at least 1.8, 2.5, and 3.5, for charge states +1, +2, and +3, respectively were used (10). In addition, a

second Sequest score, ΔCN , is used to indicate the difference between the first and second best PSM. Once peptides are identified, protein inference algorithms (DTASelect (6), ID Picker(15)) are used to filter quality PSM and organize peptides into their corresponding protein sequences. DTASelect (6) was used in this work, as part of our bioinformatics pipeline, to assemble proteins (Figure 2.8).

Label free quantification and statistical analyses: In a label free quantification approach, spectral counts, the number of times a peptide fragment ion is measured, are used as a unit of relative abundance in quantifying peptides and thus proteins (98, 99). Basically, by comparing the number of MS/MS spectra from two different proteins within a sample, the relative difference in quantification between the two is determined. Oftentimes, proteins are normalized relative to their protein length to account for the fact that larger proteins have the potential to contribute more peptides. Thus, normalized spectral abundance factors (NSAFs) (100) are used in which the spectral counts are divided by protein length, then normalized against sum of all spectra collected for a particular run. Statistical analyses are typically performed to validate differentially expressed proteins such as ANOVA or Poisson exact test. Each proteomic dataset is unique and requires careful consideration of the normalization and statistical analyses. Specific details for each dataset in this dissertation are discussed in the corresponding chapters.

Database design and considerations: Since search algorithms match theoretical spectra with experimental spectra for peptide identification, a peptide-spectral match cannot be

made if the corresponding sequence is not in the database. Therefore, careful consideration must be taken when designing a predicted protein database. This is more straightforward when working with sequenced isolates with high-quality genomes where the predicted protein sequences are wholly representative of what is in the sample. However, it becomes more complicated when working with low quality genomes, or complex metagenomes where lower abundance organisms have not been deeply sequenced. Indeed, if the genomes are not well annotated, contain modifications, or genes/organisms are missing, important information may be lost. Thus, when working with environmental samples, it is ideal to have matched metagenomes in which deep sequencing was done on the same samples. However, since it is typically not feasible to sequence every sample, due to the cost and labor intensiveness, steps need to be taken in the database design and post-processing data analysis (discussed further below). This may involve including representative sequenced isolates in the search database, which is not ideal, but sometimes necessary. However, on the upside, due to the sensitivity of the technology, high-quality metagenomic information is available, proteins from two closely related strains can be distinguished from each other, allowing biological insights which cannot be achieved by any other technology. For example, resolution at the species and strain level allows assignment of functional roles, metabolic activities, and niche partitioning between community members.

In addition, if modifications or variations such as single amino acid polymorphisms (SAPs) or post-translational modifications (PTMs) are present, spectral matches cannot be correctly assigned unless stipulations are made in the searching parameters. For example, IAA used in the SDS-TCA sample prep method causes

carbamidomethylation to Cysteine residues, so this modification (C+57) is incorporated into the search parameters.

A relative measure of the confidence of the PSMs is calculated using false discovery rates (FDRs)(21). By searching against a decoy database, typically generated by reversing the amino acid sequences in the search database, the FDR is calculated by doubling the number of reverse hits and dividing by the total number of hits (either peptide, protein, or spectra) ($\% \text{ FDR} = 2[n_{\text{rev}} / (n_{\text{rev}} + n_{\text{real}})]$).

Challenges in Protein Identification and Quantification: Database clustering and

spectral balancing: Protein and peptide assignments become more challenging in complex communities with large numbers of peptides. If the sequences are not present, the search algorithms cannot find a correct match. Likewise, when a peptide is found in multiple proteins, it cannot be easily deciphered as to which protein(s) it comes from.

This can be particularly problematic in organisms like humans and plants, which contain large paralogous gene families, splice variants, and multiple protein isoforms. And, quantification of proteins with shared peptides can be problematic in label-free shotgun proteomics that is determined by spectral counts of detected peptides. Therefore, bioinformatic clustering of functionally redundant proteins can be carried out by grouping together proteins with similar amino acid sequences. This approach has been utilized in previous plant proteomic studies (25), as well as an infant microbiome study, discussed in chapter 6, in which clustering of microbial proteins with 100% amino acid identity and human proteins with 90% amino acid identity was carried out. These newly formed clusters or ‘protein groups’ are comprised of the longest ‘seed’ sequence and

matching 'hit' sequences. In many cases, they are comprised of only a single protein. Spectral counts are assigned to each protein group and balanced between proteins containing shared peptides based on the number of unique peptides within that protein group. This method provides a more accurate quantification of the proteins while avoiding over-representation of redundant protein abundances, and possible mis-interpretation of the data.

CHAPTER THREE

Phage-Induced Expression of CRISPR-Associated Proteins is Revealed by Shotgun Proteomics in *Streptococcus thermophilus*

Text and figures were taken from: Young JC, Dill BD, Pan C, Hettich RL, Banfield JF, Shah M, Fremaux C, Horvath P, Barrangou R, and VerBerkmoes NC (2012) Phage-Induced Expression of CRISPR-Associated Proteins Is Revealed by Shotgun Proteomics in *Streptococcus thermophilus*. PLoS One 7: e38077.(26)

Jacque Young's contributions included: experimental design, performed all experiments and mass spectrometry runs, and wrote, edited and revised manuscript.

Abstract

The CRISPR/Cas system, comprised of clustered regularly interspaced short palindromic repeats along with their associated (Cas) proteins, protects bacteria and archaea from viral predation and invading nucleic acids. While the mechanism of action for this acquired immunity is currently under investigation, the response of Cas protein expression to phage infection has yet to be elucidated. In this study, we employed shotgun proteomics to measure the global proteome expression in a model system for studying the CRISPR/Cas response: infection of *S. thermophilus* DGCC7710 with phage 2972. Host and viral proteins were simultaneously measured following inoculation at two different multiplicities of infection and across various time points using two-dimensional liquid chromatography tandem mass spectroscopy. Thirty-seven out of forty predicted

viral proteins were detected, including all proteins of the structural virome and viral effector proteins. In total, 1,013 of 2,079 predicted *S. thermophilus* proteins were detected, facilitating the monitoring of host protein synthesis changes in response to virus infection. Importantly, Cas proteins from all four CRISPR loci in the *S. thermophilus* DGCC7710 genome were detected, including loci previously thought to be inactive. Many Cas proteins were found to be constitutively expressed, but several demonstrated increased abundance during peak infection, including the Cas9 proteins from the CRISPR1 and CRISPR3 loci, which are key players in the interference phase of the CRISPR/Cas response. Altogether, these results provide novel insights into the proteomic response of *S. thermophilus*, specifically CRISPR-associated proteins, upon phage 2972 infection.

Introduction

Bacteriophages (phages) are abundant and ubiquitous viruses in most natural environments and play an important role in the ecology of their bacterial hosts. In turn, bacteria have evolved various mechanisms to defend themselves against viral predation. One of these strategies involves the CRISPR/Cas system, in which acquired immunity is achieved against invading nucleic acids, providing resistance that can be passed on to future generations (101-104). Clustered regularly interspaced short palindromic repeats (CRISPRs) are loci found in approximately 46% and 87% of bacteria and archaea, respectively (105). These hypervariable regions consist of a leader sequence followed by an array of direct nucleotide repeats interspersed with non-repetitive DNA regions called spacer sequences. Immediately flanking the CRISPR loci are CRISPR-associated (*cas*)

genes (106-108). Host genomes that have acquired spacer sequences corresponding to phage sequences are rendered resistant to that particular phage and are thus termed bacteriophage insensitive mutants (BIMs) (101, 109). The mechanism of action of the CRISPR/Cas system is mediated by small interfering crRNA (CRISPR RNA) molecules (110-114) and occurs in two phases: immunization/adaptation, and immunity/interference (104). Several studies have established that the immunization process, which is based on novel spacer acquisition, and the immunity process, which is based on crRNA interference by seed sequence interactions with target DNA, rely on the Cas protein machinery, although the roles of the various Cas proteins are unknown but under investigation (110, 115, 116). The sequence and function variability across the Cas proteins of the three CRISPR/Cas types (I, II, and III) (117), along with the functional idiosyncrasies of the various core Cas proteins, have compounded the difficulty of Cas proteins characterization.

The link between CRISPR loci and phage-specific acquired immunity was first demonstrated in *Streptococcus thermophilus*, an economically important lactic acid bacterium used as a starter culture in the production of yogurt and various cheeses (101). In industrial batch cultures, *S. thermophilus* is subject to phage attack, resulting in a negative impact on the fermentation process, and thus vast economic and manufacturing losses. Therefore, many studies have monitored these phages in hopes of developing anti-viral strategies. Numerous *S. thermophilus* phages have been characterized via comparative genomics and transcriptomics, including phage 2972, a virulent *pac*-type phage composed of an isometric capsid and long non-contractile tail (47, 118). The structural proteins of phage 2972 have been characterized, including the major capsid

protein (orf9), two major (orf15 and orf17) and three minor (orf18, orf19, and orf21) tail proteins, the portal protein (orf5), and the receptor binding protein (orf20) (47). However, the complete proteome of the virus has yet to be elucidated, rendering an incomplete characterization of the functional signature for phage 2972.

The CRISPR content of various microorganisms, including numerous strains of *S. thermophilus*, have been analyzed allowing characterization of novel spacer additions and strain typing based on spacer content and hypervariability. These features reflect biogeography and provide a historical perspective of exposure to foreign genetic elements (44, 119-121). *S. thermophilus* DGCC7710, the strain used in this study, contains four CRISPR loci within its genome (103). The CRISPR1 and CRISPR3 loci, both type II CRISPR/Cas systems (Nmeni subtype) (106, 117) are known to be active, with the ability to acquire novel spacers in response to phage challenge (103, 109, 121). CRISPR2 (Type III system, Mtube subtype) and CRISPR4 (Type I system, Ecoli subtype) loci contain three and twelve spacer sequences, respectively. However, new spacer additions have never been observed at CRISPR2 or CRISPR4 loci despite multiple viral challenges.

In this study, we employed shotgun proteomics via 2D-LC MS/MS to measure the global proteomes of *S. thermophilus* DGCC7710 cells upon infection with phage 2972 at two different multiplicities of infection (MOI). Through this study we were able to simultaneously measure bacterial and phage proteins and gain insights into the phage proteins synthesized as well as the global response of the host upon phage infection. In addition, we monitored the Cas protein abundances from all four CRISPR loci in *S. thermophilus* DGCC7710 as a function of time post infection.

Materials and Methods

Bacterial cultures and phage 2972 infection: *Streptococcus thermophilus* DGCC7710

and phage 2972 were obtained from Danisco, USA Inc. (Madison, WI, USA). *S.*

thermophilus DGCC7710 was cultivated in M17 medium (Difco, Lawrence, KS, USA)

supplemented with 0.5% lactose (LM17) at 42°C. A mid-log phase culture (O.D.₆₀₀=0.4)

was spun down (10,000g for 10 minutes), resuspended in fresh LM17 medium containing

10 mM CaCl₂, then infected with phage 2972 at an M.O.I. of 0.1 or 1 and incubated at

42°C.

Cellular and viral enriched fraction preparation: At times 0, 0.5, 1, 2, 4, and 24 hours

post-infection (hpi), 10 ml aliquots were taken and separated into cellular fractions or

viral enriched fractions via PEG precipitation (for MOI=1 only). Cellular fractions were

obtained by centrifugation at 10,000g for 10 min at 4°C and retaining the pellets. The

supernatant was then PEG-precipitated (122). Briefly, DNase I and RNase were added at

a final concentration of 1 µg/ml and incubated for 30 min at room temperature. 1 M NaCl

was added to the supernatant incubated for 1 h on ice, then centrifuged at 10,000g for 10

min at 4°C. Phage particles were precipitated by the addition of PEG8000 (Sigma

Aldrich, St. Louis, MO) (10% w/v) for 1 h on ice, then centrifuged at 10,000g for 10 min

at 4°C. The pellets were resuspended in SM buffer (122) and equal volume of chloroform

(Sigma Aldrich, St. Louis, MO), then spun down and the aqueous phase recovered.

Protein denaturation and digestion: For cell lysis and protein denaturation, cellular

pellets were resuspended in 6 M guanidine HCl (Sigma Aldrich St. Louis, MO),

sonicated (Branson Sonifier; 10% amplitude, 10 seconds on/off cycles for 10 min total), and incubated at 60°C for 1 h. Protein concentrations were measured using the Pierce bicinchoninic acid assay (BCA) (Thermo Scientific, Rockford, IL) then disulfide bonds were reduced with 10 mM dithiothreitol. The protein solution was diluted to 1 M guanidine in 50 mM Tris (pH 7.6), 10 mM CaCl₂, and proteins were enzymatically digested into peptides using sequencing-grade trypsin (Promega, Madison, WI). The peptide solutions were desalted by C18 solid-phase extraction (SepPak, Waters, Milford, MA), solvent exchanged into 0.1% formic acid, concentrated, and passed through a 0.45µm filter (Millipore, Bedford, MA). Samples were frozen at -80°C until analyzed by 2D-LC-MS/MS.

Nano 2D-LC-MS/MS Analysis: Peptide mixtures were separated using on-line two-dimensional liquid chromatography with a split phase column containing reverse phase (C18) and strong cation exchange (SCX) materials (9, 123, 124). Peptides were eluted from the SCX resin by increasing ammonium acetate salt pulses followed by reverse phase resolution over two hour organic gradients as described previously (28, 29, 33), ionized via nanospray (200 nl/min) (Proxeon, Cambridge MA), and analyzed using an LTQ XL linear ion trap mass spectrometer (Thermo Fisher Scientific, San Jose, CA). Technical duplicates were run for all samples with 22 hour runs for the cellular fractions and 8 hour runs for the PEG-precipitated fractions. The LTQ was run in data-dependent mode (top 5 most abundant peptides in full MS selected for MS/MS) with dynamic exclusion enabled (repeat count=1, 60 s exclusion duration). Two microscans were collected in centroid mode for both full and MS/MS scans.

Database construction and analysis: A protein database was generated from the genome sequence of *S. thermophilus* strain DGCC7710 (<http://compbio.ornl.gov/CRISPRproteomics/>) and phage 2972 (GenBank accession no. AY699705) (47), along with other common contaminants such as trypsin and keratins. MS/MS spectra from all LC-MS/MS runs were searched with the SEQUEST algorithm (16) using the database above, and filtered with DTASelect/Contrast (17) at the peptide level with standard filters [SEQUEST Xcorrs of at least 1.8 (+1), 2.5 (+2), 3.5 (+3), $\Delta\text{CN} > 0.08$]. Only proteins identified with two fully tryptic peptides were considered for further biological study. Representative runs were calculated to have false positive rates $< 0.3\%$ at the peptide level using reversed database searching. COG (clusters of orthologous groups) assignments for each protein sequence were performed by running rpsblast against the COG database from NCBI, with an *E*-value threshold of 0.00001, and the top hit used for the assignment (42). All databases, peptide and protein results, MS/MS spectra, and supplementary tables are archived and made available as open access via (<http://compbio.ornl.gov/CRISPRproteomics/>) website.

Statistical Analysis: Spectral counts, values that can be used to approximate relative protein abundances in LC-MS/MS analyses (98), were normalized to account for technical variability among runs by equalizing the total spectral counts of all runs in the time course. First, an average of the total spectral counts of all runs in the time course experiment was calculated. Then, the normalization factor for each run was calculated as the ratio of the average total spectral count and the run's total spectral count. Finally, protein spectral counts per run were normalized by multiplying the raw spectral counts by

the run normalization factor. Normalized spectral counts of proteins were compared between two time points to identify proteins with statistically significant abundance changes. Because spectral counts follow a Poisson distribution (125, 126), spectral counts per protein were compared between two time points using the exact Poisson test. As proteins have two replicate spectral counts at every time point, p values were calculated by comparing the two closest replicate spectral counts from two time points to minimize type I errors. Proteins with a p value less than 0.05 were considered to have a significant abundance change. Pairwise comparisons were performed between each time point after infection and time zero in the three time courses (Supplemental Table 2). Comparisons were also performed between a time point early in infection and a time point during peak infection: 0.5 and 1 hpi for MOI=1, and 1 and 2 hpi for MOI=0.1.

Results

Overall results

S. thermophilus DGCC7710 cultures were infected with phage 2972 at an MOI=1 or MOI=0.1, and after 0, 0.5, 1, 2, and 24 hours post infection (hpi), cellular fractions were collected and analyzed via nano-2D-LC MS/MS. Uninfected controls were also analyzed in tandem. In addition, at the higher infection rate (MOI=1), fractions were enriched for phage 2972 via PEG precipitation of the corresponding cell supernatants collected at each time point. Two technical replicates were run per time point. High reproducibility was shown between the replicates. The overall protein, peptide, and spectral counts for each fraction and time point are summarized in Table 3.1.

Table 3.1: Number of proteins, peptides, and spectra identified by LC-MS/MS in cellular (MOI= 0.1 and 1) or PEG-enriched viral fractions (MOI=1) at each time point of infection

Sample Fraction	Time Point (hpi)	Protein Identifications (Total/Viral)	Peptide Identifications (Total/Viral)	MS/MS Spectra (Total/Viral)
Viral (MOI=1)	0	2/1	14/2	119/3
Viral (MOI=1)	0.5	6/4	70/15	143/14
Viral (MOI=1)	1	84/17	1170/266	3311/726
Viral (MOI=1)	2	86/16	1217/299	3197/824
Viral (MOI=1)	4	68/12	845/209	2104/544
Viral (MOI=1)	24	61/14	611/269	1972/1299
Cellular (MOI=1)	0	625/0	8151/0	27821/0
Cellular (MOI=1)	0.5	540/18	6503/138	26438/335
Cellular (MOI=1)	1	477/32	5046/567	12866/2341
Cellular (MOI=1)	2	616/35	6662/565	14015/1969
Cellular (MOI=1)	4	477/29	4314/366	12002/1789
Cellular (MOI=1)	24	560/31	6599/453	15628/1005
Cellular (MOI=0.1)	0	650/0	6679/0	27499/0
Cellular (MOI=0.1)	0.5	693/7	6928/34	30023/61
Cellular (MOI=0.1)	1	810/26	9618/198	35347/ 415
Cellular (MOI=0.1)	2	611/28	5525/178	20147/ 521
Cellular (MOI=0.1)	4	732/24	8217/170	28244/388
Cellular (MOI=0.1)	24	780/27	10185/228	26908/446
Cellular (Uninfected)	0	697/0	7892/0	25345/0
Cellular (Uninfected)	0.5	670/0	7370/0	30168/0
Cellular (Uninfected)	1	698/0	8132/0	30637/0
Cellular (Uninfected)	2	717/0	7898/0	28582/0
Cellular (Uninfected)	4	741/0	8940/0	30421/0
Cellular (Uninfected)	24	706/0	8979/0	34137/0

Numbers based on non-redundant identifications

All numbers are averages of two technical replicate runs

Viral Proteome Characterization

The virulent *pac*-type phage 2972 contains 44 open reading frames. Due to two group I introns, the genome encodes 40 putative proteins (47). In our study, we detected

thirty-seven out of the forty predicted proteins, including all of those from the packaging, capsid morphogenesis, tail morphogenesis, and host lysis modules (Figure 3.1 and Table 3.2). PEG precipitation was performed on the cultures infected at MOI=1 in order to enrich the viral structural proteins. However, sequence coverage of the phage structural proteins was, in most cases, better in the cellular fractions than in the virus-enriched fractions (Table 3.2). In addition, the non-structural proteins were highly detected in the whole cell fractions. Therefore, the remainder of the data analyses focused on the cellular fractions.

The capsid and tail morphogenesis modules encompass all of the structural proteins, most of which are highly represented in our samples. Specifically, capsid morphogenesis proteins account for up to 2,871 normalized spectral counts in one run, and tail morphogenesis proteins account for 1,540 (Figure 3.1). In turn, individual structural proteins in these modules contribute a high number of total spectra, with up to 2,680 normalized spectral counts for the major capsid protein (*orf 9*), 2,220 for the major tail protein (*orf 15*), and 2,804 for one head protein (*orf 8*) across all runs infected at MOI=1 (Figure 3.1). In addition, all of the proteins from the host lysis module were synthesized, including a protein of unknown function, the holin, and the lysin (*orfs 24-26*). The phage proteins that were not detected in our study were genes of unknown function from the transcriptional regulation (*orf 39 & orf 41*) and lysogeny remnant modules (*orf 30*) (Table 3.2).

The spectral count abundances of phage 2972 proteins at each time point correlate well with the phase abundance values during the period in which complete cell lysis occurred (Figure 3.1). Specifically, the highest number of spectra in the MOI=1

experiment were recorded after one hour and lysis of the cell cultures occurred after two hours. The less robust infection at MOI=0.1 yielded fewer phage proteins, however the highest number was detected at two hours post-infection, and complete lysis occurred after four hours. The peaks in phage protein abundance after 1 and 2 hours for MOI=1 and MOI=0.1, respectively, followed by lysis, indicates these were peak infection times in our study (Figure 3.1).

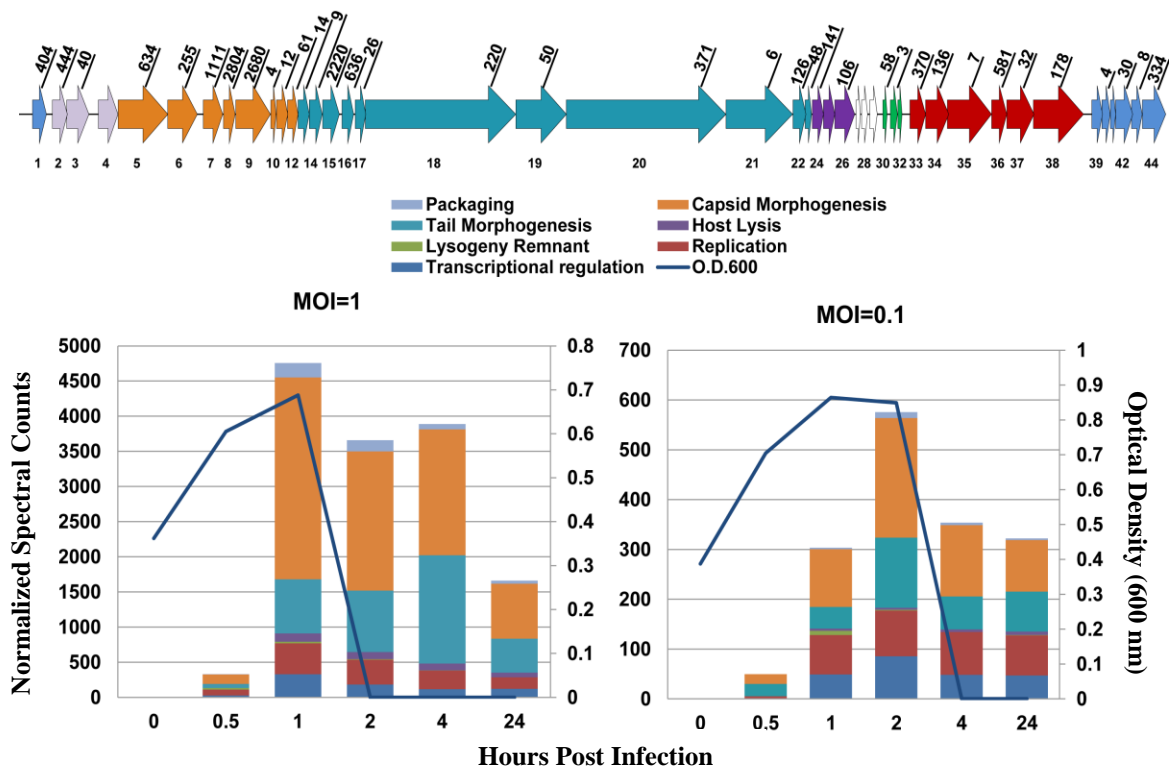


Figure 3.1: Phage 2972 spectral abundances. A.) Depiction of phage 2972, color coded according to functional modules. Each arrow represents an open reading frame and numbers on top are normalized spectral counts totaled across all MS runs at MOI=1. B.) Normalized spectral counts were added together at each time point of infection for MOI=1 (left panel) and MOI=0.1 (right panel). Optical density measurements (600 nm) (blue line) show cell lysis occurring immediately following the time points in which the highest numbers of phage spectra are detected at each MOI. Colors within each bar correspond to phage functional modules.

Table 3.2: Sequence coverages of phage 2972 proteins from virus-enriched and cellular fractions across infection time points.

Module	Orf	Description, Molecular Weight	Virus Enriched Fraction MOI=1						Cellular Fraction MOI=1						Cellular Fraction MOI=0.1					
			0	0.5	1	2	4	24	0	0.5	1	2	4	24	0	0.5	1	2	4	24
*	1	unknown function, MW:16160							65%	80%	85%	69%	71%			66%	63%	40%	35%	
Packaging	2	terminase small subunit, MW:16777			19%	24%	25%	56%	31%	85%	71%	67%	81%			19%	38%	25%	11%	
	3-4	terminase large subunit, MW: 47066								16%	43%	20%	19%							
Capsid Morphogenesis	5	portal protein, MW:57498			50%	58%	50%	51%	33%	56%	58%	55%	53%			43%	25%	24%	25%	
	6	head protein, MW:34367			21%	14%	13%	12%	19%	54%	53%	47%	85%			30%	18%	17%	24%	
	7	scaffold protein, MW:21262		34%	94%	96%	77%	85%	54%	99%	99%	73%	52%	11%	59%	47%	51%	44%		
	8	head protein, MW:12720		41%	96%	96%	98%	100%	86%	88%	91%	100%	89%	50%	77%	66%	63%	55%		
	9	major capsid protein, MW:37491	10%	20%	81%	85%	82%	82%	59%	83%	84%	81%	82%	37%	62%	61%	65%	63%		
	10	unknown function, MW:5997								89%					53%					
	11	unknown function, MW:13021			31%	61%				50%	47%	54%				37%				
	12	unknown function, MW:11470			66%	73%	68%	70%		58%	64%	39%	58%		35%	30%	30%	45%		
Tail Morphogenesis	13	unknown function, MW:12495					49%			35%	35%									
	14	unknown function, MW:14637					19%				26%	24%	32%		32%	24%	24%	28%		
	15	major tail protein, MW:18525		26%	84%	90%	86%	86%	64%	88%	86%	88%	86%	53%	66%	60%	65%	64%		
	16	unknown function, MW:13138							39%	100%	98%	93%	84%	58%	50%	66%	48%	30%		
	17	major tail protein, MW:12613								43%	42%	23%	36%							
	18	minor tail protein, MW:153506			9%	12%	8%	17%		14%	20%	14%	14%		2%	8%	5%	10%		
	19	minor tail protein, MW:57710			18%	20%	6%	22%		18%	28%	15%	23%			17%	7%	15%		
	20	antireceptor, MW:177330			18%	22%	14%	16%	5%	22%	29%	25%	34%		2%	5%	7%	18%		
	21	minor tail protein, MW:74279			11%	10%		7%			5%		4%			6%		4%		

	22	unknown function, MW:14539			73%	31%					82%	90%	77%	84%			20%	44%		
	23	unknown function, MW:5475			45%						47%	47%	47%	47%		47%				
Host Lysis	24	unknown function, MW:12328				33%					33%	38%	33%	38%				28%	23%	
	25	holin, MW:12004			31%						75%	74%	72%	59%			39%			
	26	lysine, MW:21754			49%	48%	38%				39%	49%	44%	35%			31%	20%	29%	42%
Lysogeny Remnant	30	unknown function, MW:5249																		
	31	cro-like repressor, MW:7850								92%	84%	56%	42%				54%	46%		46%
	32	unknown function, MW:5039								85%							58%			
Replication	33	unknown function, MW:18072								53%	65%	64%	52%	60%			32%	32%	43%	52%
	34	unknown function, MW:26164								28%	21%	47%	46%	66%			36%	22%	35%	44%
	35	helicase, MW:50968										9%		6%				10%		9%
	36	unknown function, MW:17290								82%	99%	99%	89%	82%		48%	88%	77%	88%	93%
	37	replication protein, MW:30474									29%	15%	28%	33%			16%			16%
	38	primase, MW:59060								9%	34%	40%	15%	36%			24%	11%	20%	30%
*Transcriptional Regulation	39	unknown function, MW:12133																		
	40	unknown function, MW:9580										25%		37%			52%			25%
	41	unknown function, MW:6311																		
	42	DNA binding protein, MW:19572								23%	41%	18%		52%			54%	54%	48%	57%
	43	unknown function, MW:12132									37%	35%					26%	24%	26%	
	44	unknown function, MW:27652								47%	86%	83%	45%	72%			31%	52%	56%	41%

Values are percent sequence coverages determined by dividing the number of amino acids detected in the mass spectrometry run by the total number of amino acids in a given protein. Numbers are averages between two technical replicates. Phage functional modules are labeled on the right (47). * Orf1 is part of the transcriptional regulation module. Intron regions are not included in the figure (*orf27-29*) and *orf3* and *orf4* encode one protein due to a splicing event (47).

***S. thermophilus* DGCC7710 Proteome Characterization**

In total, across all MS runs, 1,013 *S. thermophilus* DGCC7710 proteins were detected (Supplementary Table 1). As the genome encodes 2,079 open reading frames (<http://compbio.ornl.gov/CRISPRproteomics/>), this equates to proteomic identification of nearly half of the predicted proteins, the highest reported for any lactic acid bacterium to date (127). A global functional analysis was carried out by grouping host proteins detected at each time point by their COG (clusters of orthologous groups) categories (128) (Figure 3.2). Host proteins encompassed the range of cellular functions from energy production and conversion to defense mechanisms, with the greatest percentage of proteins in the translation, ribosomal structure and biogenesis, and carbohydrate transport and metabolism categories. The uninfected control cultures did not have any major changes in overall protein functional categories across the six time points measured. However, global changes in the host proteome were detected in phage 2972-infected cultures, including a decrease in protein abundances in the translation, ribosomal structure and biogenesis category around two hours post infection for the lower MOI =0.1 (37% at time 0 to 24% at 2 hpi), and at one hour post infection for the higher MOI =1 (33% at 0 hpi to 23% at 1 hpi). These time points correspond to peak infections of the cell populations at each MOI, as described earlier.

In addition, at the higher MOI=1, protein abundances in the carbohydrate transport and metabolism category show a considerable reduction following peak infection (22% at 0 hpi vs. 11% at 1 hpi.) (Figure 3.2). Decreased abundances of several key enzymes involved in carbohydrate transport and metabolism were detected, including

pyruvate kinase, enolase, 6-phosphofructokinase, 3-phosphoglycerate kinase, and glucose-6-phosphate isomerase (Figure 3.3 and Supplemental Table 2).

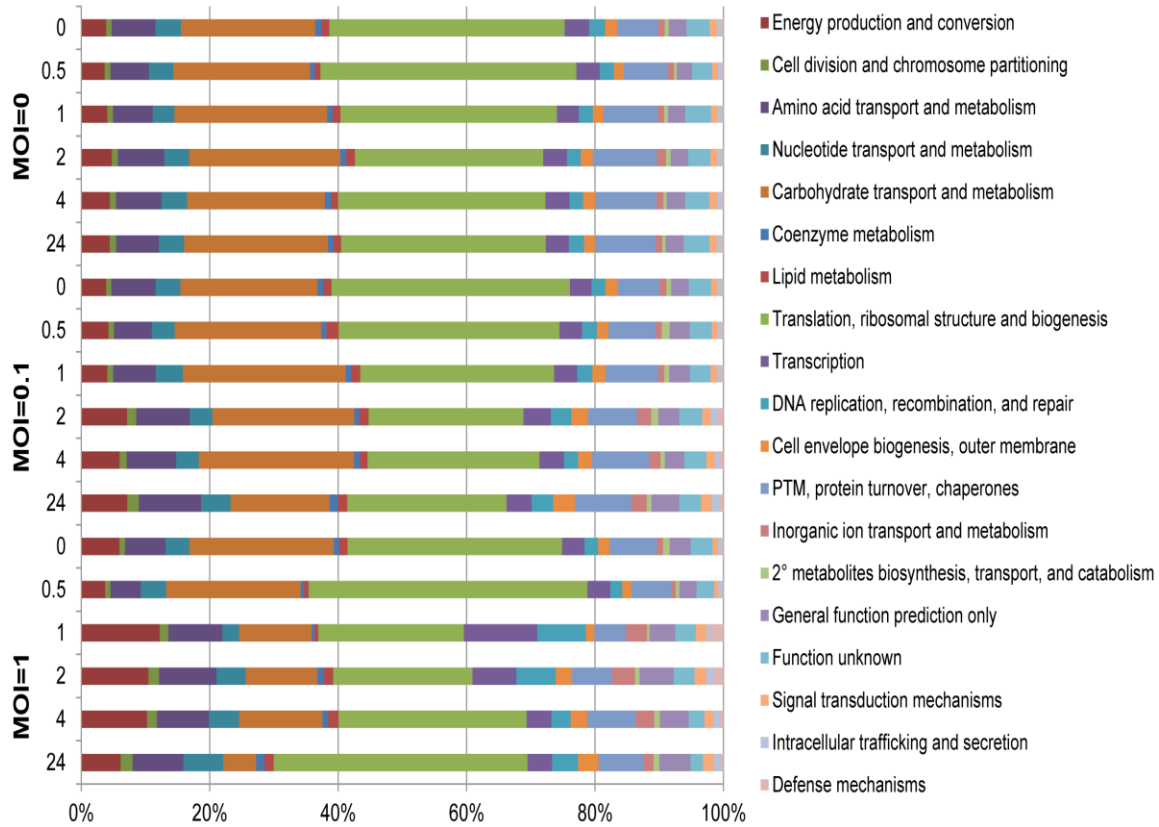


Figure 3.2: COG classification of *S. thermophilus* proteomes across infection time points. Proteins were grouped into functional categories by COG assignments. Percentages were calculated using normalized spectral counts averaged between two technical replicates.

Ribosomal protein abundances decreased during peak infection (41 ribosomal proteins decreased at 1 hpi MOI=1, 30 decreased at MOI=0.1) (Figure 3.3 and Supplemental Table 2). In contrast, abundances of ABC-type transporter proteins (28 at MOI=1, 26 at MOI=0.1), the majority of which are annotated as amino acid transporters but others include oligopeptide, metal ion, and phosphate transporters, increased (Figure 3.3 and Supplemental Table 2). The increased expression of ABC transporters is part of

the general stress response of these bacteria (129). Additionally, six subunits of the ATP synthase (α , β , δ , γ , ϵ , b) were detected and most increased in abundance in response to infection at both MOIs (Figure 3.3 and Supplemental Table 2). Interestingly, several restriction-modification protein subunits were also increased at peak infection times including two different methyltransferase subunits (HsdM) and two different endonuclease (HsdS) subunits (Figure 3.4).

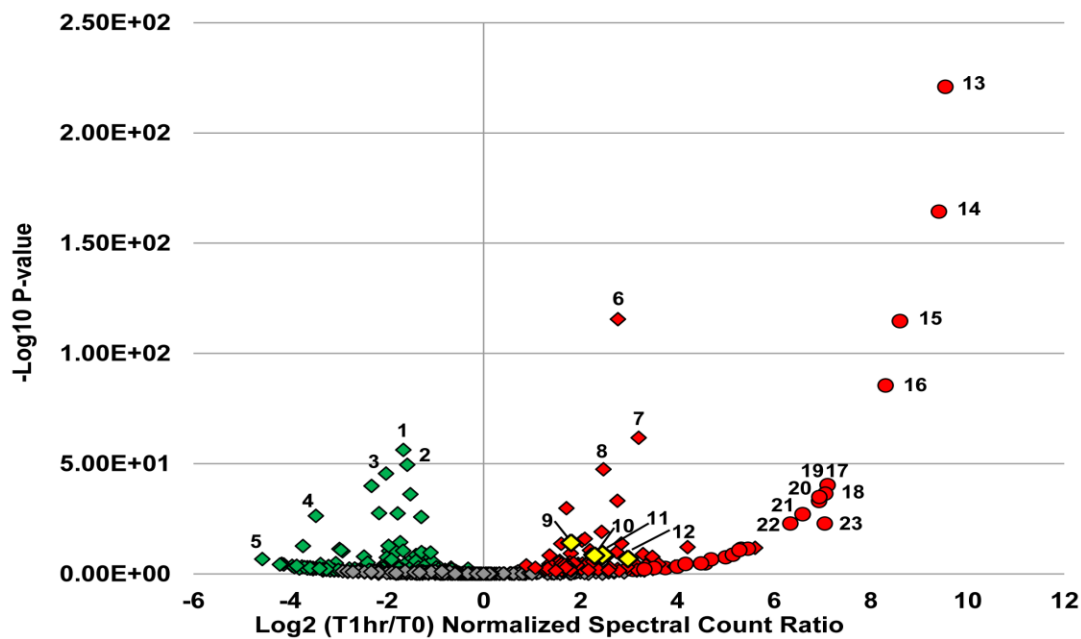


Figure 3.3: Volcano plot of protein abundance changes during peak infection at MOI=1. Normalized spectral counts were averaged between two technical replicates and the log₂ ratios taken between time 0 (pre-infection) and 1 hour post infection (peak infection). P-values were calculated using the exact Poisson test as described in the Materials and Methods section. The $-\log_{10}$ of the P-values are plotted on the y-axis. Red color indicates an increase in abundance, green a decrease in abundance, and grey, no change. Diamonds represent host proteins: 1.) glyceraldehyde -3-phosphate dehydrogenase, 2.) pyruvate kinase, 3.) 3-phosphoglycerate kinase, 4.) ribosomal protein S9, 5.) ribosomal protein S8, 6.) ATP synthase, β subunit, 7.) ABC transporter, ATPase, 8.) RNA polymerase, β -subunit. Cas proteins are highlighted in yellow: 9.) Cas6e (CRISPR4), 10.) Cas7 (CRISPR4), 11.) Cas9 (CRISPR1), and 12.) Cas9 (CRISPR3). Phage proteins are depicted in circles: 13. and 14.) head proteins, 15.) scaffold protein, 16.) tail protein, 17.) terminase small subunit, 18.) portal protein, 19.-23.) phage proteins of unknown function.

