Spring 4-16-2013

The Influence of Diet on the Mammalian Gut Microbiome

Brian David Muegge
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Molecular Microbiology and Microbial Pathogenesis

Dissertation Examination Committee:
Jeffrey Gordon, Chair
Gautam Dantas
Daniel Goldberg
Scott Hultgren
Clay Semenkovich
David Wang

The Influence of Diet on the Mammalian Gut Microbiome

by

Brian David Muegge

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2013

St. Louis, Missouri

Table of Contents

Chapter 3

Assembly of the functional gut microbiome in healthy and malnourished children

Chapter 4

Prospectus

Appendices

List of Figures

Chapter 2

Diet drives convergence of gut microbiome functions across mammalian phylogeny and within humans

Chapter 3

Assembly of the functional gut microbiome in healthy and malnourished children

List of Tables

Chapter 2

Diet drives convergence of gut microbiome functions across mammalian phylogeny and within humans

Chapter 3

Assembly of the functional gut microbiome in healthy and malnourished children

Acknowledgements

When I reflect on what I like best about scientific research, my answer is always "the people." I certainly enjoy the pursuit of discovery and the chance to work in a world of ideas. But it is the opportunity to share those ideas each day with brilliant, inspiring people from all around the planet that really motivates me. To all of my coworkers, colleagues, classmates, teachers and mentors, I sincerely thank you for your role in helping me to complete this research. Thank you for encouraging me, for pushing me, and for making me laugh around the coffee pot each afternoon. I wish you all the best in your future careers and hope that our paths will cross again, soon and often.

I am particularly indebted to my thesis mentor, Jeffrey Gordon, for giving me the opportunity to train in his lab. Jeff is an impossibly supportive and positive cheerleader for his mentees. His commitment to our career development is humbling, and I may never fully be able to appreciate all of the ways that he has helped me get started in biomedical research. In particular, I will always carry with me Jeff's deep respect for the power of words, and lessons about how to effectively communicate our research to the community.

There are a few people from the Gordon Lab who deserve particular thanks. First, thanks to lab managers Jill Manchester and Sabrina Wagoner, who somehow kept chaos at bay so that we could pursue our studies. Thanks also for the delicious birthday cobblers! Su Deng and Marty Meier were both very helpful in processing more than two thousand tubes of DNA for the Bangladesh project described in Chapter 3. Laura Kyro provided incredible assistance generating figures and managing documents, and Stephanie Amen provided tremendous administrative support. Former post-docs Justin Sonnenburg and Peter Crawford mentored me during my summer rotation and helped me get my bearing in the lab, and former graduate students Mike Mahowald and Peter Turnbaugh patiently taught me the basics of coding and informatics. I am particularly indebted to Ruth Ley, my former bay mate, who started the Mammal project described in Chapter 2 and provided great advice during my first year in the lab.

My colleagues in the lab the last four years have been inspirational. I'm sorry that I cannot thank you all individually for the role you have played in helping me to develop as a scientist. I'm

especially appreciative to the post-docs who joined the lab at the same time I did: Andy Goodman, Jeremiah Faith, and Federico Rey. I hope that I can model my scientific career on your incredible examples. I'm similarly thankful to my graduate school classmates Tanya Yatsunenko and Nate McNulty, who have been good friends and colleagues as we found our paths through graduate school. Thanks also to one of the newest students in the lab, Sathish Subramanian, who was a great partner on the Bangladesh project during his summer rotation project.

One thing I was surprised to learn during my thesis work is how much fun it is to engage in close, productive collaborations. I have been incredibly fortunate to learn from our very close collaborator Rob Knight and his group at the University of Colorado. Rob may be the only person on the planet who can match Jeff's enthusiasm for science, and he has been exceedingly generous with time and support in my numerous visits to his lab.

I cannot close without thanking other scientists who have helped me to reach this point. I have been very blessed in my life to stumble into the labs of three outstanding mentors: Jeff, Bob Cava at Princeton University, and Mark Richter at Missouri State University. I only hope that someday I can provide to my own students the support and inspiration that you have provided to me. I am also grateful to those who have financially supported my studies: the Medical Scientist Training Program, an Infectious Diseases Departmental Training Grant, and the Bill and Melinda Gates Foundation.

The acknowledgements must end with those who deserve my sincerest appreciation: my family. My parents, sister, and grandparents are blessings in my life, and have always provided unconditional love and support. About a month after graduate school began, my family expanded when I married my wife Carrie. Carrie's mother, brother, and extended family are treasured additions to my life. Of course, I am especially thankful for the support and love of Carrie. It has been a privilege and honor to grow and develop together through these last six years, and I look forward to many more to come. I cannot close without mentioning Carrie's specific support of my research. In addition to serving as an emotional support throughout graduate school, and being extremely

understanding of my strange hours and frequent time away from home, Carrie has saved me literally dozens of hours of coding by teaching me the "index-match" function in Excel. Let's hope that this is just the start of a beautiful and long-lasting collaboration.

ABSTRACT OF THE DISSERTATION

The influence of diet on the mammalian gut microbiome

by

Brian David Muegge

Doctor of Philosophy in Biology and Biomedical Sciences

(Molecular Microbiology and Microbial Pathogenesis)

Washington University in St. Louis, 2013

Professor Jeffrey I. Gordon, Chair

The mammalian gut is residence to a large microbial community whose collective set of millions of genes (microbiome) encodes a vast array of functions, including many that process dietary components. Few tools are available to change the microbiome's properties to promote host health because the factors governing community assembly and operation are poorly understood. My thesis focused on the impact of one factor, host diet.

I used a variety of experimental and computational methods to perform a comparative metagenomic study of the fecal communities of 39 diverse mammals to assess microbiome variation. These animals included herbivores, omnivores and carnivores representing different portions of the mammalian tree, Targeted sequencing of the bacterial 16S rRNA gene and shotgun gene sequencing of total fecal DNA revealed that animals with similar diets possessed similar microbiomes, both in terms of bacterial species detected and the functions that their genomes encoded. A large number of genes were found in all gut microbiomes, but the relative abundance of many traits was differentially represented in carnivorous compared to herbivorous hosts, demonstrating that diet can shape a microbiome. We next studied microbiome variation within a single host species consuming a range of diets, using a population of 18 calorie-restricted adult humans with carefully recorded dietary intake. Even in these free-living, unrelated individuals, the amount of protein and dietary fiber consumed was significantly correlated with the variation in the gut metagenomes.

My second study addressed the period of dramatic dietary variation in human life when breast fed infants are weaned. The study population was a birth cohort of 103 children born in Bangladesh, some of who developed malnutrition during the first two years of life. My results demonstrate that despite large inter-personal variation, common patterns are found in the establishment of bacterial species and functions as the children age. The data revealed a developmental pattern of human gut microbiome functional assembly. This was exemplified by the emergence of glycan degradation capacity in all children studied. Some of the taxonomic and functional changes can be correlated with the host's diet, but many properties of the developing infant microbiome cannot be explained by diet alone.

Chapter 1

From Animalicules to *Zobellia:* Development and Function of the Human Gastrointestinal Microbiota

From Animalicules to *Zobellia:* Development and Function of the Human Gastrointestinal Microbiota

"I would like to point out that we depend on more than the activity of some 30,000 genes encoded in the human genome. Our existence is critically dependent on the presence of upwards of 1000 bacterial species (the exact number is unknown because many are uncultivable) living in and on us; the oral cavity and gastrointestinal tracts contain particularly rich and active populations. Thus, if truth be known, human life depends on an additional 2 to 4 million genes, mostly uncharacterized. Until the synergistic activities between humans (and other animals) with their obligatory commensals has been elucidated, an understanding of human biology will remain incomplete."

-Julian Davies (Davies, 2001)

The report of the human genome sequence in 2001 ushered in a new era of genomic medicine, providing biology with a comprehensive resource to study the role of our genes in health and disease (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). At the same time, remarkable new DNA sequencing technologies developed for that project opened new windows to the vast collection of genes encoded by the microbes covering and colonizing every surface of the human body. The *microbiota* - the ensemble of bacteria, archaea, eukaryotes, and viruses associated with the host - influences human postnatal development, physiology, and maybe our predilection towards disease. Just as the human genome provides the blueprint for our cellular operations, the genomes of the members of the microbiota guide their interactions with our bodies and the environment. Studying the *microbiome*, or the collection of all of the genes and genomes in the microbiota, should help to reveal the molecular basis for this incredible symbiosis (Hooper and Gordon, 2001; Lederberg and McCray, 2001).

This thesis focuses on the largest microbial community in the human body, the gastrointestinal microbiota. There are more microbial cells in the gastrointestinal tract than there are human cells in the rest of the body, and the hundreds to thousands of different species living in a typical gut collectively contain at least 100 times more genes than the human genome (Qin et al., 2010). The genes encode a diverse array of functions that are absent from the human genome, including a large number of enzymes to process and utilize ingested carbohydrates. The organizing hypothesis of this thesis is that dietary variation within humans and between diverse mammals influences the functional properties of the gut microbiome. In turn, the resident microbes interact with our food and our bodies to influence normal health and possibly contribute to the pathogenesis of malnutrition. Later chapters of this thesis will describe the results of our shotgun gene sequencing studies of gut microbiomes that test these hypotheses. The remainder of this chapter reviews the history of human microbiome studies, new culture-independent technologies that have dramatically changed our views of human microbial diversity, and current knowledge about the influence diet on the gut microbiome. Particular attention will be paid to variations in the gut microbiomes of mammals with divergent diets, and in humans during the dramatic dietary changes that occur during infancy.

A brief history of the gastrointestinal microbiota

In 2008, the NIH launched the "Human Microbiome Project" (Peterson et al., 2009). With an investment of 150 million dollars, this ambitious project strives to describe microbial diversity in the mouth, nose, skin, vagina, and gut of several hundred healthy American adults. The goal is to create reference datasets and technologies for microbiome studies, analogous to the Human Genome Project. This project, and similar efforts in other parts of the world (Qin et al., 2010), sits at the cutting edge of modern technology and has been met with a great deal of anticipation. But while the technology is new and exciting, comparative studies of host associated microbial communities are as old as microbiology itself.

The first images of microbes were reported by the Dutch microscopist Anton van Leeuwenhoek in 1683. His work has always been credited as the start of microbiology, but it was simultaneously the origin of human microbiome studies. van Leeuwenhoek was curious about the white matter that persisted between his own teeth despite cleaning, so he mixed the matter with his saliva and placed it under the objective of his microscope. He was searching for a microscopic explanation of taste, but instead witnessed "very many small living animals, which moved very prettily" (Bardell, 1983). van Leeuwenhoek called these organisms "animalcules," and we now recognize by his drawings and descriptions that he was first human to see a bacterium.