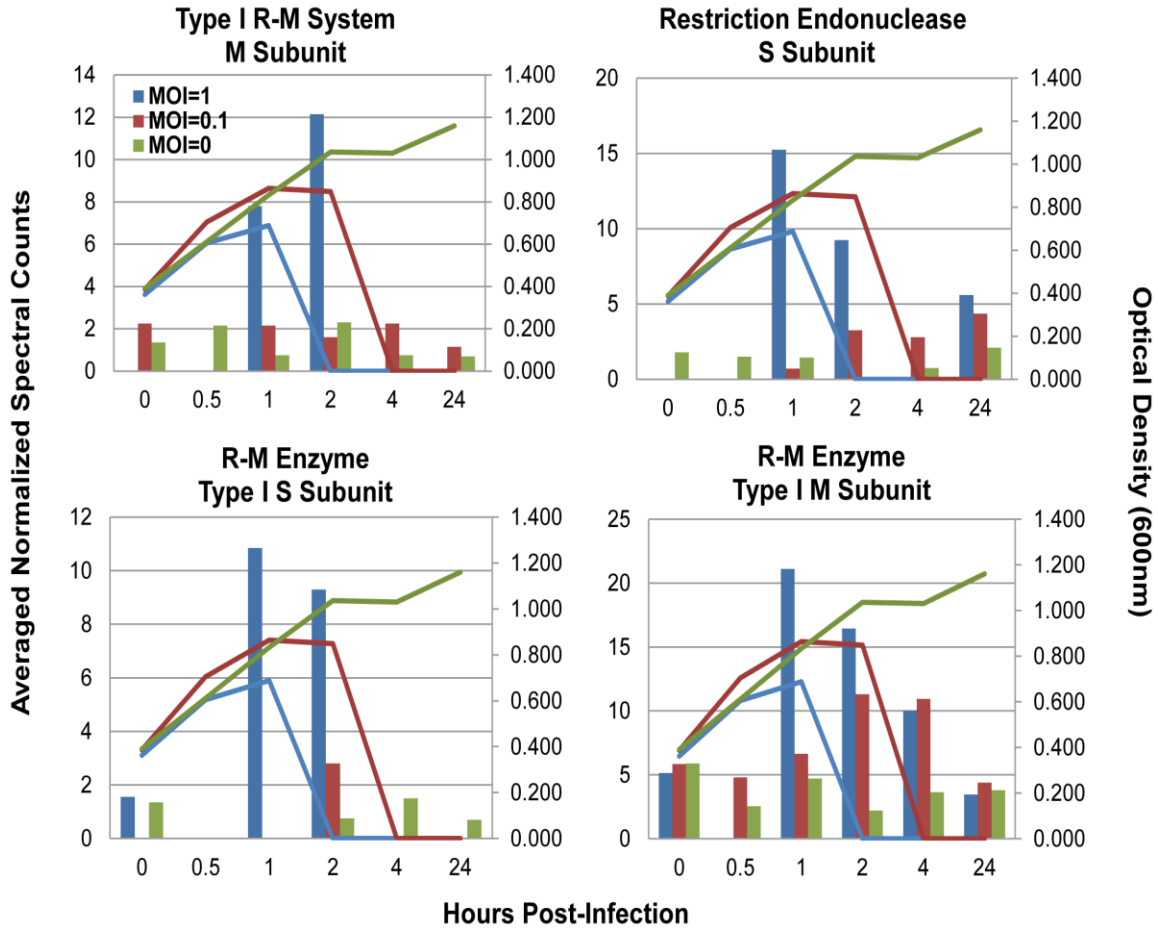


Figure 3.4: Restriction modification protein subunits increased at peak infection times. Bars indicate normalized spectral counts averaged between two technical replicates and lines are optical density measurements taken at each time point. Untreated cells MOI=0, green bars and lines, infected cells at MOI=0, maroon bars and lines, and infected cells at MOI=1, blue bars and lines. From top left to bottom right: Type I restriction-modification system methyltransferase subunit (ST89_075300), Restriction endonuclease S subunit (ST89_099800) Restriction-modification enzyme type I S subunit; specificity determinant HsdS (ST89_187033), Restriction-modification enzyme type I M subunit; type IC modification subunit HsdM (ST89_187066).

Analysis of the CRISPR/Cas response to phage infection

The most significant host response to phage 2972 was the increased production of several CRISPR-associated (Cas) proteins. Cas proteins were detected by unique peptides from each of the four loci present in *S. thermophilus* DGCC7710 (Table 3.3). Some, predominantly from CRISPR2 and CRISPR4, were constitutively expressed throughout

the time course, even in the uninfected cells. Interestingly, a clear increase in abundances of several Cas proteins corresponded to peak infections at both MOIs (1 hpi at MOI=1, 2 hpi at MOI=2) (Figure 3.5). The most marked increases were seen for the Cas9 proteins from locus CRISPR1 (ST89_070900), and locus CRISPR3 (ST89_147700), and Cas7 from locus CRISPR4 (ST89_103850).

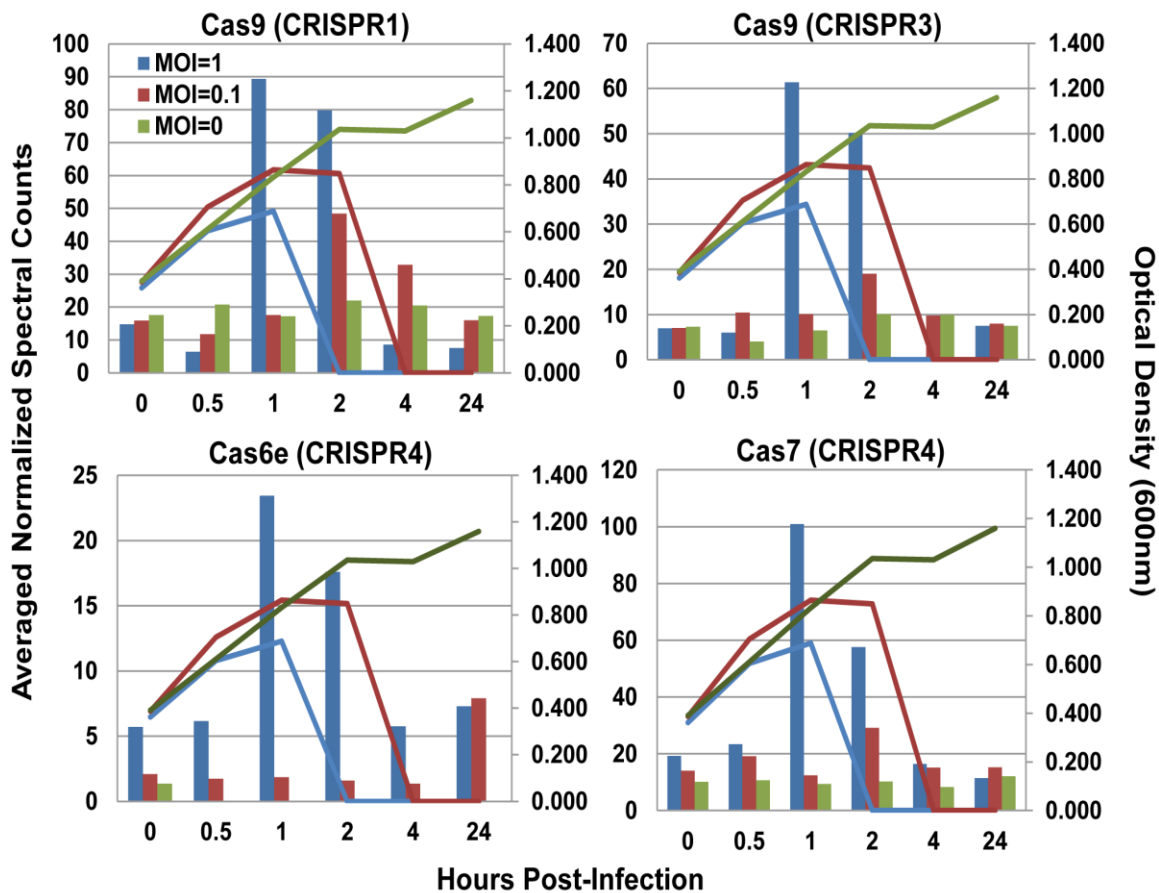


Figure 3.5: Cas proteins changing in response to phage 2972 infection. Values are normalized spectral counts averaged between two technical replicates. Untreated cells MOI=0, green bars, infected cells MOI=0, maroon bars, and infected cells at MOI=1, blue bars. Lines of the same color represent optical density measurements for each group. From top left to bottom right: Cas9 (ST89_070900) from CRISPR1 locus, Cas9 (ST89_097000) from CRISPR3 locus, Cas6e (ST89_103830) and Cas7 (ST89_103850) from CRISPR4 locus.

Table 3.3: Expression of Cas proteins from *S. thermophilus* DGCC710 across time.

Values are averaged normalized spectral counts taken at each time point from cells infected at MOI=1, MOI=0.1, and uninfected cells (MOI=0). All Cas proteins were detected by unique peptides.

Protein	Loci	MOI=1						MOI=0.1						MOI=0					
		0	0.5	1	2	4	24	0	0.5	1	2	4	24	0	0.5	1	2	4	24
Cas9	CRISPR1	15	6	89	80	9	8	16	12	18	48	33	16	18	21	17	22	21	17
Cas1																			
Cas2																			
Cas4																			
Cas1	CRISPR2									1									
Cas2																			
Cas6																			
Cas10										2							1		1
Csm2																	1		1
Csm3		4	3		3		2	4		2	2	1	4	2	4	2	4	4	4
Csm4									1	1	1		4	1				1	
Csm5										2		2	2		2	2	1	4	1
Csm6																			
Cas9	CRISPR3	7	6	61	50		8	7	10	10	19	10	8	7	4	7	10	10	8
Cas1																			
Cas2																			
Csn2																			
Cas3	CRISPR4				2				1	1	1		1		1		2		
Cse1										1		1					1		
Cse2		1			5		1	1	1	2	1		3			1	1		1
Cas7		19	23	101	58	16	11	14	19	12	29	15	15	10	11	9	10	8	12
Cas5					2			3	1	2	4	3	3	5	1	2	3	2	1
Cas6e		6	6	23	18	6	7	2	2	2	2	1	8	1					
Cas1					7	2			1										
Cas2																			

Discussion

The simultaneous measurement of phage and microbial host proteins over a time course of infection provides opportunity for novel insights into both phage protein production and the host anti-phage response. In this study, we detected nearly all the predicted phage proteins, validating the *in silico* protein predictions. In addition, expression of certain proteins within the cellular fraction and not in the viral enriched fraction suggests that the phage is utilizing the host machinery to produce these proteins, and they are likely not part of the phage structure. This is expected, given that most are encoded by the lysogeny, replication, and transcriptional regulation modules (Table 3.2). Many proteins identified in the cellular fractions were annotated as hypothetical or proteins of unknown function. Although we cannot define their specific functions, their synthesis indicates that they probably play some role in phage propagation.

Transcriptomic data for phage 2972 have been reported previously (118). Transcription of early, middle, and late genes occurs by 27 minutes after infection. However, we focused our analyses around the time of the expected phage burst (40 minutes after exposure) when viral proteins were at abundant levels to allow detection. Our inferred protein abundances correlate well with transcript abundance patterns, despite the lack of infection synchronicity and presence of cells that were phage resistant (101).

Since we were able to detect the vast majority of host proteins, we were able to characterize the overall host response upon infection with phage. The overall decrease in the translation, ribosomal structure and biogenesis COG category and in ribosomal proteins in particular, at peak infections, reflects the dramatic impact that phage infection

has on host physiology, especially immediately before lysis. Some of the changes in host proteome may be the result of phage take-over of cellular processes for transcription and translation of phage material, notably phage DNA packaging and proteins important for particle assembly.

Of particular interest was the detection of Cas proteins throughout our time course. Many Cas proteins were constitutively produced, consistent with reports indicating that crRNA is constitutively transcribed in the host, and can represent the most abundant small RNA species in the cell (130). Co-constitutive expression of both guide crRNA and Cas proteins would provide the cell with readily accessible defense against invading elements. Given the speed at which viruses can take over the host machinery, and their short replication cycle, constitutive expression of the CRISPR/Cas immune system ensures that the host immune response will be readily available upon infection.

Given that spacer addition has not been detected in CRISPR4 in prior studies (103, 131), it is notable that most of the CRISPR4 Cas proteins were constitutively expressed in uninfected cells, and that some increased in abundance in response to phage exposure. However, it is not known specifically how each locus acts and how the four loci in DGCC7710 interact. The proteins encoded by the CRISPR4 locus are homologous to the Cas proteins of *Escherichia coli* K12, and consist of: Cas1 (endonuclease), Cas2, and Cas3 as well as CasABCDE which form a complex called Cascade (CRISPR associated complex for antiviral defense) (114). The Cascade complex is composed of six copies of CasC, two copies of Cas B, and one copy each of Cas A, D, and E (128) *E. coli* Cas proteins A, B, C, D, and E are homologous to *S. thermophilus* Cas proteins Cse1, Cse2, Cas7, Cas5, and Cas6e (ST89_103870, ST89_103860, ST89_103850,

ST89_103840, ST89_103830), respectively. Cas7 (homologous to CasC in *E. coli*, the protein present in the most copies in the Cascade complex) was the most abundant *S. thermophilus* protein and dramatically increased around the time of peak phage 2972 infection. These data suggest that the CRISPR4 locus is functional (though not expanding its spacer inventory).

At peak phage infection, we detected dramatic increase in abundance of Cas9 proteins of CRISPR1 and CRISPR3, the two loci with previously demonstrated CRISPR activity. The Cas9 protein from locus CRISPR1, which is the signature protein for Type II CRISPR/Cas systems, was previously shown to be important in CRISPR-based immunity since deletion of the *cas9* gene (previously called *cas5* or *csn1*) eliminated phage-specific resistance despite the presence of matching spacer sequences (101). Cas9 was also recently shown to be necessary for the cleavage of invading plasmid and phage DNA (131). Observing an increase in Cas9 levels at the peak of infection is consistent with a prominent role of Cas9 in CRISPR-encoded immunity (130, 132). Cas9 proteins contain a HNH-like nuclease motif and are suspected to act on crRNA or foreign nucleic acids, indicating their involvement in the interference phase of the crRNA-mediated response. The increase in critical Cas protein abundance during peak infection indicates that although these proteins are constitutively produced, they can be induced following phage challenge as to increase the level of the primed CRISPR/Cas immune response. This allows the cells to readily acquire novel spacers in response to phage attack, and to mount a Cas9-dependent immune response against invading elements, notably during peak viral infection.

It is important to note that the absence of detection of the other Cas proteins does not necessarily mean a lack of expression. While there are no obvious attributes of the undetected proteins (too small, lacking sufficient tryptic peptides, or too few lysine's and arginine's) that would prohibit detection by our method, functionally, they may not need to be synthesized at high levels compared with other Cas proteins, and thus may fall below our level of detection. Notably, Cas1, which is found in nearly all genomes containing CRISPR, was not detected in our study. While it is thought that Cas1 plays an important role in the adaptation phase of the CRISPR response, it might only be synthesized by the minority of the cells in the population. In contrast, the Cas9 proteins are more highly detected and are likely expressed by the majority of the cells that take part in the interference phase.

Recently, transcription profiles of CRISPR systems in *Thermus thermophilus* HB8 upon infection with phage Φ YS40 have been reported (133, 134). However, *Thermus thermophilus* is very distant from *Streptococcus thermophilus*, and their CRISPR systems are vastly different. Actually, both the CRISPR1 and CRISPR3 systems in our model organism are idiosyncratic type II CRISPR/Cas systems, while those induced by phage in the *Thermus thermophilus* system are type I and type III systems, which have different mechanisms of action. There have also been other published works on CRISPR transcription in other systems including *E. coli* (114, 135, 136), *Sulfolobus* (137) and *P. furiosus* (112). While transcriptomic studies offer valuable information at the mRNA level, the proteomic approach used in this study is the first to quantify the final protein products, Cas proteins, over a time course of phage infection.

Interestingly, several type I restriction-modification (R-M) protein subunits were detected during our time course and some increased in abundance at peak infection (Fig 3.4). Restriction modification systems are a type of anti-viral defense in which invading foreign DNA is cleaved at target sites while host DNA is protected. Type I R-M systems utilize a multifunctional enzyme made up of three subunits encoded by different *hsd* (host specificity determinant) genes. The HsdR (restriction) subunit functions as a restriction endonuclease cleaving foreign DNA while the HsdS (specificity) and HsdM (modification) subunits are sufficient for modification activity and can form an independent methyltransferase (MTase) that specifically recognizes non-palindromic DNA sequences and cleaves at a non-specific site distant from the recognition sequence. Two Type I R-M system methyltransferase subunits (ST89_075300 and ST89_187066) were identified throughout the time course and increased in abundance during peak infection (Figure 3.4). These two related proteins were distinguishable because they have low amino acid identity and generate unique tryptic peptides upon enzymatic digestion. Similarly, two different type I R-M S subunits were identified (ST89_099800 and ST89_187033) and increased in abundance during peak infection. Detection of two distinct M subunits and two distinct S subunits suggests operation of two type I R-M systems.

This study is, to our knowledge, the first to report protein abundance increases of restriction-modification proteins, in direct correlation with time points in which Cas protein abundances are increased. While restriction-modification genes and CRISPR/Cas genes are mutually encoded in lactic acid bacterial genomes (121, 138) it is not clear whether the two anti-viral systems are working simultaneously or if they share

components. The expression of proteins from these two systems simultaneously suggests that perhaps there is a correlation between Cas proteins and R/M systems in type II CRISPR/Cas systems.

In conclusion, mass spectrometry-based proteomics studies provided insights into the protein profiles of phage 2972 and its host proteome response to viral infection. We showed that, in *S. thermophilus*, the CRISPR/Cas systems are constitutively expressed and can be induced by viral challenge.

CHAPTER FOUR

Metaproteomics of a Gutless Marine Worm and its Symbiotic Microbial Community Reveal Unusual Pathways for Carbon and Energy Use

Portions of text are adapted from: Kleiner M, Wentrup C, Lott C, Teeling H, Wetzel S, Young, J, Chang YJ, Shah M, VerBerkmoes NC, Zarzycki J, Fuchs G, Markert S, Hempel K, Voigt B, Becher D, Liebeke M, Lalk M, Albrecht D, Hecker M, Schweder T, and Dubilier N. (2012) Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proceedings of the National Academy of Sciences* 109: E1173–E1182.(139).

Jacque Young's contributions include experimental preparation of samples, nano-2D-LC-MS/MS analysis, data analysis, transposase analysis, and authorship.

Abstract

Low nutrient and energy availability has led to the evolution of numerous strategies for overcoming these limitations, of which symbiotic associations represent a key mechanism. Particularly striking are the associations between chemosynthetic bacteria and marine animals that thrive in nutrient-poor environments such as the deep-sea because the symbionts allow their hosts to grow on inorganic energy and carbon sources such as sulfide and CO₂. Remarkably little is known about the physiological strategies that enable chemosynthetic symbioses to colonize oligotrophic environments.

In this study, we used metaproteomics and metabolomics to investigate the intricate network of metabolic interactions in the chemosynthetic association between *Olavius algarvensis*, a gutless marine worm, and its bacterial symbionts. We propose novel pathways for coping with energy and nutrient limitation, some of which may be widespread in both free-living and symbiotic bacteria. These include (i) a pathway for symbiont assimilation of the host waste products acetate, propionate, succinate and malate, (ii) the potential use of carbon monoxide as an energy source, a substrate previously not known to play a role in marine invertebrate symbioses, (iii) the potential use of hydrogen as an energy source, (iv) the strong expression of high affinity uptake transporters, and (v) novel energy efficient steps in CO₂ fixation and sulfate reduction. The high expression of proteins involved in pathways for energy and carbon uptake and conservation in the *O. algarvensis* symbiosis indicates that the oligotrophic nature of its environment exerted a strong selective pressure in shaping these associations.

Introduction

Growth in nutrient-limited environments presents numerous challenges to organisms. Symbiotic and syntrophic relationships have evolved as particularly successful strategies for coping with these challenges. Such nutritional symbioses are widespread in nature and have, for example, enabled plants to colonize nitrogen-poor soils, and animals to thrive on food sources that lack essential amino acids and vitamins (140). Chemosynthetic symbioses, discovered only 35 years ago at hydrothermal vents in the deep sea, revolutionized our understanding of nutritional associations, because these symbioses enable animals to live from inorganic energy and carbon sources such as

sulfide and CO₂ (141, 142). The chemosynthetic symbionts use the energy obtained from oxidizing reduced inorganic compounds such as sulfide to fix CO₂, ultimately providing their hosts with organic carbon compounds. Chemosynthetic symbioses are thus able to thrive in habitats where organic carbon sources are rare such as the deep sea, and the symbionts are often so efficient at feeding their hosts that many have reduced their digestive system (143).

The marine oligochaete *Olavius algarvensis* is a particularly extreme example for a nutritional symbiosis: these worms are dependent on their chemosynthetic symbionts for both their nutrition and their excretion, as they have completely reduced their mouth, gut and nephridial excretory organs (144). *O. algarvensis* lives in coarse-grained coastal sediments off the island of Elba, Italy, and migrates between the upper oxidized and the lower reduced sediment layers (145). It hosts a stable and specific microbial consortium consisting of five bacterial endosymbionts in its body wall – two aerobic or denitrifying gammaproteobacterial sulfur oxidizers (γ 1- and γ 3-symbiont), two anaerobic deltaproteobacterial sulfate reducers (δ 1- and δ 4-symbiont) and a spirochaete with an unknown metabolism (146, 147). The sulfate-reducing δ -symbionts provide the sulfur-oxidizing γ -symbionts with reduced sulfur compounds as an internal energy source for autotrophic CO₂ fixation via the Calvin-Benson cycle, thus explaining how *O. algarvensis* can thrive in its sulfide-poor environment (145) (148). However, as all living organisms, the symbiosis is dependent on external energy sources, but to date it has remained unclear what these are.

Like the vast majority of symbiotic microbes, the *O. algarvensis* symbionts have so far defied cultivation attempts, making cultivation-independent techniques essential

for their analysis. A metagenomic analysis of the *O. algarvensis* symbionts yielded first insights into their potential metabolism (148). However, the incomplete genome sequences hindered the reconstruction of complete metabolic pathways, leaving many questions unanswered (149). Furthermore, as in all genomic analyses, detailed insights into the physiology and metabolism of an organism are limited as these analyses can only predict the metabolic potential of an organism, but not its actual metabolism and physiology (150). This limitation is most apparent in a multi-member community in which the interactions between the different members and between these and their environment lead to a level of metabolic complexity that can greatly exceed the predictive ability of genomic reconstructions from single species.

While metagenomic analyses reveal the metabolic potential of a microbial community, metaproteomic and metabolomic analyses provide evidence for the metabolic and physiological processes that are actually used by the community. In this study, we used metaproteomics and metabolomics as well as enzyme assays and in situ analyses of potential energy sources to gain an in-depth understanding of the intricate interactions between *O. algarvensis* and its microbial symbiont community, and between these and their environment. Our goal was to identify the compounds that provide energy for the symbiosis, the functional roles of the different partners, and their interactions within the symbiosis.

Materials and Methods

Sample collection and symbiont enrichment

Worms were removed from the sediment via decantation and either frozen immediately or symbionts were enriched via isopycnic centrifugation using a HistoDenz™ (Sigma® Saint Louis, Missouri, USA) based density gradient prior to freezing. Symbiont abundance and composition in density gradient fractions was analyzed with catalyzed reporter deposition-fluorescence *in situ* hybridization (CARD-FISH) using symbiont specific probes. Density gradient fractions in which specific symbionts were enriched were chosen for subsequent analyses.

Protein identification and proteome analyses

One dimensional polyacrylamide gel electrophoresis followed by liquid chromatography (1D-PAGE-LC) and two dimensional liquid chromatography (2D-LC) were used for protein and peptide separation as described previously (151, 152) with slight modifications. Mass spectra and tandem mass spectra were acquired with a hybrid linear ion trap-Orbitrap (Thermo Fischer Scientific) as described previously (151, 153). All MS/MS spectra were searched against two protein sequence databases composed of the symbiont metagenomes and the genomes of related organisms using the SEQUEST algorithm. For protein identification only peptides identified with high mass accuracy (maximum ± 10 ppm difference between calculated and observed mass) were considered and at least two different peptides were required to identify a protein. False discovery rates were estimated with searches against a target-decoy database as described previously (154, 155) and were determined to be between 0 - 3.27%. For relative quantitation of proteins, normalized spectral abundance factor (NSAF) values were calculated for each sample according to the method of Florens et al. (156). All identified

proteins and their relative abundance in different samples are shown in datasets S1 and S2. Protein databases, peptide and protein identifications as well as all MS/MS spectra are available at http://compbio.ornl.gov/olavius_algarvensis_symbiont_metaproteome/.

All supplemental material can be downloaded from:

<http://www.pnas.org/content/109/19/E1173/suppl/DCSupplemental>.

Proteomics-based binning

Proteins encoded on metagenome fragments that were not previously assigned to a specific symbiont were tentatively assigned (binned) to a specific symbiont if they were repeatedly detected in higher abundances in enrichments of only one specific symbiont. To validate this approach and to calculate a false assignment rate we also did proteomics binning with the proteins that had already been assigned to a specific symbiont in the metagenomic study (Table S3, SI Text).

Enzyme tests

Enzymatic activities were determined in cell extracts from either whole worms or enriched symbionts (SI Text, Table S5). Detailed methods for all enzyme activity assays are provided in the SI Text.

Measurement of hydrogen and CO concentrations in the *O. algarvensis* habitat

Seawater and pore water samples from 25 cm sediment depth were collected by research divers using a stainless steel needle and capped syringes. A total of 9 sites within an area of approx. 100 m² at the *O. algarvensis* collection site were sampled. Hydrogen and CO

concentrations were measured the same day using a RGA3 reduction gas analyzer (Trace Analytical Inc., Menlo Park, CA, USA).

Metabolite identification and quantification in whole worms and pore water

Whole worms were extracted using ice cold ethanol based solvent mixture and ultrasonication. Metabolites were measured with GC-MS, LC-MS and ¹H-NMR as described previously (157). Detected metabolites are shown in table S2. Relative quantification of metabolites was performed on the basis of complete spectrum/chromatogram intensities (SI Text). Pore water was sampled at different sediment depths in the *O. algarvensis* habitat by scuba divers with RHIZON® MOM 10 cm soil water samplers (F. Meijboom, Wageningen, NL) and measured using GC-MS as described in Liebeke et al. (2011) (157).

Results/Discussion

High coverage of the symbiosis metaproteome and metabolome

We identified and quantified a total of 2,819 proteins and 97 metabolites in *O. algarvensis* and its symbiotic community (Datasets S1 and S2, Tables S1 and S2) using different methods for both the metaproteomic and the metabolomic analyses to overcome the intrinsic biases inherent in a single detection method (SI Text). For host proteins, sequences from related annelids enabled the cross-species identification of 530 *O. algarvensis* proteins, thus providing the first insight into the metabolism of a marine oligochaete, a group of annelid worms for which no genomic data is available. For symbiont proteins, the published *O. algarvensis* symbiont metagenome, which contains

only sequences assigned to specific symbionts through binning analyses (148), led to the identification of 1,586 proteins. The addition of unassigned sequences from the unbinned *O. algarvensis* symbiont metagenome allowed us to identify a total of 2,265 symbiont proteins, a 43% increase compared to the published metagenome alone. No proteins were found that could be unambiguously assigned to the spirochaete symbiont of *O. algarvensis*, due to the lack of metagenomic information for the spirochaete (148).

To further improve coverage of the metaproteome we developed a method using density gradient centrifugation for physical separation of the *O. algarvensis* symbionts from each other and host tissues (SI Text, Fig. S1). This greatly enhanced the number of identified symbiont proteins, particularly for those present in lower abundances (Fig. 1, Table S1). An additional advantage of symbiont enrichments was that we were able to assign proteins from the unbinned metagenomic sequences to a specific symbiont if they were detected in high abundances in enrichment fractions of the given symbiont (SI Text). This “proteomics-based binning” allowed us to assign 544 previously unassigned proteins to a specific symbiont, thus extending our understanding of the symbionts’ metabolism significantly (Dataset S3, Table S3).

Energy sources for the symbiosis

One of the major unresolved questions in the *O. algarvensis* symbiosis is what the sources of energy from the environment are that fuel the association. Earlier studies found that reduced sulfur compounds are supplied internally as an energy source to the aerobic sulfur-oxidizing γ -symbionts by the anaerobic sulfate-reducing δ -symbionts. In return, the δ -symbionts are supplied with oxidized sulfur compounds as electron

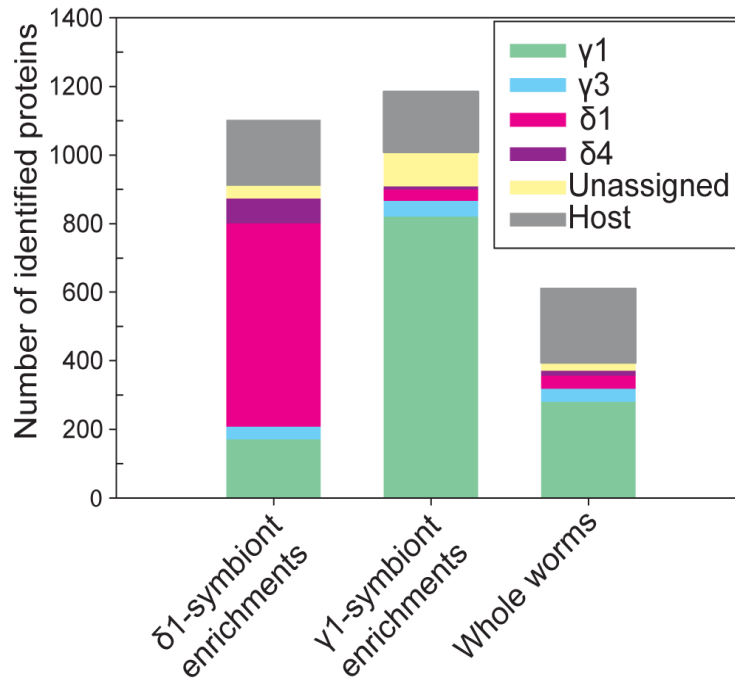


Figure 4.1: Positive effect of symbiont enrichment using density gradient centrifugation. Enrichments considerably increased the number of identified proteins in a given symbiont compared to analyses of whole worms. Average protein numbers were calculated for 2D-LC-MS/MS experiments (Table S1). γ 1-symbiont enrichments (n=2), δ 1-symbiont enrichments (n=2) and whole worm samples (n=4). For assignment of proteins to a symbiont, both metagenomic and proteomic binning information was used.

acceptors (145, 148). Our metaproteomic analyses confirmed this model of syntrophic sulfur cycling with the detection of abundantly expressed sulfur oxidation proteins in the γ -symbionts and sulfate reduction proteins in the δ -symbionts (Fig. 2, S2 and S3a, SI Text). However, for net growth and compensation of thermodynamic losses, external energy sources are required. Most chemosynthetic symbioses are fueled by an external supply of reduced sulfur compounds, but concentrations of reduced sulfur compounds are

extremely low in the habitat of *O. algarvensis* (145). This indicates that other energy sources play an important role in the symbiosis.

Carbon monoxide may be used by three symbionts

Our metaproteomic analyses indicate that three of the *O. algarvensis* symbionts use carbon monoxide (CO) as an energy source. CO is not known to be used as an electron donor by chemosynthetic symbionts. Its toxicity for all aerobic forms of life precluded the assumption that this reductant could play an important role in animal symbioses. We detected both aerobic and anaerobic CO dehydrogenases in the *O. algarvensis* consortium, the aerobic type in the γ 3-symbiont and the anaerobic type in the deltaproteobacterial symbionts (for reviews of aerobic and anaerobic CO oxidation see King and Weber (158) and Oelgeschläger and Rother (159)). The deltaproteobacterial symbionts express two versions of the anaerobic CO dehydrogenase, one that oxidizes free CO generating a proton gradient across the membrane, and one that is likely involved in the Wood-Ljungdahl pathway and oxidizes enzyme bound CO (SI Text).

To support the metaproteomic prediction that CO could be an energy source for the symbiosis we measured CO concentrations in the *O. algarvensis* habitat. CO concentrations in the sediment pore waters ranged from 17 to 46 nM (Fig. S4). These concentrations are sufficient to support marine CO oxidizers, which can use concentrations of 2-10 nM in surface sea water (160, 161) and about 100 nM at hydrothermal vents (158). Pore water CO concentrations were well above the concentrations in the seawater overlaying the sediment (8 to 16 nM) (Fig. S4), indicating the presence of a CO source in the sediment. CO can be produced through abiotic and

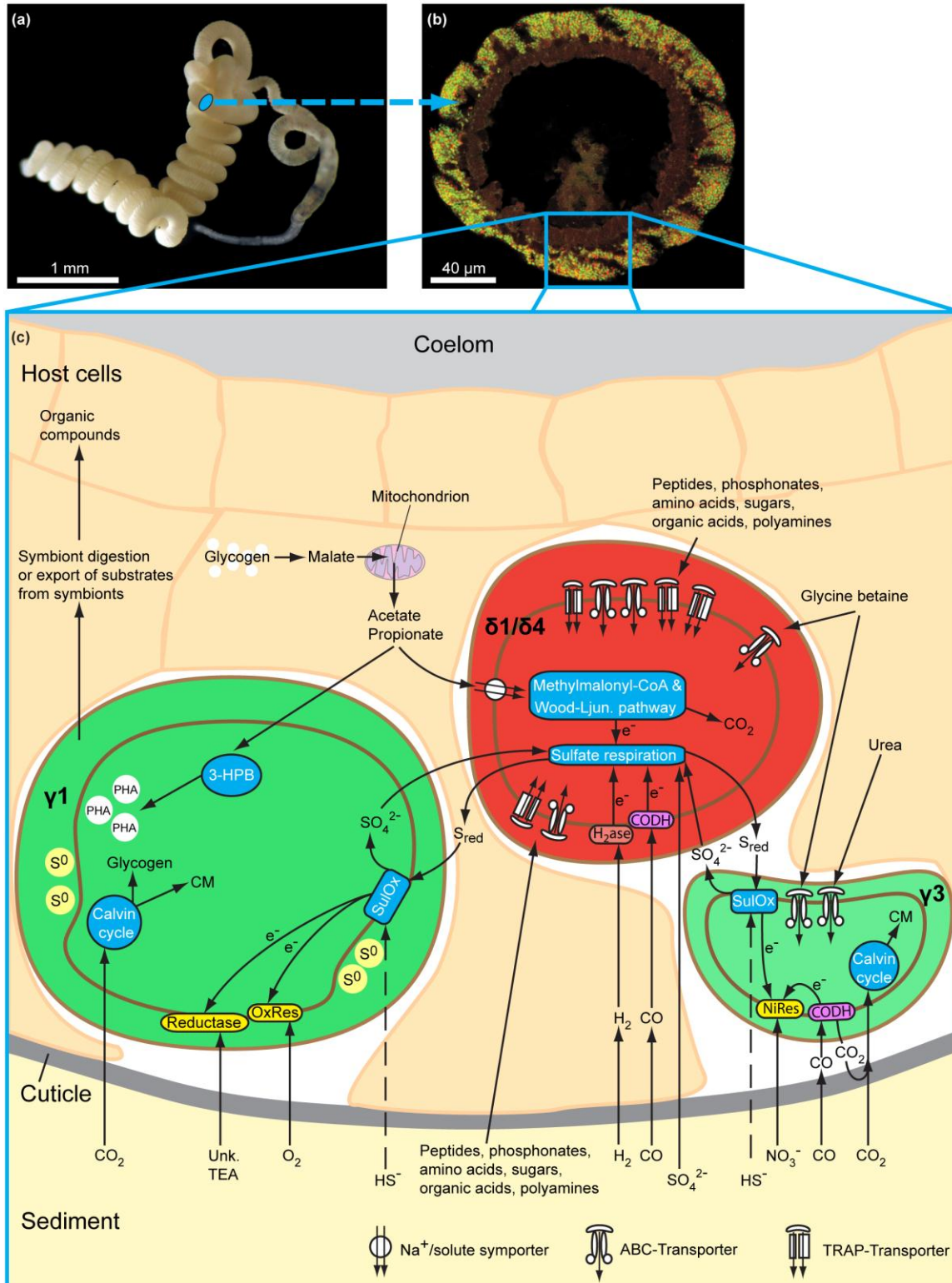


Figure 4.2: Overview of symbiotic metabolism based on metaproteomic and metabolomic analyses. (a) Live *Olavius algarvensis* specimen. (b) Cross section through *O. algarvensis* showing the symbionts just below the worm's cuticle with specific

fluorescence in situ hybridization probes (sulfur-oxidizing symbionts in green, sulfate-reducing symbionts in red). (c) Metabolic reconstruction of symbiont and host pathways. The $\delta 1$ - and $\delta 4$ -symbionts are shown as a single cell, because most metabolic pathways were identified in the $\delta 1$ -symbiont and only a small fraction of the same pathways were identified in the $\delta 4$ -symbiont due to the low coverage of its metaproteome. 3-HPB, partial 3-hydroxypropionate bi-cycle; CM, cell material; CODH, carbon monoxide dehydrogenase (aerobic or anaerobic type); NiRes, nitrate respiration; OxRes, oxygen respiration; PHA, polyhydroxyalkanoate granule; S0, elemental sulfur; Sred, reduced sulfur compounds; SulOx, sulfur oxidation; Unk. TEA, unknown terminal electron acceptor (56).

biotic processes from plant roots (162) and from decaying seagrass-derived organic matter (163), which are abundant at the collection site of the worms.

The CO_2/CO couple has a very negative redox potential E° of -520 mV (164) making CO an excellent electron donor, whose electrons can be transferred to a variety of terminal electron acceptors such as oxygen, nitrate, elemental sulfur and sulfate (158, 159, 165). CO could therefore be used as an energy source by the *O. algarvensis* symbionts under all redox conditions as the worm shuttles between sediment layers. In the reduced sediment layers the δ -symbionts could use sulfate for the anaerobic oxidation of CO, thereby producing reduced sulfur compounds for the γ -symbionts, while in the oxic and suboxic sediment layers the $\gamma 3$ -symbiont could oxidize CO with nitrate as a terminal electron acceptor (SI Text on terminal electron acceptors).

Hydrogen may be used by the sulfate-reducing symbionts

Our metaproteomic analyses revealed that hydrogen may play an important role as an energy source in the *O. algarvensis* symbiosis, based on the abundant expression of periplasmic uptake [NiFeSe] hydrogenases in both δ -symbionts ($\delta 1$: SP088, $\delta 4$: SP089).

These [NiFeSe] hydrogenases have high affinities for hydrogen (166), which is consistent with the low hydrogen concentrations reported for oligotrophic sediments (<10 nM) (167) and marine sediments in general (<60 nM) (168). We were therefore surprised to measure unusually high concentrations of hydrogen ranging from 438 to 2147 nM in the sediment pore waters at the *O. algarvensis* collection site (Fig. S5). These high concentrations could be a result of biological H₂ production by anaerobic CO oxidizers, and are consistent with the elevated CO concentrations at the collection site. The hydrogen concentrations in the worms' habitat are much higher than those needed by common hydrogen oxidizing microorganisms for growth (169), indicating that the δ -symbionts could use the hydrogen present in the Elba sediment as an energy source.

The use of hydrogen as an energy source by chemoautotrophic sulfur-oxidizing symbionts was recently shown for deep-sea *Bathymodiolus* mussels from hydrothermal vents (170). Our study indicates that hydrogen could also play a role as an energy source in shallow-water chemosynthetic symbioses. As with carbon monoxide, the use of externally supplied hydrogen might be another adaptation of the *O. algarvensis* symbiosis to life in the sulfide-depleted sediments of Elba.

Highly abundant uptake transporters for organic substrates in the δ -symbionts

The sulfate-reducing δ -symbionts expressed extremely high numbers and quantities of high affinity uptake transport related proteins, which enable them to take up organic substrates at very low concentrations (Datasets S2 and S4). In the δ 1-symbiont, between 89 and 116 transport proteins were detected per sample, which corresponds to an average of 29% of all identified δ 1-symbiont proteins. In terms of abundance the δ 1-

symbiont transport proteins amounted to over 38% of the total δ 1-symbiont protein (Table S4). Higher abundances of these types of transporters have to our knowledge only been found in the metaproteome of the α -proteobacterium *Pelagibacter ubique* (SAR11) from the Sargasso Sea during extremely low nutrient conditions (171) (Table S4). Most of the identified transport proteins in the δ 1-symbionts were periplasmic binding proteins of high-affinity ABC or TRAP type transporters, which actively transport substrates against a large concentration gradient while using energy in the form of ATP or an ion gradient (172) (173). The vast majority of the detected δ 1-symbiont transport proteins are used for the uptake of a variety of substrates such as amino acids, peptides, di- and tricarboxylates, sugars, polyamines and phosphonates, with amino acid and peptide transporters the most dominant ones (Dataset S4). The remarkable abundance of transport-related proteins in the δ -symbionts suggests that these symbionts use organic substrates not only as an energy source, but also as source for preformed building blocks, thus saving resources by not having to synthesized these *de novo*.

The organic substrates used by the δ -symbionts could be supplied internally from within the worms or externally from the environment. Our metabolomic analyses of whole worms revealed considerable amounts of dicarboxylates and some amino acids (in the low mM range), making an internal source of the organic substrates possible (Fig. S7, Table S2). However, the relatively high concentrations of these substrates contradict the expression of energy consuming high-affinity transporters by the δ -symbionts. In cultured bacteria (174-176) as well as in environmental communities (171, 177) ABC/TRAP transporters are induced at low substrate concentrations, while under nutrient-rich conditions, less energy-consuming transporters are used. Most likely the metabolites that

we measured in homogenized worms are not easily accessible to the δ -symbionts *in situ* as they are enclosed in host or symbiont cells.

To examine if organic substrates are supplied externally from the *O. algarvensis* environment, we analyzed sediment pore waters from the worm's collection site with GC-MS for the presence of a large range of di- and tricarboxylates, amino acids and sugars. None of these metabolites were measurable, with detection limits at about 10 nM (Fig. S8). Such oligotrophic conditions are consistent with the high expression of ABC/TRAP transporters that have extremely high affinities for substrates at concentrations far below the detection limits of our method (178, 179). The worm's cuticle is permeable for small negatively charged compounds as well as substrates up to 70 kDa (144) and thus the δ -symbionts would have access to both small organic compounds such as di- and tricarboxylates as well as larger organic substrates such as sugars and polyamines from the environment. The expression by the δ -symbionts of transporters for a very broad range of substrates would allow them to quickly respond to and take up many different substrates that could be either consistently present at low concentrations in their environment or fluctuate over time and space as the worm migrates through the sediment.

Regardless of whether the organic substrates come from the environment or internally from within the symbiosis, the high abundances of high-affinity uptake transporters in the δ -symbionts indicate that they experience nutrient limitation, forcing them to dedicate a major part of their resources to the acquisition of substrates. Despite their endosymbiotic location, the lifestyle of these bacteria thus appears to most closely resemble that of planktonic SAR11 bacteria from low-nutrient extremes in the Sargasso Sea (171).

Recycling and waste management

Given the extremely low concentrations of nutrients in the *O. algarvensis* habitat, the conservation of substrates and energy should be highly advantageous for the symbiosis. Our metaproteomic and metabolomic analyses revealed several pathways, some of which have not been previously described, that could enable the symbionts to recycle waste products of their hosts and conserve energy.

Proposed pathways for the recycling of host fermentative waste in multiple symbionts

Cross-species identification of host proteins enabled us for the first time to gain insight into the metabolism of *O. algarvensis*. Our analyses revealed that *O. algarvensis* expressed proteins for an anaerobic metabolism that produces large amounts of acetate, propionate, malate, and succinate as fermentative waste products, when living in deeper anoxic sediment layers (180, 181) (SI Text, Fig. S6, Dataset S2). Correspondingly, we detected considerable amounts (1 to 8 mM) of malate, succinate, and acetate in the worm metabolome (Fig. S7, Table S2). Aquatic invertebrates without symbionts must excrete these fermentative waste products to keep their internal pH stable, thereby losing large amounts of energy-rich organic compounds. In *O. algarvensis*, the ability of the sulfate reducing δ -symbionts to use their host's fermentative waste as substrates recycles and preserves considerable amounts of energy and organic carbon within the symbiotic system (SI Text).

Surprisingly, the dominant γ 1-symbiont, previously assumed to only fix carbon autotrophically, may also function heterotrophically by assimilating acetate, propionate, succinate, and malate, thus also contributing to host waste recycling. We detected

abundantly expressed enzymes for an almost complete 3-hydroxypropionate bi-cycle (3-HPB) in the γ 1-symbiont (Dataset S2, Fig. 3 and S3b). The 3-HPB is used for autotrophic CO₂ fixation in *Chloroflexus aurantiacus*, a filamentous anoxygenic phototroph (182), but parts of the 3-HPB pathway can also be used for the heterotrophic assimilation of acetate, propionate, succinate and malate (183).

In retrospect, it is clear why the 3-HPB was not discovered in the metagenomic analyses of the *O. algarvensis* symbionts: many of its genes occurred on sequence fragments that could not be assigned to a specific symbiont and were therefore not included in the annotation analyses (148). Here, we used our 'proteomics-based binning' method described above to assign abundantly expressed 3-HPB enzymes encoded on unassigned metagenomic fragments to the γ 1-symbiont (SI Text, Materials and Methods, Dataset S3). This enabled us to identify nearly all enzymes required for the complete 3-HPB, with the exception of two diagnostic enzymes of the 3-HPB - malonyl-CoA reductase and propionyl-CoA synthase - that were missing in both the metagenome and the metaproteome (Fig. 4.3).

To better understand how the 3-HPB might function in the symbionts, we performed enzyme assays with extracts from whole worms and enriched γ 1-symbionts. Activities of all 3-HPB enzymes were detected, except the two diagnostic enzymes that were also absent from the metaproteome (Table S5). We therefore propose a modified incomplete 3-HPB as shown in Fig. 4.3, which the γ 1-symbiont could use to assimilate the host's fermentative waste products acetate, propionate, succinate, and malate. The abundant expression of the modified 3-HPB suggests that it plays an important role in the central carbon metabolism of the γ 1-symbionts. The net fixation of CO₂ is unlikely

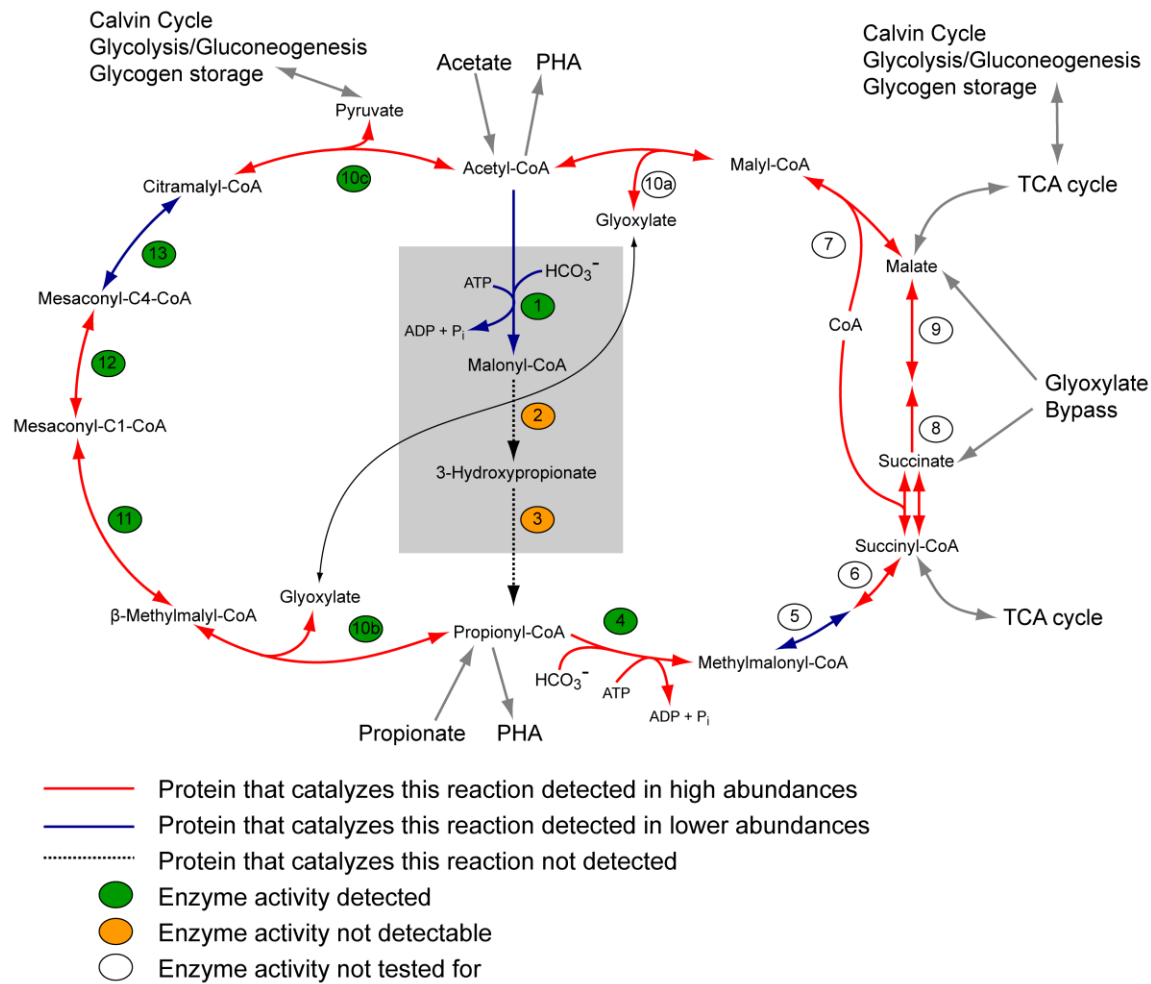


Figure 4.3: Modified version of the 3-hydroxypropionate bi-cycle (3-HPB) in the γ 1-symbiont. Reactions not needed for the assimilation of propionate and acetate are shown in the grey box; reaction 1 can also play a role in fatty acid metabolism. [1] Acetyl-CoA carboxylase (2004223475), [2] malonyl-CoA reductase, [3] propionyl-CoA synthase, [4] propionyl-CoA carboxylase (2004223080), [5] methylmalonyl-CoA epimerase (RASTannot_91923), [6] methylmalonyl-CoA mutase (RASTannot_20798), [7] succinyl-CoA:(S)-malate-CoA transferase (RASTannot_529, RASTannot_48547), [8] succinate dehydrogenase (2004223104, 2004223105), [9] fumarate hydratase (2004223692), [10a,b,c] (S)-malyl-CoA/ β -methylmalyl-CoA/(S)-citramalyl-CoA (MMC) lyase (RASTannot_91504), [11] mesaconyl-C1-CoA hydratase (β -methylmalyl-CoA dehydratase) (2004222675), [12] mesaconyl-CoA C1-C4 CoA transferase (RASTannot_38616), [13] mesaconyl-C4-CoA hydratase ((S)-citramalyl-CoA dehydratase) (RASTannot_6738)

because of the absence of the two diagnostic enzymes and the low activities of the carboxylases involved in the 3-HPB (Table S5). The pathway could also be linked to the synthesis and/or mobilization of the storage compound polyhydroxyalkanoate (PHA). A putative PHA synthase (2004222379) and a phasin protein (PHA granule protein, 6frame_RASTannot_14528) are highly expressed in the γ 1-symbiont metaproteome, showing the importance of PHA synthesis for this symbiont. Under anaerobic conditions, PHA synthesis would not only produce a valuable storage compound, but also relieve the symbiont of superfluous reducing equivalents.

Intriguingly, one of the closest free-living relatives of the γ 1-symbiont - *Allochromatium vinosum* - whose genome was recently sequenced, does not possess the genes needed for the 3-HPB or its modified version (<http://genome.jgi-psf.org/allvi/allvi.home.html>). This suggests that the genes for this pathway were gained through lateral transfer. Certainly, there is a strong selective advantage for this pathway in the γ 1-symbionts. The γ 1-symbionts are present in almost all gutless oligochaete species and therefore assumed to be the ancient primary symbionts that first established a mutualistic relationship with the oligochaetes (144). The ability to recycle organic host waste would have been a considerable advantage during the early stages of the symbiosis before the establishment of associations with other bacteria such as the heterotrophic sulfate-reducing symbionts.

Uptake and recycling of nitrogenous compounds

Since sources of nitrogen are extremely limited in the habitat of *O. algarvensis* (184), efficient strategies for dealing with nitrogen limitation represent a selective

advantage. Our metaproteomic and metabolomic analyses of the *O. algarvensis* association indicate two major strategies for dealing with nitrogen limitation: (i) the use of high affinity systems for the uptake of nitrogenous compounds from the environment and (ii) conservation of nitrogen within the symbiosis through recycling.

Environmental nitrogen is most likely assimilated by the symbionts using glutamine synthetases as well as high affinity uptake transporters. The $\gamma 1$ -, $\gamma 3$ - and the $\delta 1$ -symbiont abundantly expressed glutamine synthetases (Dataset S2). This enzyme assimilates ammonia into glutamine with high affinity at very low ammonia concentrations and is only expressed in cultured organisms under low nitrogen conditions (185, 186). Uptake of organic compounds from the environment is presumably a further source of nitrogen, given the abundant expression of high affinity amino acid and peptide uptake transporters in the $\delta 1$ -symbiont that enable these to acquire nitrogen-containing substrates at extremely low concentrations.

The second proposed strategy of the *O. algarvensis* association for dealing with low nitrogen availability is the internal recycling of nitrogenous host osmolytes and waste products by the symbionts. In many invertebrates, these compounds are removed through excretory organs called nephridia. Gutless oligochaetes are the only known annelid worms without nephridia and their reduction indicates that their symbionts have taken over the role of waste and osmolyte management. Our metabolomic analyses revealed high concentrations of two nitrogenous osmolyte and waste compounds in *O. algarvensis*, glycine betaine and urea (Table S2), with glycine betaine the most abundant metabolite detected in NMR measurements (~60 mM) (Fig. S7). Glycine betaine is a well-known osmolyte in all kingdoms of life (187) and most likely also serves this

function in *O. algarvensis*. The relatively high amounts of urea in *O. algarvensis* are unusual, as this nitrogenous waste compound and osmolyte is not commonly found in aquatic animals (187). The *O. algarvensis* symbionts abundantly expressed proteins for glycine betaine and urea uptake and for the pathways required to use them as carbon and nitrogen sources (Fig. 4.2, SI Text).

Energy conservation with proton-translocating pyrophosphatases

We propose several novel pathways for energy conservation in the *O. algarvensis* symbiosis. Both the γ -symbionts and the δ 1-symbiont expressed pyrophosphate-dependent enzymes that could conserve energy in as yet undescribed modifications of classical metabolic pathways. Our analyses of published genomes indicate that these pathways may be common in sulfate reducers and chemoautotrophic bacteria.

The key enzyme for the proposed energy conservation pathways is a membrane bound proton-translocating pyrophosphatase (H^+ -PPase), which was abundantly expressed in both the γ - and the δ 1-symbionts (SI Text). H^+ -PPases are widespread in all three domains of life. Despite their pervasiveness, remarkably little is known about the metabolic pathways in which they are used (188). H^+ -PPases are proton pumps that use the hydrolysis of inorganic pyrophosphate (PPi) instead of ATP to generate a proton-motive force through translocation of protons across biological membranes (Fig. 4.4).

They can also work reversibly as proton-translocating pyrophosphate synthases (H^+ -PPi synthase) and produce PPi using a proton-motive force (188).

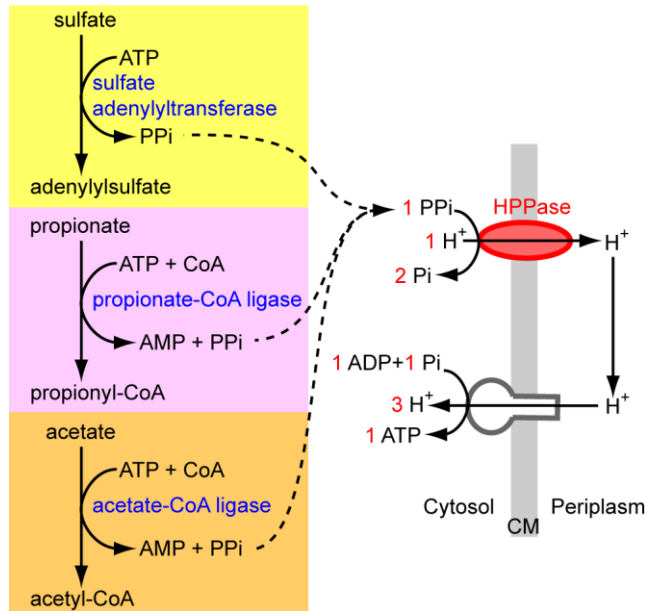


Figure 4.4: Suggested role of the proton-translocating pyrophosphatase (H⁺-PPase) in the $\delta 1$ -symbiont. Energy is conserved through the use of a membrane bound proton-translocating pyrophosphatase instead of a cytosolic pyrophosphatase. Inorganic pyrophosphate (PPi) is produced by abundantly expressed enzymes, which catalyze the initial steps of sulfate reduction, propionate oxidation and acetate oxidation. Red numbers show the stoichiometry.

H⁺-PPase energy conservation in sulfate reducers

Sulfate-reducing bacteria produce large amounts of PPi as a by-product of the first step of sulfate reduction (Fig. 4.4). This PPi has to be immediately removed to pull the reaction in the direction of sulfate reduction (189). For most sulfate reducers the mechanism of PPi removal is unknown. In some it occurs through a 'wasteful' hydrolysis of PPi by a soluble inorganic pyrophosphatase (190). In others the energy from PPi hydrolysis may be conserved with a H⁺-PPase (191), but to date this has not yet been proven. Our metaproteomic analyses provide support for the conclusion that the sulfate-reducing $\delta 1$ -symbiont uses the H⁺-PPase to conserve energy from PPi based on the

abundant expression of a H^+ -PPase and the absence of a soluble pyrophosphatase. The stoichiometry of the H^+ -PPase yields one ATP per hydrolysis of three PPi (192), which would provide the $\delta 1$ -symbiont with a considerable energy gain of one additional ATP per three molecules of sulfate reduced.

Other sources of PPi besides sulfate reduction also appear to play an important role in the metabolism of the $\delta 1$ -symbiont. In addition to expressing PPi-producing enzymes found in all organisms such as aminoacyl-tRNA synthetases, RNA and DNA polymerases, the $\delta 1$ -symbiont abundantly expressed at least two other PPi-producing enzymes - the acetate-CoA ligase (2004210485) and the propionate-CoA ligase (2004210481). We therefore postulate, based on the abundant expression of numerous PPi-producing enzymes in the $\delta 1$ -symbiont, that H^+ -PPase plays a key role in energy conservation in its metabolism (Fig. 4.4 and S3a).

To examine how widespread H^+ -PPases are in sulfate reducers, we analyzed the genomes of sulfate reducers available in the databases. These revealed H^+ -PPases in several sulfate reducers from two bacterial divisions, *Desulfatibacillum alkenivorans* AK-01 and *Desulfococcus oleovorans* Hxd3 from the Deltaproteobacteria, and *Candidatus Desulforudis audaxviator* MP104C and *Desulfotomaculum reducens* MI-1 from the division Clostridia. This suggests that the use of H^+ -PPases for energy conservation may be widely distributed among phylogenetically diverse sulfate-reducing bacteria.

Energy-efficient PPi-dependent pathways in sulfur oxidizers

We propose that the γ -symbionts use novel energy-saving modifications of the Calvin cycle, glycolysis and gluconeogenesis pathways. The key enzymes for the

proposed modifications are the H⁺-PPase and a closely coupled PPI-dependent 6-phosphofructokinase (PPI-PFK). We show that these enzymes could save as much as 30% energy over ATP-dependent pathways and that this energy saving pathway may be widespread in chemoautotrophic bacteria.

Metagenomic analyses of the *O. algarvensis* consortium showed that the γ 1-symbiont lacks two key enzymes of the classical Calvin cycle, fructose-1,6-bisphosphatase and sedoheptulose-1,7-bisphosphatase (the γ 3-symbiont lacks only the latter) (Fig. 4.5c). Interestingly, the chemoautotrophic symbionts of the hydrothermal vent tubeworm *Riftia pachytila* and the vesicomid clams *Calymene magnifica* and *C. okutanii* also lack the genes for these two enzymes, even though all of them fix CO₂ via the Calvin cycle (193-195). For the *C. magnifica* symbiont, Newton et al. (2008) (194) hypothesized that a PPI-PFK might replace fructose-1,6-bisphosphatase, but no enzyme was found that could replace sedoheptulose-1,7-bisphosphatase. It therefore remained unclear how the Calvin cycle could function in these chemoautotrophic symbionts.

We found that both γ -symbionts of *O. algarvensis* possess a gene for a PPI-PFK that is highly similar to that of the methane-oxidizer *Methylococcus capsulatus* (amino acid identities, γ 1: 71%; γ 3: 69%). The *M. capsulatus* PPI-PFK catalyzes three reactions: i) the reversible, phosphate dependent transformation of fructose-1,6-bisphosphate to fructose-6-phosphate and PPI, ii) the reversible, phosphate dependent transformation of sedoheptulose-1,7-bisphosphate to sedoheptulose-7-phosphate and PPI, and iii) the PPI dependent phosphorylation of ribulose-5-phosphate to ribulose-1,5-bisphosphate. Thus, PPI-PFK can replace the enzymes involved in these three reactions: fructose-1,6-

bisphosphatase, sedoheptulose-1,7-bisphosphatase and phosphoribulokinase (196) (SI Text, Fig. 4.5b and c). The P_{Pi}-PFK was abundantly expressed in the γ 1-symbiont (the low coverage of the γ -3-symbiont proteome could explain why it was not detected in this symbiont). We propose that in the *O. algarvensis* γ -symbionts and possibly other chemoautotrophs (see below), the P_{Pi}-PFK has multiple functions in the Calvin Cycle, glycolysis and gluconeogenesis, and that this leads to considerable energy savings as described below (Fig. 4.5a and b).

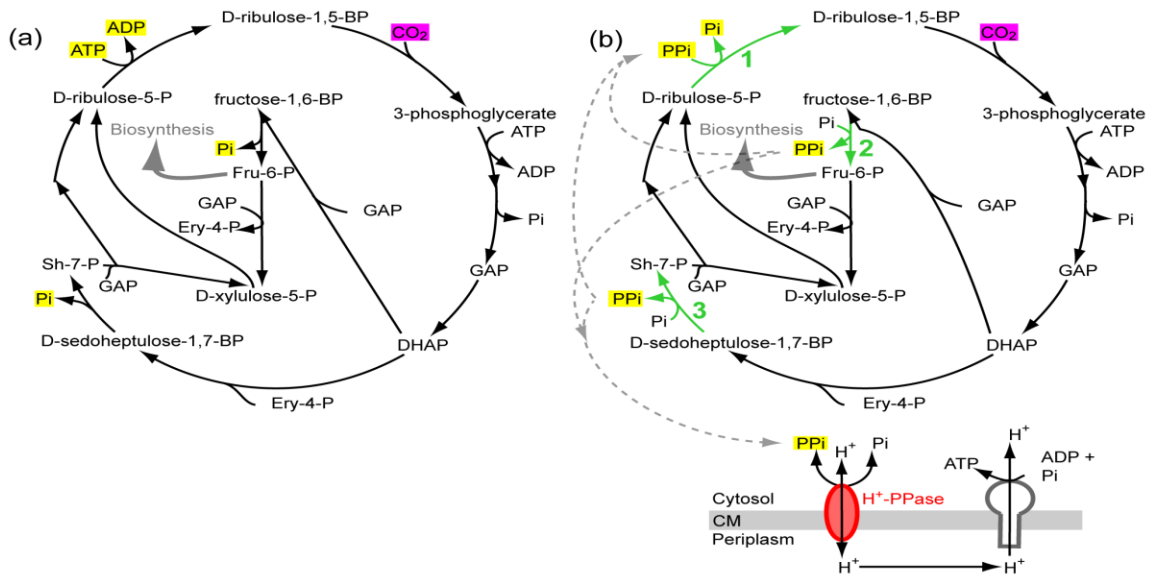
In the classical Calvin cycle, the reactions catalyzed by fructose-1,6-bisphosphatase and sedoheptulose-1,7-bisphosphatase produce phosphate ions that cannot be used for energy gain. In contrast, if P_{Pi}-PFK replaces these enzymes, energy-rich pyrophosphates are produced in both reactions. Interestingly, in the genomes of both γ -symbionts, the genes for P_{Pi}-PFK are located in the immediate neighborhood of H⁺-PPases, indicating a close metabolic relationship between these two enzymes and their co-transcription (Fig. 5d), as shown for *M. capsulatus*, in which these genes also co-occur (Fig. 5d) (196). We propose that the pyrophosphate produced by the P_{Pi}-PFK in the Calvin cycle is used to conserve energy via the proton-motive force generated by the H⁺-PPase (Fig. 5b). This metabolic coupling between the P_{Pi}-PFK and H⁺-PPase would lead to energy savings of at least 9.25% (1 ²/₃ ATP less per 6 molecules of fixed CO₂ in comparison to the 'classical' Calvin cycle in which 18 ATP are used for the fixation of six CO₂). An even higher energy gain of 31.5% is possible if P_{Pi}-PFK also replaces ATP-dependent phosphoribulokinase in the last step of the Calvin cycle: the conversion of ribulose-5-phosphate to ribulose-1,5-bisphosphate could be energized with P_{Pi} from the

two other Calvin cycle reactions and/or the H⁺-PPase working in PPi synthesis direction, so that a total of 5 ²/₃ ATP (31.5%) would be saved per 6 CO₂ fixed.

In addition to their proposed role in the Calvin cycle, we hypothesize that the PPi-PFK and H⁺-PPase also provide considerable energy savings in glycolysis and gluconeogenesis as well as through several additional enzymes (SI Text). We thus conclude that PPi-PFK and H⁺-PPase could play a key role in energy conservation in the γ 1-symbiont, and most likely also in the γ 3-symbiont.

Widespread occurrence of co-localized H⁺-PPase/PPi-PFK genes in chemoautotrophic bacteria

To examine if other microorganisms could also use the PPi-PFK and H⁺-PPase for the pathways we propose above, we analyzed all bacterial [1354] and archaeal [58] genomes available in the NCBI genomic database on January 29th 2009 (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi). We discovered co-localized H⁺-PPase/PPi-PFK genes, indicating close metabolic coupling and co-transcription, in the chemoautotrophic sulfur-oxidizing symbionts of *C. magnifica* and *C. okutanii* as well as in 8 free-living bacterial species (Gamma- and Betaproteobacteria and Thermotogae), all of which possess ribulose-1,5-bisphosphate carboxylase/oxygenase genes for autotrophic CO₂ fixation (Fig.4.5d). This broad distribution of co-localized H⁺-PPase/PPi-PFK genes in bacteria for which genomes are available suggests that H⁺-PPase/PPi-PFK dependent pathways for energy conservation are widespread in both symbiotic and free-living chemoautotrophic bacteria.



No. in (b)	Enzyme	Pathway	γ 1-symbiont		γ 3-symbiont		M. c.*	E. p.*	R. m.*	V. o.*
			Gene	Proteome	Gene	Proteome				
1, 2 and 3	PPI-PFK (PPI dependent phosphofructokinase)	Glycolysis / Calvin cycle / Gluconeogenesis	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
2	Fructose-1,6-bisphosphatase	Gluconeogenesis / Calvin cycle	No	No	Yes	No	No	No	No	No
3	Sedoheptulose-1,7-bisphosphatase	Calvin cycle	No	No	No	No	No	No	No	No
1	Phosphoribulokinase	Calvin cycle	No	Yes**	Yes	Yes	Yes	Yes	Yes	Yes
Not shown	ATP dependent phosphofructokinase	Glycolysis	No	No	No	No	No	No	No	No

* M. c. *Methylococcus capsulatus*, E. p. *Riftia pachyptila* symbiont, R. m. *Calyptogena magnifica* symbiont, V. o. C. *okutanii* symbiont
 ** Detected with gene of *Thiobacillus denitrificans*

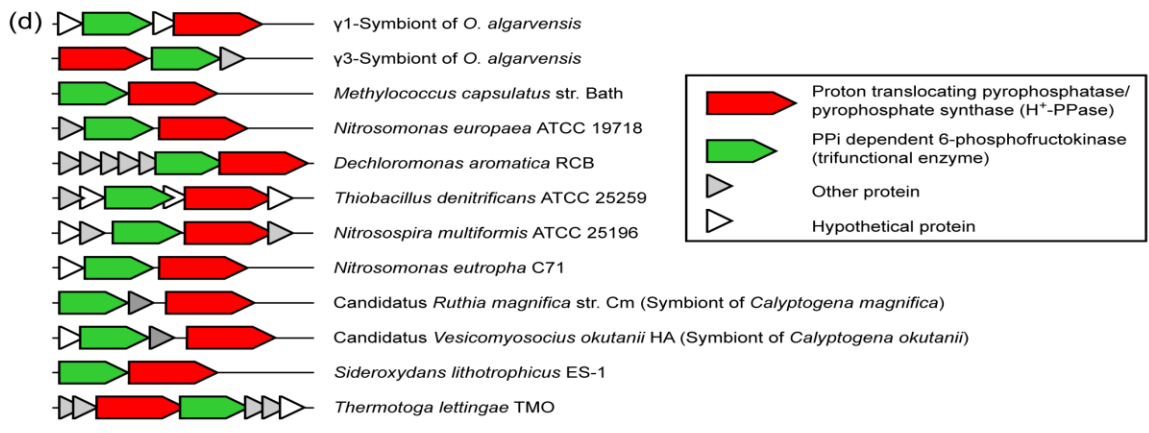


Fig. 4.5: Comparison of the “classical” Calvin cycle with a proposed more energy efficient version. (a) The text book version of the Calvin cycle. (b) More energy-efficient version of the Calvin cycle in the gamma1-symbionts through the use of a pyrophosphate (PPI)-dependent trifunctional 6-phosphofructokinase/sedoheptulose-1,7-bisphosphatase/phosphoribulokinase (green arrows) and a proton-translocating pyrophosphatase/proton-translocating pyrophosphate synthase (H^+ -PPase/ H^+ -PPI synthase in red). The main differences between the cycles are highlighted in yellow. CM, cell membrane; DHAP, dihydroxyacetone phosphate; GAP, D-glyceraldehyde-3-

phosphate; PPi, inorganic pyrophosphate; Sh-7-P, D-sedoheptulose-7-phosphate. (c) Overview of genes that are replaced by the trifunctional PPi-dependent enzyme in different organisms (d) Co-localized H⁺-PPase/PPi-PFK genes in the γ -symbionts and other symbiotic as well as free-living bacteria.

The discovery of these pathways in chemoautotrophic bacteria is particularly interesting in light of evidence that H⁺-PPases may have an ancient origin (188). H⁺-PPase is the simplest known primary proton pump. It is the only known alternative to ATP synthases for the production of energy-rich phosphoanhydride bonds, and the only primary pump that is preserved in all three domains of life (188, 197). Given mounting evidence that the earliest forms of life were chemoautotrophic (198), the apparent pervasiveness of energy conserving H⁺-PPase pathways in chemoautotrophs adds further weight to the hypothesis that PPi preceded ATP as the central energy carrier in the early evolution of life (197, 199, 200).

Conclusions

Our metaproteomic and metabolomic analyses of the *O. algarvensis* symbiosis provide strong indirect evidence for a number of novel and unexpected metabolic pathways and strategies that were not identified in the metagenomic analysis of the symbiotic consortium (148). We gained further functional insights by using proteomics-based binning. This method allowed us to include an additional 9 Mb of sequences in our analyses that could not be mined for genomic information by Woyke et al. (2006) (148), because their lengths were too short to enable a clear assignment to a specific symbiont. One of the key questions in the metagenomic analyses of complex symbiotic consortia including those of the human gut is why there is so much functional redundancy (201,

202). The selective advantage for *O. algarvensis* of harboring two sulfur-oxidizing γ -symbionts with apparent functional redundancy was so far not clear. Our study shows that the only physiological traits shared by these two symbionts are their common use of reduced sulfur and carbon fixation via the Calvin cycle. Otherwise, they show very marked differences in their use of additional energy and carbon sources as well as electron acceptors. The γ 3-symbionts most likely use carbon monoxide and the host-derived osmolyte glycine betaine as additional energy and carbon sources, while the γ 1-symbionts may use fermentative waste products from their hosts as additional carbon sources. Furthermore, the γ 1-symbionts appear to rely heavily on storage compounds such as sulfur and polyhydroxyalkanoates, while storage compounds do not appear to play a dominant role in the metabolism of the γ 3-symbionts. Resource partitioning is also visible in the differences in electron acceptors used by the two symbionts. The γ 1-symbionts depend predominantly on oxygen for their respiration while the γ 3-symbionts are not able to use this electron acceptor and instead use the energetically less favorable nitrate (SI Text). Our metaproteomic analyses thus indicate functional differences in the metabolism of these two symbionts despite their genetic similarities in key metabolic pathways for chemosynthesis. This appears to be a common theme in microbial communities, as several recent proteomic and metaproteomic studies have shown that ecological differences between microorganisms with similar genomes are due to major differences in their protein expression (203-205).