He was also the first investigator to pursue a comparative study of host-associated microbiota. His report continued by studying tooth scrapings from four additional people; an adult woman, a young girl, a man who didn't drink or smoke, and an older man who cleaned his few remaining teeth by rinsing his mouth with wine each day. van Leeuwenhoek reported that some of the rods he detected in his own mouth were seen in all of the samples, but the sober man also had spiral shaped "animals" not seen anywhere else, and the older man lacked certain long rods seen in the other samples.

Microbiology and bacteriology developed in the centuries after van Leeuwenhoek's drawings, and by the early 1900s many principles had been established that still guide our under-

standing of the gut microbial community. The gastrointestinal tract was described as "a singularly perfect incubator" because it represented a warm, protected habitat for microorganisms with a nearly continuous supply of nutrients from the host diet (Kendall, 1909). The importance of diet in influencing bacterial communities had been established in animal models, where defined diets were used in monkeys and dogs to identify elements of the microbiota responsive to specific food ingredients (Kendall, 1909; Torrey, 1919). Multiple studies had identified differences in the bacterial communities of nursing children compared to weaned children and adults, and Tissier's isolation of *Bifidobacterium bifidus* from an infant's stool led to the first "probiotic" trials using ingested live bacteria to ameliorate diarrheal disease (Schrezenmeir and de Vrese, 2001). Scientists at the turn of the century had already begun creating "germfree" animal models using laboratory animals born and maintained in sterile conditions lacking any microorganisms. Early studies in germfree guinea pigs, chickens, tadpoles, and turtles hinted at the role of the microbiota in shaping host physiology because most of the animals became sick after several weeks of life without microbes (Kendall, 1915).

In the ensuing century, biologists have created increasingly detailed catalogs of the bacteria and other organisms in the gut microbiota, but an understanding of how those communities are controlled remains elusive. The typical study in this era relied on enumeration of bacterial species with (increasingly sophisticated) anaerobic culture techniques. In some cases, these counts could be related to in vitro observations of how those organisms behaved, but in general these species catalogs provided very little insight into the biological properties of the microbes. Culture-dependent counting studies carried several important limitations. First, it was well established that many organisms from the environment were not readily cultured in standard media and conditions, giving rise to the "plate count anomaly" (Staley and Konopka, 1985). These culture studies thus capture only a fraction of the naturally occurring microbial community. Second, the classification of organisms by biochemical properties and visual appearance does not readily correlate to taxonomy. Closely related microorganisms can look and behave very differently, while distantly related organisms may be functionally and visually similar in a Petri dish (Eisen, 2007). One hun-

dred years of culturing yielded much data but no overarching principles of how communities take root in the gut or respond to perturbations (Freter, 1983).

The metagenomic revolution

In the last thirty years, the study of microbial communities has been revolutionized by new technologies that allow the molecular characterization of organisms based on their genomes rather than growth in culture or visual inspection (Handelsman, 2004; Riesenfeld et al., 2004). The foundation for this shift towards molecular phylogeny was made when Woese and colleagues invented an approach to describe all organisms systematically using sequence similarity (Woese and Fox, 1977). They developed their phylogeny using the ribosomal RNA molecule because it is found in all living organisms and mutates slowly without selective pressure. Microbial ecologists, most notably led by Norm Pace and his trainees, leveraged this approach to characterize complex communities of microorganisms from environmental habitats like soil and ocean. They sequenced genes encoding bacterial small subunit ribosomal RNAs (16S rRNAs) without having to culture the component members (Lane et al., 1985; Stahl et al., 1984). The same technologies also allowed researchers to sequence cloned DNA fragments from total community DNA, revealing the encoded metabolic traits of the organisms (Schmidt et al., 1991; Stein et al., 1996). The general approach of using DNA isolated from an environment to assess community membership and function without culturing came to be known as *metagenomics* (Handelsman et al., 1998).

Medical researchers interested in the complex microbial communities associated with the human body quickly adopted the culture-independent methods pioneered by environmental microbiologists. Seminal advances for host-associated studies include the first major descriptions of species (Eckburg et al., 2005) and functions (Gill et al., 2006) found in healthy human gut microbial communities using metagenomic methods. These initial reports generated data from only two or three human individuals, but since then sequencing costs have plummeted while sequencing capacity has dramatically increased. The field has progressed to the stage where now it is possible to sequence hundreds to thousands of samples with current highly parallel instruments (Caporaso et al., 2011; Turnbaugh et al., 2008; Qin et al., 2010).

Culture-independent methods are changing the nature of microbiome research. For the first time, the entire diversity and functional potential of the gut microbial community can be measured through targeted amplification of phylogenetic markers (i.e., rRNA genes) or direct shotgun gene sequencing of community DNA. These datasets provide resources to expand our knowledge of how host factors like diet influence the operation of the gut microbes, and in turn how microbes may be influencing the host's nutritional status. For example, work from our lab has demonstrated that compared to lean controls, obese mice and humans have altered abundances of members of major bacterial phyla in their gut microbiota as well as reduced diversity (Ley et al., 2005, 2006), while obese microbiomes have more encoded potential in their genomes for energy harvest (Turnbaugh et al., 2006).

Metagenomics is critical for a top-down systems level analysis of community structure and function. The same types of data can also be used to discover very specific interactions between ingested dietary components and the members of the gut community. A shotgun gene sequencing survey of Japanese human gut microbiomes revealed that a gut Bacteroidetes species possessed the genetic machinery to degrade seaweed polysaccharide. This gene function was not detected in other gut bacterial genomes or in the microbiomes of USA residents, but instead had homology to a Bacteroidetes normally found in the ocean, *Zobellia galactanivorans*. It seems likely that the marine bacterium entered the gut along with dietary seaweed and transferred its seaweed degrading gene to the gut specialist through horizontal gene transfer (Hehemann et al., 2010).

Culture-independent metagenomic methods can reveal associations and community operations that are impossible to detect with culture dependent methods. However, unless these new technologies are used to develop and test hypotheses of factors controlling microbiome operation, they will not be much more useful than the culture-dependent studies of the past. The remainder of this chapter will focus on what is currently known about how one such factor, host diet, influences the mammalian gut microbiome.

Diet and the microbiome: variation across mammals

Mammals have evolved a stunning range of digestive strategies and gastrointestinal structures, and in many cases these anatomic variations are directly linked to microbial communities (Stevens and Hume, 1998). The original mammals were small carnivores, and mammalian genomes ever since have encoded peptidases and proteases needed to chemically digest a protein diet. These ancestors, along with most modern carnivores, have a relatively simple gut anatomy. As mammals radiated into herbivorous niches, they lacked the encoded enzymatic repertoire to degrade complex plant polysaccharides like cellulose. Herbivorous animals instead came to rely on microorganisms in their GI tract to generate useable energy and nutrients from plants by anaerobic fermentation producing short chain fatty acids (SCFA). These types of mammals evolved specialized anatomic structures to house their microbial bioreactors.

Ruminants such as cattle and sheep use a strategy of "foregut fermentation." They have multi-chambered stomach that harbor diverse fermentative microbial communities; SCFA produced by these microbes account for more than 60% of the daily energy requirements for some ruminants and are completely essential for life (Stevens and Hume, 1998; Wolin, 1981). The alternative evolutionary strategy used by herbivores ranging from rabbits to elephants is "hindgut fermentation." In these animals, enlarged cecums and/or colons are the site for the largest portion of the gut microbiota. Humans ingest plants as part of an omnivorous diet, but our gastrointestinal tracts are relatively simple and lack the major fermentation structures found in strict herbivores. Some calculations suggest that microbial production of SCFA could contribute up to 5-10% of daily energy needs in humans, although this number would vary according to host diet (McNeil, 1984). Aside from their role in host energy balance, it is certain that SCFA in the mammalian gut help to regulate intestinal physiology and immune responses through interactions with specific receptors expressed on the epithelial surface (Fukuda et al., 2011; Maslowski et al., 2009; Pryde et al., 2002).

With the exception of domesticated ruminants, the microbial diversity in most mammalian guts has been poorly described. This changed with the first broad survey of mammalian gut microbes, using stool samples from 60 different mammalian species (Ley et al., 2008). The surveyed animals lived in two zoos and in the wild and represented a wide range of mammalian orders, gut physiologies, and diets. Bacterial 16S rRNA sequences were amplified by PCR from each mammal's fecal sample: the results showed that mammals sharing a diet had bacterial communities that were more phylogenetically similar to each other than the communities from mammals with different diets. Communities were also more similar in hosts from the same order. Even in very divergent mammals, this result shows a convergence of microbial species found in the gut as a function of host diet.

The open question remaining from this study was whether or not there is a similar diet-driven convergence of microbial functions in mammalian guts. Additionally, it is unknown if the functional microbial communities of mammals have co-evolved with the host through evolutionary time. This would be reflected by the detection of bacterial functions that are highly associated with certain branches of the mammalian evolutionary tree. These questions could not be answered from the existing literature, because beyond humans and laboratory mice, very few mammalian gut microbiomes have been characterized by shotgun sequencing. All are reports of communities from a single host species; cattle (Brulc et al., 2009; Hess et al., 2011), pigs (Lamendella et al., 2011), dogs (Swanson et al., 2011), and the Tamar wallaby (Pope et al., 2010). The goal in most of these studies was to describe biological systems for biomass conversion, not to understand how different mammalian microbiomes responded functionally to similar diets. To parse the contributions of host phylogeny and diet in shaping the functional microbiome, a broad comparison of the microbiomes from many mammals was required, inspiring the work described in Chapter 2 of this thesis.

Diet and the microbiome: successional changes during suckling and weaning

As discussed previously, changes in the gut bacterial community after birth have been studied

for more than a century. The topic has been comprehensively reviewed many times (Cooperstock and Zedd, 1983; Mackie et al., 1999). However, there are few culture-independent studies, and there are essentially no reports of comparative shotgun metagenomic studies of a large number of infants. This section of Chapter 1 briefly addresses what is known about bacterial succession during the dietary changes of infancy, with special emphasis on changes in bacterial taxonomy and relation to life events. The next section will review what is known about how the gut microbiome matures functionally during the same time period.

Successional patterns have been much more clearly delineated in lab mice than in human infants, primarily in a series of elegant reports from the Rockefeller Institute in the 1960s using anaerobic culture techniques. The group initially reported the species composition of the mouse gut over the first 20 days of life, demonstrating the early appearance of lactobacilli and streptococci in all regions of the alimentary canal, and the eventual replacement of those groups by anaerobic Bacteroides species and coliforms in the large intestine (Schaedler et al., 1965). They went on demonstrate that different bacterial populations are associated with the intestinal epithelium compared to the lumen (Savage et al., 1968) and introduced the concept of an "*autochthonous biota*" of species that seem specialized for stable colonization of the host (Dubos et al., 1965).

The most significant report moved beyond description to experimental manipulation of the environment, trying to test which environmental factors drove the observed changes in microbial population levels (Lee et al., 1971). They reared a group of pups in a cage where the only available source of food was whole milk powder devoid of complex polysaccharides found in normal mouse diets. Unlike the control animals weaned onto standard chow, the emergence of the "Fusiform" group of Gram variable tapered rods (likely Firmicutes) and decline in coliforms did not occur at the weaning transition in the milk weaned pups. However, Bacteroides organisms expanded at the time of weaning in both groups with similar kinetics. This data showed that some changes in the bacterial community at weaning can be specifically attributed to diet, but other changes cannot be related solely to the ingested food. To this day, few if any infant succession studies have seriously

considered the non-nutritive factors associated with weaning (e.g., maturation of the gut and withdrawal of passive immunity) that may also be driving bacterial community changes.

At the same time, Smith was characterizing the successional pattern in a range of infant mammals. A survey of the bacterial communities in the stools of young cows, lambs, pigs, and humans demonstrated a similar pattern of succession in all of the mammalian hosts; *Escherichia coli*, *Clostridium welchii*, and streptococci were detected in stool within one day of birth, followed a few days later by the emergence of Bacteroides and lactobacilli species, which later became dominant (Smith and Crabb, 1961). While the bacterial communities were very similar in the different hosts immediately after birth and as long as the animals were consuming mother's milk, the microbiome patterns diverged dramatically once the animals were weaned onto their adult diets. The researchers went on to show that all of the mammals surveyed possessed bacterial communities along the entire length of the alimentary canal, that each anatomic site harbored a unique collection of bacterial species compared to other parts of the gut, and that the kinetics of succession varied in the different mammals studied (Smith, 1965).

In human studies, patterns of infant microbiome maturation are less clear. One chief problem is that most studies are anecdotal case reports, studying a few infants with a sporadic sampling schedule. Additionally, metadata on life events such as diet changes or diarrheal episodes are generally imprecise. This makes it very difficult to compare studies or to generate firm, testable hypotheses about causes of microbiome changes. A corollary to this problem is a trend in the literature to casually attribute all changes in aging infants to the dietary changes of food supplementation and weaning. Diet may be the driving factor for many or even most microbiome changes in the developing gut, but as the whole milk powder experiment showed in mice, there are likely some bacterial changes that are independent of diet. There are essentially no dose-response studies temporally linked to diet switches that prove which components of the microbiome are responding to the new substrates (Mackie et al., 1999). In general, the field lacks an ecological understanding linking the environmental changes in infancy with species or functional responses. Similarly,

there have not been enough large studies with well-characterized samples to describe the variation across many humans or within a single community over time.

In a meta-analysis of culture-dependent human succession studies, Cooperstock and Zedd argued that there are at least four phases of microbiome assembly in human infants (Cooperstock and Zedd, 1983). Aerotolerant organisms not typically seen in adult guts initially colonize the human gut. This phase lasts for only 1 to 2 weeks, with tremendous interpersonal variation. The model argues that the next phases are all governed by host diet; phase 2 is the remaining time with exclusive breast feeding, phase 3 begins with the introduction of cereal or other supplements, and phase 4 marks the end of breast milk feeding and complete weaning onto an adult like diet. The bacteria detected in phase 2 are gut specialists adapted to lactose-rich diets; with the gradual replacement of breast milk by complex polysaccharides, bacteria capable of anaerobic fermentation become dominant.

Two culture-independent studies from the recent literature are especially noteworthy. The largest study of infants to date used quantitative real-time PCR (qPCR) to measure the abundance of bacterial phyla in the stool of 1,036 children at 1 month of age (Penders et al., 2006). The authors demonstrated that children born by Caesarian-section had significantly lower bifidobacteria and *Bacteroides fragilis* group counts than vaginally delivered counterparts, and that children exclusively fed formula instead of breast milk had more *E. coli*, *Clostridum difficile*, *Bacteroides*, and lactobacilli in their gut microbiota. The study has significant limitations, notably a single time point per child and the use of qPCR for bacterial abundance detection, but it is still a significant contribution because of the very large number of children sampled, and the quantitative data analysis used to correlate host features with microbiota signatures.

The most comprehensive report of species changes in the gut used a microarray designed to resolve bacterial genuses and species using conserved regions of the 16S rRNA gene (Palmer et al., 2007). The study population included 13 infants, sampled intensively for the first three months of life with less frequent sampling thereafter until 1 year. The report demonstrated a remarkable

degree of fluctuation in the fecal microbiota of each child over time, and that there was also a great deal of variation between the children. Early in life, the microbiota of babies are not only different from adults, but also very different from the other babies.

Diet and the microbiome: functional changes during suckling and weaning

Compared to the numerous studies of bacterial species succession, there are very few community level studies of the functional properties of the developing gut microbiome. The studies that do exist fall into two main categories. The first group measure community metabolism, exemplified by the work of Tore Midtvedt characterizing "microflora-associated characteristics," or MACs (Mackie et al., 1999; Midtvedt et al., 1987, 1988). These characteristics are metabolic or chemical processes, including SCFA production and mucin degradation, which are completely absent in germ-free animals but present in conventionally-raised animals. In a series of papers studying these properties in human infants and children, they described the emergence of these functions in small numbers of individuals; all MACs became more abundant in stools as the children aged (Midtvedt and Gustafsson, 1981; Midtvedt and Midtvedt, 1992; Midtvedt et al., 1994). The limitation of these biochemical studies is that the organism (or organisms) responsible for the function can't be determined. There is also no way to determine if the gene responsible for a characteristic is in the community unless it is being actively expressed.

The second and more recent group of studies used shotgun gene sequencing to interrogate community functional structure. All studies to date have been limited by a small number of infants under the age of 5 (mean=2, maximum=5) and relatively few time points with shotgun gene sequencing (mean=3.6, median=1, maximum=15) (Gupta et al., 2011; Koenig et al., 2011; Kurokawa et al., 2007; Morowitz et al., 2011; Vaishampayan et al., 2010). In a comparative study of Japanese infants, children, and adults, all human gut microbiomes were enriched with genes for carbohydrate metabolism compared to a database of all environmental and host associated bacterial genomes. Compared to adult microbiomes, infant microbiomes had a relative enrichment in transporter systems and decreased abundance in anaerobic metabolism pathways. Some of the spe-

cific carbohydrate metabolism suites enriched in adult microbiomes were also enriched in the infants, but each age group also had a unique suite of glycoside hydrolases (Kurokawa et al., 2007).

In the only time series functional analysis of a single child over several months, samples obtained in the first week after birth were relatively enriched in genes for the utilization of the simple sugars lactose, galactose, and sucrose. Samples from the third and fourth months of life, when the infant was still exclusively breastfed, had a greater diversity of carbohydrate degradation genes. These included pathways for sialic acid and mucin degradation, and also the breakdown of complex polysaccharides not yet introduced to the diet. Samples obtained after weaning had additional enrichment for "adult" -like functions including xenobiotic metabolism and a full array of vitamin biosynthetic pathways (Koenig et al., 2011).

Overview of the Thesis

The guiding hypothesis of my research is that the structure and function of the human gut microbiome is modulated by the consumption of food, and that this modulation is reproducible across diverse mammalian and human hosts consuming similar diets. In turn, we believe that the operation of the gut microbiome may interact with diet and other factors to influence human nutritional status. This interplay could predispose the host to health or disease, and we particularly focus on the important problem of infant malnutrition. Our goal was to determine general principles and mechanisms that mediate the response of gut microbial communities to ingested foodstuffs, as a precursor to future efforts to intentionally manipulate bacterial communities to improve human health.

To study these hypotheses, I leveraged recent advances in next-generation sequencing technology to conduct comparative metagenomic analyses of several populations with well-characterized diets. We deployed novel computational methods to study the relationships between bacterial membership and functional profile of these fecal samples. This thesis describes the results of these studies.

In Chapter 2, I present the results of a study addressing the influence of diet on the functional microbiomes of adult humans and mammals. We extended the survey of bacterial species in diverse mammalian hosts by asking if there was a functional convergence of microbiomes from animals with similar diets. Fecal DNA was extracted from the stools of 39 adult mammals representing 33 distinct mammalian species. Animals spanned 10 orders of mammals and included carnivores, omnivores, and herbivores. We found that all mammalian gut microbiomes share a large core of genes, but host diet strongly influences the relative abundance of many of these functions. In particular, we demonstrated that carnivores and herbivores possessed distinguishable microbiomes, and that amino acid and pyruvate metabolic pathway abundances were significantly different in the two groups. We then turned to a population of 18 adult humans with well-characterized diets to ask if the dietary variation within a single host species could be correlated to microbiome functional profiles. The data showed that total protein intake was significantly correlated with the functional profile of the community, while fiber intake was significantly related to the species composition. Together, these experiments demonstrated that diet causes reproducible changes in the gut microbiomes of diverse mammalian hosts. It also established the analytic tools necessary for large-scale comparative metagenomic analysis.

The second part of my thesis focused on the period of greatest dietary change in human life, the transition from maternal milk to a complex diet in early infancy. Chapter 3 describes our cohort study of 103 infants born in Bangladesh, with monthly stool sampling from birth through two years of life. The study was greatly enhanced by the precise clinical metadata collected by our collaborators in Bangladesh related to diet, nutritional status, and diarrhea. Bacterial community membership was determined by 16S rRNA gene sequencing for all samples, and for a subset of the children, the functional microbiome was characterized by shotgun gene sequencing. We found a stunning level of intrapersonal variation in the infant microbiota, with essentially a complete replacement of the bacterial species found in early life by two years of age. However, there were common patterns of development across the cohort including reproducible taxonomic shifts and the establishment of specific functional metabolic pathways. Increasing age and diet switches were

both significantly associated with these changes, but age appears to be the stronger predictor of community structure across all children. We characterized the emergence of a particular bacterial function for sialic acid degradation that is completely absent from the microbiomes in early life but is found in all children by the end of the study. This timing of the appearance is not clearly related to diet or other external stimuli, and hints that shifts in endogenous glycan expression during infancy may be an important means for the host to shape the development of the gut microbiome.

References

Bardell, D. (1983). The roles of the sense of taste and clean teeth in the discovery of bacteria by Antoni van Leeuwenhoek. Microbiol. Rev *47*, 121–126.

Brulc, J. M., Antonopoulos, D. A., Miller, M. E. B., Wilson, M. K., Yannarell, A. C., Dinsdale, E. A., Edwards, R. E., Frank, E. D., Emerson, J. B., Wacklin, P., et al. (2009). Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. Proc. Natl. Acad. Sci. U.S.A *106*, 1948–1953.

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N., et al. (2011). Moving pictures of the human microbiome. Genome Biol *12*, R50.

Cooperstock, M. S., and Zedd, A. J. (1983). Intestinal flora of infants. In Human intestinal microflora in health and disease, Hentes, David J., ed. (New York: Academic Press), pp. 79–100.

Davies, J. (2001). In a Map for Human Life, Count the Microbes, Too. Science *291*, 2316.

Dubos, R., Schaedler, R. W., Costello, R., and Hoet, P. (1965). Indigenous, normal, and autochthonous flora of the gastrointestinal tract. J. Exp. Med *122*, 67–76.