While resource partitioning provides the association versatility and the ability to harvest a wide spectrum of energy and carbon sources, in one key aspect all four symbionts appear to share a remarkably similar metabolic strategy. They all express

proteins involved in highly efficient pathways for the uptake, recycling, and conservation of energy and carbon sources. These include (i) multiple strategies for the recycling of host waste products, (ii) the possible use of additional inorganic energy sources besides reduced sulfur compounds, such as hydrogen and carbon monoxide, (iii) the extremely abundant expression of high-affinity uptake transporters that would allow the uptake of a wide range of substrates at very low concentrations and (iv) novel energy efficient steps in sulfate reduction and CO₂ fixation. Given the oligotrophic, nutrient-poor nature of the worm's environment in which organic compounds were below detection limits and reduced sulfur compounds barely detectable, the selective pressure for metabolic pathways that maximize energy and carbon acquisition and conservation appears to have been very strong in the shaping of these symbioses.

CHAPTER FIVE

Abundant Transposase Expression in Mutualistic Endosymbionts is Revealed by Metaproteomics

Text is adapted from: *Kleiner M, *Young, JC., Shah, M, Verberkmoes, NC., and Dubilier, N. Metaproteomics Reveals Abundant Transposase Expression in Mutualistic Endosymbionts. Submitted to *mBio* in October, 2012. (*these authors contributed equally to this work) (206).

Jacque Young's contributions included performing all nano-2D-LC-MS/MS runs, data analysis, manuscript writing and editing.

Abstract

Transposases, enzymes that catalyze the movement of mobile genetic elements, are the most abundant genes in nature. While many bacteria encode a wealth of transposases in their genomes, the current paradigm is that transposase gene expression is tightly regulated and generally low due to its severe mutagenic effects. In the current study, we detected the highest number of transposase proteins ever reported in bacteria, in symbionts of the gutless marine worm *Olavius algarvensis* using metaproteomics. At least 26 different transposases from 12 different families were detected and genomic and proteomic analyses suggest many of these are active. This high expression of transposases suggests the mechanisms for their tight regulation may have been disabled or destroyed.

Introduction

The expansion of transposable elements (TE) within genomes of host-restricted symbionts and pathogens plays an important role in their emergence and evolution, and might be a key mechanism for adaptation to the host environment. However, little is known so far about the underlying causes and evolutionary mechanisms of this TE expansion. The current model of genome evolution in host-restricted bacteria explains TE expansion within the confines of the paradigm that transposase expression is always low. However, recent work by Plague *et al.* (207) failed to verify this model. Our data suggests that increased transposase expression, which has not previously been described, may play a role in TE expansion, and could be one explanation for the sometimes very rapid emergence and evolution of new obligate symbionts and pathogens from facultative ones.

Transposases are enzymes that catalyze the movement of mobile genetic elements in and between genomes, and are the most abundant and ubiquitous genes in nature (208). Most often transposases are part of transposable elements, which only encode the transposase gene and some short flanking sequences necessary for transposition: these basic transposable elements are called insertion sequence (IS) elements. Classically, transposable elements are considered to be selfish genetic elements or parasitic DNA with no other purpose than reproducing themselves (38, 209, 210). However, in more recent years it has become clear that transposable elements are not always parasites, but can also have beneficial effects increasing host fitness (For reviews of the ongoing debate see (2, 38, 210)). Transposable elements (especially IS elements) are involved in gene deletions, gene duplications, genome rearrangements, and horizontal gene transfer (for

reviews see (211) and (212)), all of which can have beneficial effects on the host population by generating genomic diversity and thus enabling adaptation to environmental changes (213-215). However, transposable elements can also be detrimental if they disrupt important functional genes. Therefore, transposase expression and thus transpositional activity are usually very low (84), because the mutagenic effects of transposases would drive their hosts into extinction, thereby also eradicating their own existence (216). Accordingly, a large variety of mechanisms for the tight regulation of transposase expression exist both at the transcriptional and translational level (84).

Transposable elements (TE) and thus transposase genes are particularly enriched in the genomes of some pathogen and mutualistic symbionts (just called ‘symbionts’ in the following) that have recently transitioned to an obligate host-associated lifestyle (2), and it has been shown that this TE expansion plays a crucial role in the emergence and early evolution of new pathogens (213, 217-219) and mutualistic symbionts (220, 221). Currently there is much uncertainty about the factors that lead to high TE loads in host-restricted bacteria, however several hypotheses have been put forth (reviewed in (38)). The two main ones, which represent opposing views, are: (i) TE expansion is beneficial for symbionts and pathogens transitioning to an obligate lifestyle, for example, by providing enhanced genomic plasticity for faster adaptation to the host environment (213, 214, 222). (ii) Temporary TE expansion in the genomes of host restricted bacteria is due to a reduced effectiveness of natural selection against deleterious transpositions (2). The relaxed natural selection according to this hypothesis is caused by genetic drift due to small population sizes and transmission bottlenecks and the fact that in the host many genes become superfluous and thus can act as neutral integration sites for TEs. A recent

study that tested this hypothesis by subjecting *Escherichia coli* for 4000 generations to simulated conditions of relaxed natural selection raised doubts if relaxed natural selection alone can account for TE expansion (222) because no TE expansion occurred under the tested conditions. Based on their negative result Plague et al. (222) hypothesized that other factors including increased transposase activity might be necessary to allow for the massive TE expansion observed in host restricted bacteria. These two main hypotheses explain TE expansion in host-restricted bacteria within the confines of the paradigm that transposase expression and thus transpositional activity is always low.

The symbionts of the gutless marine worm *Olavius algarvensis* possess high numbers of transposase genes in their genomes. *O. algarvensis* inhabits shallow water sediments in the Mediterranean and lacks both a digestive and an excretory system, relying instead for nutrition and waste recycling on a symbiotic community of two gammaproteobacterial sulfur oxidizers (γ 1- and γ 3-symbiont), two deltaproteobacterial sulfate reducers (δ 1- and δ 4-symbiont) and a spirochaete (148, 223). A previous metagenomic analysis of the *O. algarvensis* showed that the γ 1-symbiont has a remarkably high number of transposases in its genome at nearly 21% of all genes, followed by the γ 3-symbiont with 7.5% and the δ 1-symbiont with 2.3% (148, 149). In the current study, we demonstrate that the γ 1- and δ 1-symbionts express a surprisingly high number of these transposase proteins, and many of these are intact and possibly active.

Materials and Methods

Worms collection, symbiont enrichment, and protein identification via two-dimensional liquid chromatography followed by tandem mass spectrometry on a hybrid

linear ion trap-Orbitrap (Thermo Fischer Scientific) are described in detail in chapter four and Kleiner et al. (139). Protein databases, peptide and protein identifications, as well as all MS/MS spectra are available at http://compbio.ornl.gov/olavius_algarvensis_symbiont_metaproteome.

Results and Discussion

Transposase abundance in *O. algarvensis* symbionts

We detected the highest number of transposase proteins ever reported in bacteria in the γ 1-symbiont and the δ 1-symbiont of *O. algarvensis*. The two symbionts expressed at least 26 different transposase proteins (Table 5.1) with the majority originating from the γ 1-symbiont (22 proteins). The remaining ones originated from the δ 1-symbiont (2 proteins) or were identified with unassigned metagenome fragments (2 proteins). Transposases comprised up to 1.95% of the total protein expressed by the γ 1-symbiont (Table 5.2) and up to 0.084% of the total δ 1-symbiont protein (Table 5.3). The abundance of transposases in the γ 1-symbiont is comparable to some of its most abundant housekeeping genes such as the ATPase B subunit (1.15%), the malate dehydrogenase (0.38%) and the 6-phosphofructokinase (0.35%) (139). Within the context of natural microbial communities, highly abundant transposase protein expression has, so far, only been reported from a microbial biofilm found in acid mine drainage, however, relative abundances were less than half of the amounts observed in this study (28).

Transposase genes are present in multiple, nearly identical, copies in the symbiont metagenomes (139, 224). We detected peptides belonging to transposase proteins encoded by 134 gene sequences in the metagenome (Table S3). Given that proteins

encoded by almost identical sequences with identical tryptic peptides cannot be distinguished from each other by mass spectrometry based proteomic analyses, we could not identify which of the 134 transposase genes were the expressed ones and if multiple copies of identical genes were expressed. Therefore, we assigned the transposases that were identified with similar sets of peptides to 12 different groups (Tables 5.1 and S4) and used this grouping to identify the minimal set of transposases that need to be expressed to explain all transposase related peptides. This non-redundant set consisted of the above-mentioned 26 transposases.

We classified the 26 non-redundant transposases into IS element families using BLASTp against the curated IS-finder database (225) (<http://www-is.biotoul.fr/>) and found that they belong to at least 10 different IS-element families in the γ 1-symbionts and 2 families in the δ 1-symbiont (Table 5.1). This clearly shows that the expressed transposase genes originated from multiple unrelated IS elements.

Table 5.1: Overview of all expressed transposases grouped according to shared peptide matches

Accession number ⁶	Symbiont	Transposase group	Transposases/Group	Transposases/group required to explain all peptide matches	Transposase classification	Intact	Premature translation abortion	Length (amino acids)
2004221906	γ 1	1	38	4	IS630	No	No	304
2004222763	γ 1	1			IS630	No	No	169
2004222856	γ 1	1			IS630	Yes	No	345
2004222872	γ 1	1			IS630	No	No	268
2004223397	γ 1	2	19	3	IS110	Yes	No	355
2004222027	γ 1	2			IS110	No	No	263
2004222937	γ 1	2			IS110	No	No	205
Symbiont_37746	Unknown	3	7	1	Unclassified	Unknown	?	163
2004223511	γ 1	4	1	1	Unclassified	Yes	?	344
Symbiont_28062	Unknown	5	1	1	IS1634	Yes	No	440
2004221868	γ 1	6	37	7	IS481	Yes	No	419
2004221830	γ 1	6			IS3	No	Yes	241
2004222310	γ 1	6			IS481	Yes	No	333
2004222544	γ 1	6			IS481	No	No	238
2004222874	γ 1	6			IS481	No	No	115
2004223428	γ 1	6			IS5	No	No	192
2004223468	γ 1	6			IS481	No	No	227
2004207437	δ 1	7	2	1	IS4	Yes	No	454
2004223208	γ 1	8	1	1	IS1595	Yes	No	282
2004212411	δ 1	9	3	1	IS21	Yes	Yes	228
2004222138	γ 1	10	16	2	IS1	Yes	Yes	245
2004222867	γ 1	10			IS1	Yes	Yes	245
2004222325	γ 1	11	7	3	IS5	Yes	No	345
2004223125	γ 1	11			IS630	No	No	134
2004223697	γ 1	11			IS630	No	No	165
2004221963	γ 1	12	3	1	Unclassified	Yes	?	313
SUM			135	26				

Table 5.2: Organism normalized NSAF values (relative abundance) for the γ 1-symbiont transposases

Minimal number of γ 1-symbiont transposase proteins needed to explain detected peptides. Normalized spectral abundance factor (NSAF) values indicate the relative abundance of the respective protein in relation to all identified γ 1-symbiont proteins.

Symbionts enriched with density gradient centrifugation				Whole worms frozen directly				
Protein Accession #	Delta1 Run2	Delta1 Run3	Gamma1 Run1	Gamma1 Run2	WholeWorm Run1	WholeWorm Run2	WholeWorm2 Run3	WholeWorm4 Run4
2004221830			0.00051	0.00070				
2004221868	0.00257	0.00152	0.00158	0.00136			0.00160	
2004221906			0.00125	0.00054	0.00191	0.00188		
2004222027			0.00087	0.00044				
2004222138			0.00035					
2004222310			0.00125	0.00165			0.00147	
2004222325			0.00074	0.00048				
2004222544			0.00306	0.00272			0.00220	
2004222763			0.00100					0.00168
2004222856	0.00298	0.00444	0.00122	0.00048	0.00167	0.00166		0.00205
2004222867			0.00051	0.00089				
2004222872			0.00116	0.00089				
2004222874			0.00109	0.00143				
2004222937	0.00339		0.00061	0.00108	0.00113			0.00342
2004223125			0.00158	0.00244		0.00171		
2004223208			0.00029					
2004223397			0.00023					
2004223428			0.00064	0.00114		0.00182		
2004223468			0.00055					
2004223511				0.00032				
2004223697			0.00103	0.00298		0.00138		
2004221963						0.00111		0.00087
Total NSAF	0.00894	0.00596	0.01953	0.01951	0.00470	0.00956	0.00527	0.00803
Relative abundance (%)	0.89370	0.59645	1.95290	1.95103	0.47032	0.95633	0.52719	0.80319

Table 5.3: Organism normalized NSAF values for the δ 1-symbiont transposases

This table only contains the δ 1-symbiont transposases that are needed to explain all δ 1-symbiont transposase related peptides according to the grouping in table S4 (non-redundant set) . The NSAF values in this table give the relative abundance of the respective protein among all identified δ 1-symbiont proteins

Symbionts enriched with density gradient centrifugation				Whole worms frozen directly				
Protein Accession #	Delta1 Run2	Delta1 Run3	Gamma1 Run1	Gamma1 Run2	Whole Worm Run1	Whole Worm Run2	Whole Worm2 Run3	Whole Worm4 Run4
2004212411	0.000564725	0.000506781						
2004207437	0.000282363	0.000265457						
Total NSAF	0.000847088	0.000772238						
Relative abundance (%)	0.084708781	0.077223804						

Could abundant transposase expression be caused by stress?

Previous studies have reported increased transposase expression in response to stressful conditions (85, 86). However, the relative abundances reported were much lower compared to those detected in our study. To exclude the possibility that transposase expression in the *O. algarvensis* symbionts was caused by stressful conditions during the one hour long symbiont enrichment procedure, we compared their proteomes with those of symbionts that were frozen in whole worms immediately following removal from the sediment. As was the case in enriched symbionts, high transposase expression was seen in the immediately frozen symbionts (Tables 5.2 and S3). Thus, we concluded that the observed transposase expression was not due to stressful sampling conditions, but actually reflects expression under environmental conditions.

We inferred that some of the expressed transposases are likely active, by excluding the two main reasons for potential inactivity: (i) transposase genes could be in the process of gene degradation and their expression thus would lead to incomplete and potentially inactive transposases and (ii) the expression of some transposase genes is regulated via programmed translational frameshifting, which can lead to the translation of a truncated, non-functional version of the transposase protein (84, 211, 226).

Thus, to test whether the expressed transposase genes are intact or in a state of gene degradation, we compared their gene sizes and sequences to closely related transposases in the IS finder database and then compared their protein domain structures with those of similar transposases using the Pfam ‘domain organization’ feature (<http://pfam.sanger.ac.uk/search>). We found that around half of the expressed transposase genes are intact, whereas the other half are in various states of gene degradation (Table 5.1). Second, we checked in the literature for which IS element families regulation via programmed translational frameshifting is known to occur (84) and found that premature translation abortion could only be shown for four out of the 26 IS families of the detected transposases (Table 5.1). Thus it is likely that the majority of the expressed transposases are translated to full-length proteins. This is supported by our proteomic data for some of the transposases which we detected peptides from the beginning, middle, and end of the protein, indicating that they were translated from start to finish (Figure 5.1) (http://compbio.ornl.gov/cgi-bin/mspipeline/seqcvg/contrastprtns_xcorr_SeqCvg.cgi?contrastdir=mspipeline/dubilier/Deep_Sea_Worms_Set2_Greifswald/analysis/Deep_Sea_Worms_Set2_Greifswald/contrast/tryp_20090729_fr/final/p2_verbose_transposase).

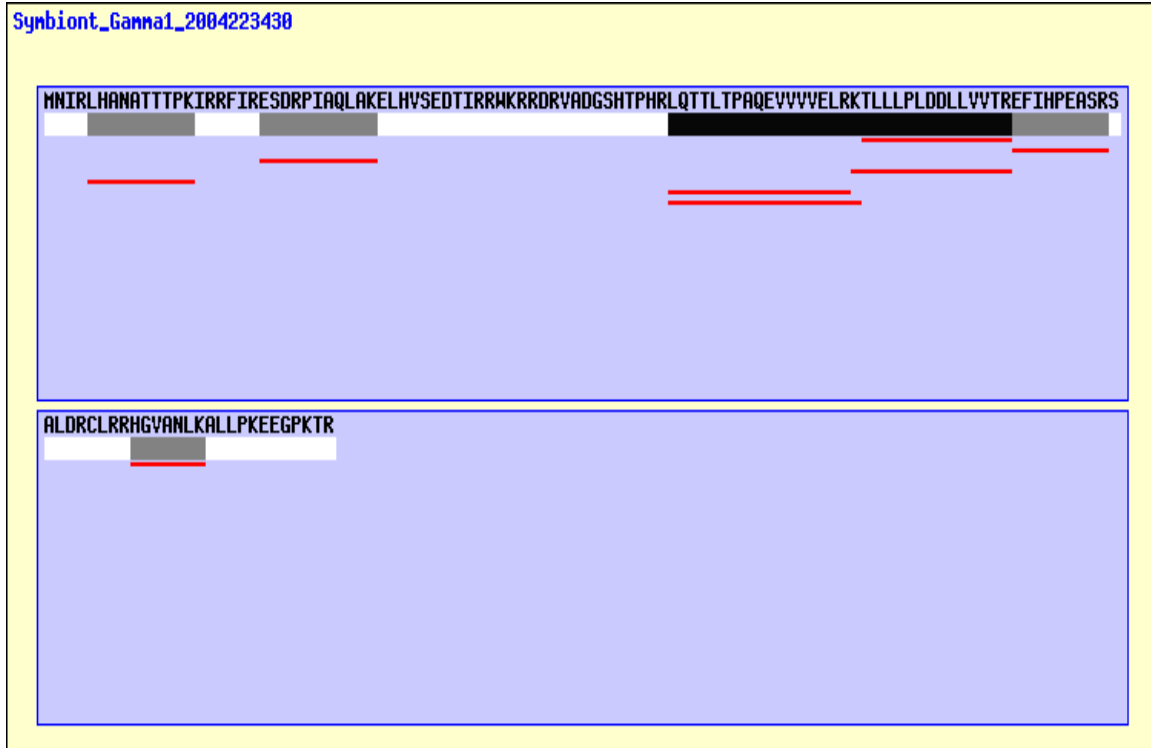


Figure 5.1: Sequence coverage of a representative transposase detected from the γ 1-symbiont. Depth of coverage is shown below the amino acid sequence (white to black, by increased depth). Red lines indicate detected peptide

Conclusions

Our results show that the paradigm that transposase expression in bacteria must be kept to a minimum to prevent the host population from going extinct does not hold true in all cases. We present the first evidence at the protein level that transposases are abundantly expressed in beneficial symbionts with high TE numbers in their genomes. This high expression of transposases indicates that the mechanisms for their tight regulation have been disabled or destroyed, for example by mutations in proteins that are involved in transposase regulation (41, 48, 221). The fixation of such mutations could be enabled by the relaxed purifying selection suggested by Moran et al. (2004) for

symbionts and pathogens that recently transitioned to an obligate host-associated lifestyle (221).

Currently, it is not possible to determine if abundant transposase protein expression is present in other symbionts and pathogens with high TE numbers, because no comparable proteomic datasets exist for these bacteria. Nevertheless, many recent studies have shown high transcription of transposase genes in symbionts and pathogens (227-232). However, the presence of these transcripts does not represent conclusive evidence for transposase expression, because it is possible that they are not translated to proteins due to the above mentioned regulatory mechanisms (84, 228, 229).

Based on the abundance of transposase proteins in the *O. algarvensis* symbionts and abundant transposase transcription in other symbionts and pathogens, we speculate that high transposase expression may be the missing factor for explaining TE expansion in host-restricted bacteria. As discussed above, an experiment that simulated conditions of relaxed natural selection failed to cause TE expansion after 4000 generations in *E. coli*. If, as we speculate, high transposase protein expression is the major cause of TE expansion, it could be that the necessary mutations affecting transposase expression had simply not occurred yet in the Plague et al. (2011) experiment. Additional studies that investigate transposase expression in pathogens and symbionts are needed, because high transposase expression could be an important factor in the sometimes very rapid emergence and evolution of new obligate symbionts and pathogens from facultative ones.

The thing that strikes us the most about the high transposase expression in the *O. algarvensis* symbionts is the question how the symbionts can function over evolutionary time periods, if, as suggested by the high transposase abundance, potentially deleterious

transpositions happen with high frequency. Currently we have no answer to this question, however, for other organisms that deal with frequent transpositions, genome rearrangements and disruptions, it has been suggested that polyploidy buffers against the detrimental effects of several factors that lead to these genome disruptions including transposable elements, introns, heat and ionizing radiation (233-236). Since polyploidy has been recently shown for several symbionts including sulfur-oxidizing symbionts of clams (237) it seems possible that it also plays a role in the *O. algarvensis* symbionts.

CHAPTER SIX

Metaproteomics Reveals Time-Dependent Functional Shifts in Microbial and Human Proteins in the Premature Infant Gut

Text was adapted from: Young JC, Pan C, Adams R, Brooks B, Banfield JF, Morowitz MJ, and Hettich RL. Metaproteomics Reveals Time-Dependent Functional Shifts in Microbial and Human Proteins in the Premature Infant Gut, *submitted to Molecular Systems Biology, November 2012*.

Jacque Young's contributions were: sample preparation, performed all MS runs, data analysis, and manuscript preparation.

Abstract

Microbial colonization of the human gastrointestinal tract plays an important role in the establishing overall health and homeostasis, however this process is incompletely understood. While prior studies have investigated the succession of gut microbiota in newborn infants at the genome level, the time-dependent functional signatures of microbial and human proteins have yet to be determined. In this study, we employed shotgun proteomics to simultaneously monitor microbial and human proteins from fecal samples obtained from a healthy preterm infant from days 7, 13, 15, 17, 18, 20, and 21 after birth. Approximately 800-3,700 proteins were detected from each run, with the microbial protein abundances and community complexity increasing over time. Changes in microbial community compositions were consistent with metagenomic data from

matched samples comprising three distinct colonization phases. Despite microbial compositional changes, overall community functions were established relatively early in development and remained stable throughout the time course. Detected human proteins consisted of those responsible for homeostatic functions, including epithelial barrier function and antimicrobial activity. Many of these proteins were expressed at relatively constant levels throughout the time course; however some neutrophil-derived proteins were increased in abundance early in the study period suggesting activation of the innate immune system. Likewise, abundances of cytoskeletal and mucin proteins increased later in the time course, suggestive of subsequent adjustment to the increased microbial load. This study provides the first elucidation of human and microbial proteins in the infant gut during early development.

Introduction

Microbial communities in the gastrointestinal tract play important roles in human health by processing essential nutrients, protecting against pathogenic bacteria, promoting angiogenesis, and regulating host immune responses (58-61). Initially sterile, the infant gastrointestinal tract assembles a microbial community of over 1,000 different symbiotic species in the first 2.5 years of life. This symbiotic relationship requires a careful balance, and it is believed that disruption of the host-microbe relationship in the gut can lead to diseases such as inflammatory bowel disease and neonatal necrotizing enterocolitis (NEC). Initial temporal colonization patterns and species distributions vary between individual infants and may be influenced by environmental exposures, delivery mode, diet, and health (71, 75). In general, the gut microbial communities of newborn

infants are far less complex than those of older children and adults. This lack of complexity provides a powerful opportunity to study the microbiota at high resolution.

Recently, the microbial compositional patterns of a healthy preterm infant during the first month of life were characterized in a metagenomic study (7). Through 16S rRNA gene-based analysis, the dominant taxa were identified, and community compositional changes revealed three distinct colonization phases. Specifically, days five through nine of the infant's life were dominated by *Leuconostoc*, *Weisella*, and *Lactococcus* species while days ten through fourteen, consisted primarily of *Pseudomonas* and *Staphylococcus*. The third phase was primarily composed of members of the *Enterobacteriaceae* family including *Citrobacter* and *Serratia* species and occurred during days fifteen through twenty-one. This pattern is consistent with dietary adjustments at days nine and fifteen and was similar to premature infants from other studies (7, 238, 239). Metagenomic sequencing followed by reconstruction and intensive curation of population genomic datasets of the dominant microbial members from days 10, 16, 18, and 21 revealed three major strains from these later time points: a *Serratia* strain (*UCISER*), an *Enterococcus* strain (*UCIENC*), and two closely related *Citrobacter* strains (*UCICITi* and *UCICITii*). Also present were plasmids *UCICITp*, *UCIENCp*, and bacteriophage *UCIENCv*. While metagenomic information provides a blueprint for possible gene products, we employed shotgun proteomics via nanospray-two dimensional liquid chromatography coupled with tandem mass spectrometry (nano-2D-LC-MS/MS) to elucidate *functional signatures* of translated gene products (i.e. proteins) from matched samples of the same preterm infant.

The use of mass spectrometry-based proteomics allows characterization and quantification of thousands of proteins within a microbial community (28, 33, 240, 241).

While early attempts using metaproteomics for the characterization of the infant microbiome demonstrated feasibility, protein identifications were limited due to insufficient genome information (76). Since, the number of sequenced microbial isolates in the human gut has increased dramatically, and more importantly, the use of community genomic sequences from matched samples has allowed confident identification of proteins at the species and strain level (29). Additionally, the advancement of high performance shotgun mass spectrometry-based proteomics has enabled a measurement depth not previously possible (9). While prior studies have focused on characterizing microbial genes and proteins, most current methodologies prohibit global analyses of microbial proteins in conjunction with human proteins. Thus far, there have been no studies to our knowledge that have investigated the composite relationships and interplay between gut microbial proteins simultaneously with human host proteins. In this study, we utilized shotgun proteomics to simultaneously monitor microbial and human proteins from a preterm infant during the first month of life.

Materials and Methods

Description of Preterm Infant: A female infant born at 28 weeks gestation due to premature rupture of membranes was delivered by cesarean section and given antibiotics for the first 7 days of life (7). Enteral feedings with breast milk were given on days 4-9, and then on days 9-13, feedings were withheld due to abdominal distension. After day 13, enteral feedings were readministered in the form of artificial infant formula. The baby also received supplemental parenteral nutrition until day 28. The baby had no major anomalies or comorbidities and was discharged to home on day of life 64. Fecal material

was collected on days 5-21 as available, with institutional approval. Metagenomic and 16S rRNA data was analyzed in a companion study from matched samples (7). Based upon sample availability, proteomic measurements were performed on fecal samples from days 7, 13, 15, 17, 18, 20, and 21 after birth.

Protein Extraction and Enzymatic Digestion of Fecal Samples: Approximately 250 μ g fecal material was boiled for five minutes in 1 ml 100 mM Tris-Cl containing 4% w/v SDS and 10 mM DTT, then underwent bead beating for 30 minutes, in order to lyse cells, and denature and reduce proteins. The supernatant was collected, boiled again, spun down (14,000 g), and precipitated with 20% trichloroacetic acid at 80°C overnight. Protein pellets were washed in ice-cold acetone, re-solubilized in 8 M urea diluted in 100mM Tris-HCl pH 8, and then sonicated using a Branson sonic disruptor in order to break up the pellet (5 minutes at 20%; 10 seconds on, 10 seconds off). Iodoacetamide (IAA) was added to block disulfide bond reformation. Proteins were quantified using Bicinchronic assay (BCA) and between 1-3 mg protein were diluted to 4 M urea in 100 mM Tris-HCl pH 8, and enzymatically digested into peptides using sequencing grade trypsin (Promega) for four hours at room temperature. Peptides were diluted to 2 M urea, a second dose of trypsin added, and digestion continued overnight. An acidic salt solution (200mM NaCl, 0.1% formic acid), was used to clean up the peptides which were then spun through a 10kDa cutoff spin column filter (VWR). Peptides were quantified by BCA assay and stored at -80°C until further use.

Nano-2D-LC-MS/MS: A 150µg peptide mixture was loaded onto a split-phase fused silica column containing reverse phase (C18) and strong cation exchange (SCX) materials. Samples were washed offline with solvent A (95% HPLC grade water, 5% Acetonitrile, 0.1% formic acid) for 30 minutes to desalt the column, followed by five gradients from 100% solvent A to 100% solvent B (70% acetonitrile, 30% HPLC grade water, 0.1% formic acid). Peptides were placed in line with a nanospray emitter (New Objective) packed with reverse phase material then separated on-line using high performance two-dimensional liquid chromatography (9, 123, 124). Peptides were eluted from the SCX resin by increasing ammonium acetate salt pulses followed by reverse phase resolution over two hour organic gradients as described previously (28, 29, 33), ionized via nanospray (200 nl/min) (Proxeon, Cambridge MA), and analyzed using an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, San Jose, CA). Technical duplicates were run for all samples. The LTQ was run in data-dependent mode with the top 10 most abundant peptides in full MS selected for MS/MS, and dynamic exclusion enabled (repeat count=1, 60 s exclusion duration). Two microscans were collected in centroid mode for both full and MS/MS scans.

Database Construction and Searching: A search database was generated from the predicted protein sequences of dominant members reconstructed from metagenomic sequences collected on days 10, 16, 18, and 21 from matched samples. These included a *Serratia* species *UCISER*, two closely related *Citrobacter* strains, *UCIi* and *UCIii*, an *Enterococcus* species *UCIENC*, and associated virus and plasmids *UCIENCp*, *UCIENCv*, and *UCICITp*. Since samples from early time points were not represented in

the metagenomic sequences, additional isolate sequences were included in the search database. Using 16S rRNA information, closely related species were chosen as representative organisms including: *Arcobacter butzleri* RM4018, *Acinetobacter junii* SH205, *Bacteroides fragilis* NCTC 9343, *Bifidobacterium adolescentis* ATCC 15703, *Bifidobacterium longum infantis* ATCC 15697, *Campylobacter concisus* 13826, *Clostridium sporogenes* ATCC 15579, *Enterobacter cancerogenus* ATCC 35316, *Escherichia coli* K12 DH10B, *Eubacterium rectale* ATCC 33656, *Fusobacterium* sp. 1_1_41FAA, *Klebsiella* sp. 1_1_55, *Lactococcus lactis* subsp. *lactis* KF147, *Lactobacillus reuteri* 100-23, *Leuconostoc mesenteroides cremoris* ATCC 19254, *Pseudomonas aeruginosa* PAO1, *Staphylococcus aureus* 04-02981, *Streptococcus* sp. 2_1_36FAA, *Weissella paramesenteroides* ATCC 33313. (acquired from JGI: http://www.hmpdacc-resources.org/cgi-bin/img_hmp/main.cgi). In addition, human protein sequences (NCBI RefSeq_2011) and common contaminants (i.e. trypsin) were appended to the database (Table S2). All MS/MS spectra were searched against the concatenated database with the SEQUEST algorithm (97), and filtered with DTASelect version 1.9 (6) at the peptide level with standard filters [SEQUEST Xcorr_s of at least 1.8 (+1), 2.5 (+2) 3.5 (+3)] organizing identified peptides to their corresponding protein sequences (Table 1). Due to carbamidomethylation effects of IAA, a static cysteine modification (+57) was included in all searches. Only proteins identified with two fully tryptic peptides were considered for further biological study. Reversed protein sequences were appended to the database in order to calculate false discovery rates (21).

Database clustering & spectral balancing: Protein sequences in the predicted protein database were clustered into protein groups based on sequence homology. Using the publically-available software, USEARCH v.5.0 (242), microbial proteins were clustered into a protein group if they shared 100% amino acid identity, and human proteins were clustered into a protein group if they contained $\geq 90\%$ amino acid similarity. These differing similarity thresholds were chosen based on the higher numbers of paralogous proteins present within the human genome, and were supported by plotting similarity thresholds ranging from 0.5-1 against the percent proteome reduction via clustering. Spectral counts were assigned, balanced, normalized, and adjusted according to methods previously described (25, 34, 243).

Statistical Analyses: Hierarchical clustering of individual proteins based on trends in abundance changes, and principal component analysis (PCA) plots were performed using JMP Genomics software (JMP Version 7. SAS Institute Inc., Cary, NC, 1989-2007). Statistical analysis of human proteins and pathways was performed using Ingenuity.

Results

Overall proteome measurement:

The fecal microbiome of a preterm infant was examined on days 7, 13, 15, 16, 17, 18, 20, and 21 after birth via nano-2D-LC-MS/MS. Up to 67,471 spectra, 15,226 peptides, and 3,732 proteins were detected per run (Table 6.1), providing deep proteomic coverage of these complex fecal samples. Technical duplicates were run for each sample with comparable reproducibility between replicates (Figure 6.1).

Table 6.1: Number of proteins, peptides, and MS/MS spectra identified across all time points.

<i>Sample Timepoint</i>	<i>Protein Identifications</i>	<i>Peptide Identifications</i>	<i>MS/MS Spectra</i>
Day 7	804	4750	32986
Day 13	1603	8581	64008
Day15	2646	10918	63093
Day 16	1373	4428	40752
Day 17	3732	15226	65593
Day 18	3410	13830	62142
Day 20	3280	15178	67471
Day 21	2936	12392	51635

Protein, peptide, and spectra values are based on non-redundant identifications from Sequest using a p2 filter

All numbers are averages of two technical replicate runs

Microbial and human proteins were simultaneously measured throughout the time course, yielding an in-depth view of complex fecal samples. Since mass spectrometry based proteomics identifies proteins by their corresponding peptide sequences, consideration must be taken when designing a proteome reference database to accurately reflect the species composition in the sample. In turn, data analysis must take into consideration the high levels of protein redundancy within and between species to avoid inflating the total number of proteins identified or misinterpretation of the biological conclusions by over-representing proteins with the same function. We have designed a proteome reference database to most accurately reflect the sample composition by including reconstructed metagenomes from matched samples from days 16, 18, and 21, along with bacterial isolates selected based on 16S rRNA data from earlier time points,

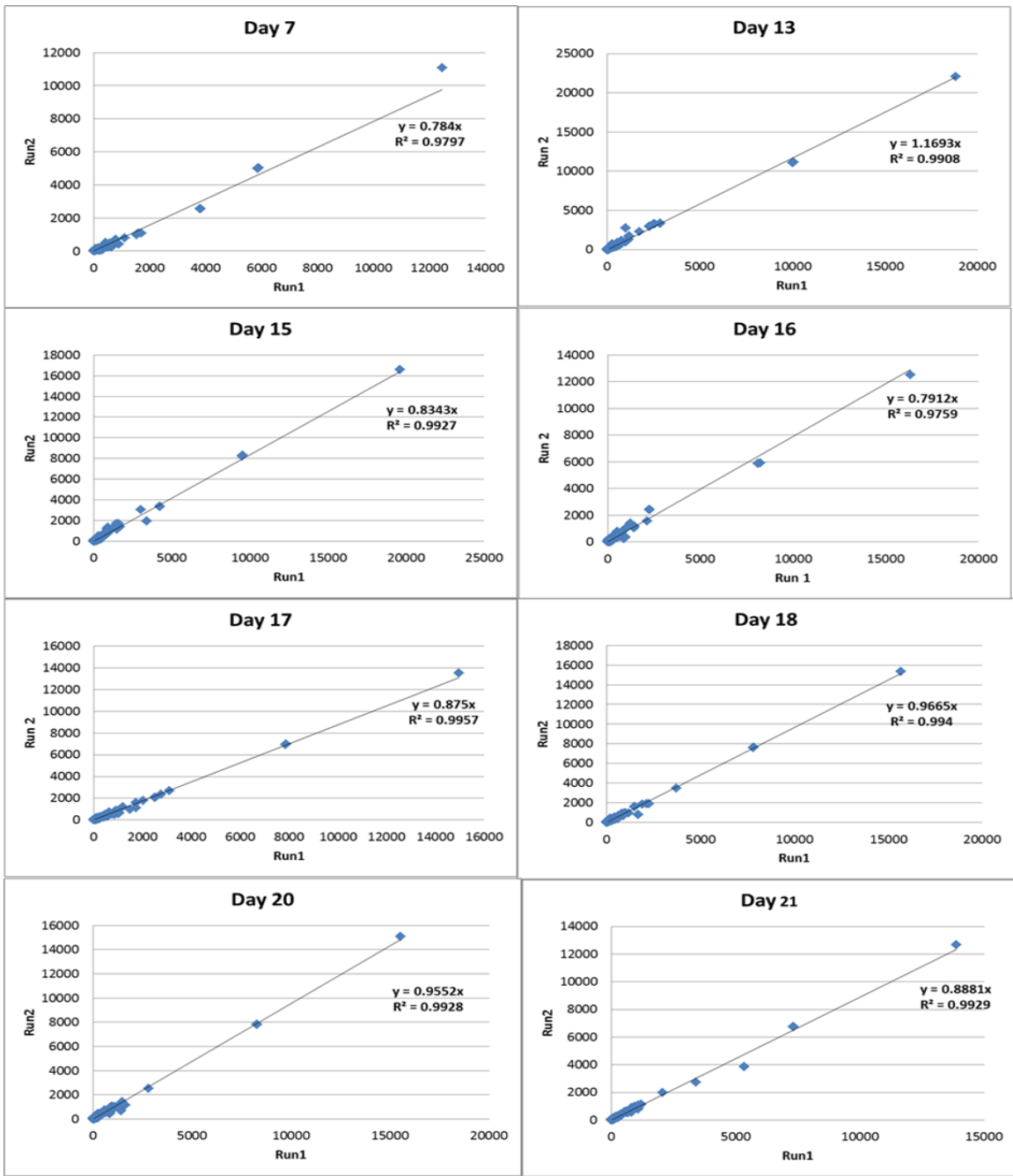


Figure 6.1: Reproducibility Between Technical Replicates. Spectral counts from two technical replicates were plotted against each other, with replicate 1 on the x-axis and replicate 2 on the y-axis. A linear regression was performed, and the slope of the line (m), and R^2 values calculated providing a statistical measure (a value between zero to one) indicating how well one term predicts another term. All values were >0.97 , confirming the technical reproducibility across replicates.

and the human genome. We then applied a bioinformatic clustering algorithm to the database in order to improve confidence in protein identification and quantification.