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E., and Relman, D. A. (2005). Diversity of the Human Intestinal Microbial Flora. Science *308*, 1635–1638.

Eisen, J. A. (2007). Environmental Shotgun Sequencing: Its Potential and Challenges for Studying the Hidden World of Microbes. PLoS Biol *5*, e82.

Freter, R. (1983). Mechanisms that control the microflora in the large intestine. In Human intestinal microflora in health and disease, D. J. Hentges, ed. (New York: Academic Press), pp. 33–54.

Fukuda, S., Toh, H., Hase, K., Oshima, K., Nakanishi, Y., Yoshimura, K., Tobe, T., Clarke, J. M., Topping, D. L., Suzuki, T., et al. (2011). Bifidobacteria can protect from enteropathogenic infection through production of acetate. Nature *469*, 543–547.

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006). Metagenomic Analysis of the Human Distal Gut Microbiome. Science *312*, 1355–1359.

Gupta, S., Mohammed, M., Ghosh, T., Kanungo, S., Nair, G., and Mande, S. (2011). Metagenome of the gut of a malnourished child. Gut Pathogens *3*, 7.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol *5*, R245–R249.

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol. Mol. Biol. Rev. *68*, 669–685.

Hehemann, J.-H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., and Michel, G. (2010). Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature *464*, 908–912.

Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., Luo, S., Clark, D. S., Chen, F., Zhang, T., et al. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science *331*, 463–467.

Hooper, L. V., and Gordon, J. I. (2001). Commensal Host-Bacterial Relationships in the Gut. Science *292*, 1115–1118.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Kendall, A. I. (1915). The bacteria of the intestinal tract of man. Science *42*, 209–212.

Kendall, A. I. (1909). Some observations on the study of the intestinal bacteria. Journal of Biological Chemistry *6*, 499–507.

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., and Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut

microbiome. Proc. Natl. Acad. Sci. U.S.A *108 Suppl 1*, 4578–4585.

Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., Sharma, V. K., Srivastava, T. P., et al. (2007). Comparative Metagenomics Revealed Commonly Enriched Gene Sets in Human Gut Microbiomes. DNA Research *14*, 169–181.

Lamendella, R., Domingo, J. W. S., Ghosh, S., Martinson, J., and Oerther, D. B. (2011). Comparative fecal metagenomics unveils unique functional capacity of the swine gut. BMC Microbiol *11*, 103.

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., and Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc. Natl. Acad. Sci. U.S.A *82*, 6955–6959.

Lederberg, J., and McCray, A. T. (2001). 'Ome Sweet 'Omics-- A Genealogical Treasury of Words. The Scientist *15*, 8.

Lee, A., Gordon, J., Lee, C. J., and Dubos, R. (1971). The mouse intestinal microflora with emphasis on the strict anaerobes. J. Exp. Med *133*, 339–352.

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., and Gordon, J. I. (2005). Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences of the United States of America *102*, 11070–11075.

Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., Schlegel, M. L., Tucker, T. A., Schrenzel, M. D., Knight, R., et al. (2008). Evolution of Mammals and Their Gut Microbes. Science *320*, 1647–1651.

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: Human gut microbes associated with obesity. Nature *444*, 1022–1023.

Mackie, R. I., Sghir, A., and Gaskins, H. R. (1999). Developmental microbial ecology of the neonatal gastrointestinal tract. Am. J. Clin. Nutr *69*, 1035S-1045S.

Maslowski, K. M., Vieira, A. T., Ng, A., Kranich, J., Sierro, F., Di Yu, Schilter, H. C., Rolph,

M. S., Mackay, F., Artis, D., et al. (2009). Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. Nature *461*, 1282–1286.

McNeil, N. (1984). The contribution of the large intestine to energy supplies in man. The American Journal of Clinical Nutrition *39*, 338–342.

Midtvedt, A. C., and Midtvedt, T. (1992). Production of short chain fatty acids by the intestinal microflora during the first 2 years of human life. J. Pediatr. Gastroenterol. Nutr *15*, 395–403.

Midtvedt, A. C., Carlstedt-Duke, B., and Midtvedt, T. (1994). Establishment of a mucin-degrading intestinal microflora during the first two years of human life. J. Pediatr. Gastroenterol. Nutr *18*, 321–326.

Midtvedt, A. C., Carlstedt-Duke, B., Norin, K. E., Saxerholt, H., and Midtvedt, T. (1988). Development of five metabolic activities associated with the intestinal microflora of healthy infants. J. Pediatr. Gastroenterol. Nutr *7*, 559–567.

Midtvedt, T., and Gustafsson, B. E. (1981). Microbial conversion of bilirubin to urobilins in vitro and in vivo. Acta Pathol Microbiol Scand B *89*, 57–60.

Midtvedt, T., Carlstedt-Duke, B., Höverstad, T., Midtvedt, A. C., Norin, K. E., and Saxerholt, H. (1987). Establishment of a biochemically active intestinal ecosystem in ex-germfree rats. Appl Environ Microbiol *53*, 2866–2871.

Morowitz, M. J., Denef, V. J., Costello, E. K., Thomas, B. C., Poroyko, V., Relman, D. A., and Banfield, J. F. (2011). Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. Proc. Natl. Acad. Sci. U.S.A *108*, 1128–1133.

Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., and Brown, P. O. (2007). Development of the Human Infant Intestinal Microbiota. PLoS Biology *5*, e177 EP -.

Penders, J., Thijs, C., Vink, C., Stelma, F. F., Snijders, B., Kummeling, I., van den Brandt, P. A., and Stobberingh, E. E. (2006). Factors influencing the composition of the intestinal

microbiota in early infancy. Pediatrics *118*, 511–521.

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., Deal, C., et al. (2009). The NIH Human Microbiome Project. Genome Res *19*, 2317–2323.

Pope, P. B., Denman, S. E., Jones, M., Tringe, S. G., Barry, K., Malfatti, S. A., McHardy, A. C., Cheng, J.-F., Hugenholtz, P., McSweeney, C. S., et al. (2010). Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. Proc. Natl. Acad. Sci. U.S.A *107*, 14793–14798.

Pryde, S. E., Duncan, S. H., Hold, G. L., Stewart, C. S., and Flint, H. J. (2002). The microbiology of butyrate formation in the human colon. FEMS Microbiology Letters *217*, 133–139.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. (2004). METAGENOMICS: Genomic Analysis of Microbial Communities. Annu. Rev. Genet. *38*, 525–552.

Savage, D. C., Dubos, R., and Schaedler, R. W. (1968). The gastrointestinal epithelium and its autochthonous bacterial flora. J. Exp. Med *127*, 67–76.

Schaedler, R. W., Dubos, R., and Costello, R. (1965). The development of the bacterial flora in the gastrointestinal tract of mice. J. Exp. Med *122*, 59–66.

Schmidt, T. M., DeLong, E. F., and Pace, N. R. (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J. Bacteriol *173*, 4371–4378.

Schrezenmeir, J., and de Vrese, M. (2001). Probiotics, prebiotics, and synbiotics--approaching a definition. Am. J. Clin. Nutr *73*, 361S-364S.

Smith, H. W. (1965). The development of the flora of the alimentary tract in young animals. The Journal of Pathology and Bacteriology *90*, 495–513.

Smith, H. W., and Crabb, W. E. (1961). The fæcal bacterial flora of animals and man: Its development in the young. The Journal of Pathology and Bacteriology *82*, 53–66.

Stahl, D. A., Lane, D. J., Olsen, G. J., and Pace, N. R. (1984). Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. Science *224*, 409–411.

Staley, J. T., and Konopka, A. (1985). Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. Annu. Rev. Microbiol. *39*, 321–346.

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H., and DeLong, E. F. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J. Bacteriol *178*, 591–599.

Stevens, C. E., and Hume, I. D. (1998). Contributions of Microbes in Vertebrate Gastrointestinal Tract to Production and Conservation of Nutrients. Physiological Reviews *78*, 393–427.

Swanson, K. S., Dowd, S. E., Suchodolski, J. S., Middelbos, I. S., Vester, B. M., Barry, K. A., Nelson, K. E., Torralba, M., Henrissat, B., Coutinho, P. M., et al. (2011). Phylogenetic and gene-centric metagenomics of the canine intestinal microbiome reveals similarities with humans and mice. ISME J *5*, 639–649.

Torrey, J. C. (1919). The Regulation of the Intestinal Flora of Dogs through Diet. J Med Res *39*, 415–447.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., et al. (2008). A core gut microbiome in obese and lean twins. Nature. Available at: http://dx.doi.org/10.1038/nature07540 [Accessed January 13, 2009].

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature *444*, 1027–1131.

Vaishampayan, P. A., Kuehl, J. V., Froula, J. L., Morgan, J. L., Ochman, H., and Francino, M. P.

(2010). Comparative Metagenomics and Population Dynamics of the Gut Microbiota in Mother and Infant. Genome Biol Evol *2*, 53–66.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The Sequence of the Human Genome. Science *291*, 1304–1351.

Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U.S.A *74*, 5088–5090.

Wolin, M. J. (1981). Fermentation in the rumen and human large intestine. Science *213*, 1463–1468.

Chapter 2

Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans

Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans

"It's probably not a good idea to take too personal an interest in your microbes."

Bill Bryson, *A Short History of Nearly Everything*, (2003), p. 302.

This chapter has been published as:

Muegge B.D., Kuczynski J., Knights D., Clemente J.C., González A., Fontana L., Henrissat B., Knight R., Gordon J.I. "Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans." *Science*. 2011. 332(6032): 970-4.

Abstract

Coevolution of mammals and their gut microbiota has profoundly affected their radiation into myriad habitats. We used shotgun sequencing of microbial community DNA and targeted sequencing of bacterial 16S ribosomal RNA genes to gain an understanding of how microbial communities adapt to extremes of diet. We sampled fecal DNA from 33 mammalian species and 18 humans who kept detailed diet records, and we found that the adaptation of the microbiota to diet is similar across different mammalian lineages. Functional repertoires of microbiome genes, such as those encoding carbohydrate-active enzymes and proteases, can be predicted from bacterial species assemblages. These results illustrate the value of characterizing vertebrate gut microbiomes to understand host evolutionary histories at a supraorganismal level.

Introduction

Comparative culture-independent metagenomic studies of the microbial species assemblages that compose mammalian gut microbiota, and the functions that these communities encode in their aggregate genomes (microbiomes), can provide a complementary perspective to comparative studies of host genomes. A previous bacterial 16S ribosomal RNA (rRNA)-based study of 59 mammalian species revealed that their fecal microbiota clustered according to diet rather than host phylogeny (*1*). This finding raises several questions: What is the functional evolution of the gut microbiome in relation to diet? Is the process unique to each mammalian lineage? To what extent does microbial phylogeny predict function within microbial communities? Could analysis of interspecific differences among mammals create a pipeline for deciphering intraspecific differences among humans in response to varied diets or other factors? Therefore, we have extended our 16S rRNA studies to a broader sampling of microbial genes in total fecal community DNA prepared from herbivores, omnivores, and carnivores.

Methods

We generated shotgun pyrosequencing data sets from 33 mammalian species, along with newly collected bacterial 16S rRNA data. These adult animals represent 10 Orders and varied digestive physiologies (hindgut-fermenters, foregut-fermenters, and simple-gut). In some cases, free-living and captive representatives of a given species were sampled (**Table 1,Table S1**). Methods for classifying diets and for collecting and processing fecal samples for metagenomic analyses have been described (*1*). Multiplex pyrosequencing of amplicons generated from the V2 region of bacterial 16S rRNA genes yielded 149,675 high-quality de-noised reads (average = 3838 ± 1080 per sample) (**Table S2**) (*2*). After chimera removal, 8541 operational taxonomic units (OTUs) were identified in the combined data set (an OTU was defined as reads sharing ≥97% nucleotide sequence identity). Shotgun sequencing of the same fecal DNA preparations produced 2,163,286 reads [mean = 55,469 ± 28,724 (SD) per sample; 261 ± 83 nt per read] (**Table S3**) (*2*). Shotgun reads were functionally annotated using three databases: KEGG [for KEGG Orthology (KO) groups and Enzyme Commission (E.C.) numbers], CAZy (for carbohydrate-active enzymes), and MEROPS (for peptidases) (*3*-*5*). When shotgun reads were assigned to phylogenetic bins using the program MEGAN (*6*), the results revealed that fecal microbiomes were dominated by members of Bacteria, had low levels of Eukarya (0.15 to 5.35% of identifiable reads), and archaeons were variably represented (0 to 1.77% of assignable reads, with none detected in any carnivore microbiome). Seventeen samples had reads assigned to known viruses (**Table S4**) (*2*).

Results and Discussion

Concordance of microbiome structure and function

Procrustes analysis (least-squares orthogonal mapping) was used to test whether the functional properties of a microbiome can be predicted from the bacterial species that compose it (*2*). Procrustes analysis attempts to stretch and rotate the points in one matrix, such as points obtained by principal coordinates analysis (PCoA), to be as close as possible to points in the other matrix, thus

preserving the relative distances between points within each matrix (*7, 8*) (**Fig. 1A**). We first took the 16S rRNA data set and used the UniFrac metric to compare the overlap between each pair of communities in terms of their evolutionary distance (*9)*. The similarity in functional profiles was then determined using the Bray-Curtis distance metric applied to KO groups, E.C.s, CAZYmes, or peptidases. Principal-coordinates reduction was performed separately on the 16S rRNA and annotated shotgun (microbiome) data sets, and the point clouds were aligned using Procrustes. For each comparison, the goodness of fit, or M^2 value, of the transformed data sets was measured over the first three dimensions. The statistical significance of the goodness of the fit was measured by a Monte Carlo label permutation approach (*2*).

The agreement between phylogenetic and functional measurements was remarkable for all mammals, regardless of their diet, host lineage, or gut physiology. Figure 1, B to E, shows how the goodness of fit was robust to different functional databases. The analysis was also robust to taxon- or phylogenetic-based species classification, weighted or unweighted metrics, and whether one or more member of each mammalian species was considered (Fig. S1). For both bacterial 16S rRNA and whole community gene data sets, the PCoA plots separated carnivores and omnivores from herbivores, emphasizing the importance of diet in differentiating gut microbial communities ($P < 0.05$) (*2*). Our previous study using full-length 16S rRNA sequences revealed that the fecal microbiota of conspecifics were significantly more similar than the communities of different host species (*1*). The V2 16S rRNA data generated in this study confirmed this result, using both weighted and unweighted UniFrac distances ($P < 0.05$ by 1000 Monte Carlo permutations) (*2*).

The Procrustes results prompted us to use a nearest-neighbor model to test whether the functional configuration of a microbiome could be predicted from its 16S rRNA sequences. Using a fecal sample's nearest neighbor, as defined by unweighted or weighted UniFrac, to predict the sample's functional profile generated a significantly better prediction than a random neighbor; this was true for KOs, E.C.s, CAZymes, and peptidases ($P < 0.0001$, 10^6 Monte Carlo permutations) (*2*).

Testing for coevolution between mammalian phylogeny and microbiomes

The concordance of diet and microbiome structure and function raises the question of whether it is caused primarily by coevolution between mammals and their gut microbiota and microbiome, or by the many parallel dietary shifts that have occurred over the course of mammalian evolution (*10*). We tested which of these hypotheses, which have traditionally been viewed as competing but need not be mutually exclusive, were supported by looking for congruence between mammalian phylogeny and subsets of bacterial species, KOs, CAZymes, peptidases, or other enzymatic activities. Briefly, the mammalian phylogenetic tree defines sets of organisms that are monophyletic; that is, groups containing all and only the descendants of a common ancestor. We reasoned that if bacterial taxa or functions originated rarely, then these taxa or functions should be vertically transmitted during mammalian speciation. Therefore, there should be more cases in which a given taxon or function occurred in all members of a monophyletic mammalian group than chance would predict. Using this analytic approach (*2*), we found that the overall distribution of microbial species and microbiome functions in the gut does not mirror mammalian phylogeny. 198 different named bacterial genera were detected in our data set; of these, only 3 were significantly associated with the mammalian phylogenetic tree more than would be expected by chance (Prevotella, Barnesiella, and Bacteroides). No CAZymes or peptidases and only 18 of the 3866 KOs tested were associated with host phylogeny. We repeated the analysis using a more relaxed constraint that a taxon or function occurs in a given monophyletic group more frequently than expected by chance rather than requiring strict presence or absence agreement (*2*). The relaxed definition gave similar results; only three additional genera and a total of 90 KOs were detected as having a significant association with the mammalian tree. We concluded that bacterial taxa and functions are evolutionarily labile and do not explain the concordance between bacterial communities and microbiome functions.

Mammalian gut microbiomes share a functional core

Bipartite network analysis provided an additional tool for exploring the interrelationship between host diet, host lineage, gut physiology, and shared and unique bacteria taxa *(1)*. Mammalian hosts

and bacterial OTUs were used as nodes in a bipartite graph, with edges connecting OTU nodes to the hosts in which they are found (*2*). Using 1900 V2 16S rRNA sequences from each mammalian host, the network shows clear separation of fecal communities by host diet (**Fig. 2A**), mirroring our earlier results based on smaller numbers of full-length 16S rRNA sequences (*1*).

We reasoned that the bipartite graph approach could also be used to connect mammalian samples to individual microbial gene functions from shotgun reads. The power of the bipartite graph approach is to represent both genes and mammalian species explicitly as nodes, thus visualizing which genes connect with which species. The clear separation by diet disappears when we consider gene functions (Fig. 2B, Fig. S2), suggesting that rather than a diet- or physiology-specific set of genes, the relationship among mammalian gut microbiomes is that they share a large core repertoire of functions. We confirmed this result by plotting the frequency of shared taxa in the 39 mammalian fecal samples, and also species- and genus-level OTU bins (*2*). All of the curves demonstrate an essentially exponential decay as successive samples are added, with no OTUs found in more than 30 samples (Fig. 2C). However, the plot of KO frequency flattens out, with 35 KOs found in all samples. This effect cannot be due to differences in the number of OTUs relative to KOs: There are more OTUs than KOs, and fewer assigned species or assigned genera, yet all the taxonomic curves show the same rapid decay, unlike the KOs.

Metabolic differences in microbiomes from carnivores and herbivores

This result does not imply that there are no differences among the functional configurations of microbiomes of host species having different diets. Rather, it suggests that the differences between microbiomes probably stem from differing abundances of shared functions, such as enzymes that break down chemical substrates in the host diet. We identified 495 E.C.s with significantly different proportional abundance in the 7 carnivorous and 21 herbivorous mammalian microbiomes, using the program Shotgun FunctionalizeR (adjusted $P < 0.001$ after multiple hypothesis correction) (**Table S5**) (*11*). Many of the enzymes distinguishing carnivorous and herbivorous fecal mi-

crobiomes are involved in amino acid metabolism. Microbiomes from herbivores were enriched in enzymes that map to biosynthetic reactions for 12 amino acids, whereas no carnivore samples were enriched in amino acid biosynthetic enzymes (**Table S6**). In contrast, nine amino acid degradation pathways contained enzymes that were significantly increased in carnivores, whereas the only degradative enzymes whose representation was significantly greater in herbivores were those involved in the breakdown of branched-chain amino acids (Val, Leu, and Ile). Glutamate metabolism is particularly illustrative of these trends. Both the adenosine triphosphate (ATP)-dependent and ATP-independent pathways for glutamate biosynthesis are significantly increased in herbivore microbiomes, whereas the catabolic reactions to break down glutamate and glutamine are increased in carnivores (**Fig. 3A**). These results suggest that carnivorous microbiomes have specialized to degrade proteins as an energy source, whereas herbivorous communities have specialized to synthesize amino acid building blocks.

The distinctiveness of carnivorous and herbivorous microbiomes was also revealed at a central anaplerotic node (Fig. 3B). When gluconeogenesis is required, oxaloacetate (OAA) can be converted to phosphoenolpyruvate (PEP) and pyruvate. When tricarboxylic acid cycle intermediates are withdrawn for biosynthesis, they are replenished by converting PEP and pyruvate directly to OAA (*12*). All of the genes encoding enzymes catalyzing OAA production from pyruvate or PEP are significantly increased in the carnivore microbiomes, whereas the reverse reactions are catalyzed by enzymes whose representation is increased in herbivore microbiomes.

Association of human diet with microbiome variation

Our studies comparing mammalian species revealed a relationship between host diet and gut microbial community structure and function. We next asked whether similar trends could be detected using diet variation within a single free-living host species, namely humans. Quantitative studies of diet in most human populations are complicated by the known inaccuracy of self-reported data (*13*), so we turned to a group of adults known to keep meticulous records about their daily food composition and consumption. The selected cohort consisted of 18 lean members of the Calorie

Restriction Society who typically measure and record all components of their diets on a daily basis with computer software to insure optimal nutrition despite reduced energy intake (*14, 15*). We collected their dietary records for a 4-day period (conservatively encompassing at least one complete intestinal transit time) before obtaining a single fecal sample, and analyzed macro- and micronutrient consumption using a validated protocol (*2, 16*). An average of 3642 ± 3826 bacterial V2 16S rRNA reads and 54,295 ± 28,086 shotgun reads were obtained per sample (**tables S7-S10**).

Procrustes analysis revealed a significant association between the bacterial phylogenetic structure of their fecal communities (16S rRNA) and the functions encoded in their microbiomes [$P < 0.05$ for KOs, E.C.s, and CAZymes (glycoside hydrolases)]; not significant for peptidases ($P = 0.061$) (Fig. S3). These results suggest that the processes that drive the functional differentiation of microbiomes within an individual host species may be fundamentally similar to those that drive their differentiation across mammalian evolution.