Proteins were clustered into groups based on 100% amino acid similarity for microbial proteins and 90% for human proteins. Different similarity thresholds were chosen to reflect the higher level of redundancy in the human genome due to gene duplications, splice variants, and multiple protein isoforms. Microbial proteins were clustered using more stringent criteria in order to preserve species information and distinguish functional contributions of different community members. In total, 4,413 microbial and 3,062 human protein groups were detected across the dataset. Protein groups range from singletons to groups that contain multiple protein isoforms.

Both microbial and human proteins were measured simultaneously in each run, revealing an increased complexity of the microbial composition and a decrease in the ratio of total human/microbial proteins with time (Figure 6.2). At the earliest time points, when the initial microbial communities were being established, human proteins comprised ~96% of the total proteins identified on day 7, possibly reflecting the administration of antibiotics for the first week of life, and ~72% on day 13. By day 15, the percent of human proteins decreased to ~30% due to an increase in the number of microbial proteins detected. The ratio of human to microbial proteins remained at this level for the remainder of the times measured, with the exception of day 20, when an unexplained rise in human proteins occurred (as detailed below). The number of total spectra collected on day 20 was comparable to adjacent days (Figure 6.3), so the variance was likely not due to a technical issue related to the mass spectrometry run.

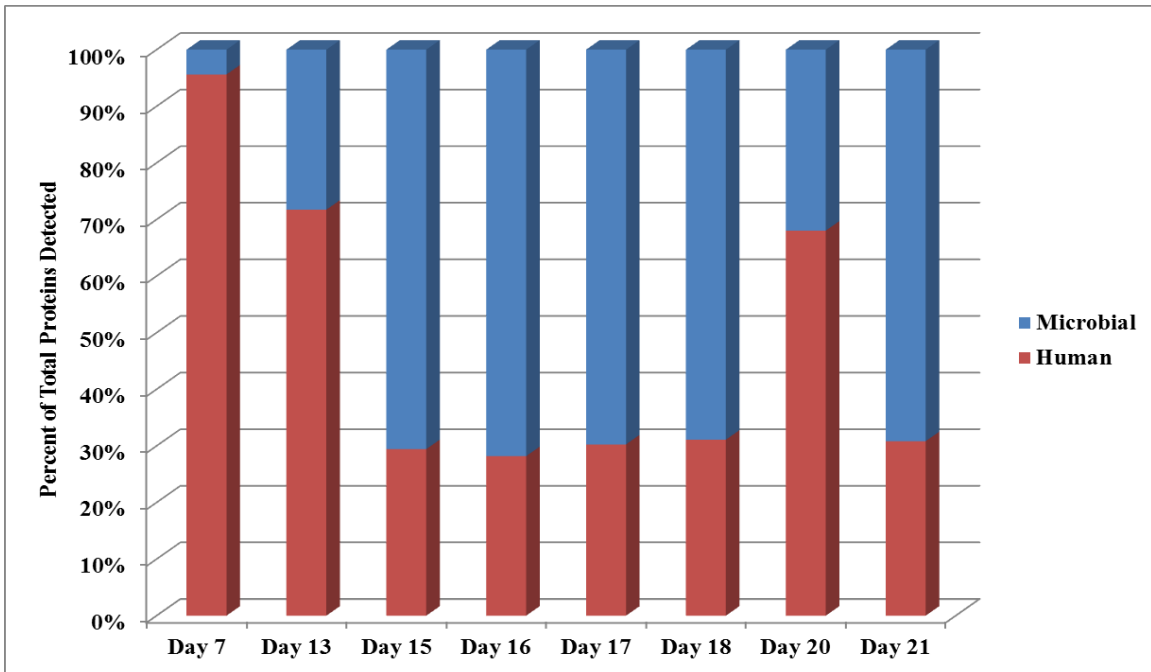


Figure 6.2: Distribution of Human and Microbial Proteins. Microbial (blue) and human (red) protein groups were averaged between two technical replicates, summed for each time point (x-axis), and plotted as a percent of the total proteins detected for each day (y-axis).

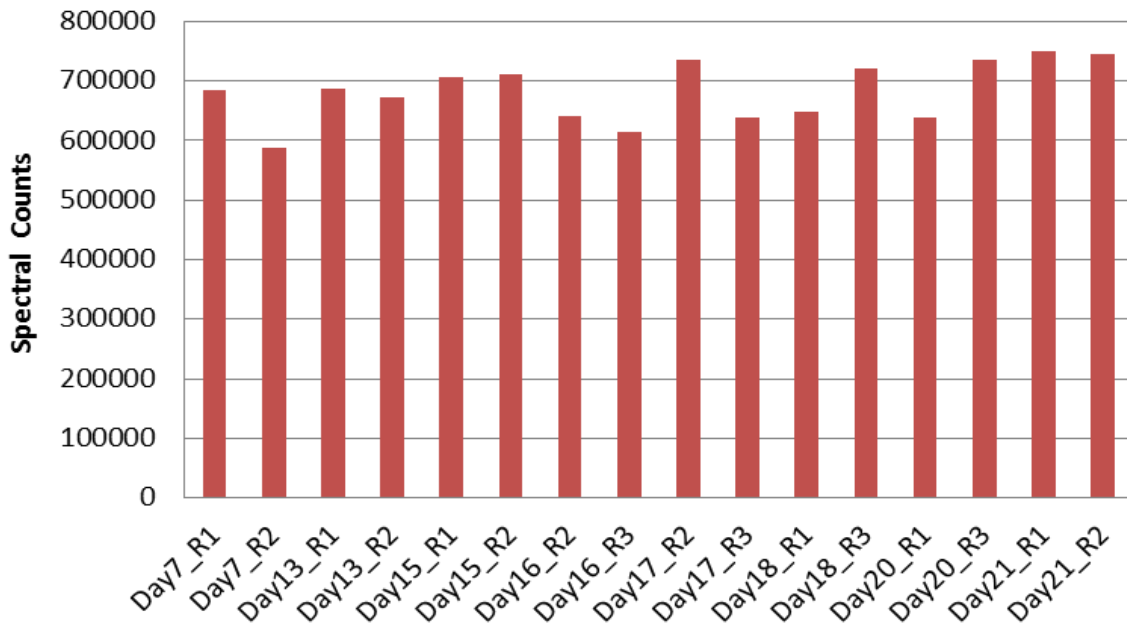


Figure 6.3: Total unfiltered MS2 spectra collected for each run.

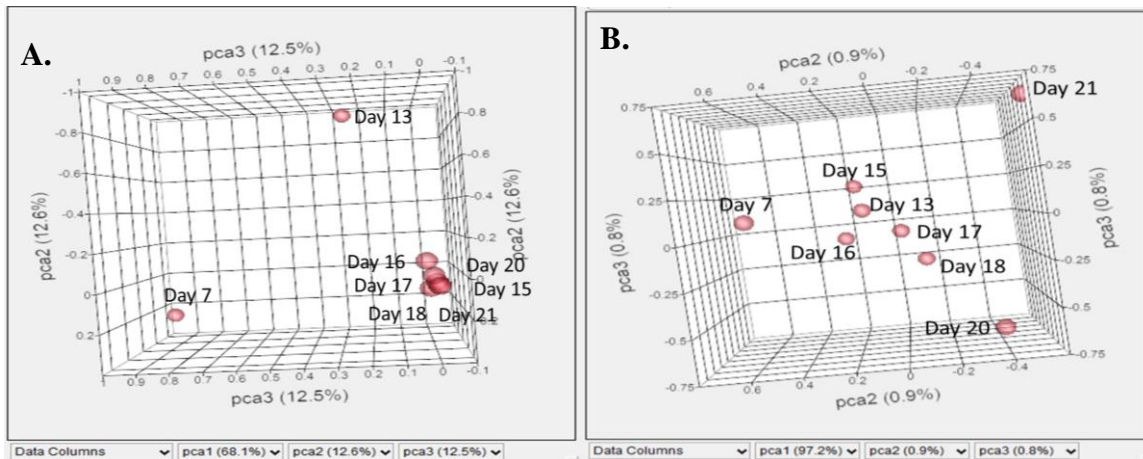


Figure 6.4: Variation among Human and Microbial Proteins Across Time. Principal coordinate analysis (PCA) was performed on A.) human, and B.) microbial proteins for all time points using JMP Genomics software (JMP Version 7. SAS Institute Inc., Cary, NC, 1989-2007). Adjusted spectral counts averaged between two technical replicates were input for each day.

Principal component analysis (PCA) was applied to the data to determine the amount of variance between time points. When looking only at the microbial proteins, three distinct clusters are apparent: days 15, 16, 17, 18, 20, and 21 cluster closely together, while day 7 and day 13 are individually distinct from these time points (Figure 6.4). This is consistent with 16S rRNA data from matched samples that also indicate three distinct microbial colonization phases at these times (7). In contrast, clustering of human proteins does not appear to follow the same pattern. None of the days appeared to be closely associated with each other, and days 7 and 21 were the most distant from the other time points, and from each other.

Microbial Protein Distribution and Functional Categorization:

In concordance with an increase in microbial load, the abundance of microbial proteins increased across time (Figure 6.5). When comparing the trends in microbial protein groups from different species across time, the contribution/distribution follows

patterns similar to that seen in 16S rRNA and metagenomic data. At day 7, microbial proteins were very low in abundance whereas abundance levels increased by day 13, with this time point dominated by *Pseudomonas* and *Staphylococcus* proteins. However, by day 15 we began to see the emergence of *Serratia* (*UC1SER*) and *Citrobacter* (*UC1CIT*) proteins, which dominated the samples in days 16-21. This corresponds well with previous metagenomic data from matched samples showing distinct community memberships in colonization phase I (days 5-9), phase II (days 10-15), and phase III (days 16-21) (7). Proteomic data also suggests *UC1SER* and *UC1CIT* were the functionally dominant members of the community during the third colonization phase, as demonstrated by the highest contribution of microbial proteins from these species during these time points.

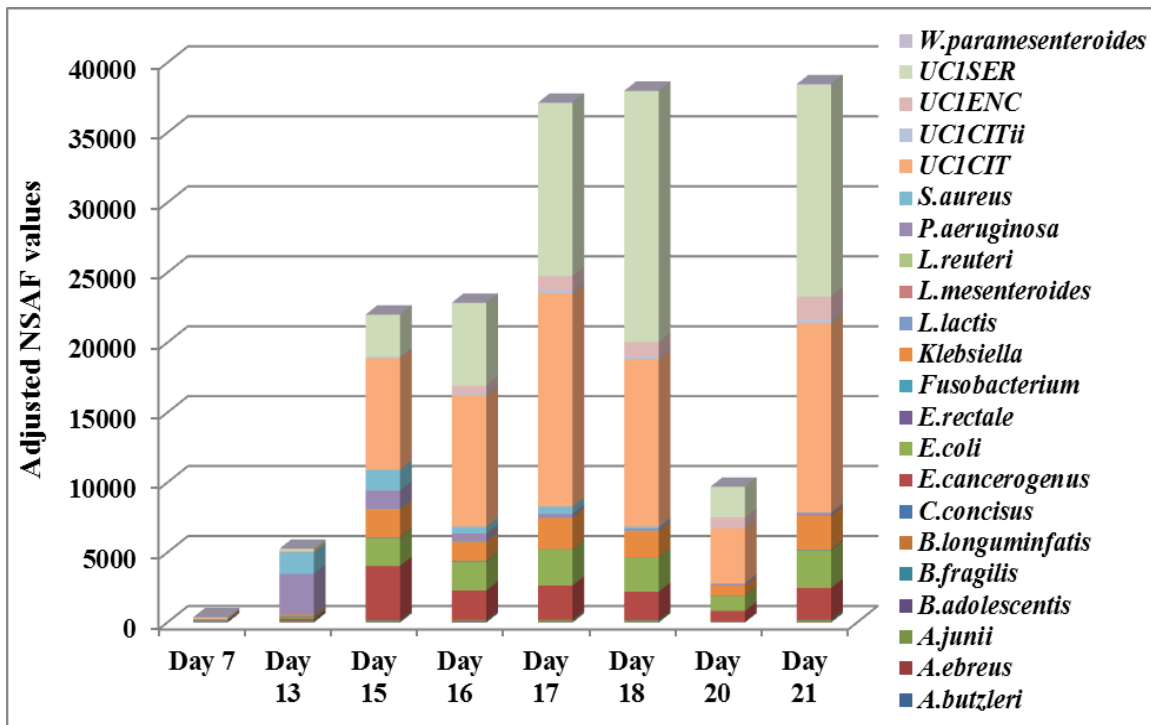


Figure 6.5: Microbial proteins detected across time. Species distributions were calculated for each time point by summing adjusted NSAF values belonging to each species. Protein groups belonging to multiple species were negligible and thus removed from the analyses.

Microbial community functions were analyzed by grouping and quantifying proteins by clusters of orthologous groups (COG) categories (Figure 6.6). At day 7, while the overall microbial protein abundance was low, the majority of spectra came from an aspartokinase I-homoserine dehydrogenase protein belonging to the amino acid metabolism and transport COG category. This enzyme catalyzes a reaction in the aspartate pathway, and may aid in providing essential amino acids from dietary sources to the infant at this early stage of development. On day 13, the community appears to be spending most of its resources producing proteins involved in energy production and conversion, as well as translation. Compared to subsequent time points, the post-translational modification, protein turnover, and chaperone category is significant, while the carbohydrate metabolism and transport category is minimal. Thus, as might be anticipated, this early community is focusing its resources on biomass growth, protein production, and protein folding at this establishment stage, and then switches to more complex metabolism at later times once the community is more established. By days 15-21 the distribution of functions were relatively similar, except for a slight increase in proteins involved in energy production and conversion at day 20. In general, it appears the overall functions of the microbial community are established relatively early (by day 13), persist, and remain relatively stable for the remainder of the time course. This is independent of taxonomic flux, suggesting functional redundancy among early gut colonizers.

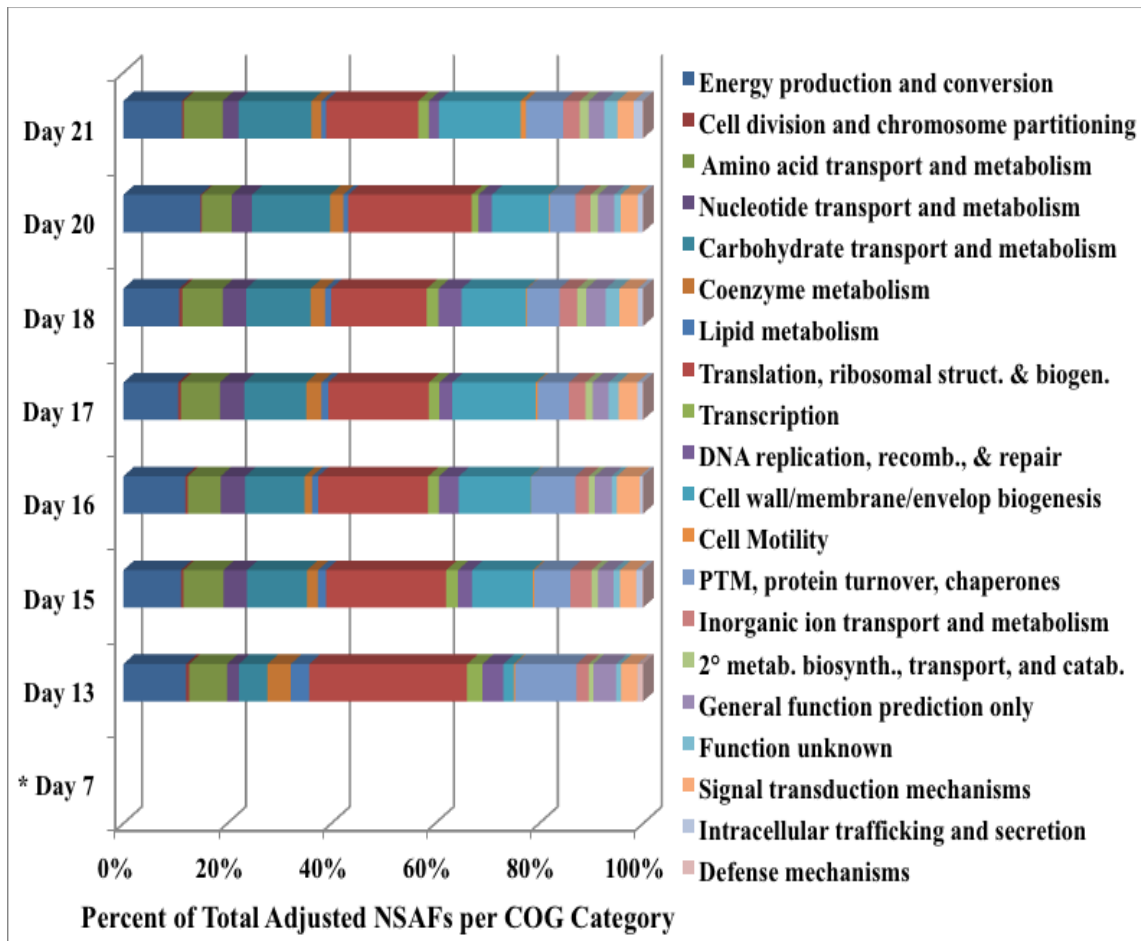


Figure 6.6: Analysis of microbial proteins by COG category classifications. Microbial proteins were clustered based on 100% amino acid identity generating protein groups. Spectral counts were balanced between protein groups and normalized to generate adjusted normalized spectral abundance factors (NSAFs). NSAFs from all microbial protein groups were summed and grouped into their respective clusters of orthologous group (COG) categories. COG (clusters of orthologous groups) assignments for each protein sequence were performed by running rpsblast against the COG database from NCBI, with an *E*-value threshold of 0.00001, and the top hit used for the assignment (42). Day seven is removed from the analysis due to low abundance values of microbial proteins from that time point.

Human Proteins in the Preterm Infant Fecal Microbiome

Global analysis of human proteins detected:

Human proteins detected across all time points were mapped to canonical pathways using Ingenuity Pathway Analysis (IPA) software (Ingenuity® Systems, www.ingenuity.com). The topmost abundant categories based on the number of proteins per category were those related to basic cellular functions such as glycolysis, oxidative phosphorylation, and elongation factor 2 signaling (Figure 6.7). In addition, proteins were categorized by molecular function, cellular compartment, and biological function using Gene Ontology (GO) classifications. A wide variety of biological processes were represented in the dataset, highlighting the significant depth of the proteomic measurements. Other categories, such as inflammatory response, were not in the list since the numbers of proteins detected in this category were not in the top 20 overall. However, it is worthwhile to note that we did detect over 30 inflammatory proteins, and that some of them were among the most abundant proteins detected in our samples (S100A8, LTF, MPO) (Table 6.2). Also, not surprising was that the most represented GO categories for molecular function consisted of general functions such as protein, ATP, and nucleotide binding. And, the cellular compartment GO category shows that most detected proteins were from the cytoplasm.

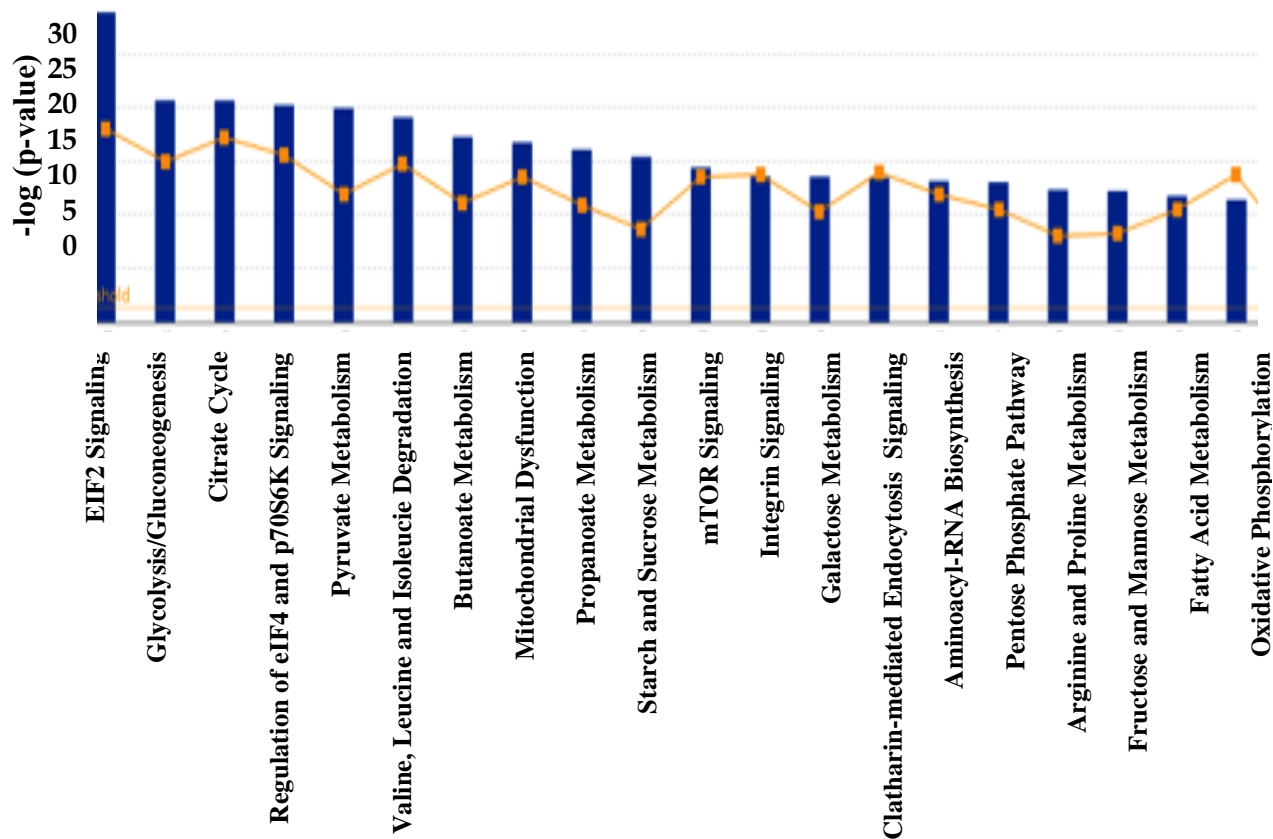


Figure 6.7: Top Canonical Pathways. The major canonical pathways for human proteins detected across the dataset were determined using Ingenuity Pathway Analysis (IPA) software (Ingenuity® Systems, www.ingenuity.com). The significance of the association was measured by calculating the ratio of number of *detected* proteins that map to the pathway divided by the *total* number of proteins from that pathway (orange boxes). The Fisher's exact test was used to calculate a *p*-value determining the probability that the association between the proteins in the dataset and the canonical pathway is explained by chance alone (y-axis).

Proteins Involved in Intestinal Barrier Function and Integrity:

Throughout our dataset, we found numerous proteins involved in intestinal barrier formation and functions (Table 6.2). The intestinal barrier is composed of enterocytes, absorptive epithelial cells held together by tight junctions, which serve as a physical

barrier. Breakdown of this barrier or incomplete formation, as is oftentimes seen in premature infants, can contribute to bacterial translocation and disease states such as NEC (244). We detected numerous tight junction proteins including occluding (OCLN), claudins (CLDN18, CLDN23, CLDN3, CLDN7), and tight junction proteins 1, 2, and 3 (TJP1, TJP2, TJP3, or zona occludens 1, 2 and 3,). In addition, proteins involved in the tight junction signaling pathway were expressed (Figure 6.8)

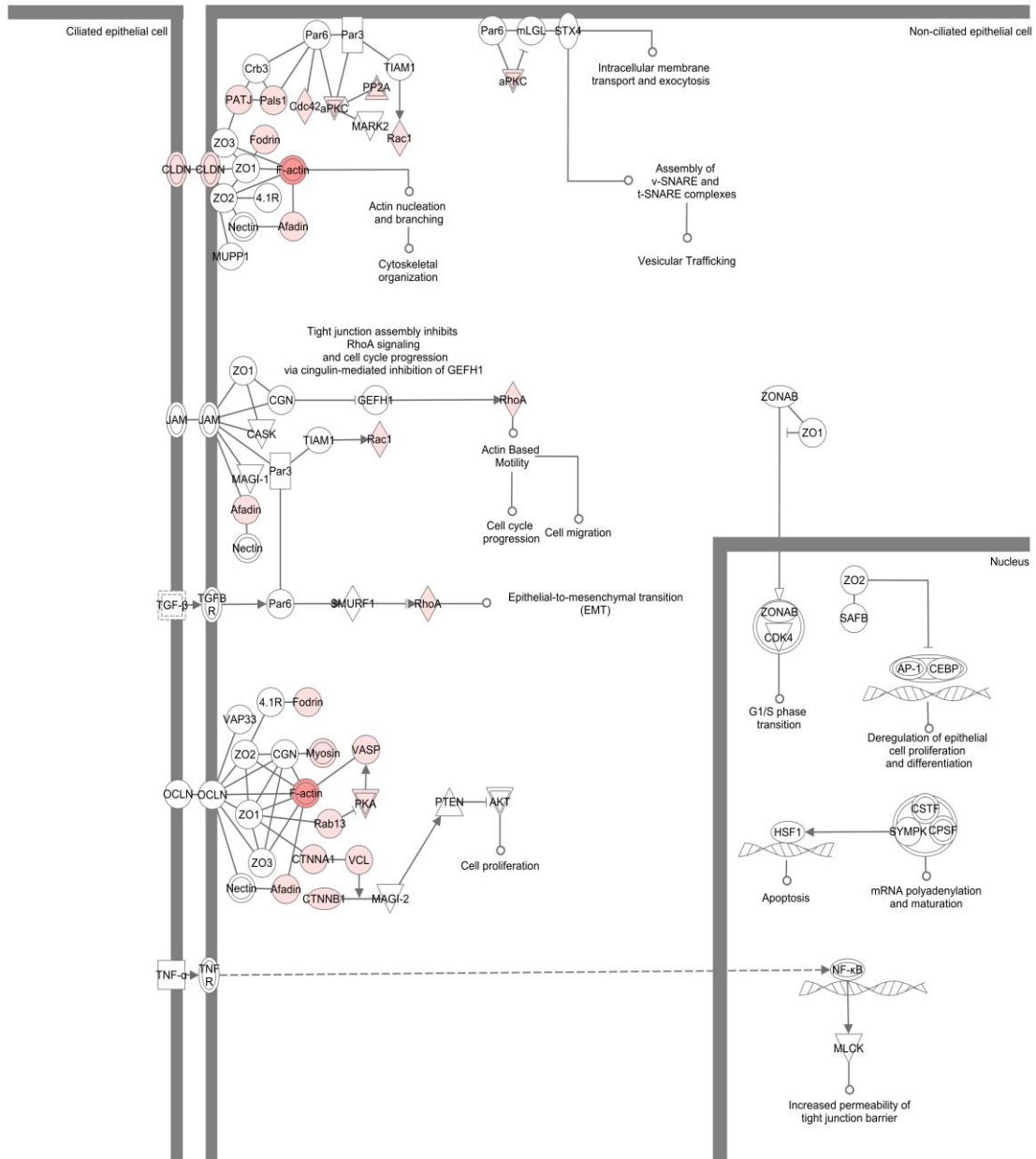
The mucus layer is a major component of the intestinal barrier, which helps maintain homeostasis between the gut microbiota and their host by minimizing physical contact between the microbes and intestinal epithelial cells (63, 245-248). The mucus layer is comprised of mucins, glycoconjugates of a polypeptide core covered in O-linked carbohydrate side chains, secreted by specialized intestinal epithelial cells called goblet cells. We detected numerous mucin proteins, including both secretory gel-forming mucins (MUC2, MUC5AC, MUC5B, and MUC6) and membrane-bound mucins (MUC1, MUC3B, and MUC4). O-linked glycans from mucins provide an energy source for bacteria in the outer mucus layer (61). Several enzymes in the o-glycan biosynthesis pathway were detected including those involved in synthesizing core 3 type glycans, the major type found associated with MUC2 (245).

Mucin 2 (MUC2) has been shown to bind the Fc fragment of the IgG binding protein (FCRPB/Fcgbp), a protein expressed by placental and colonic epithelial cells which plays a role in immune protection and inflammation, and was by far, the most dominant protein detected in our samples [16,17] (Table 6.2). Interestingly, the second most abundant protein we detected was the calcium-activated chloride-channel 1 (CLCA1) protein, another protein involved in mucus secretion by goblet cells [18]. In

addition, all three trefoil factor family peptides TFF1, TFF2, and TFF3, a family of proteins which play an important role in maintenance and repair of the intestinal mucosa, were detected (249).

Table 6.2: Topmost Abundant Human Fecal Proteins with Relevance to Host-Microbe Interactions

Protein Symbol	Protein Name	Function/ Relevance
FCGBP	IgGfc-binding protein	Mucin-binding protein expressed by placental and gut epithelial cells
CLCA1	Calcium-activated Cl channel regulator	Involved in the regulation of mucus production and secretion by goblet cells
DMBT1	Deleted in malignant brain tumors 1	Secreted glycoprotein known to bind broad range of bacterial and viruses; believed to confer mucosal protection; upregulated in IBD
ANPEP	Aminopeptidase N	Brush border protease that can serve as receptor for bacterial toxins and viral particles; known to be upregulated during inflammation
SERPINA3	Alpha-1-antichymotrypsin	Protease inhibitor known to be overexpressed during inflammation
ALPI	Intestinal-type alkaline phosphatase	Modulates inflammation by dephosphorylating bacterial lipopolysaccharide; also a marker of postnatal maturation
LTF	Lactotransferrin	Abundant whey protein found in breast milk and other mucosal secretions; potent antimicrobial properties
MUC2	Mucin-2	Dominant secreted mucin glycoprotein; MUC2 knockout animals develop spontaneous colitis
ITLN1	Intelectin	Soluble lectin that contributes to innate immune response by binding bacterial sugars
OLFM4	Olfactomedin	Highly expressed by neutrophils and gut epithelial cells; upregulated in IBD



© 2000-2012 Ingenuity Systems, Inc. All rights reserved.

Figure 6.8: Tight Junction Signaling Pathway. Proteins in the tight junction signaling pathway as determined by Ingenuity Pathway Analysis (IPA) software. Proteins colored in pink are those detected by proteomics.

Table 6.3: Detected Proteins Involved in Epithelial Barrier Functions

Symbol	Protein Name
MUC1	mucin 1, isoform 10, cell surface associated
MUC1	mucin 1, isoform 11, cell surface associated
MUC1	mucin 1- isoform 12, cell surface associated
MUC1	mucin 1-isoform 20, cell surface associated
MUC17	mucin 17, cell surface associated
MUC2	mucin 2, oligomeric mucus/gel-forming
MUC3A	mucin 3A, cell surface associated
MUC4	mucin 4, cell surface associated
MUC5AC/MUC5B	mucin 5AC, oligomeric mucus/gel-forming
MUC5AC/MUC5B	mucin 5AC, oligomeric mucus/gel-forming
MUC6	mucin 6, oligomeric mucus/gel-forming
MUC12	mucin 12, cell surface associated
MUC13	mucin 13, cell surface associated
GALNT1	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 1 (GalNAc-T1)
B3GNT6	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 6
GCNT3	glucosaminyl (N-acetyl) transferase 3, mucin type
FCGBP	Fc fragment of IgG binding protein
CLCA1	chloride channel accessory 1
CDH1	cadherin 1, type 1, E-cadherin (epithelial)
CDH17	cadherin 17, LI cadherin (liver-intestine)
CDH2	cadherin 2, type 1, N-cadherin (neuronal)
CDHR2	cadherin-related family member 2
CDHR5	cadherin-related family member 5
EPCAM	epithelial cell adhesion molecule
TFF1	trefoil factor 1
TFF2	trefoil factor 2
TFF3	trefoil factor 3 (intestinal)
CLDN18	claudin 18
CLDN23	claudin 23
CLDN3	claudin 3
CLDN7	claudin 7
OCLN	occludin
TJP1	tight junction protein 1 (zona occludens 1)
TJP2	tight junction protein 2 (zona occludens 2)
TJP3	tight junction protein 3 (zona occludens 3)
IGA	Immunoglobulin A
DEFA5	defensin, alpha 5, Paneth cell-specific
DEFB4A/DEFB4B	defensin, beta 4A
LYZ	lysozyme
LCN15	lipocalin 15
LCN2	lipocalin 2
CEACAM5	carcinoembryonic antigen-related cell adhesion molecule 5
ITGB1	integrin, beta 1
ERLEC1	endoplasmic reticulum lectin 1
ITLN1	intelectin 1 (galactofuranose binding)
ITLN2	intelectin 2

Secretory IgA is an important component of the epithelial barrier that specifically binds bacteria, limiting their association with the epithelial cell surface and restricting penetration across the gut epithelia (63, 250-252). We detected components of secretory Immunoglobulin A, including the two IgA heavy chain constant regions, (IgA1 and IgA2) which form a dimer held together by the J chain (15 kDa polypeptide), which was also detected. Secretory IgA contains a secretory component that is a portion of the polymeric immunoglobulin receptor (pIgR: 130kDa). The poly Ig-receptor is expressed by epithelial cells, binds to the IgA oligomers and allows transport across the mucosal epithelium. The majority of the pIgA receptor is proteolyzed, except the secretory component, which is secreted and diffuses along with dimeric IgA through the lumen. This protein was also detected throughout our samples.

Antimicrobial proteins are secreted by gut epithelial cells and kill bacteria through a variety of mechanisms (22, 23, 27). We detected several antimicrobial proteins including defensins (DEFA1, DEFA5), lysozyme (LYZ), lipocalin (LCN). Detection of these proteins suggests the premature infant gut, even at early stages of development, is adjusting to the introduction of microbial inhabitants, and to changing community compositions in order to carefully maintain homeostasis.

Abundant human proteins detected in fecal samples:

In addition to FCRPB and CLCA1, some of the most abundantly detected human proteins in our dataset have known relevance to host-microbe interactions (Table 6.2). In particular, antimicrobial and innate immune proteins including lactoferrin (LTF), intelectin (ITLN1), and olfactomedin (OLFM4) were among the most abundant proteins

detected. Lactoferrin (aka lactotransferrin), an iron-binding glycoprotein, is a key player in the innate immune system and is abundant and ubiquitous in human secretions such as breast milk. It has been shown to attenuate pathogenic bacteria, interfering with colonization (253) and biofilm formation (254) (255). Also, among the most abundant proteins were several that modulate or are upregulated by inflammation, like aminopeptidase N (ANPEP), alpha-1-antichymotrypsin (SERPINA3) and intestinal-type alkaline phosphatase (ALPI).

Human proteins changing over time:

Overall, human proteins, summed across all samples, contributed mostly to generalized maintenance functions (Figure 6.7). However, when human proteins were clustered based on shared trends in spectral count abundance changes (Figure 6.9), time shifts were apparent. Several neutrophil derived proteins such as neutrophil elastase (ELANE), calprotectin (S100A8), and myeloperoxidase (MPO) were most abundant at day 7 (Figure 6.9, cluster #6) suggesting activation of the innate immune system occurs early in correspondence with the early establishment of the microbiome. Cytoskeletal proteins (KRT8, KRT13, KRT18, KRT19, and KRT20) and mucins (MUC2, MUC5B) were more predominant in later time points (days 20-21) (Figure 6.9, clusters #7 and #10), suggesting structural and epithelial barrier proteins are compensating for the increased microbial load.