Documentation of the weight of each ingredient in each meal consumed by these individuals (Table S7) allowed us to perform a follow-up analysis examining the impact on fecal bacterial community configuration of three dietary components (total protein, carbohydrate, and insoluble fiber intake). We chose these diet categories because protein intake is markedly different between carnivores and herbivores, and because an extensive literature exists about the impact of ingested polysaccharides and fiber on the gut microbiota (*17*). Linear regression of the three dietary categories against the position of each individual's microbiome along principal coordinate 1 of the PCoA plots revealed that total protein intake was significantly associated with KO data [adjusted linear regression coefficient of determination (R^2) value = 0.307, adjusted P value = 0.030] (*2*). In contrast, insoluble dietary fiber was significantly associated with bacterial OTU content (Bray-Curtis metric; adjusted R^2 value = 0.371; adjusted P value = 0.013) (Table S11). These results confirm that within a single free-living species, both the structure and function of the gut microbiome are significantly associated with dietary intake.

Prospectus

Taken together with our prior work *(1),* these results teach us that even fecal samples from mammals living in zoos and human samples from a single self-selected population can provide insights into the factors driving the evolution of the gut microbiome. They also compel us, at a time when complete genomes are to be sequenced for 10,000 vertebrates (*18*), to take the next step and perform systematic studies that rigorously test specific mechanisms that drive the evolution of hosts and their (gut) microbial symbionts. These studies should be guided by experts who can choose taxa that radiated at different points in their evolutionary history, with parallel shifts in their diet, morphology, biogeography, or other key factors known or hypothesized to influence evolution. The results should help address questions such as what functional features in host intestinal environments (including the biochemical characteristics of mucosal surfaces) are related to the representation of specific bacterial taxa and microbiome functions, and how readily microbial populations have been acquired and reacquired during the course of vertebrate evolution. Additionally, our findings emphasize the need to sample humans across the globe with a variety of extreme diets and lifestyles, including relatively ancestral hunter-gatherer lifestyles, in order to provide new insights into the limits of variation within a host species and the possibility that our microbes, in coevolving with our bodies and our cultures, have helped shaped our physiological differences and environmental adaptations.

Acknowledgements

We thank Jill Manchester and Sabrina Wagoner for technical assistance, Brandi Cantarel, Vincent Lombard, Corinne Rancurel, and Pedro Coutinho for CAZyme annotation, Ruth Ley and members of the Gordon lab for their suggestions; and Stephen Bircher Rob Ramey, Michael Schlegel, Mark Schrenzel, Tammy Tucker, and Peter Turnbaugh for past help in procuring mammalian fecal samples. This work was supported by grants from NIH (DK30292, DK70977, DK078669, UL1 RR024992), the Crohn's and Colitis Foundation of America, and NIH Institutional Training Grant T32-A1007172 (to B.D.M.). The National Center for Biotechnology Information Sequence Read

Archive accession number for the bacterial 16S rRNA and shotgun data sets related to our studies of mammalian and Calorie Restriction Society member human fecal microbiomes is SRA030940. Sequence data has also been deposited in MG-Rast with project accessions qiime:625, qiime:626, qiime:627, and qiime:628.

References

1. R. E. Ley *et al*., Evolution of mammals and their gut microbes. *Science* 320, 1647 (2008).

2. See supplemental information.

3. M. Kanehisa, S. Goto, KEGG: Kyoto encylocpedia of genes and genomes. *Nucleic Acids Res*. **28**, 27 (2000).

4. B. L. Cantarel *et al*., The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res*. **37**, (database issue), D233 (2009).

5. N. D. Rawlings, A. J. Barrett, A. Bateman, MEROPS: the peptidase database. *Nucleic Acids Res*. **38**, (database issue), D227 (2010).

6. D. H. Huson, A. F. Auch, J. Qi, S. C. Schuster, MEGAN analysis of metagenomic data. *Genome Res*. **17**, 377 (2007).

7. J. R. Hurley, R. B. Cattell. The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behav. Sci*. **7**, 258 (1962).

8. J.C. Gower. Generalized Procrustes analysis. *Psychometrika*, **40**, 33 (1975).

9. C. Lozupone, R. Knight, UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol*. **71**, 8228 (2005).

10. T. E. Cerling, J. R. Ehleringer, J. M. Harris, Carbon dioxide starvation, the development of C4 ecosystems, and mammalian evolution. *Philos. Trans. R. Soc. London Ser. B* **353**, 159 (1998).

11. E. Kristiansson, P. Hugenholtz, D. Dalevi, ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* **25**, 2737 (2009).

12. O. E. Owen, S. C. Kalhan, R. W. Hanson, The key role of anaplerosis and cataplerosis for citric acid cycle function. *J. Biol. Chem*. **277**, 30409 (2002).

13. K. Poslusna, J. Ruprich, J. H. M. de Vries, M. Jakubikova, P. van't Veer, Misreporting of

energy and micronutrient intake estimated by food records and 24 hour recalls, control and adjustment methods in practice. *Br. J. Nutr.* **101**, S73 (2009).

14. L. Fontana, S. Klein, Aging, adiposity, and calorie restriction. *JAMA* **297**, 986 (2007).

15. L. K. Heilbronn *et al*., Pennington CALERIE Team., Effect of 6-month calorie restriction on biomarkers of longevity, metabolic adaptation, and oxidative stress in overweight individuals: a randomized controlled trial. *JAMA* **295**, 1539 (2006).

16. S. F. Schakel, Y. A. Sievert, I. M. Buzzard, Sources of data for developing and maintaining a nutrient database. *J. Am. Diet. Assoc.* **88**, 1268 (1988).

17. H. J. Flint. Polysaccharide breakdown by anaerobic microorganisms inhabiting the Mammalian gut. *Adv. Appl. Microbiol.*. **56**, 89 (2004).

18. Genome 10K Community of Scientists, Genome 10K: a proposal to obtain whole-genome sequences for 10,000 vertebrate species. *J. Hered.* **100**, 659 (2009).

Figure Legends

Figure 1. Procrustes analysis shows that mammalian gut bacterial lineages and microbiome gene content give similar clustering patterns. (A) Cartoon illustrating Procrustes analysis. The Procrustes transformation of the blue and red data types (i) results in a good fit, while the transformation of the green and red data (ii) yields a worse fit with large distances separating data from samples B and C. **(B-E)** Procrustes analysis of 16S rRNA sequences (weighted UniFrac) against KEGG Orthology (KO) groups, CAZymes (glycoside hydrolases), MEROP peptidases, and E.C.s. Every sphere represents a single mammalian fecal community and is colored by host diet. The black end of each line connects to the 16S rRNA data for the sample, while the orange end is connected to the functional annotation data. The fit of each Procrustes transformation over the first three dimensions is reported as the M^2 value.

Figure 2. Mammalian gut bacterial communities share a functional core. (A-B) Bipartite network diagrams of evenly sampled bacterial 16S rRNA-derived OTUs (**A**) or KOs (**B**). Edges connecting mammalian nodes (circles) to species-level OTUs or KOs found in that sample are colored by host diet. Sample labels are removed from the KO diagram for legibility (high-resolution image of removed labels presented in **Fig. S2)**. **(C)** Mammalian gut communities share a core suite of KOs. Using evenly subsampled OTU or KO datasets, the distribution of counts is plotted as a function of the number of mammalian host microbiomes where the KO or phylotype was detected. The results demonstrate exponential decay for the 16S rRNA data (OTU, species, and genus), with no OTU or bacterial species found in all samples, although a "core" set of KOs is detectable in all fecal communities sampled.

Figure 3. Differences in metabolic features encoded in fecal microbiomes among herbivores versus carnivores. (A) Carnivorous and herbivorous microbiomes indicate opposing directionality for amino acid metabolism. Colored arrows denote enzyme functions whose representation is signficiantly ($P < 0.001$) greater in the fecal microbiomes of herbivores (green) or carnivores (red). **(B)** Carnivorous and herbivorous microbiomes suggest opposing directionality at the central

PEP-Pyruvate-Oxaloacetate node. Coloring scheme as in panel A. Abbreviations: 2-OG, alpha-ketoglutarate; Ck, carboxykinase; Cx, carboxylase; DH, dehydrogenase; Dx, decarboxylase; GABA, γ–aminobutyrate; GDH, Glu DH; OAA, oxaloacetate; ODx, OAA Dx; PEP, phosphoenolpyruvate; PPDk, Pyr-Phosphate Dikinase; Pyr, Pyruvate; SSA, Succinate-Semialdehyde.

Figures

Figure 1.

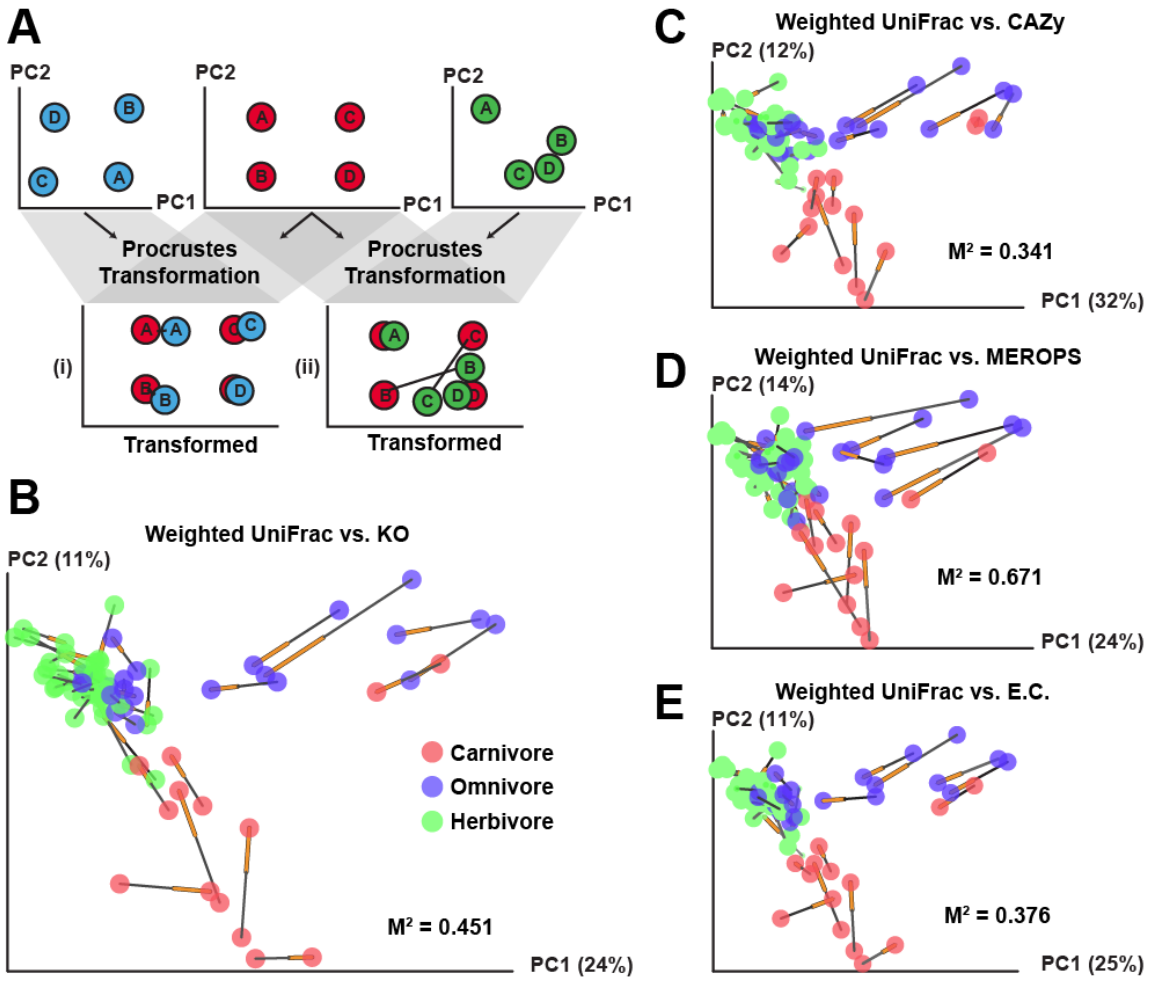


Figure 2.

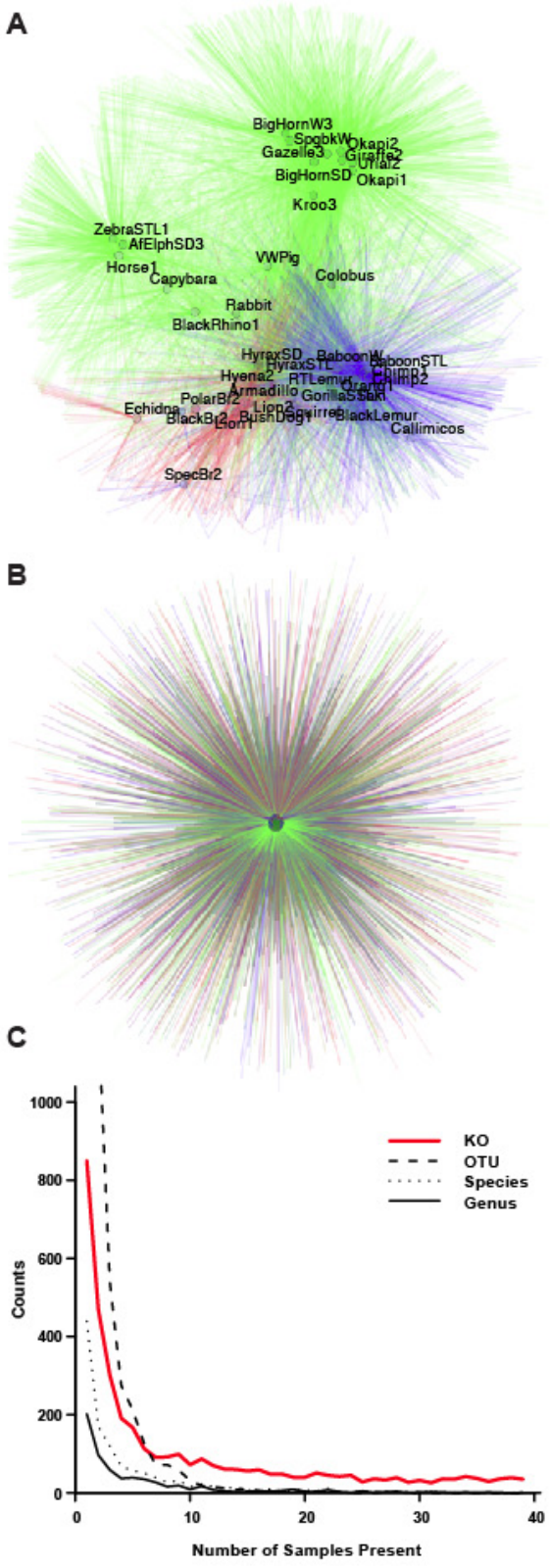
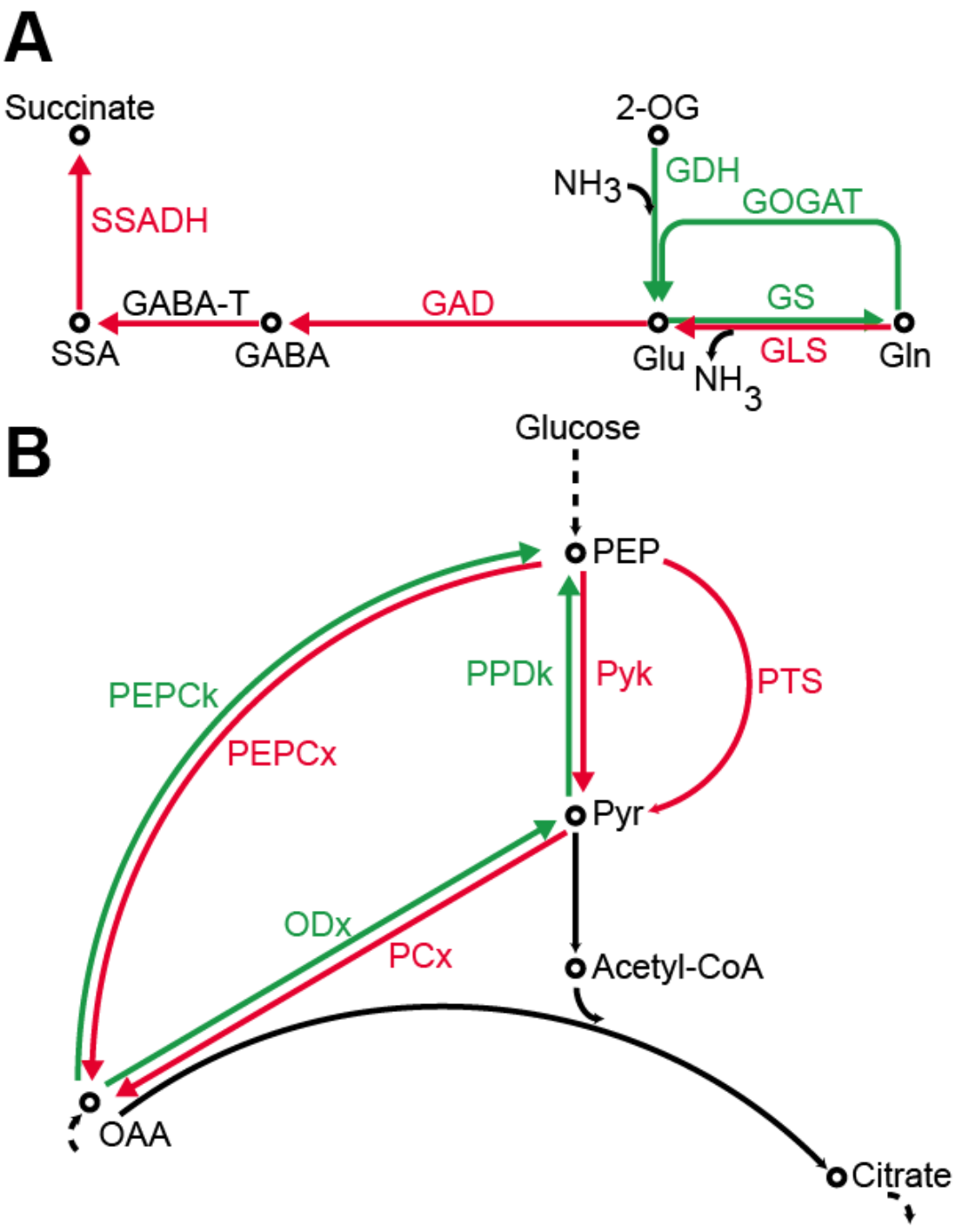

A
BigHornW3
SpgbkW
Okapi2
Gazelle3
Giraffe2
Urial2
BigHornSD
Okapi1
Kroo3
ZebraSTL1
AfElphSD3
VWPig
Horse1
Colobus
Capybara
Rabbit
BlackRhino1
HyraxSD
BaboonW
HyraxSTL
BaboonSTL
Chimp2
Hyena2
RTLemur
Armadillo
PolarBr2
Echidna
BlackLemur
Callimicos
SpecBr2
B
C
1000
800
600
400
200
0
Counts
0
10
20
30
40
Number of Samples Present
KO
OTU
Species
Genus

Figure 3.


A
Succinate
2-OG
NH3
GDH
GOGAT
SSADH
GABA-T
GAD
GS
SSA
GABA
Glu
NH3
GLS
Gln
B
Glucose
PEP
PEPCk
PEPCx
PPDk
Pyk
PTS
Pyr
ODx
PCx
Acetyl-CoA
OAA
Citrate

Tables

Table 1. Overview of mammals in this study. Sample abbreviations used in figures and tables are noted in parentheses. See **Table S1** for additional details.

Foregut-fermenting herbivores	**Hindgut-fermenting herbivores**
Bighorn sheep 1 (BigHornSD)*	African elephant (AfElphSD3)*
Bighorn sheep 2 (BigHornW)‡	Black rhinoceros (BlackRhino1)†
Colobus (Colobus)†	Capybara (Capybara)†
Gazelle (Gazelle3)†	Gorilla (GorillaSTL)†
Giraffe (Giraffe2)†	Horse (Horse1)‡
Rock Hyrax 1 (HyraxSD)*	Orangutan (Orang1)†
Rock Hyrax 2 (HyraxSTL)†	European rabbit (Rabbit)†
Kangaroo (Kroo3)†	Zebra (ZebraSTL1)†
Okapi 1 (Okapi1)†	
Okapi 2 (Okapi2)†	
Springbok (SpgbkW)‡	**Omnivores**
Transcaspian Urial sheep (Urial2)†	Baboon 1 (BaboonSTL)†
Visayan warty pig (VWPig)*	Baboon 2 (BaboonW)‡
	Black bear (BlackBr2)†
Carnivores	Black lemur (BlackLemur)†
Armadillo (Armadillo)†	Callimicos (Callimicos)†
Bush dog (BushDog1)†	Chimpanzee 1 (Chimp1)†
Echidna (Echidna)†	Chimpanzee 2 (Chimp2)†
Hyena (Hyena2)†	Ringtailed lemur (RTLemur)†
Lion 1 (Lion1)†	Saki (Saki)†
Lion 2 (Lion2)†	Spectacled bear (SpecBr2)†
Polar bear (PolarBr2)†	Squirrel (Squirrel)†

*=San Diego Zoological Park. †=St. Louis Zoo. ‡=Wild, free-living.