As noted above, there was a dramatic increase in the numbers of human proteins identified on day 20. Since many of these proteins were keratins, components of skin cells and gut epithelial cells, there are two possible explanations: 1) human contamination

at any point in the sample handling, including preparation of samples for MS measurements, or 2) an increased sloughing event in the GI tract at this time. 134 human proteins were exclusively detected at this time point, and these contribute to a wide range of biological functions (Figure 6.10). Consequently, we consider that contamination during sample handling not the most likely explanation. The most highly expressed canonical pathways on this day were those of basic metabolic functions including: EIF2 signaling, pyruvate metabolism, glycolysis (Figure 8A). Additional proteins from the pathways for pyruvate metabolism, glycolysis, and granzyme A signaling were detected day 20 (Figure 6.10). And, in particular, several HLA class I histocompatibility antigen proteins were up-regulated at day 20. HLA class I molecules, along with β 2-microglobulin (also detected in our samples), make up the major histocompatibility class I complex (MHC I) which present antigens to CD8⁺ T cells, suggesting there may have been an activation of the adaptive immune response on this day.

Discussion

Initial microbial colonization of the gastrointestinal tract is a crucial process in development. The process educates the innate immune system and initiates the establishment of a delicate homeostasis between human host and resident microbes. In premature infants, the host-microbe relationship is probably impacted significantly by underdevelopment of the intestinal barrier, an immature innate immune system, antibiotic administration, and exposure to pathogenic organisms in the intensive care unit (248). While prior studies have investigated the succession of gut microbiota at the gene level, the functional signatures of microbial and human proteins early in life have yet to be determined. Thus, this study provides the first report of simultaneous measurement of

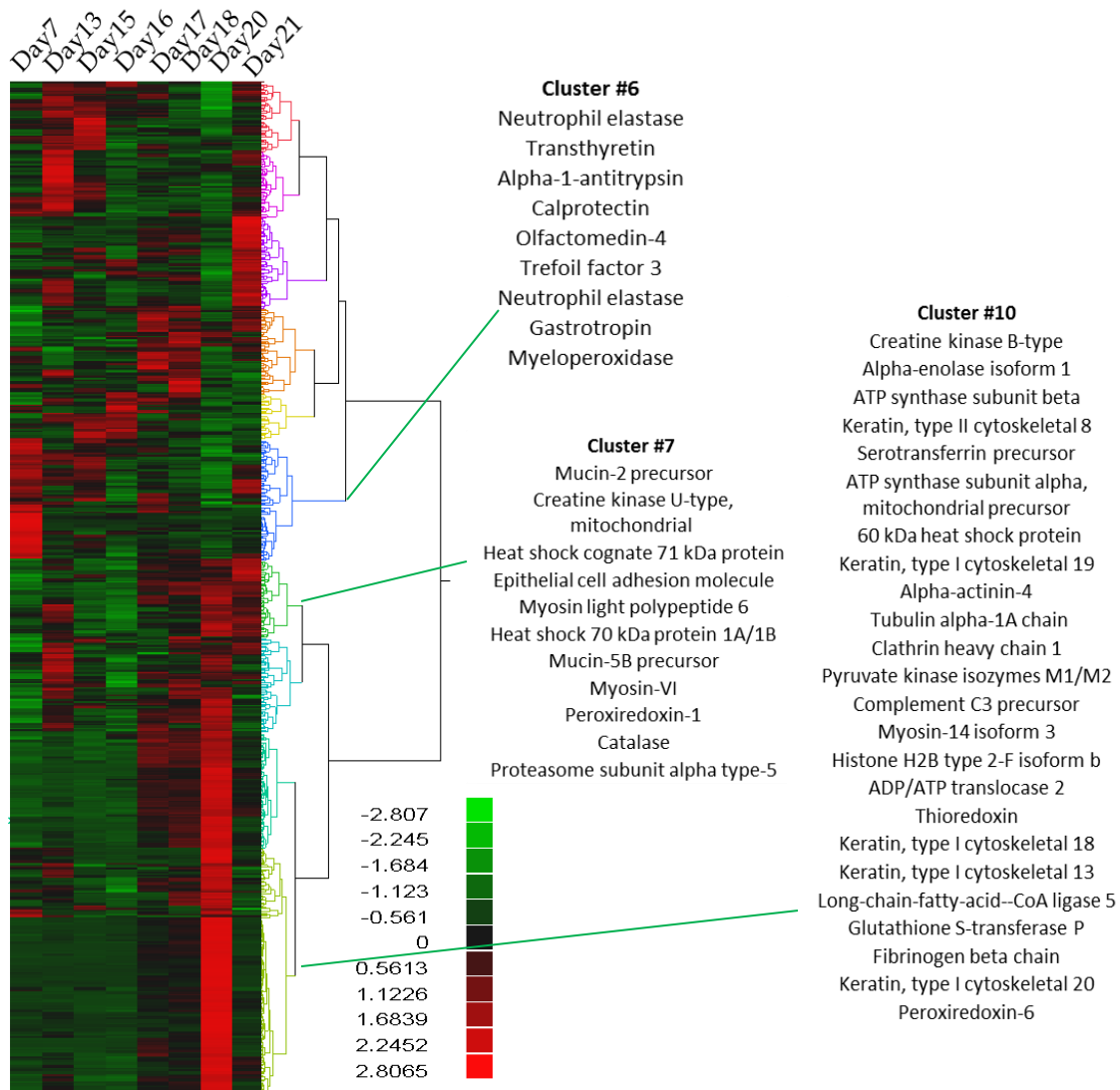
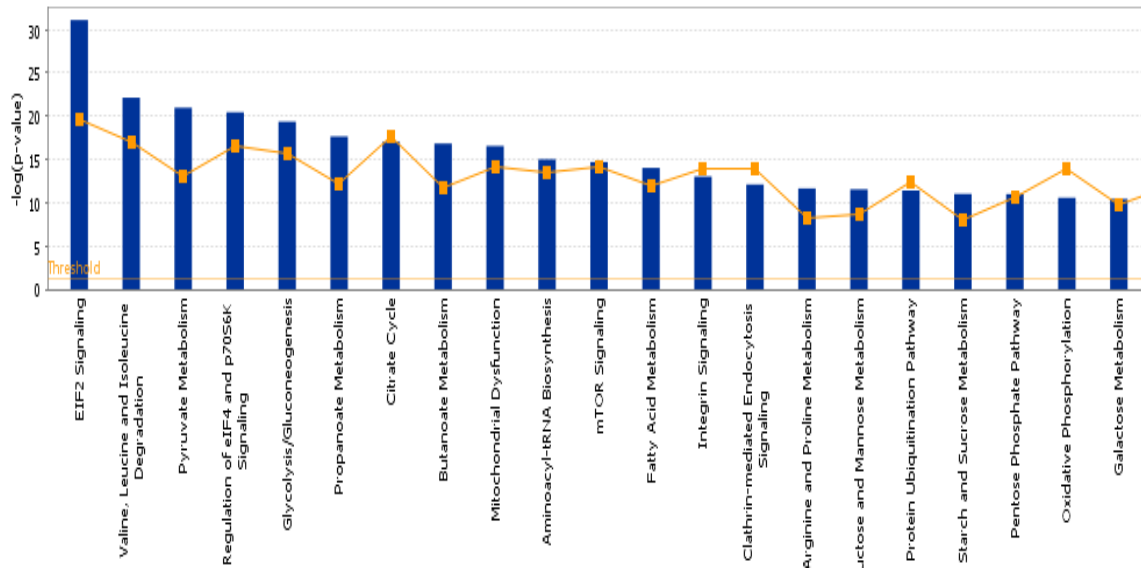
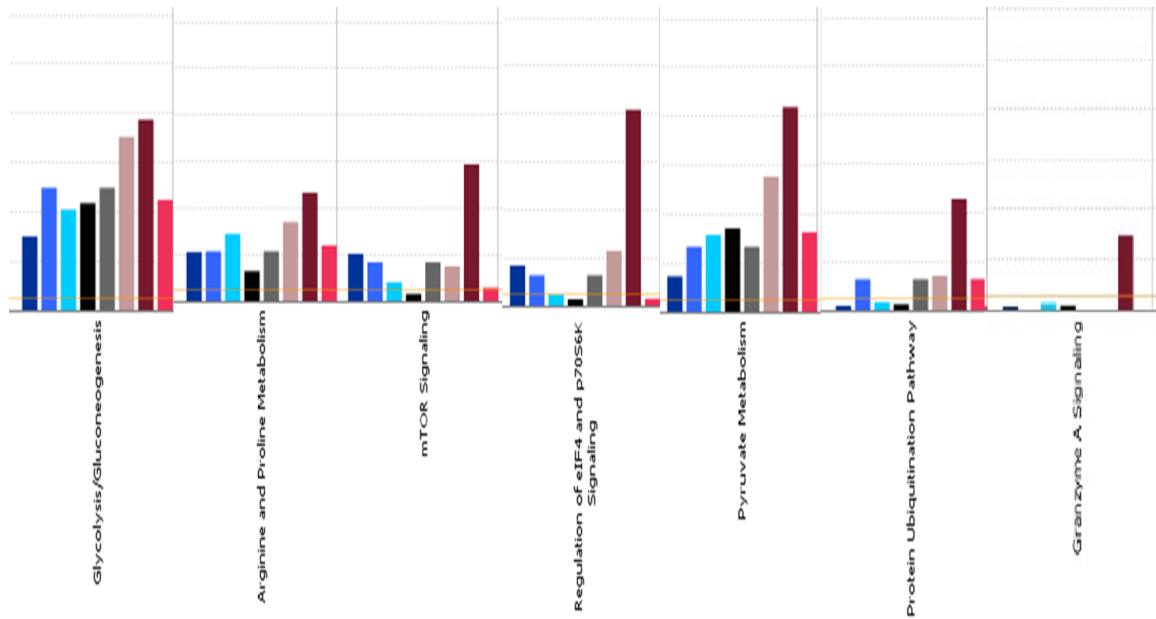


Figure 6.9: Human Proteins Changing Across Time. Proteins were clustered based on abundance changes across time. The mean of the normalized spectral counts across all time points for each protein was taken. The scale reflects the log transformed value above and below the median.

A.

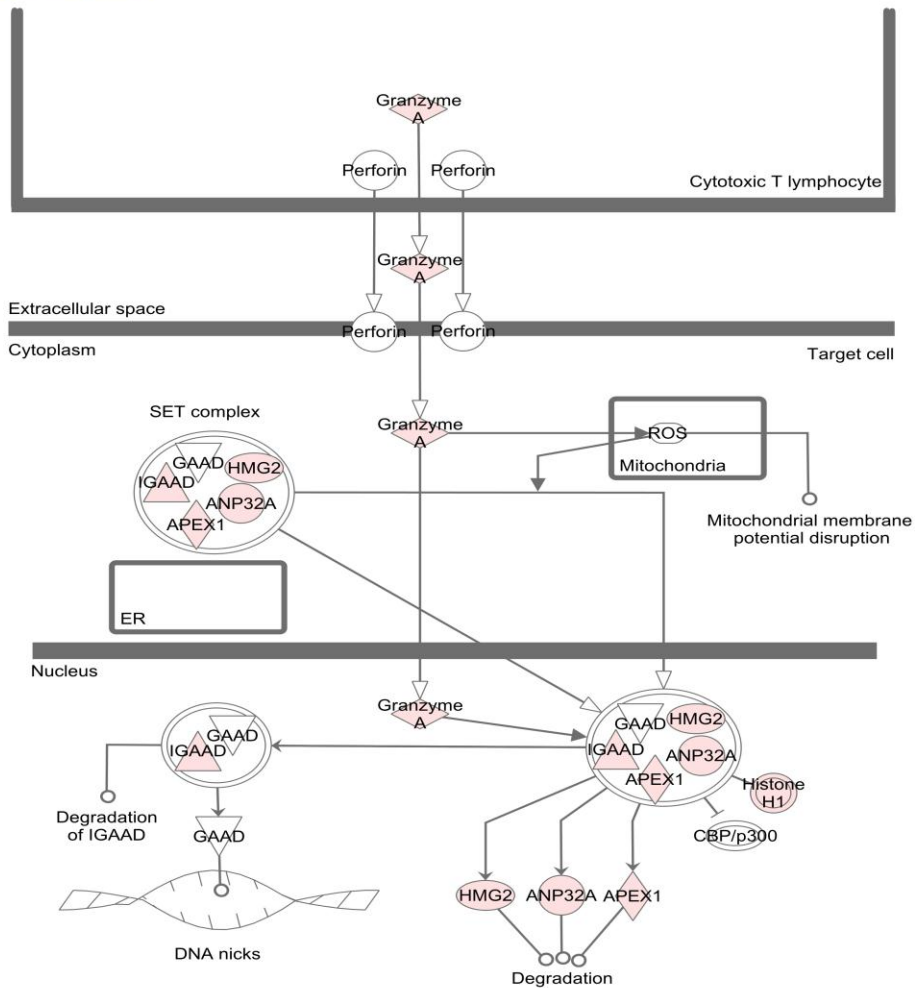


B.



C.

Granzyme A Signaling



© 2000-2012 Ingenuity Systems, Inc. All rights reserved.

Figure 6.10: Top Canonical Pathways Expressed at Day 20:

A.) The major canonical pathways were determined using Ingenuity Pathway Analysis (IPA) software (Ingenuity® Systems, www.ingenuity.com). The significance of the association was measured by calculating the ratio of number of *detected* proteins that map to the pathway divided by the *total* number of proteins from that pathway (, orange boxes). The Fisher's exact test was used to calculate a p-value determining the probability that the association between the proteins in the dataset and the canonical pathway is explained by chance alone (y-axis). B.) Canonical pathways increased in abundance at day 20. C.) Granzyme A signaling pathway.

microbial and human proteins in fecal samples from a newborn premature infant during the first month of life. Microbial proteins detected in our time course are consistent with metagenomic inference of three distinct colonization phases with vastly different species composition. Despite temporal changes in microbial community composition, the overall functions of the community stabilize relatively early and remain conserved thereafter.

Predominant throughout our sampling were human proteins involved in intestinal barrier formation. The development of the intestinal barrier involves formation of a thick mucus layer, which covers and protects intestinal epithelial cells. In the colon, the outer mucus layer harbors commensal bacteria while the thicker, impenetrable inner layer offers protection by providing a physical barrier as well as containing antimicrobial compounds and secretory IgA (69, 256). The small intestine is composed of only one mucus layer, but still provides a physical barrier with a 50 μm area separating the bacteria from the epithelia (257). The mucus layer is composed of mucins, glycoconjugates of a polypeptide core covered in O-linked carbohydrate side chains that are secreted by goblet cells. The O-linked glycans provide an energy source for bacteria in the outer mucus layer (61) (258). In our proteomic analyses, we detected numerous mucin proteins. Most of these were detected at relatively constant abundances throughout all the time points. However, some like the mucin 2 precursor, increased in abundance during the third colonization phase. Mucin 2 (MUC2) is the most abundant mucin in the intestine, has been directly linked to protecting the colonic epithelium from enteric pathogens, and is down regulated in patients with ulcerative colitis and Crohn's disease (245, 259, 260).

Several neutrophil proteins were predominant at day 7, possibly highlighting the importance of the innate immune system at this early time in development. In contrast, later in the time course, many epithelial barrier proteins increased in abundance in conjunction with the increased microbial load. Overall, these data suggest an adaptation of the host in response to the changing microbiome, resulting in a dynamic interplay between the host and its resident microbes.

CHAPTER SEVEN

Applying a Metaproteomics Approach Unravels Intra- and Inter-Individual Variation in the Preterm Infant Gut Microbiome

While chapter six discusses an in-depth analysis of the fecal microbiome of one preterm infant (the UC1 baby) over a time course of early neonatal development, the goal of this chapter is to demonstrate the feasibility of the metaproteomics approach to investigate fecal microbiomes from multiple infants. Careful consideration was taken in the experimental design and methodology used in these studies to improve this method in order to “dig deeper” into the proteomes from fecal samples. This chapter discusses the progress of method development, including some experimental trials that did and did not improve the results. Importantly, we demonstrate the value of the current method, and what we have learned about the inter- and intra-individual variability in microbial and human proteins from multiple infant fecal microbiomes.

Using Metaproteomics to Measure Fecal Microbiomes from Multiple Infants

In addition to infant #64 (the UC1 baby) described in chapter six, the fecal microbiomes of five other infants were measured using metaproteomics including: infant #74 (the Carrol baby), #6502, #7702, #Un091609, and Unlabeled#2. Three different samples from infant #74 (Carrol) were measured, which corresponded to days 15, 22, and 23 after birth. The Carrol baby was a preterm infant, and the gestational ages and dates of sample collection from the remaining infants are unknown. The experimental design used

in this section was described in detail in chapter 6 (for the UC1 baby), and in chapter two of this dissertation. Briefly, samples were prepared using the SDS-TCA protocol and measured with shotgun proteomics via nano-2D-LC-MS/MS. An overview of the proteomic measurements is shown in Table 7.1. While the values of protein, peptide, and spectra vary between samples (reasons for which are discussed in detail below), these data demonstrate that this experimental design, including the chosen sample prep method and MS instrumental setup, is capable of efficiently and reproducibly providing deep proteomic measurements from multiple infants.

Table 7.1: Overview of proteomic results from multiple infant fecal microbiomes

Infant #	Proteins	Peptides	Spectra
#74 (Carrol) Day 15	2895	11546	89895
#74 (Carrol) Day 22	2230	10489	89601
#74(Carrol) Day 23	2300	9771	93990
#6502	710	3416	19664
#7702	827	3775	24110
#Un091609	2544	8351	31914

Protein, peptide, and spectral values are all non-redundant. Searches were performed using the DBDigger algorithm (261), against the isolate database (Infant_Isolate_db_010611) for infants #6502, #7702, #Un091609, and a metagenome database (carrol_metagenome_42312_HRefseq2011_IgA_contams) for infant #74 (Values for day 15 are averages of two technical replicates, and day 23 averages from three technical replicates). Values for the UC1 baby (infant #64) are shown in chapter six.

Microbial Species Distribution in the Carrol Baby Fecal Microbiome

In addition to demonstrating the feasibility of our metaproteomic approach with multiple infants, preliminary analyses are currently underway to gain a more in-depth perspective of the inter-individual vs. intra-individual variabilities of these fecal

microbiomes. In addition to the UC1 baby (infant #64) described in chapter six, the fecal microbiome of a second preterm infant, #74, also referred to as the Carrol baby, was analyzed over multiple days after birth. In an initial companion study, fecal samples were collected from the infant over multiple days during the first month of life (days 15-24) and community genomic analysis was performed (262). Specifically, whole genomes were reconstructed from metagenomic sequences, including eight near-complete bacterial genomes and three phage genomes. New metagenomic methods, including improved binning and genome reconstruction, allowed highly efficient resolution of the community members at not only the species level, but also at the strain level (an improvement over binning methods used for the UC1 baby (7)). These genome sequences, along with the human genome and common contaminants, were used to generate the predicted protein database (carrol_metagenome_42312_HRefseq2011_IgA_contams) for metaproteomic analyses.

Interestingly, the microbial species composition was drastically different in the Carrol baby, compared to that of the UC1 baby previously described (see chapter 6). Specifically, dominant microbial members from the UC1 infant microbiome, as determined by 16S rRNA, metagenomics, were *Citrobacter*, *Serratia*, and *Enterococcus* species (7) (Young et al. *in preparation*). In contrast, the dominant members of the Carrol baby's microbial community, as determined by metagenomics, were *Enterococcus faecalis*, *Propionibacterium* and nine species of *Staphylococcus*, including four different *Staphylococcus epidermidis* strains (Figure 7.1) (262).

Metaproteomic analyses were carried out on the same samples from the Carrol baby collected at the beginning and end of the time course, on days 15 and 23,

respectively. Proteins were detected from most microbial species in the community (Table 7.2). The highest number of microbial proteins detected from day 15 belonged to *Enterococcus faecalis*, *Staphylococcus epidermidis* strain 3, and *Staphylococcus epidermidis* strain 1, while on day 23 most proteins were detected from *Peptoniphilus carrol*, and *Enterococcus faecalis* (Table 7.2). Likewise, the highest abundance of proteins, as determined by spectral counts, belonged to *E. faecalis* at both time points. However, the abundance of *Peptoniphilus carrol* proteins increased significantly from day 15 to day 23 (Figure 7.2).

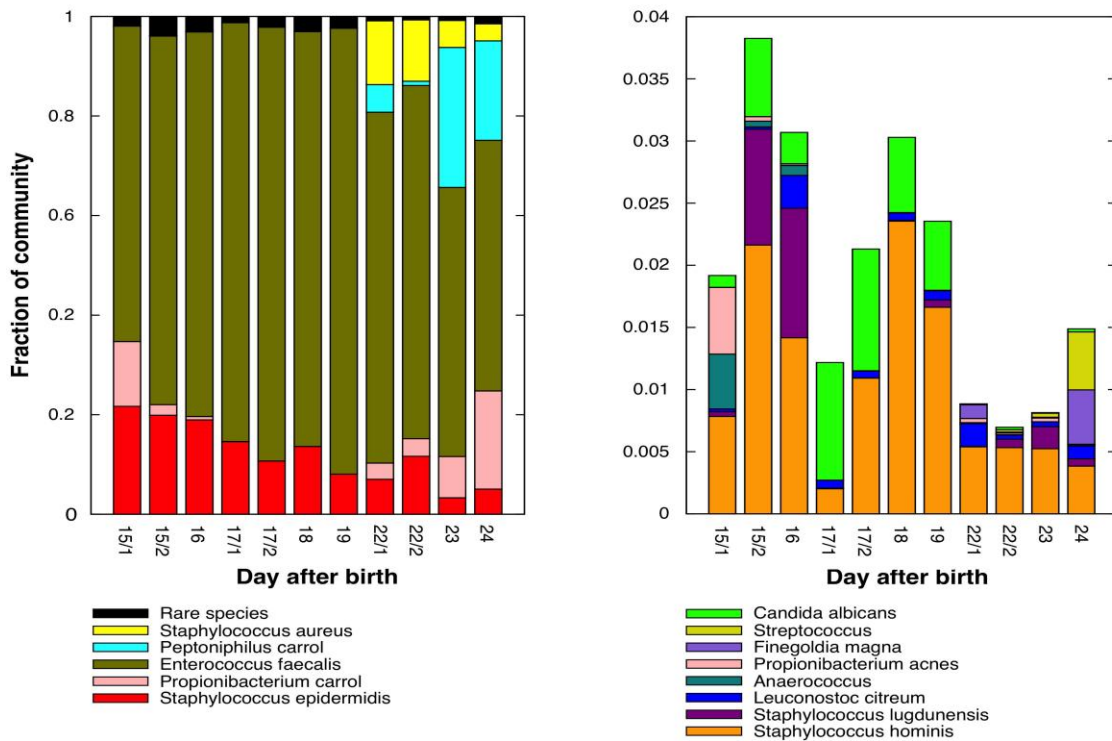


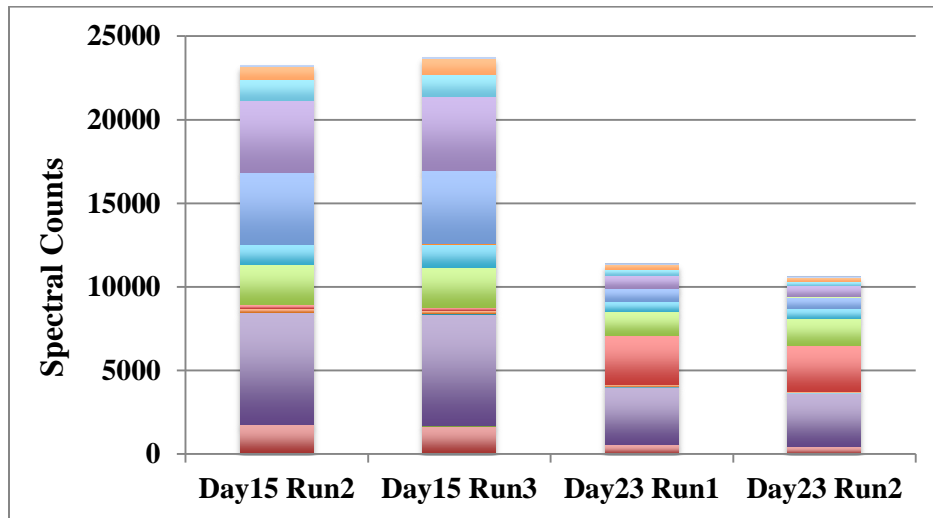
Figure 7.1: Microbial species distribution from infant #74 as determined by metagenomic sequencing. Taken from Sharon et al., *Genome Research*, 2012 (262)

Table 7.2 Microbial Proteins Detected per Species from Baby Carrol (Infant #74)

Organism	Day15 Run1	Day15 Run2	Day23 Run1	Day23 Run2
<i>Anaerococcus</i>	5	5	2	3
<i>Candida albicans</i>	109	123	51	54
<i>Enterococcus faecalis plasmid</i>	2	1	6	1
<i>Enterococcus faecalis</i>	299	304	198	195
<i>Finnegoldia magna</i>	5	8	10	8
<i>Leuconostoc citreum</i> ⁴	8	11	14	8
<i>Propionibacterium acnes</i> ⁵	1	1	0	2
<i>Peptoniphilus carrol</i>	35	40	168	186
<i>Propionibacterium carrol</i>	164	187	112	136
<i>Staphylococcus aureus strain 23</i>	0	1	1	0
<i>Staphylococcus aureus strain 11</i>	109	120	77	87
<i>Staphylococcus epidermidis misc.</i>	2	3	1	2
<i>Staphylococcus epidermidis strain 3</i>	230	228	72	72
<i>Staphylococcus epidermidis strain 4</i>	0	1	0	2
<i>Staphylococcus epidermidis phage</i>	0	0	3	5
<i>Staphylococcus epidermidis strain 1</i>	228	228	76	71
<i>Staphylococcus hominis</i>	109	120	62	53
<i>Staphylococcus lugdunensis</i>	69	84	47	48
<i>Streptococcus</i>	6	8	3	8
Total	1381	1473	903	941

Data searched against the metagenome database (carrol_metagenome_42312_HRrefseq201_IgA_contams) using the DBDigger algorithm. Values shown in the table are the number of proteins detected per each species. (Redundant values, no clustering, no normalization).

A.



B.

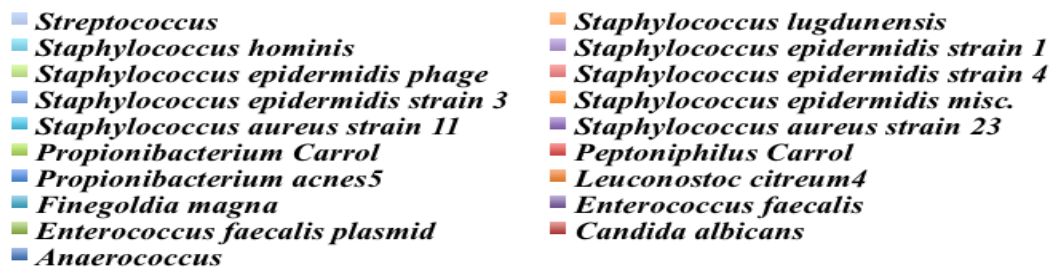
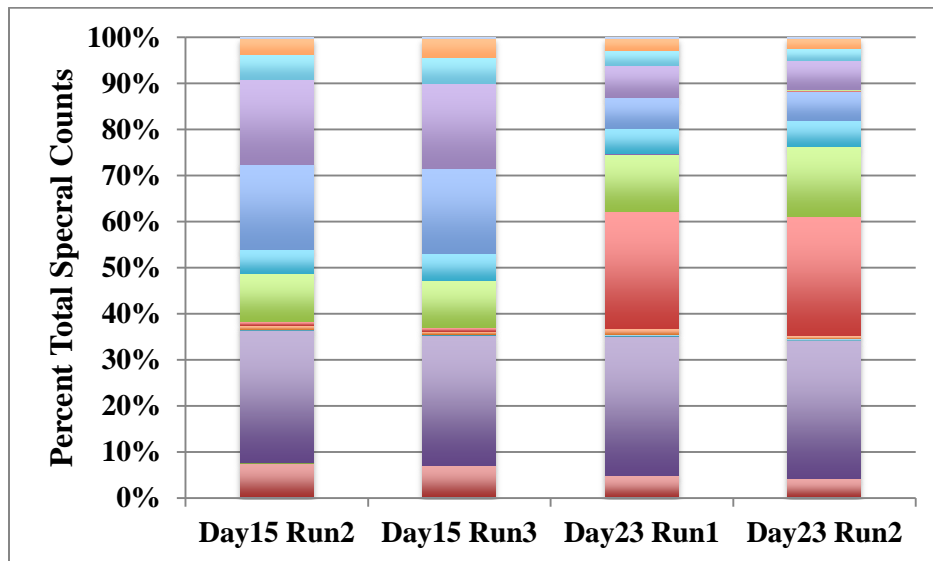


Figure 7.2: Microbial Species Distribution from Baby Carrol (Infant #74) as Determined by Metaproteomics. A.) Total microbial spectra detected per species, B.) Percent distribution of microbial spectra per species. Data searched against the metagenome database (carrol_metagenome_42312_HRefseq2011_IgA_contams) using the DBDigger algorithm. (Redundant values, no clustering, no normalization).

Human proteins in the Carrol baby

A key advantage of using a metaproteomics approach to study the human microbiome is the ability to simultaneously monitor microbial and human proteins. While the initial metagenomic study on baby Carrol (262) provided valuable information on the changing microbiota over the first month of the infant's life, the addition of metaproteomic data provides direct *in situ* measurements of the functional signatures carried out by these microbes, as well as analysis of the human proteins which may be contributing/adapting to microbial colonization and fluctuations. Similar data was shown for infant #64 (UC1 baby) in chapter six, while the current chapter demonstrates that this method is successful at measuring microbial *and* human proteins from additional infants, despite the variation between and within these individuals (discussed in more detail below). Table 7.3 lists the topmost abundant human proteins measured from the Carrol baby from days 15 and 23 after birth. Lactoferrin (aka lactotransferrin) was the most abundant human protein detected, as determined by total spectral counts. Lactoferrin is an iron binding glycoprotein and a major component of innate immune system. It is a ubiquitous and abundant constituent of human external secretions, including breast milk (Legrand, 2008). Interestingly, this protein increased in abundance from the first time point measured (day 15 after birth) to the last time point measured (day 23 after birth). This was in contrast to the UC1 baby whose most abundant protein was the IgG Fc-receptor binding protein (FCRBP) (Table 7.4 and chapter 6). The FCRBP protein was also abundant in the Carrol samples but decreased from day 15 to day 23 (Table 7.3), and only comprised a fraction of the spectra compared to the UC1 baby on comparable days of the baby's lives (days 21 and 22 respectively) (Table 7.4). It is also striking that other





















































proteins such as IgA, and alpha-2 macroglobulin were more dominant in the Carrol baby, whereas some like mucin 2 and Ca-activated Cl channel regulator were more dominant in the UC1 baby.

Table 7.3: Abundant Human Proteins from Baby Carrol (Infant #74)

Day15 Run2	Day15 Run3	Day23 Run1	Day23 Run2	Abundance Trends	Description
6411	8345	26931	24658	— — ■ ■	lactotransferrin isoform 1 precursor
12012	11085	8132	6585	■ ■ — —	IgGFc-binding protein precursor
4082	3967	5187	4378	— — ■ —	alpha-1-antitrypsin precursor
4687	5167	2798	2193	■ ■ — —	polymeric immunoglobulin receptor precursor
3689	4160	2235	2040	■ ■ — —	IgA C region
1288	883	5220	5006	— — ■ ■	alpha-2-macroglobulin precursor
2117	2123	2091	1318	■ ■ ■ —	chymotrypsin-like elastase family member 3A
1932	2421	999	990	■ ■ — —	Ig M C region
1766	2057	1896	1578	■ ■ ■ —	calcium-activated chloride channel regulator 1
2572	2930	554	805	■ ■ — —	alpha-1-antichymotrypsin precursor
1220	1320	2559	1933	— — ■ ■	complement C3 precursor
1450	1611	515	596	■ ■ — —	galectin-3-binding protein
332	241	1856	2481	— — ■ ■	serum albumin preproprotein
976	797	1045	1246	■ — ■ ■	transthyretin precursor
1011	907	706	849	■ ■ — —	chymotrypsin-C preproprotein
797	859	751	657	■ ■ ■ —	aminopeptidase N precursor
828	593	493	566	■ — — —	mucin-2 precursor
716	939	387	382	■ ■ — —	deleted in malignant brain tumors 1 protein
711	657	496	420	■ ■ — —	neprilysin
648	708	384	517	■ ■ — —	sucrase-isomaltase, intestinal

Values displayed are spectral counts (non-normalized, no clustering) listed in order of abundance, determined by total spectral counts per protein. Data were searched using DBDigger against the metagenome database. Technical MS duplicate runs are shown to display consistency between replicate runs.

Table 7.4: Comparison of Abundant Human Proteins in the UC1 baby and the Carrol baby at Similar Time Points

UC1baby Day 21	Carrol baby Day 22	Comparative Abundances		Description
22755	5767			IgGFc-binding protein
6788	1394			calcium-activated chloride channel regulator 1
5132	3744			alpha-1-antitrypsin
4856	854			deleted in malignant brain tumors 1
3108	696			aminopeptidase N
2968	749			intestinal-type alkaline phosphatase
2702	738			mucin-2
2651	10			lithostathine-1-alpha
2541	629			neprilysin
2230	409			meprin A subunit alpha
2109	1764			chymotrypsin-like elastase family member 3A
1664	1161			alpha-1-antichymotrypsin
1352	362			intelectin-1
1306	577			chymotrypsin-like elastase family member 3B
1293	82			actin, cytoplasmic 2
1288	25144			lactotransferrin isoform 1
1160	1216			chymotrypsin-C preproprotein
1160	373			xaa-Pro aminopeptidase 2
1020	950			transthyretin
766	786			sucrase-isomaltase, intestinal
706	673			galectin-3-binding protein
315	2327			alpha-2-macroglobulin
138	246			tenascin
107	924			complement C3
93	2411			polymeric immunoglobulin receptor
59	855			PREDICTED: complement C3-like, partial

Values are non-normalized spectral counts listed in order of most abundant protein in the UC1 sample. Searches were performed using Myrimatch and IDPicker 2.0 (40, 263) against the metagenome database for the Carrol sample, and the metagenome + isolate database for the UC1 sample.

Variability Among Ratios of Human and Microbial Proteins Across Multiple Infant Fecal Microbiome Samples

During the course of method development for this study, fecal microbiomes from six different infants were measured, with two of the six infants monitored over multiple time points (UC1, and Carrol baby, as described above and in chapter six). Interestingly, among the total proteins detected, the numbers of proteins belonging to microbial species compared to the number of human proteins varied dramatically between different samples: including those from different infants, as well as ratios within the same baby at different times after birth (Figure 7.3). Specifically, among all the samples measured, the percent of detected human proteins ranged from 17%-98% of the total proteins detected per run (Figure 7.3). Within the UC1 infant, the ratio of human proteins ranged from 40-98% across a series of times (Figure 7.3 and chapter six), and in the Carrol baby, the ratio of human proteins varied from 59-69% from the beginning to the end of the sample collection times. In addition, when taking into account abundance levels of total human proteins by comparing spectral counts summed for all the microbial proteins, there is also a wide range in the values. In particular, the abundance of human proteins is between 20%-99% of the total spectra collected from that sample, with microbial spectra ranging from 1%-80% of the total spectra. And, again the abundance ratios changed within the same infants over multiple times (Figure 7.4).

Samples from infants 6502 and 7702 contained higher numbers and abundances of human proteins (thus lower microbial, and overall protein identifications) (Table 7.1). We noticed that these two samples were very black and viscous with a tar-like consistency compared with other samples, which were more greenish or brownish in

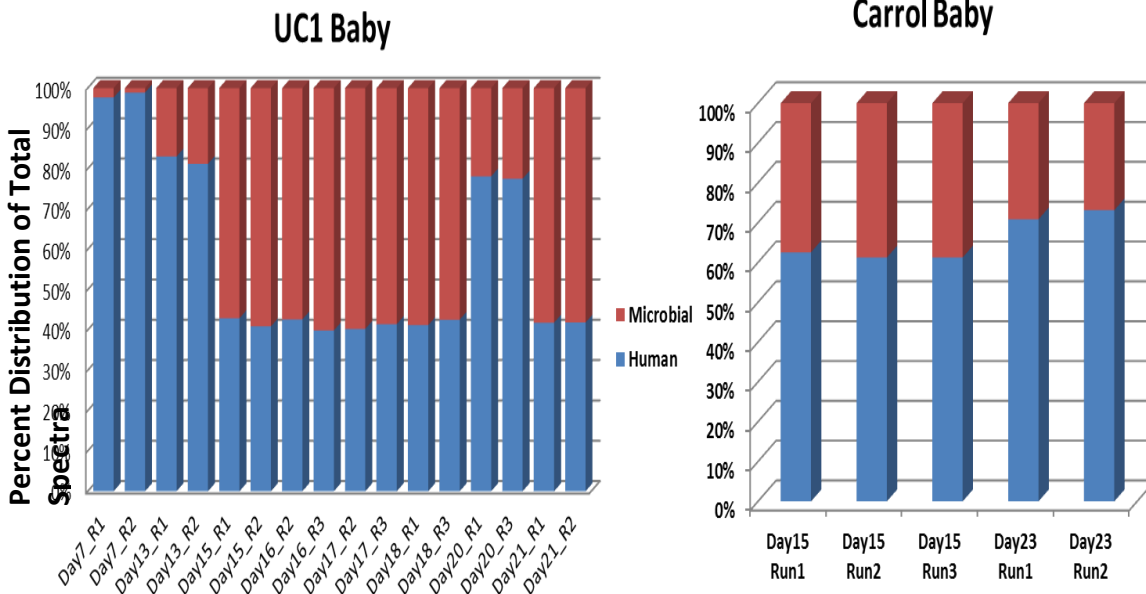
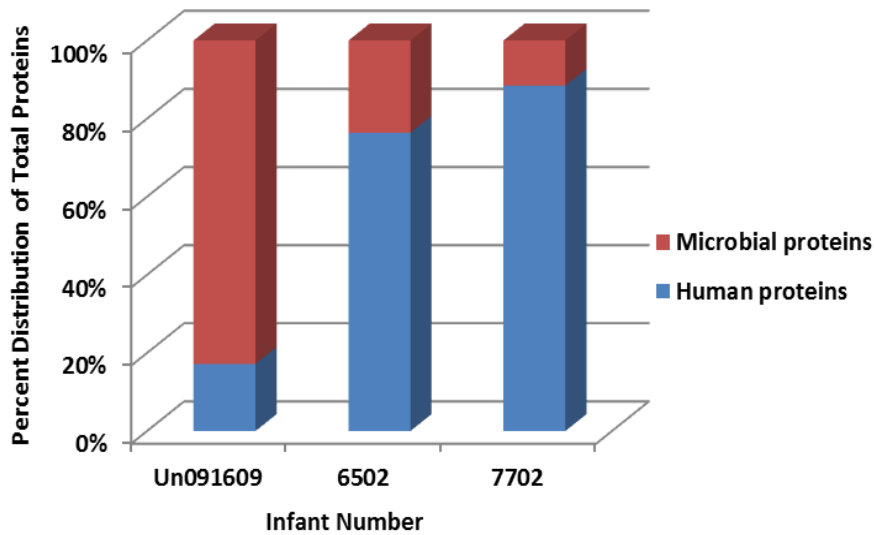


Figure 7.3: Distribution of Microbial and Human Proteins: Samples from infants Un091609, 6502, and 7702 were searched against the isolate database (Infant_Isolate_db_010611) using the DBDigger algorithm. Samples from the UC1 baby were searched using the Sequest algorithm against the Isolate_UC1_HrefSeq2011_IgAM_20 database. Carrol baby samples were searched using the DBDigger algorithm against the database: carrol_metagenome_42312_HRefseq2011_IgA_contams. Proteins identified were based on redundant values (no protein clustering)

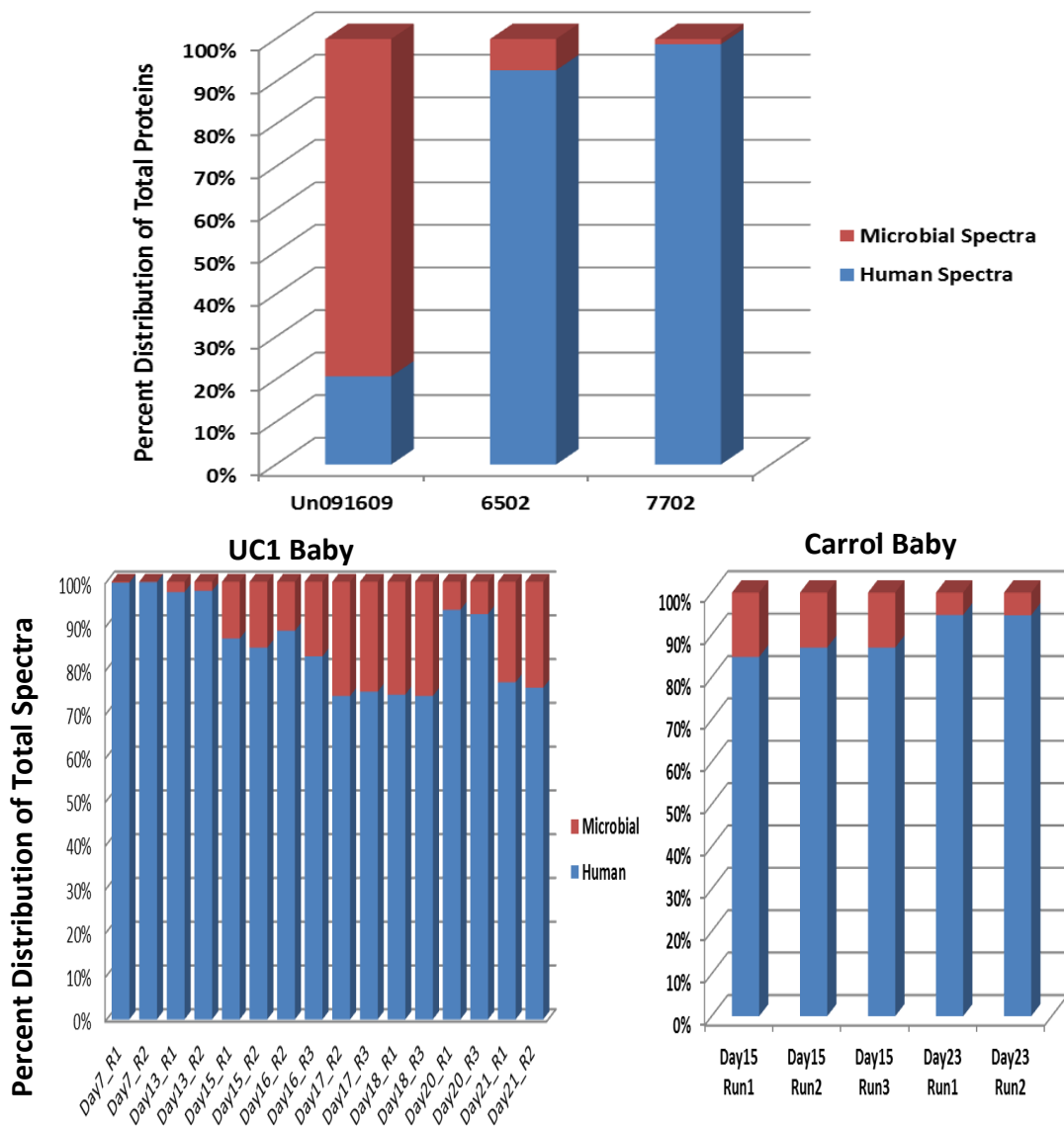


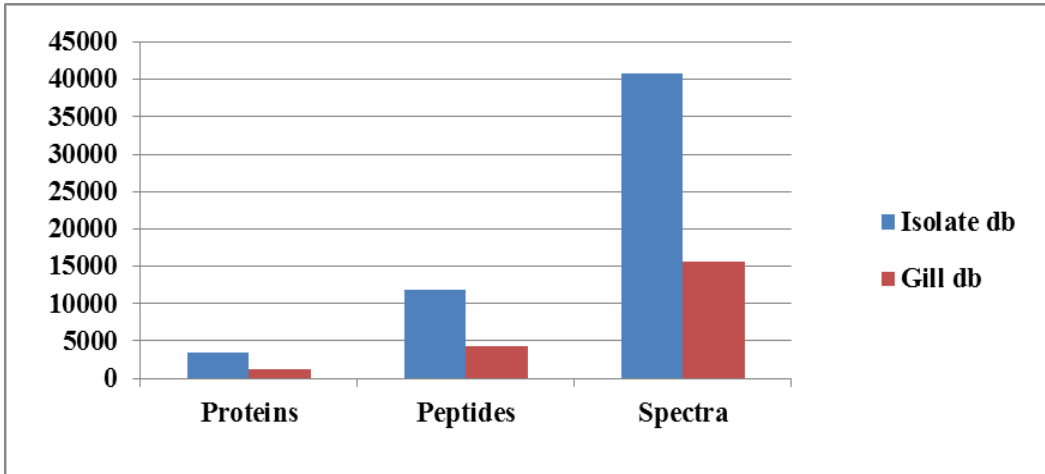
Figure 7.4: Distribution of Microbial and Human Spectra. Comparative abundances were calculated by adding all spectral counts from microbial proteins and human proteins and plotting the percent ratio of the two. For the UC1 baby, Sequest searches were performed against the database: Isolate_UC1_HrefSeq2011_IgAM_20. Data from the Carrol baby was searched using DBDigger searches against the database: carrol_metagenome_42312_HRefseq2011_IgA_contams. No protein clustering or normalization was performed on either dataset.

color and chunkier. We think that these two samples may have been collected within the first few days of the infant's life and could possibly be meconium samples. Meconium contains substances ingested by the infant *in utero* such as epithelial cells, mucus, and bile, and importantly only contains very few microbes. However, since we do not know (these samples were sent as test samples for method development, and the details of the infant or date of collection were not recorded), we cannot make definitive conclusions about this. However, if these samples were meconium, it would make sense that they would contain very few microbial proteins, but rather be dominated by human proteins.

Initial Method Development: Building the Optimal Search Database

As discussed in chapter two (materials and methods), designing a proper protein search database is crucial to yielding optimal PSMs and thus improving the overall results. During initial method development, matched metagenomic data was not available for the UC1 samples. Thus, a search database was generated from a published metagenomic sequences from an adult fecal microbiome (264) (Gill db). However, searches against this database resulted in very few protein, peptide, and spectra matches (Figure 7.5). Therefore, a second database (Isolate db) was generated in which information from 16S rRNA data from the same sample was used to choose twenty-one closely related bacterial isolate sequences (obtained from JGI) (; Infant_Isolate_db_01061; species are listed in chapter six). Searching against this database improved the results significantly, due to the database more closely representing the species in the sample. Subsequently, once the metagenomic data became available, which consisted of whole-genome reconstruction of the four dominant members from that community (see chapter six or Morowitz et al), a third and final database (Isolate_UC1_HrefSeq2011_

A.



B.

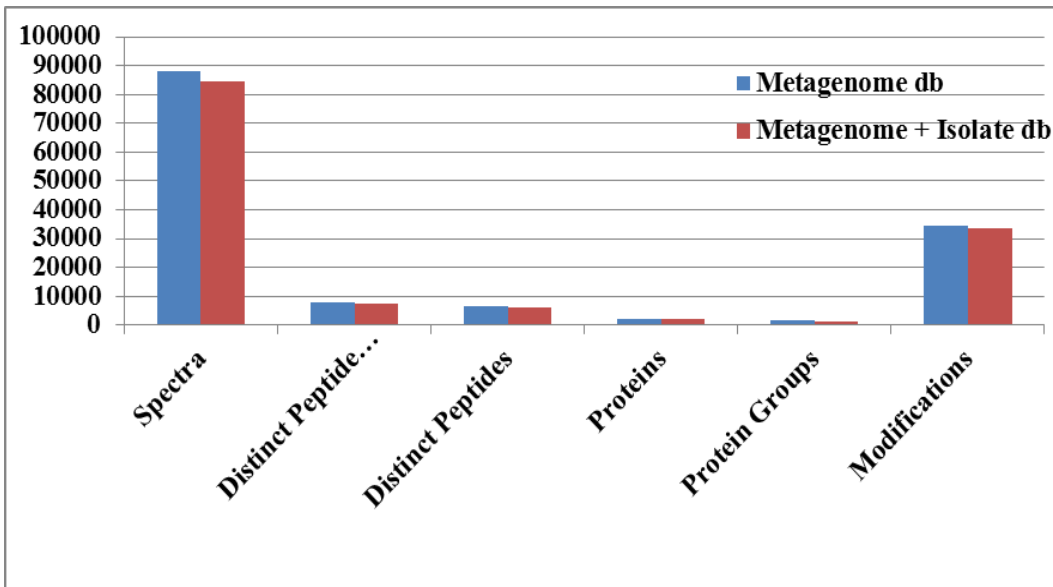


Figure 7.5: Choosing Optimal Database Design. A.) Sample 91609Un run on the the LTQ-Orbitrap XL and searched using DBDigger against the isolate database (Infant_Isolate_db_010611) and unmatched metagenome database (Gill db) (265) B.) Carrol samples: Myrimatch searches against metagenome only database and metagenome with isolate sequences appended. Twenty-one isolate sequences previously identified in the human gastrointestinal tract were obtained from JGI (see materials and methods section of chapter 5 for details).

_IgAM_20) was constructed (results shown in chapter six) and used as the final search database. This highlights the importance of using metagenomic sequences from matched samples (with the inclusion of less dominant organisms). Likewise, for the Carrol samples, a metagenomic database was generated which contained more highly resolved microbial species and strains (compared to the UC1 metagenome). When these samples were searched against a database containing the metagenomes only, then results compared with searches against a database containing the metagenome plus twenty-one isolate sequences, the results were not improved upon addition of the isolate sequences (Figure 7.5B). Again, this highlights the importance of using high-quality matched genomes for generating predicted protein database.

Improvement with Next-generation mass spectrometers

Also during the course of method development for this project, we were fortunate to be able to purchase a next-generation, high-performance mass spectrometer: the LTQ-Orbitrap Velos. The same sample, from infant #64 (UC1 baby) collected on day 21 after birth, was ran on two different instruments for comparison: the LTQ-Orbitrap XL and the LTQ-Orbitrap Velos. Not surprisingly, protein identifications and abundances improved dramatically with the use of the higher-end Velos instrument (Figure 7.6). (For details on the differences between the LTQ-Orbitrap XL and the LTQ-Orbitrap Velos, see materials and methods chapter two.)

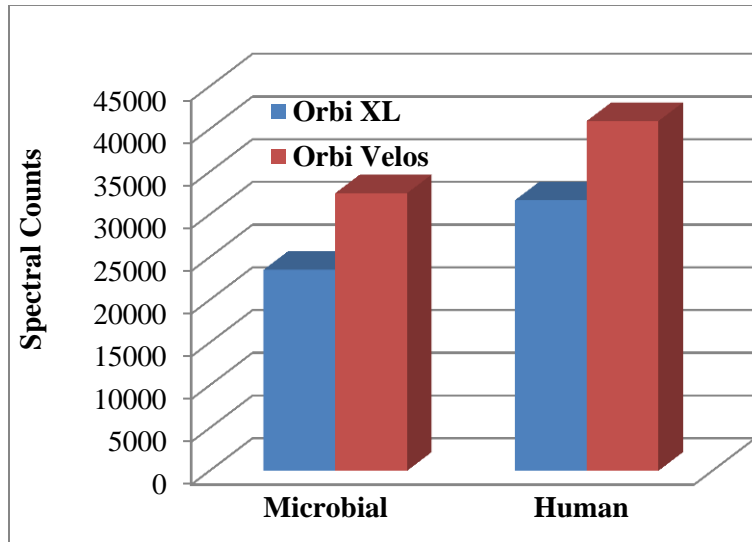


Figure 7.6: Greater proteome Depth Achieved with Next-Generation Velos Instrument. Sample #6421 (infant #64 on day 21 after birth) was run on both instruments and total human and microbial spectra are shown for comparison. Searched using DBDigger against the isolate database.

Attempts to Deplete Abundant Human Proteins: Mass Exclusion List & Differential Centrifugation

As described previously, some infant fecal samples contained a high abundance of human proteins, and initially there was some concern as to whether the human proteins would dominate the samples so much that we would not be able to sufficiently measure the microbial proteins at a deep enough level to extract useful biological information. In addition to the number of proteins, the abundance of human proteins is a concern since the mass spectrometer could be spending all of its time measuring spectra from an abundant human protein and may be missing a less dominant microbial protein buried underneath the peak. For example, in the UC1 baby, the most abundant human protein is the IgG Fc receptor binding protein (FCGBP), with 20,000 spectral counts collected in

one run. In contrast, the most abundant microbial protein only comprised around 500 spectra in the same run. Therefore, several experimental trials were attempted to circumvent this problem. The first entailed applying a mass exclusion list such that peptide masses from the most dominant human proteins (FCGBP, Calcium activated chloride channel regulator, alpha-1 antitrypsin) were programmed into the mass spectrometer excluding them from being measured. Secondly, a differential centrifugation was applied during sample preparation in attempts to deplete the human proteins. A similar method was used for metaproteomics in adult fecal samples (33), however infant fecal are limited by the amount raw material available, so modifications must be made for these significantly lower sample amounts. Unfortunately, however, neither applying a mass exclusion list, nor differential centrifugation improved the overall results in terms of significantly reducing the percent of human proteins or spectra measured (and in turn did not increase the percent of microbial protein identifications or spectra) (Table 7.5).

Table 7.5: Experimental Attempts at Depleting Abundant Human Proteins

	SDS-TCA Prep	Mass Exclusion List	Differential Centrifugation
Total microbial spectra	31599	23302	33240
Total human spectra	38903	26901	41643
Percent microbial spectra	45%	46%	44%
Percent human spectra	55%	54%	56%
Total microbial proteins	3025	2001	2287
Total human proteins	1830	1089	1167
Percent microbial proteins	62%	65%	66%
Percent human proteins	38%	35%	34%

Sample 6421(UC1 baby from day 21) run on the OrbiXL and searched with DBDigger against the isolate database (Infant_Isolate_db_010611).

Additional Attempts to Deplete Abundant Human Proteins in Fecal Samples

Next, we tried to fractionate the sample by protein size using a 50kDa molecular weight cutoff (MWCO) filter, based on the idea that many of the abundant human proteins were very large, especially in comparison to the microbial proteins. For example, the molecular weight of the abundant human proteins FCGBP and CLCA2 are 572kDa and 103kDa, respectively. Thus, following extraction, the proteins were passed through the 50kDa MWCO filter and the top and bottom fractions measured on the LTQ-Orbitrap Elite. We were hoping that the smaller proteins would pass through the filter, enriching the microbial proteins in the bottom fraction. However, this was not the case, since most of the proteins were identified from the top fraction (Figure 7.7B) and proteins above and below 50kDa were detected at comparable levels in both the top and bottom fractions

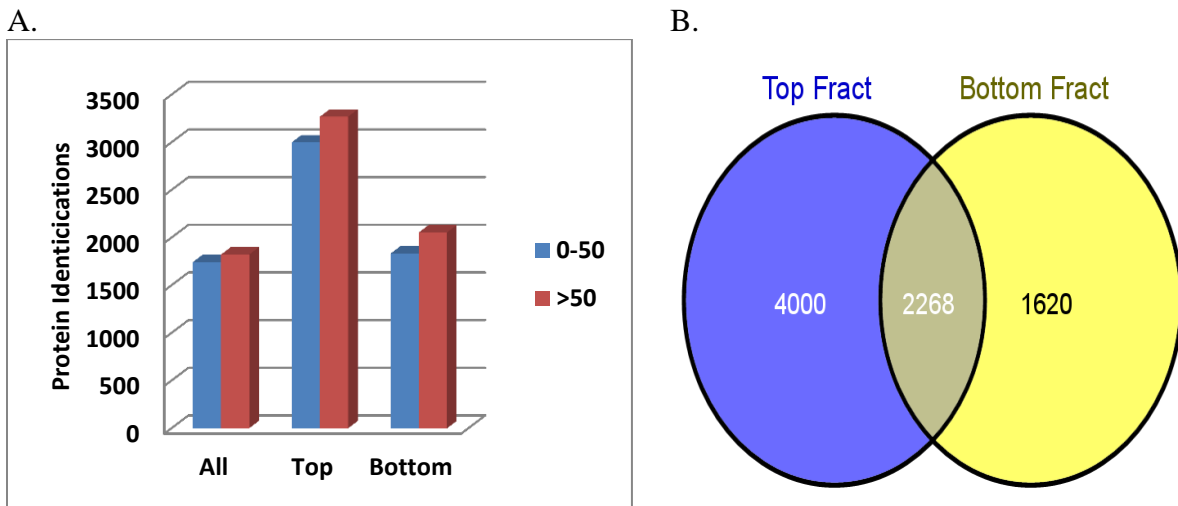


Figure 7.7: Attempt to Deplete Abundant Human Proteins using a 50kDa MWCO filter. A.) Protein identifications from an unfractionated sample (all), the top fraction from the filter, and bottom fraction. B.) Venn diagram showing the number of proteins only identified in the top or bottom fractions as well as those which overlap in both fractions.

(Figure 7.7A). Results from this experiment were confounding since a significantly higher number of the proteins were detected in the top fraction. This led us to speculate that there may possibly be some small molecules interfering with MS signal that are passing through during filtration.

So, our next idea was to try a different type of sample preparation, filter aided sample preparation (FASP) (266), in which proteins are captured on a filter, washed with buffer allowing small molecules to pass through, then tryptically digested on the filter. However, when three different samples were prepared using this method then compared with the SDS-TCA method, the number of protein identifications was not improved, nor was the ratio of human:microbial proteins. Thus, we reaffirmed that the SDS-TCA method was the optimal choice for processing infant fecal microbiome samples.

Table 7.6: Comparison of SDS-TCA and FASP Sample Prep Methods

	FILTERED SPECTRA	PEPTIDES	PROTEINS	% HUMAN PROTEINS
Unlabeled_SDSTCA	100703	11954	5133	11%
Unlabeled_FASP	74651	9510	4892	13%
UC1_Day21_SDSTCA	146388	11464	4031	43%
UC1_Day21_FASP	115557	8334	3403	47%
CARROL_Day23_SDSTCA	87314	5428	1416	58%
CARROL_Day23_FASP	54240	5039	1343	61%

Data searched using Myrimatch and ID Picker against the isolate database for Unlabeled sample, the metagenome database for the Carrol sample, and the UC1metagenome + isolates database.

In conclusion, utilizing a mass spectrometry- based metaproteomic approach to measure fecal microbiomes from multiple preterm infants revealed significant variability between and within individuals. This was apparent in terms of microbial memberships as well as for the types and relative abundances of human proteins. While attempts were made to improve the methodology and deplete abundant human proteins, the existing method proved robust and applicable to multiple infants.

CHAPTER EIGHT

Conclusions: Insights into microbial symbiotic interactions gained through metaproteomic investigations

Throughout this dissertation work, we have demonstrated that a mass spectrometry-based metaproteomic approach provides a robust and broadly applicable platform for studying microbial symbiotic interactions. This was highlighted in several symbiotic systems ranging from the basic genetic level with parasitic transposable elements, to single phage-bacteria interactions. We expanded this to microbial co-symbionts within an invertebrate host, then on to a more complex microbial commensalism in the human gut. From each of these studies, we were able to monitor the functional signatures of all members taking part in the symbioses. Acquiring deep proteomic measurements within each of these systems allowed extraction of valuable biological information on symbioses in general, as well as specific features from the various symbiotic relationships.

Characterization of microbial symbiotic interactions in past work has consisted of looking at one or two genes/proteins at a time, and provided useful, yet limited, information. Currently, in the systems biology era, whole genomes can now be characterized and their functional signatures measured via genomics and proteomics, respectively. However, while characterizing functions of a cell under defined laboratory conditions is valuable, the level of complexity is compounded when characterizing multiple symbiotic members whose functions are dictated and changing according to their role within the symbiotic relationship. Oftentimes, multiple members may have the

encoded genetic potential for carrying out certain metabolic activities, but the symbiotic system has evolved such that only one member carries out this function, because it is unnecessary for multiple members to do so while another member will contribute a different needed function. This information cannot be obtained exclusively from the genomic level, but rather requires functional measurements, which can be obtained through proteomics. In addition, obtaining *in situ* measurements from symbiotic microbes in their natural environment is crucial to unraveling true biological/functional roles of each member. Thus, metaproteomics fits this niche well. However, due to instrument cost and the level of technical expertise required, only limited metaproteomic studies on symbiotic systems have been undertaken. Through this dissertation work, we have been able to push the field forward by utilizing a robust metaproteomic platform to gain biological insights into various symbiotic interactions.

A primary contributing factor for this advancement has been the rapid developments in the field of mass spectrometry-based proteomics within recent years. In particular, within the time frame of this dissertation work, the innovation of new mass spectrometry instruments has progressed dramatically. In the beginning of this work, we started out measuring microbial samples and environmental samples on a basic linear trapping quadrupole (LTQ-XL). Then, we progressed to higher- end instruments such as the LTQ-Orbitrap, which was developed in 2005, to afford high mass accuracy measurements crucial to analyzing complex environmental samples. With the introduction of the next-generation instruments like the LTQ-Orbitrap-Velos in 2009, and the LTQ-Orbitrap Elite, just released in 2012, the increase in dynamic range and robustness allowed deeper proteome measurements, which would not have been possible

with earlier generation instruments, especially for complex community samples like those from the infant fecal microbiome (as demonstrated in previous chapters). Due to the available instrument capabilities in our laboratory, we have been able to generate valuable proteomic data necessary to categorize complex dynamic relationships between multiple members forming symbiotic associations.

Adding to the advancement of this field, genome sequencing capabilities within the last five years has increased prolifically, such that the number of genomes sequenced has increased exponentially, allowing better search databases for proteomics. Also, bioinformatic algorithms for metagenomic assembly and proteome data analysis tools have contributed to driving the field forward.

Information revealed about phage-bacteria symbiotic interactions:

To investigate phage-bacterial symbiosis, we set out to characterize the global anti-viral proteomic response across an infection time course. To this end, we obtained information about which Cas proteins are important in the CRISPR/Cas response, as well as the overall anti-viral proteomic response. Importantly, this was the first simultaneous measurement of phage and host proteins during infection. Many proteomic studies have characterized viral structural proteins (such as capsid and tail proteins) following virus purification or enrichment away from the host, but this only provides limited information about the virus-host symbiotic interaction.

While we were able to move the field forward by measuring not only phage structural proteins, but also phage proteins produced within the bacterial cell upon infection, our system only consisted of infection of a bacterial isolate with a single phage.

In future studies, it would be valuable to study multiple bacteria and co-infecting phage, especially when studying the CRISPR/Cas system where both the host and phage are constantly evolving. There have been genomic studies looking at SNPs in the CRISPR/Cas system (266, 267), and it would be interesting to see how this translates to the proteomic level. Since we have proven that a proteomics approach is technically robust, using this platform to investigate phage co-infections is the next logical step. However, this would require next-generation instruments to obtain the dynamic range necessary for investigating these complex interactions.

In addition, it would be interesting to see how viral and host proteomes are changing in environmental ecosystems. Again, this would require looking at both phage and host proteins during active infection cycles, in order to not only characterize the structural viral proteins, but also those produced within the bacterial cell. This raises the issue of adequately detecting those viral proteins, which are probably low in abundance, and contained in a complex matrix from the environment. While the field of viral metagenomics has rapidly progressed in recent years, obtaining viral metaproteomes has its own unique challenges. Viral metagenomic sequences are obtained from purifying/enriching virus like particles (VLPs) from environmental samples. However, measuring enriched VLPs by metaproteomics would only yield information about the phage structural proteins. One could compare these data with metaproteomic data collected from the whole microbial community, but detecting virus proteins from a community is difficult due to the dynamic range issue of dominant microbial proteins masking lower abundant viral proteins. Even though viral sequences make up a large portion of genomic data collected from environmental samples (oceans, AMD, etc.), their

proteins may be in lower abundance relative to the microbes, and there is no protein amplification method (like PCR for DNA) to help with this issue. Again, next-generation MS instruments would be necessary for these metaproteomics experiments and continued improvements to achieve deeper dynamic range are necessary (via sample preparation, instrumentation, etc.).

In addition, for more successful metaproteomics measurements of viruses, there needs to be improvement in building virus reference genome databases. With the majority of viruses yet undiscovered, and most virus genes not matching any sequences in the NCBI nonredundant database, metaproteomics data is virtually impossible to acquire without some type of *de novo* protein sequencing. This is also compounded by the fact that viruses mutate very rapidly. In metaproteomics measurements, the accuracy of peptide mass and sequences information is so specific that a single amino acid change would prevent identification of that protein. Another disadvantage of metagenomic analyses of enriched VLPs from the environment is that by extracting the DNA, this technique neglects the RNA viruses. It could be valuable to use a proteomics approach to look for these RNA viruses; however one would need a complete RNA virus reference genome database to search against. Currently, developments of such databases are underway, but they still have a long way to go.

Insights into symbioses in the human gut microbiome

Through-out this dissertation work, we have shown that metaproteomics can unravel insights into the commensalistic relationship between microbes and their human host. We used preterm infants as a model system, since among other factors the lower

microbial complexity (compared to adult fecal microbiomes) allowed reconstruction of high-quality microbial genomes. This, in combination with high-end MS instrumentation and bioinformatic analyses, allowed us to characterize not only the microbial proteins, but also human proteins in the preterm infant fecal microbiome. Thus, we were able to advance the knowledge base by obtaining insights into this symbiotic relationship between the human gut and its resident microbes. In addition, we demonstrated that this metaproteomics approach is applicable to multiple infants, and that a significant variability exists within and between different infants at the proteomic level.

Many questions remain, such as; what is a ‘normal’ or ‘healthy’ microbial colonization pattern in newborn infants, and given the drastic variability, how many infants will need to be measured to determine this? We have begun to answer the first question in this dissertation work; however, we will need to increase the number of infants studied in order to determine a baseline for healthy infants. We plan to use this metaproteomics approach to simultaneously monitor the microbial and human functional signatures. In addition, experimental plans are underway to study preterm infants who have developed necrotizing enterocolitis (NEC). Comparing the fecal metaproteomes of these infants with those of healthy infants may aid in determining factors leading to disease onset or pathogenicity.

A major challenge remaining in obtaining metaproteomic measurements from fecal microbiome samples is the dominance of human proteins (discussed in chapters six and seven). Even though the current SDS-TCA method proved successful at measuring microbial proteins in conjunction with human proteins, experimental methods for “digging deeper” into the microbial proteomes and measuring those lower abundance

proteins, is necessary for future experiments. Continued efforts are underway to explore options for depleting human proteins/ enriching the microbial proteins from fecal samples, which may be the major bottleneck in obtaining functional information about the microbial community members.

Investigators involved in the Human Microbiome Project (HMP) are rapidly collecting metagenomic data from multiple body sites including the mouth, vagina, lung, and skin. However, even though most of the current HMP studies are gene based, and will identify microbial community members or potentially important genes, there will still be a gap in the knowledge about how these members are functioning, which could be highly variable from the encoded potential of the microbes. Metaproteomics can fill this void. In addition, a key component of this commensalism, which is missing from most studies, is how the resident microbes are affecting their host. As discussed in earlier chapters, the enormous diversity between and within individuals at the microbial species level complicates the ability to determine what a ‘normal’ or ‘healthy’ microbiome looks like and this makes comparing microbiomes from ‘sick’ individuals and determining causality difficult. Again, a missing key component may not be what the members are, but rather what they are doing, and most importantly, how the human host is responding to them. This highlights the increasing need for metaproteomic studies in adult samples from multiple body sites.

A key to improving metaproteomic analyses in the human microbiome (gut and other body sites) will be getting high-quality genomic information to compile search databases. While high-throughput sequencing provides metagenomic data at a rapid pace, the complexity of the microbial communities in the gut (especially the adult gut),

and limited bioinformatics tools to bin and reconstruct whole genomes reconstruction from metagenomic data, limit the ability to achieve optimal resolution of microbial species and strains within a community. Without binning and assembling genomes, only information at community level but not at species and strain level (such as niche partitioning, ecological divergence, etc.) can be obtained. While this is challenging in more complex samples such as the adult fecal microbiome, the stage has been set in starting with lower complexity communities like the infant gut, and is beginning to translate to larger communities where an increasing number of whole genomes are being reconstructed (80+ genomes from an environmental sample: data unpublished from the Banfield lab).

Outlook

The field of mass spectrometry-based proteomics/ metaproteomics has experienced rapid dramatic advances within recent years. However, we have just begun to scratch the surface of what can be done. With continued improvements in bioinformatic tools, sample preparation methods, and rapidly developing mass spectrometers (with improved performance, sensitivity, robustness, dynamic range, and throughput), added to this already robust platform, the level of measurements will allow thorough and insightful biological information to be gained. Thus, metaproteomics can and should continue to be utilized as a vital approach to looking at symbiotic interactions. I fully anticipate that within the next five to ten years, metaproteomics will allow complete characterization of the majority of the members within complex microbial consortia, uncovering ground breaking insights into microbial symbiotic associations.

LIST OF REFERENCES

1. Paracer S & Ahmadjian V (2000) *Symbiosis: an introduction to biological associations* (Oxford University Press, USA).
2. Moran NA & Plague GR (2004) Genomic changes following host restriction in bacteria. *Curr. Opin. Genet. Dev.* 14(6):627-633.
3. Wang Z, Gerstein M, & Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10(1):57-63.
4. Roux A, Lison D, Junot C, & Heilier JF (2011) Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review. *Clinical biochemistry* 44(1):119-135.
5. Yates JR, Ruse CI, & Nakorchevsky A (2009) Proteomics by mass spectrometry: approaches, advances, and applications. *Annual review of biomedical engineering* 11:49-79.
6. Tabb DL, McDonald WH, & Yates III JR (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *Journal of proteome research* 1(1):21-26.
7. Morowitz MJ, *et al.* (2011) Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proc Natl Acad Sci U S A* 108(3):1128-1133.
8. Fenn JB, Mann M, Meng CK, Wong SF, & Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246(4926):64-71.
9. Washburn MP, Wolters D, & Yates JR, 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 19(3):242-247.
10. VerBerkmoes NC, *et al.* (2002) Integrating “top-down” and “bottom-up” mass spectrometric approaches for proteomic analysis of *Shewanella oneidensis*. *Journal of proteome research* 1(3):239-252.
11. Tabb DL, Narasimhan C, Strader MB, & Hettich RL (2005) DBDigger: reorganized proteomic database identification that improves flexibility and speed. *Analytical chemistry* 77(8):2464-2474.
12. Tabb DL, Fernando CG, & Chambers MC (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of proteome research* 6(2):654-661.
13. Hu Q, *et al.* (2005) The Orbitrap: a new mass spectrometer. *Journal of mass spectrometry* 40(4):430-443.
14. VerBerkmoes NC, *et al.* (2006) Determination and Comparison of the Baseline Proteomes of the Versatile Microbe *Rhodospseudomonas palustris* under Its Major Metabolic States. *Journal of proteome research* 5(2):287-298.
15. Ma ZQ, *et al.* (2009) IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *Journal of proteome research* 8(8):3872.
16. Eng JK, McCormack, A.L., Yates III J.R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Mass Spectrom* 5:976-989.
17. 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(1):21-26.

18. Wolters DA, Washburn MP, & Yates III JR (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Analytical chemistry* 73(23):5683-5690.
19. VerBerkmoes NC, *et al.* (2002) Intact protein analysis for site-directed mutagenesis overexpression products: plasmid-encoded R67 dihydrofolate reductase. *Anal Biochem* 305(1):68-81.
20. VerBerkmoes NC, *et al.* (2006) Determination and comparison of the baseline proteomes of the versatile microbe *Rhodospseudomonas palustris* under its major metabolic states. *J Proteome Res* 5(2):287-298.
21. 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. *Journal of proteome research* 2(1):43-50.
22. Flo TH, *et al.* (2004) Lipocalin 2 mediates an innate immune response to bacterial infection by sequestering iron. *Nature* 432(7019):917-921.
23. Vaishnav S, Behrendt CL, Ismail AS, Eckmann L, & Hooper LV (2008) Paneth cells directly sense gut commensals and maintain homeostasis at the intestinal host-microbial interface. *Proceedings of the National Academy of Sciences* 105(52):20858-20863.
24. Thompson MR, *et al.* (2007) Dosage-dependent proteome response of *Shewanella oneidensis* MR-1 to acute chromate challenge. *J Proteome Res* 6(5):1745-1757.
25. Abraham P, *et al.* (2012) Defining the Boundaries and Characterizing the Landscape of Functional Genome Expression in Vascular Tissues of *Populus* using Shotgun Proteomics. *Journal of proteome research*.
26. Young JC, *et al.* (2012) Phage-Induced Expression of CRISPR-Associated Proteins Is Revealed by Shotgun Proteomics in *Streptococcus thermophilus*. *PLoS One* 7(5):e38077.
27. Pütsep K, *et al.* (2000) Germ-free and colonized mice generate the same products from enteric prodefensins. *Journal of Biological Chemistry* 275(51):40478-40482.
28. Ram RJ, *et al.* (2005) Community proteomics of a natural microbial biofilm. *Science* 308(5730):1915-1920.
29. Lo I, *et al.* (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* 446(7135):537-541.
30. Deneff VJ, *et al.* (2010) Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proc Natl Acad Sci U S A* 107(6):2383-2390.
31. Mueller RS, *et al.* (2010) Ecological distribution and population physiology defined by proteomics in a natural microbial community. *Mol Syst Biol* 6:374.
32. Goltsman DS, *et al.* (2009) Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing "*Leptospirillum rubrum*" (Group II) and "*Leptospirillum ferrodiazotrophum*" (Group III) bacteria in acid mine drainage biofilms. *Appl Environ Microbiol* 75(13):4599-4615.
33. Verberkmoes NC, *et al.* (2009) Shotgun metaproteomics of the human distal gut microbiota. *ISME J* 3(2):179-189.

34. Giannone RJ, *et al.* (2011) Proteomic characterization of cellular and molecular processes that enable the Nanoarchaeum equitans-Ignicoccus hospitalis relationship. *PLoS One* 6(8):e22942.
35. Chourey K, *et al.* (2010) A Direct Cellular Lysis/Protein Extraction Protocol for Soil Metaproteomics. *Journal of proteome research*.
36. Oliver KM, Degan PH, Hunter MS, & Moran NA (2009) Bacteriophages encode factors required for protection in a symbiotic mutualism. *Science* 325(5943):992-994.
37. Boyd EF & Brüssow H (2002) Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *TRENDS in Microbiology* 10(11):521-529.
38. Touchon M & Rocha EPC (2007) Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.* 24(4):969-981.
39. Brussow H (2001) Phages of dairy bacteria. *Annu Rev Microbiol* 55:283-303.
40. Sanger F, *et al.* (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265(5596):687-695.
41. Valle J, Vergara-Irigaray M, Merino N, Penadés JR, & Lasa I (2007) σ^B regulates IS256-mediated *Staphylococcus aureus* biofilm phenotypic variation. *Journal of Bacteriology* 189(7):2886-2896.
42. Tatusov RL, Galperin MY, Natale DA, & Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* 28(1):33-36.
43. Minot S, Wu GD, Lewis JD, & Bushman FD (2012) Conservation of gene cassettes among diverse viruses of the human gut. *PLoS One* 7(8):e42342.
44. Tyson GW & Banfield JF (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* 10(1):200-207.
45. Maxwell KL & Frappier L (2007) Viral proteomics. *Microbiology and Molecular Biology Reviews* 71(2):398-411.
46. Naryshkina T, *et al.* (2006) Thermus thermophilus bacteriophage phiYS40 genome and proteomic characterization of virions. *J Mol Biol* 364(4):667-677.
47. Lévesque C, *et al.* (2005) Genomic organization and molecular analysis of virulent bacteriophage 2972 infecting an exopolysaccharide-producing *Streptococcus thermophilus* strain. *Appl Environ Microbiol* 71(7):4057-4068.
48. Roberts D, Hoopes BC, McClure WR, & Kleckner N (1985) IS10 transposition is regulated by DNA adenine methylation. *Cell* 43(1):117-130.
49. Chaston J & Douglas A (2012) Making the Most of “Omics” for Symbiosis Research. *The Biological Bulletin* 223(1):21-29.
50. Reyes A, Semenkovich NP, Whiteson K, Rohwer F, & Gordon JI (2012) Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat Rev Microbiol* 10(9):607-617.
51. Cavanaugh CM, McKiness Z, Newton ILG, & Stewart FJ (2006) Marine chemosynthetic symbioses. *The prokaryotes* 1:475-507.
52. Kleiner M, Petersen JM, & Dubilier N (2012) Convergent and divergent evolution of metabolism in sulfur-oxidizing symbionts and the role of horizontal gene transfer. *Current Opinion in Microbiology* in press.

53. Dubilier N, Bergin C, & Lott C (2008) Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. *Nat Rev Microbiol* 6(10):725-740.
54. Dubilier N, Blazejak A, & Ruhland C (2006) Symbioses between bacteria and gutless marine oligochaetes. *Progress in molecular and subcellular biology* 41:251-275.
55. Woyke T, *et al.* (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* 443(7114):950-955.
56. Kleiner M, *et al.* (2012) Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proc Natl Acad Sci U S A* 109(19):E1173-1182.
57. Markert S, *et al.* (2007) Physiological proteomics of the uncultured endosymbiont of *Riftia pachyptila*. *Science* 315(5809):247-250.
58. Stappenbeck TS, Hooper LV, & Gordon JI (2002) Developmental regulation of intestinal angiogenesis by indigenous microbes via Paneth cells. *Proceedings of the National Academy of Sciences* 99(24):15451.
59. Hooper LV, Midtvedt T, & Gordon JI (2002) How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annual review of nutrition* 22(1):283-307.
60. MacDonald TT & Pettersson S (2000) Bacterial regulation of intestinal immune responses. *Inflammatory Bowel Diseases* 6(2):116-122.
61. Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, & Gordon JI (2005) Host-bacterial mutualism in the human intestine. *Science* 307(5717):1915.
62. Turnbaugh PJ, *et al.* (2008) A core gut microbiome in obese and lean twins. *Nature* 457(7228):480-484.
63. Hooper LV & Macpherson AJ (2010) Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nature Reviews Immunology* 10(3):159-169.
64. Ley RE, Turnbaugh PJ, Klein S, & Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. *Nature* 444(7122):1022-1023.
65. Eckburg PB, *et al.* (2005) Diversity of the human intestinal microbial flora. *Science* 308(5728):1635.
66. Frank DN, *et al.* (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A* 104(34):13780-13785.
67. Ley RE, *et al.* (2005) Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America* 102(31):11070.
68. Vijay-Kumar M, *et al.* (2010) Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science Signalling* 328(5975):228.
69. Johansson MEV, Larsson JMH, & Hansson GC (2011) The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. *Proceedings of the National Academy of Sciences* 108(Supplement 1):4659.
70. Cantarel BL, *et al.* (2011) Strategies for Metagenomic-Guided Whole-Community Proteomics of Complex Microbial Environments. *PLoS One* 6(11):e27173.
71. Palmer C, Bik EM, DiGiulio DB, Relman DA, & Brown PO (2007) Development of the human infant intestinal microbiota. *PLoS biology* 5(7):e177.

72. Yatsunenکو T, *et al.* (2012) Human gut microbiome viewed across age and geography. *Nature* 486(7402):222-227.
73. Marques TM, *et al.* (2010) Programming infant gut microbiota: influence of dietary and environmental factors. *Current Opinion in Biotechnology* 21(2):149-156.
74. Palmer C, Bik EM, DiGiulio DB, Relman DA, & Brown PO (2007) Development of the human infant intestinal microbiota. *PLoS biology* 5(7):e177.
75. Koenig JE, *et al.* (2011) Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences* 108(Supplement 1):4578.
76. Klaassens ES, De Vos WM, & Vaughan EE (2007) Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. *Applied and environmental microbiology* 73(4):1388-1392.
77. CLAUD EC & WALKER WA (2001) Hypothesis: inappropriate colonization of the premature intestine can cause neonatal necrotizing enterocolitis. *The FASEB Journal* 15(8):1398.
78. de la Cochetière MF, *et al.* (2004) Early intestinal bacterial colonization and necrotizing enterocolitis in premature infants: the putative role of Clostridium. *Pediatric research* 56(3):366-370.
79. Morowitz MJ, Poroyko V, Caplan M, Alverdy J, & Liu DC (2010) Redefining the role of intestinal microbes in the pathogenesis of necrotizing enterocolitis. *Pediatrics* 125(4):777-785.
80. Lin PW & Stoll BJ (2006) Necrotising enterocolitis. *The Lancet* 368(9543):1271-1283.
81. Hunter CJ, Upperman JS, Ford HR, & Camerini V (2008) Understanding the susceptibility of the premature infant to necrotizing enterocolitis (NEC). *Pediatric research* 63(2):117-123.
82. Doolittle WF & Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284:601-603.
83. Werren JH (2011) Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proceedings of the National Academy of Sciences* 108(Supplement 2):10863-10870.
84. Nagy Z & Chandler M (2004) Regulation of transposition in bacteria. *Res. Microbiol.* 155(5):387-398.
85. Schmid AK, *et al.* (2005) Global whole-cell FTICR mass spectrometric proteomics analysis of the heat shock response in the radioresistant bacterium *Deinococcus radiodurans*. *Journal of Proteome Research* 4(3):709-718.
86. Goodchild A, Raftery M, Saunders NFW, Guilhaus M, & Cavicchioli R (2004) Biology of the cold adapted archaeon, *Methanococoides burtonii* determined by proteomics using liquid chromatography-tandem mass spectrometry. *Journal of Proteome Research* 3(6):1164-1176.
87. Thompson MR, *et al.* (2008) Experimental approach for deep proteome measurements from small-scale microbial biomass samples. *Analytical chemistry* 80(24):9517-9525.
88. Lochner A, *et al.* (2011) Use of Label-Free Quantitative Proteomics To Distinguish the Secreted Cellulolytic Systems of *Caldicellulosiruptor bescii* and

- Caldicellulosiruptor obsidiansis. *Applied and environmental microbiology* 77(12):4042-4054.
89. Yang S, *et al.* (2012) Clostridium thermocellum ATCC27405 transcriptomic, metabolomic and proteomic profiles after ethanol stress. *BMC Genomics* 13(1):336.
 90. Wi & sacute JR (2009) Universal sample preparation method for proteome analysis. *Nature methods* 6(5):359-362.
 91. Smith PK (1989) Measurement of protein using bicinchoninic acid. (Google Patents).
 92. Horvath CG & Lipsky SR (1967) Peak capacity in chromatography. *Analytical chemistry* 39(14):1893-1893.
 93. Karas M, Bahr U, & Dülcks T (2000) Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine. *Fresenius' Journal of Analytical Chemistry* 366(6):669-676.
 94. Schwartz JC, Senko MW, & Syka JEP (2002) A two-dimensional quadrupole ion trap mass spectrometer. *Journal of the American Society for Mass Spectrometry* 13(6):659-669.
 95. Olsen JV, *et al.* (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Molecular & Cellular Proteomics* 8(12):2759-2769.
 96. Michalski A, *et al.* (2012) Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Molecular & Cellular Proteomics* 11(3).
 97. Eng JK, McCormack AL, & Yates JR (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(11):976-989.
 98. 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(14):4193-4201.
 99. Zhang B, *et al.* (2006) Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res* 5(11):2909-2918.
 100. Florens L, *et al.* (2006) Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* 40(4):303-311.
 101. Barrangou R, *et al.* (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315(5819):1709-1712.
 102. van der Oost J, Jore MM, Westra ER, Lundgren M, & Brouns SJ (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem Sci* 34(8):401-407.
 103. Horvath P & Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science* 327(5962):167-170.
 104. Deveau H, Garneau, J.E and Sylvain Moineau (2010) CRISPR/Cas System and Its Role in Phage-Bacteria Interactions. *Annual Review of Microbiology* 64:475-493.
 105. Grissa I, Vergnaud G, & Pourcel C (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* 8:172.

106. Haft DH, Selengut J, Mongodin EF, & Nelson KE (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol* 1(6):e60.
107. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, & Koonin EV (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* 1:7.
108. Jansen R, van Embden JD, Gaastra W, & Schouls LM (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* 43(6):1565-1575.
109. Deveau H, *et al.* (2008) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J Bacteriol* 190(4):1390-1400.
110. Marraffini LA & Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* 11(3):181-190.
111. Marraffini LA & Sontheimer EJ (2010) Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* 463(7280):568-571.
112. Hale CR, *et al.* (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* 139(5):945-956.
113. Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, & Doudna JA (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* 329(5997):1355-1358.
114. Brouns SJ, *et al.* (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* 321(5891):960-964.
115. Semenova E, *et al.* (2011) Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc Natl Acad Sci U S A*.
116. Wiedenheft B, *et al.* (2011) RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proceedings of the National Academy of Sciences* 108(25):10092.
117. Makarova KS, *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* 9(6):467-477.
118. Duplessis M, Russell WM, Romero DA, & Moineau S (2005) Global gene expression analysis of two *Streptococcus thermophilus* bacteriophages using DNA microarray. *Virology* 340(2):192-208.
119. Andersson AF & Banfield JF (2008) Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* 320(5879):1047-1050.
120. Horvath P, *et al.* (2009) Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int J Food Microbiol* 131(1):62-70.
121. Horvath P, *et al.* (2008) Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol* 190(4):1401-1412.
122. Sambrook J, Russell, D.W. ed (2001) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), 3rd Ed.
123. Washburn MP, Ulaszek R, Deciu C, Schieltz DM, & Yates JR, 3rd (2002) Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Anal Chem* 74(7):1650-1657.

124. McDonald W.H. OR, Miyamoto D.T., Mitchison T.J., & JR. YI (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 Journal of Mass Spectrometry* 219(1):245-251.
125. Li M, *et al.* (2010) Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. *Journal of proteome research*.
126. Thompson D, *et al.* (2010) Proteomics reveals a core molecular response of *Pseudomonas putida* F1 to acute chromate challenge. *BMC Genomics* 11(1):311.
127. Gagnaire V, Jardin J, Jan G, & Lortal S (2009) Invited review: Proteomics of milk and bacteria used in fermented dairy products: from qualitative to quantitative advances. *J Dairy Sci* 92(3):811-825.
128. Jore MM, *et al.* (2011) Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat Struct Mol Biol* 18(5):529-536.
129. Azcarate-Peril MA, *et al.* (2005) Microarray analysis of a two-component regulatory system involved in acid resistance and proteolytic activity in *Lactobacillus acidophilus*. *Applied and environmental microbiology* 71(10):5794.
130. Deltcheva E, *et al.* (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471(7340):602-607.
131. Garneau JE, *et al.* (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* 468(7320):67-71.
132. Sapranaukas R, *et al.* (2011) The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Research*.
133. Agari Y, *et al.* (2010) Transcription profile of *Thermus thermophilus* CRISPR systems after phage infection. *Journal of molecular biology* 395(2):270-281.
134. Shinkai A, *et al.* (2007) Transcription activation mediated by a cyclic AMP receptor protein from *Thermus thermophilus* HB8. *Journal of bacteriology* 189(10):3891.
135. Pul U, *et al.* (2010) Identification and characterization of *E. coli* CRISPR-cas promoters and their silencing by H-NS. *Mol Microbiol* 75(6):1495-1512.
136. Pougach K, *et al.* (2010) Transcription, processing and function of CRISPR cassettes in *Escherichia coli*. *Mol Microbiol* 77(6):1367-1379.
137. Lillestol RK, *et al.* (2009) CRISPR families of the crenarchaeal genus *Sulfolobus*: bidirectional transcription and dynamic properties. *Mol Microbiol* 72(1):259-272.
138. Van de Guchte M, *et al.* (2006) The complete genome sequence of *Lactobacillus bulgaricus* reveals extensive and ongoing reductive evolution. *Proceedings of the National Academy of Sciences* 103(24):9274.
139. Kleiner M, *et al.* (2012) Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proceedings of the National Academy of Sciences* 109(19):E1173–E1182.
140. Baumann P (2005) Biology of bacteriocyte-associated endosymbionts of plant sap-sucking insects. *Annu. Rev. Microbiol.* 59(1):155-189.
141. Felbeck H (1981) Chemoautotrophic potential of the hydrothermal vent tube worm, *Riftia pachyptila* Jones (Vestimentifera). *Science* 213(4505):336-338.

142. Cavanaugh CM, Gardiner SL, Jones ML, Jannasch HW, & Waterbury JB (1981) Prokaryotic cells in the hydrothermal vent tube worm *Riftia pachyptila* Jones: Possible chemoautotrophic symbionts. *Science* 213(4505):340-342.
143. Dubilier N, Bergin C, & Lott C (2008) Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. *Nat. Rev. Microbiol.* 6(10):725-740.
144. Dubilier N, Blazejak A, & Rühland C (2006) Symbiosis between bacteria and gutless marine oligochaetes. *Molecular Basis of Symbiosis*, Progress in Molecular and Subcellular Biology, ed Overmann J (Springer, Berlin Heidelberg), Vol 41, pp 251-275.
145. Dubilier N, *et al.* (2001) Endosymbiotic sulphate-reducing and sulphide-oxidizing bacteria in an oligochaete worm. *Nature* 411(6835):298-302.
146. Giere O & Erséus C (2002) Taxonomy and new bacterial symbioses of gutless marine Tubificidae (Annelida, Oligochaeta) from the Island of Elba (Italy). *Org. Divers. Evol.* 2(4):289-297.
147. Rühland C, *et al.* (2008) Multiple bacterial symbionts in two species of co-occurring gutless oligochaete worms from Mediterranean sea grass sediments. *Environ. Microbiol.* 10(12):3404-3416.
148. Woyke T, *et al.* (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* 443(7114):950-955.
149. Kleiner M, Woyke T, Rühland C, & Dubilier N (2011) The *Olavius algarvensis* metagenome revisited: lessons learned from the analysis of the low diversity microbial consortium of a gutless marine worm. *Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats*, ed Bruijn FJd (John Wiley & Sons, Inc., Hoboken, NJ, USA), Vol 2, pp 321-334.
150. Warnecke F & Hugenholtz P (2007) Building on basic metagenomics with complementary technologies. *Genome Biol.* 8(12):231.
151. Otto A, *et al.* (2010) Systems-wide temporal proteomic profiling in glucose-starved *Bacillus subtilis*. *Nat Commun.* 1(137):137. 10.1038/ncomms1137.
152. Washburn MP, Wolters D, & Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotech.* 19(3):242-247.
153. Verberkmoes NC, *et al.* (2009) Shotgun metaproteomics of the human distal gut microbiota. *ISME J.* 3:179-189.
154. Peng J, Elias JE, Thoreen CC, Licklider LJ, & Gygi SP (2002) 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(1):43-50. 10.1021/pr025556v.
155. Elias JE & Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Meth.* 4(3):207-214.
156. Florens L, *et al.* (2006) Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* 40(4):303-311.
157. Liebeke M, *et al.* (2011) A metabolomics and proteomics study of the adaptation of *Staphylococcus aureus* to glucose starvation. *Molecular BioSystems* 7(4):1241-1253.

158. King GM & Weber CF (2007) Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. *Nat. Rev. Microbiol.* 5(2):107-118.
159. Oelgeschläger E & Rother M (2008) Carbon monoxide-dependent energy metabolism in anaerobic bacteria and archaea. *Arch. Microbiol.* 190(3):257-269.
160. Conrad R, Meyer O, & Seiler W (1981) Role of Carboxydobacteria in Consumption of Atmospheric Carbon Monoxide by Soil. *Appl. Environ. Microbiol.* 42(2):211-215.
161. Moran MA, *et al.* (2004) Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature* 432(7019):910-913.
162. King GM (2007) Microbial carbon monoxide consumption in salt marsh sediments. *FEMS Microbiol. Ecol.* 59(1):2-9.
163. Moran JJ, House CH, Vrentas JM, & Freeman KH (2008) Methyl Sulfide Production by a Novel Carbon Monoxide Metabolism in *Methanosarcina acetivorans*. *Appl. Environ. Microbiol.* 74(2):540-542.
164. Thauer RK, Stackebrandt E, & Hamilton WA (2007) Energy metabolism and phylogenetic diversity of sulphate-reducing bacteria. *Sulphate-reducing Bacteria: Environmental and Engineered Systems*, eds Barton LL & Hamilton WA (Cambridge University Press, New York), pp 1-37.
165. Mörsdorf G, Frunzke K, Gadkari D, & Meyer O (1992) Microbial growth on carbon monoxide. *Biodegradation* 3:61-82.
166. Caffrey SM, *et al.* (2007) Function of Periplasmic Hydrogenases in the Sulfate-Reducing Bacterium *Desulfovibrio vulgaris* Hildenborough. *J. Bacteriol.* 189(17):6159-6167.
167. Goodwin S, Conrad R, & Zeikus JG (1988) Influence of pH on microbial hydrogen metabolism in diverse sedimentary ecosystems. *Appl. Environ. Microbiol.* 54(2):590-593.
168. Novelli PC, *et al.* (1988) Hydrogen and acetate cycling in two sulfate-reducing sediments: Buzzards Bay and Town Cove, Mass. *Geochimica et Cosmochimica Acta* 52(10):2477-2486.
169. Karadagli F & Rittmann B (2007) Thermodynamic and kinetic analysis of the H₂ threshold for *Methanobacterium bryantii* M.o.H. *Biodegradation* 18(4):439-452.
170. Petersen JM, *et al.* (2011) Hydrogen is an energy source for hydrothermal vent symbioses. *Nature* 476(7359):176-180.
171. Sowell SM, *et al.* (2009) Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J.* 3(1):93-105.
172. Rees DC, Johnson E, & Lewinson O (2009) ABC transporters: the power to change. *Nat. Rev. Mol. Cell Biol.* 10(3):218-227.
173. Forward JA, Behrendt MC, Wyborn NR, Cross R, & Kelly DJ (1997) TRAP transporters: a new family of periplasmic solute transport systems encoded by the dctPQM genes of *Rhodobacter capsulatus* and by homologs in diverse gram-negative bacteria. *J. Bacteriol.* 179(17):5482-5493.
174. Wick LM, Quadroni M, & Egli T (2001) Short- and long-term changes in proteome composition and kinetic properties in a culture of *Escherichia coli* during transition from glucose-excess to glucose-limited growth conditions in continuous culture and *vice versa*. *Environ. Microbiol.* 3(9):588-599.

175. Mauchline TH, *et al.* (2006) Mapping the *Sinorhizobium meliloti* 1021 solute-binding protein-dependent transportome. *Proc. Natl. Acad. Sci. USA* 103(47):17933-17938.
176. Ferenci T (1999) Regulation by nutrient limitation. *Curr. Opin. Microbiol.* 2(2):208-213.
177. Sowell SM, *et al.* (2010) Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. *ISME J.* Advance online publication. 10.1038/ismej.2010.168.
178. Kelly DJ & Thomas GH (2001) The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea. *FEMS Microbiol. Rev.* 25(4):405-424.
179. Ames GFL (1986) Bacterial Periplasmic Transport Systems: Structure, Mechanism, and Evolution. *Annu. Rev. Biochem.* 55(1):397-425.
180. Grieshaber MK, Hardewig I, Kreutzer U, & Pörtner H-O (1994) Physiological and metabolic responses to hypoxia in invertebrates. *Rev. Physiol. Biochem. Pharmacol.* 125:43-147.
181. Hellemond JJv, Klei Avd, Weelden SHv, & Tielens AGM (2003) Biochemical and evolutionary aspects of anaerobically functioning mitochondria. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 358(1429):205-215.
182. Zarzycki J, Brecht V, Müller M, & Fuchs G (2009) Identifying the missing steps of the autotrophic 3-hydroxypropionate CO₂ fixation cycle in *Chloroflexus aurantiacus*. *Proc. Natl. Acad. Sci. USA* 106(50):21317-21322.
183. Zarzycki J & Fuchs G (2011) Coassimilation of Organic Substrates via the Autotrophic 3-Hydroxypropionate Bi-Cycle in *Chloroflexus aurantiacus*. *Appl. Environ. Microbiol.* 77(17):6181-6188.
184. Nemecky SN-M (2008) Benthic degradation rates in shallow subtidal carbonate and silicate sands. Thesis for german Diplom in Biology (University of Bremen, Bremen).
185. Hua Q, Yang C, Oshima T, Mori H, & Shimizu K (2004) Analysis of Gene Expression in *Escherichia coli* in Response to Changes of Growth-Limiting Nutrient in Chemostat Cultures. *Appl. Environ. Microbiol.* 70(4):2354-2366.
186. Voigt B, *et al.* (2007) The glucose and nitrogen starvation response of *Bacillus licheniformis*. *Proteomics* 7(3):413-423.
187. Yancey PH (2005) Organic osmolytes as compatible, metabolic and counteracting cytoprotectants in high osmolarity and other stresses. *J. Exp. Biol.* 208(15):2819-2830.
188. Serrano A, Pérez-Castiñeira JR, Baltscheffsky M, & Baltscheffsky H (2007) H⁺-PPases: yesterday, today and tomorrow. *IUBMB Life* 59(2):76-83.
189. Liu CL & Peck HD, Jr. (1981) Comparative bioenergetics of sulfate reduction in *Desulfovibrio* and *Desulfotomaculum* spp. *J. Bacteriol.* 145(2):966-973.
190. Liu M-Y & Le Gall J (1990) Purification and characterization of two proteins with inorganic pyrophosphatase activity from *Desulfovibrio vulgaris* : Rubrerythrin and a new, highly active, enzyme. *Biochem. Biophys. Res. Commun.* 171(1):313-318.
191. Thebrath B, Dilling W, & Cypionka H (1989) Sulfate activation in *Desulfotomaculum*. *Arch. Microbiol.* 152(3):296-301.

192. Schöcke L & Schink B (1998) Membrane-bound proton-translocating pyrophosphatase of *Syntrophus gentianae*, a syntrophically benzoate-degrading fermenting bacterium. *Eur. J. Biochem.* 256(3):589-594.
193. Robidart JC, *et al.* (2008) Metabolic versatility of the *Riftia pachyptila* endosymbiont revealed through metagenomics. *Environ. Microbiol.* 10(3):727-737.
194. Newton ILG, *et al.* (2007) The *Calyptogenia magnifica* chemoautotrophic symbiont genome. *Science* 315(5814):998-1000.
195. Kuwahara H, *et al.* (2007) Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, *Calyptogenia okutanii*. *Curr. Biol.* 17(10):881-886.
196. Reshetnikov AS, *et al.* (2008) Characterization of the pyrophosphate-dependent 6-phosphofructokinase from *Methylococcus capsulatus* Bath. *FEMS Microbiol. Lett.* 288(2):202-210.
197. Baltscheffsky H (1996) Energy conversion leading to the origin and early evolution of life: Did inorganic pyrophosphate precede adenosine triphosphate? *Origin and Evolution of Biological Energy Conversion*, ed Baltscheffsky H (Wiley-VCH, New York), pp 1-9.
198. Say RF & Fuchs G (2010) Fructose 1,6-bisphosphate aldolase/phosphatase may be an ancestral gluconeogenic enzyme. *Nature* 464(7291):1077-1081.
199. Baltscheffsky H (1967) Inorganic pyrophosphate and the evolution of biological energy transformation. *Acta Chem. Scand.* 21(7):1973-1974.
200. Miller SL & Parris M (1964) Synthesis of Pyrophosphate Under Primitive Earth Conditions. *Nature* 204(4965):1248-1250.
201. Turnbaugh PJ, *et al.* (2007) The Human Microbiome Project. *Nature* 449(7164):804-810.
202. Yin B, Crowley D, Sparovek G, De Melo WJ, & Borneman J (2000) Bacterial Functional Redundancy along a Soil Reclamation Gradient. *Appl. Environ. Microbiol.* 66(10):4361-4365.
203. Denev VJ, *et al.* (2010) Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proc. Natl. Acad. Sci. USA* 107(6):2383-2390.
204. Wilmes P, *et al.* (2008) Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J.* 2:853-864.
205. Konstantinidis KT, *et al.* (2009) Comparative systems biology across an evolutionary gradient within the *Shewanella* genus. *Proc. Natl. Acad. Sci. USA* 106(37):15909-15914.
206. Kleiner M, Young, Jacques C., Shah, Manesh, Verberkmoes, Nathan C., and Dubilier, Nicole (Metaproteomics Reveals Abundant Transposase Expression in Mutualistic Endosymbionts. *mBio* submitted.
207. Plague GR, *et al.* (2011) Relaxed natural selection alone does not permit transposable element expansion within 4,000 generations in *Escherichia coli*. *Genetica* 139(7):895-902.
208. Aziz RK, Breitbart M, & Edwards RA (2010) Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* 38(13):4207-4217.

209. Doolittle WF & Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284(5757):601-603.
210. Werren JH (2011) Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proc. Natl. Acad. Sci. USA.* 108(Supplement 2):10863-10870. 10.1073/pnas.1102343108.
211. Mahillon J & Chandler M (1998) Insertion Sequences. *Microbiol. Mol. Biol. Rev.* 62(3):725-774.
212. Frost LS, Leplae R, Summers AO, & Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* 3(9):722-732.
213. Leavis HL, *et al.* (2007) Insertion sequence-driven diversification creates a globally dispersed emerging multiresistant subspecies of *E. faecium*. *PLoS Pathogens.* 3(1):e7. 10.1371/journal.ppat.0030007.
214. Edwards RJ & Brookfield JFY (2003) Transiently beneficial insertions could maintain mobile DNA sequences in variable environments. *Mol. Biol. Evol.* 20(1):30-37.
215. Chao L & McBroom SM (1985) Evolution of transposable elements: an IS10 insertion increases fitness in *Escherichia coli*. *Mol. Biol. Evol.* 2(5):359-369.
216. Doolittle WF, Kirkwood TBL, & Dempster MAH (1984) Selfish DNAs with self-restraint. *Nature* 307(5951):501-502.
217. Song H, *et al.* (2010) The early stage of bacterial genome-reductive evolution in the host. *PLoS Pathog.* 6(5):e1000922. 10.1371/journal.ppat.1000922.
218. Yang F, *et al.* (2005) Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res.* 33(19):6445-6458.
219. Ochiai H, Inoue Y, Takeya M, Sasaki A, & Kaku H (2005) Genome sequence of *Xanthomonas oryzae* pv. *oryzae* suggests contribution of large numbers of effector genes and insertion sequences to its race diversity. *Japan Agricultural Research Quarterly* 39(4):275-287.
220. Thomas T, *et al.* (2010) Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *ISME J.* 4:1557–1567.
221. Plague GR, Dunbar HE, Tran PL, & Moran NA (2008) Extensive proliferation of transposable elements in heritable bacterial symbionts. *J. Bacteriol.* 190(2):777-779.
222. Plague G, *et al.* (2011) Relaxed natural selection alone does not permit transposable element expansion within 4,000 generations in *Escherichia coli*. *Genetica* 139(7):895-902.
223. Kleiner M, *et al.* (2012) Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proc. Natl. Acad. Sci. USA* 109(19):7148-7149.
224. Woyke T, *et al.* (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* 443(7114):950-955.
225. Siguier P, Perochon J, Lestrade L, Mahillon J, & Chandler M (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucl. Acids Res.* 34(suppl_1):D32-36.
226. Chandler M & Fayet O (1993) Translational frameshifting in the control of transposition in bacteria. *Mol. Microbiol.* 7(4):497-503.

227. Williams D, *et al.* (2009) Implications of high level pseudogene transcription in *Mycobacterium leprae*. *BMC Genomics*. 10(1):397. 10.1186/1471-2164-10-397.
228. Schmitz-Esser S, Penz T, Spang A, & Horn M (2011) A bacterial genome in transition - an exceptional enrichment of IS elements but lack of evidence for recent transposition in the symbiont *Amoebophilus asiaticus*. *BMC Evol. Biol.* 11(1):270. 10.1186/1471-2148-11-270.
229. Bickhart D & Benson D (2011) Transcriptomes of *Frankia* sp. strain CcI3 in growth transitions. *BMC Microbiol.* 11(1):192. 10.1186/1471-2180-11-192.
230. Mitchell HL, *et al.* (2010) *Treponema denticola* biofilm-induced expression of a bacteriophage, toxin-antitoxin systems and transposases. *Microbiology* 156(3):774-788.
231. Egas C, Pinheiro M, Gomes P, Barroso C, & Bettencourt R (2012) The transcriptome of *Bathymodiolus azoricus* gill reveals expression of genes from endosymbionts and free-living deep-sea bacteria. *Marine Drugs* 10(8):1765-1783.
232. Radax R, *et al.* (2012) Metatranscriptomics of the marine sponge *Geodia barretti*: tackling phylogeny and function of its microbial community. *Environ. Microbiol.* 14(5):1308–1324.
233. Soltis DE & Soltis PS (1999) Polyploidy: recurrent formation and genome evolution. *Trends Ecol. Evol.* 14(9):348-352.
234. Salman V, Amann R, Shub DA, & Schulz-Vogt HN (2012) Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria. *Proc. Natl. Acad. Sci. USA* 109(11):4203-4208.
235. Slade D, Lindner AB, Paul G, & Radman M (2009) Recombination and replication in DNA repair of heavily irradiated *Deinococcus radiodurans*. *Cell* 136(6):1044-1055.
236. Ohtani N, Tomita M, & Itaya M (2010) An extreme thermophile, *Thermus thermophilus*, is a polyploid bacterium. *J. Bacteriol.* 192(20):5499-5505.
237. Caro A, Gros O, Got P, De Wit R, & Troussellier M (2007) Characterization of the population of the sulfur-oxidizing symbiont of *Codakia orbicularis* (Bivalvia, Lucinidae) by single-cell analyses. *Appl. Environ. Microbiol.* 73(7):2101-2109.
238. Mshvildadze M, *et al.* (2010) Intestinal microbial ecology in premature infants assessed with non-culture-based techniques. *The Journal of pediatrics* 156(1):20-25.
239. Wang Y, *et al.* (2009) 16S rRNA gene-based analysis of fecal microbiota from preterm infants with and without necrotizing enterocolitis. *The ISME journal* 3(8):944-954.
240. Knief C, *et al.* (2012) Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. *ISME J* 6(7):1378-1390.
241. Hettich RL, Sharma R, Chourey K, & Giannone RJ (2012) Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. *Current Opinion in Microbiology*.
242. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*.

243. Zybilov B, *et al.* (2006) Statistical Analysis of Membrane Proteome Expression Changes in *Saccharomyces cerevisiae*. *Journal of proteome research* 5(9):2339-2347.
244. Anand RJ, Leaphart CL, Mollen KP, & Hackam DJ (2007) The role of the intestinal barrier in the pathogenesis of necrotizing enterocolitis. *Shock* 27(2):124.
245. Dharmani P, Srivastava V, Kissoon-Singh V, & Chadee K (2008) Role of intestinal mucins in innate host defense mechanisms against pathogens. *Journal of Innate Immunity* 1(2):123-135.
246. Sheng YH, *et al.* (2011) The MUC13 cell-surface mucin protects against intestinal inflammation by inhibiting epithelial cell apoptosis. *Gut*.
247. Denson LA (2010) Immune Development and Inflammatory Bowel Disease. *The Changing World of Inflammatory Bowel Disease: Impact of Generation, Gender, and Global Trends*:33.
248. Barbosa T & Rescigno M (2010) Host-bacteria interactions in the intestine: homeostasis to chronic inflammation. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 2(1):80-97.
249. Mashimo H, Wu DC, Podolsky DK, & Fishman MC (1996) Impaired defense of intestinal mucosa in mice lacking intestinal trefoil factor. *Science* 274(5285):262.
250. Macpherson AJ & Uhr T (2004) Induction of protective IgA by intestinal dendritic cells carrying commensal bacteria. *Science's STKE* 303(5664):1662.
251. Macpherson AJ, *et al.* (2000) A primitive T cell-independent mechanism of intestinal mucosal IgA responses to commensal bacteria. *Science* 288(5474):2222-2226.
252. Suzuki K, *et al.* (2004) Aberrant expansion of segmented filamentous bacteria in IgA-deficient gut. *Proceedings of the National Academy of Sciences of the United States of America* 101(7):1981.
253. Qiu J, *et al.* (1998) Human milk lactoferrin inactivates two putative colonization factors expressed by *Haemophilus influenzae*. *Proceedings of the National Academy of Sciences* 95(21):12641.
254. Singh PK, Parsek MR, Greenberg EP, & Welsh MJ (2005) A component of innate immunity prevents bacterial biofilm development. *development* 417(6888):552-555.
255. Legrand D, *et al.* (2008) Lactoferrin structure and functions. *Bioactive Components of Milk*:163-194.
256. Rodriguez-Pineiro AM, *et al.* (2012) Proteomic study of the mucin granulae in an intestinal goblet cell model. *Journal of proteome research*.
257. Vaishnav S, *et al.* (2011) The Antibacterial Lectin RegIII {gamma} Promotes the Spatial Segregation of Microbiota and Host in the Intestine. *Science's STKE* 334(6053):255.
258. Fu J, *et al.* (2011) Loss of intestinal core 1-derived O-glycans causes spontaneous colitis in mice. *The Journal of Clinical Investigation* 121(4):1657.
259. Moehle C, *et al.* (2006) Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease. *Journal of Molecular Medicine* 84(12):1055-1066.
260. Johansson MEV, Thomsson KA, & Hansson GC (2009) Proteomic analyses of the two mucus layers of the colon barrier reveal that their main component, the

- Muc2 mucin, is strongly bound to the Fcgbp protein. *Journal of proteome research* 8(7):3549-3557.
261. Tabb DL, Narasimhan C, Strader MB, & Hettich RL (2005) DBDigger: reorganized proteomic database identification that improves flexibility and speed. *Anal Chem* 77(8):2464-2474.
 262. Sharon I, *et al.* (2012) Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.*
 263. Tabb DL, Fernando CG, & Chambers MC (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* 6(2):654-661.
 264. Gill SR, *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312(5778):1355-1359.
 265. Gill SR, *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312(5778):1355.
 266. Pride DT, *et al.* (2011) Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. *Genome Research* 21(1):126-136.
 267. Weinberger AD, *et al.* (2012) Persisting viral sequences shape microbial CRISPR-based Immunity. *PLoS computational biology* 8(4):e1002475.

Vita

Jacque Caprio Young was born in Lock Haven, PA, but grew up in East Tennessee. She attended college at the University of Tennessee, Knoxville where she received her Bachelors of Science in Biology, Masters of Science in Microbiology, and will receive her Ph.D. in Genome Science and Technology in December of 2012. Her most prized accomplishment is raising her three wonderful children Cody, Camryn, and Colin.