Supplemental Information

Interactive Plots on the Web

The kinemage files used to generate Figure 1 have been posted on our lab website (*19*), together with links for panels in that figure. Clicking the links will open java applets that contain an interactive, labeled PCoA plot of the Procrustes analysis used to generate the panel. The viewer can rotate the plot in 3-dimensions, and can also view the first 10 principal coordinate axes. The coloring and naming conventions online are consistent with Figure 1.

Materials and Methods

Subjects

Table S1 lists the mammalian species included in this study and their diet group. Further details about their diets, dietary group classification, plus methods used to recover and store fecal samples have been published previously (*1*). Eighteen members of the Calorie Restriction Society were recruited for the present study using a procedure approved by the Washington University Human Studies Committee. The volunteers had practiced calorie restriction for an average of 7.6 yrs (range 3.5-21 years). The average age of the study cohort was 59.6±10.6 years and their body mass index was 19.4±1.3 kg/m^2 (mean±S.D.). None of these individuals had consumed antibiotics in the four months prior to enrollment and none had a history of gastrointestinal disorders. Each participant provided a fecal specimen that was frozen at -20^{o}C within 30 min after its production. Samples were maintained at this temperature until they were received (within 24h) at a biospecimen repository where they were anonymized and stored at -80^{o}C prior to metagenomic analyses. The participants kept detailed diet records for four days prior to fecal collection. A dietician analyzed these diet records to quantify macro- and micronutrient content using the Nutrition Data System for Research (NDS-R; version 4.03_31) (*16*).

Isolation of fecal DNA and multiplex pyrosequencing

All mammalian and human samples were subjected to a common protocol for DNA extraction. Fecal samples were pulverized with a mortar and pestle at -80°C. A 500 mg aliquot of each frozen pulverized sample was re-suspended in a solution containing 500µL of extraction buffer [200mM Tris (pH 8.0), 200mM NaCl, 20mM EDTA], 210µL of 20% SDS, 500µL of phenol:chloroform:isoamyl alcohol (25:24:1) and 500µL of a slurry of 0.1-mm diameter zirconia/silica beads. Cells were then mechanically disrupted using a bead beater (Biospec, maximum setting; 3 min at room temperature), followed by extraction with phenol:chloroform:isoamyl alcohol and precipitation with isopropanol. An aliquot of the DNA was used for PCR amplification and sequencing of bacterial 16S rRNA genes. ~330bp amplicons, spanning variable region 2 (V2) of the gene were generated by using (i) modified primer 8F (5′- GCCTTGCCAGCCCGCTCAG**TC***AGAGTTTGATCCTGGCTCAG*-3') which consists of 454 FLX Amplicon primer B (underlined), a two base linker (bold) and the universal bacterial primer 8F (italics) and (ii) modified primer 338R (5' GCCTCCCTCGCGCCATCAGNNNNNNNNNNNN**CA***TGCTGCCTCCCGTAGGAGT* 3') which contains 454 FLX Amplicon primer A (underlined), a sample specific, error correcting 12-mer barcode (N's), a two base linker (bold), and the bacterial primer 338R (italics). Three replicate polymerase chain reactions were performed for each fecal DNA sample: each 20-mL reaction contained 100 ng of gel-purified DNA (Qiaquick, QIAGEN), 8 ml 2.5X HotMaster PCR Mix (Eppendorf) and 0.2 µM of each primer. PCR conditions consisted of an initial denaturation step performed at 95 °C for 2 min, followed by 30 cycles of denaturation (95°C, 20 s), annealing (52°C, 20 s) and amplification (65°C, 1 min). Amplicons generated from each set of three reactions were subsequently pooled and purified using Ampure magnetic purification beads (Agencourt). The amount of DNA was quantified using Picogreen (Invitrogen), and equimolar amounts of barcoded samples were pooled for each subsequent multiplex 454 FLX pyrosequencer run.

For multiplex shotgun 454 FLX pyrosequencing, each fecal community DNA sample was randomly fragmented by nebulization to 400-800 bp (FLX standard) or 500-800 bp (FLX Tita-

nium), and then labeled with a distinct MID (Multiplex IDentifier; Roche) using the MID manufacturer's protocol (general library preparation for FLX standard, Rapid Library preparation for FLX Titanium). Equivalent amounts of up to 12 MID-labeled samples were pooled prior to each pyrosequencer run (454 FLX and Titanium chemistry).

16S rRNA data processing and analysis

16S rRNA amplicon sequences were processed using the QIIME (v1.1) suite of software tools (*20*); fasta files, quality files and a mapping file indicating the sequence of the 12 nt barcode that corresponded to each sample were used as inputs. QIIME bins pyrosequencer reads by samples according to their barcode, de-noises pyrosequencer data (*21*), and classifies reads into OTUs on the basis of sequence similarity [e.g., species level phylotypes share ≥97% identity (ID)]. QIIME builds a *de novo* taxonomic tree of the sequences based on their similarity and creates a table of samples versus OTUs that can be used, together with the tree, to calculate alpha and beta diversity. Reads were aligned using PyNAST, and chimeric OTUs were removed using the ChimeraSlayer program (*22*). Procrustes analysis and network analysis were also performed in QIIME, using code made available in release 1.2.0. The network analysis was visualized using Cytoscape v2.6.3 (*23*), and the nodes and edges were placed using Cytoscape's spring-embedded algorithm.

Taxonomic assignments were made using SILVA-VOTE, an algorithm designed for improved accuracy in taxonomic assignments of V2 16S rRNA reads. Briefly, a non-redundant reference database of 34,181 bacterial 16S rRNA V2 regions was created by clustering the Silva database (release 102). Taxonomic assignment was made at each taxonomic level if more than 75% of the sequences in the cluster had an identical designation at that level (otherwise, the level was designated "unknown"). A representative sequence from each QIIME-identified OTU was compared by BLAST against this custom database, retaining hits with e-value $\leq 10^{-30}$. All BLAST hits within 10% of the best score (up to 100 hits) were used to generate a taxonomic assignment at each taxonomic level. If greater than 50% of the hits shared a designation, this annotation was assigned to the OTU. Otherwise, the OTU was noted as "nonidentified" at that level. Note that

sequences were binned into OTUs with ≥97% sequence identity in lane-masked V2 regions. In accord with common convention (*24*), when 97%ID OTUs had identical taxonomic assignments, they were binned as a single 'species-level phylotype'. When 97%ID OTUs were assigned to the same bacterial genus, they were treated as a single 'genus-level phylotype'.

Shotgun sequencing data processing and annotation

Metagenomic data was annotated using custom Perl scripts. Shotgun reads were filtered using custom Perl scripts and publicly available software to remove (i) all reads less than 60 bases in length, (ii) reads with degenerate bases (N's), (iii) all duplicates (a known artifact of pyrosequencing), defined as sequences whose initial 20 nucleotides are identical and that share an overall identity of >97% throughout the length of the shortest read (*25*) and (iv) in the case of human fecal DNAs from the calorically restricted individuals, all sequences with significant similarity to human reference genomes (BLASTN with e-value $\leq 10^{-5}$, bitscore ≥ 50, percent identity $\geq$75%) to ensure the continued anonymity of samples.

Searches against the KEGG (version 52) and MEROPS (release 9.1) databases were carried out with BLASTX. A sequence read was annotated as the best hit in the database if (i) the E-value was $\leq 10^{-5}$, (ii) the bit score was ≥ 50, and (iii) the alignment was at least 50% identical between query and subject. In the event that two entries in the database had equivalent BLAST scores as the best hit, the read was annotated with both entries. The KO, E.C., and KEGG Pathways associated with each KEGG sequence were determined using the "ko" file provided by KEGG. Reads were annotated against the CAZy database using procedures described previously (*26*). For each functional annotation schema, statistical analysis was performed on a matrix containing the count of annotated reads in each sample; e.g., for the KEGG KO data, the matrix contained the list of all possible KOs in the rows of the first column and the sample names in the column headers. The value in each "cell" of the matrix was the number of times that KO was detected as the best BLAST hit of a shotgun read from the sample. All dissimilarity metrics and related calculations were generated with QIIME. For every functional data type, an evenly rarefied matrix of func-

tional assignments was created, a distance matrix using the Bray-Curtis metric was calculated, and results were visualized with Principal Coordinates and Procrustes Analysis.

Taxonomic composition of shotgun sequencing data

The taxonomic distribution of metagenomic reads was determined using version 3.9 of MEGAN (*6*). Metagenomic reads were searched against the NCBI non-redundant protein database with BLASTX. We used the additional BLAST parameter *–F "m S"* as suggested by the authors of the software. The search results were processed in MEGAN with default parameters to generate the taxonomic profile for each sample.

Detecting differences in E.C. abundance and amino acid metabolism

E.C. analysis comparing the microbiomes of herbivores and carnivores was implemented with version 1.0.3 of Shotgun Functionalize R (*11*). As described in the main text, 495 E.C.'s were identified with significantly different relative abundance between carnivore and herbivore microbiomes, using a Benjamini-Hochberg adjusted p-value of 0.001 as our threshold for significance. These E.C.'s were mapped onto KEGG metabolic pathways and inspected visually. For every amino acid, biosynthetic and degradative reaction pathways were identified using both KEGG annotations and the experimentally confirmed metabolic information collated in the MetaCyc database (27, 28). For every significantly different E.C. that was detected in a major amino acid biosynthetic or catabolic reaction, the relative abundance of the E.C. in the microbiomes was plotted. The objective was to eliminate statistically significant results from the final analysis if there were not a sufficient number of annotations for biological confidence. The following statistically significant E.C.s were not included in the summary analysis: (i) E.C.2.1.4.1 (glycine amidinotransferase) was found in only two microbiomes with low abundance (four total assignments in Bush Dog, two total assignments in Hyena), and (ii) E.C.2.6.1.57 (aromatic amino-acid transferase) which was found in only 7 microbiomes with low abundance (maximum of 15 assignments in Polar Bear, median of two assignments). The results of this analysis are summarized in **Table S6**. For every amino acid,

E.C.s detected in biosynthetic or degradative reactions are noted, as well as reversible reactions where likely direction cannot be determined based on available information.

Using Procrustes analysis to test whether the functional properties of a microbiome can be predicted from the bacterial species

Procrustes analysis, named after the son of the Greek god Poseidon who fit unsuspecting travelers to a fixed-size bed by stretching them or removing their feet, is a technique for comparing the relative positions of points in two multivariate datasets. The method was first introduced by Hurley and Cattell (*7*), then generalized by Gower (*8*), for comparing psychometric datasets. The method has since been employed in macro-ecology (*29, 30*), and in a recent report used to compare datasets obtained from different regions of the same 16S rRNA sequence (*31*). We used the implementation of Procrustes analysis in the open-source QIIME microbial community analysis pipeline (*20*), built using the PyCogent libraries (*32*) and the Python programming language. As noted in the main text, this procedure yields a measure of fit, M^2, which is the sum of squared distances between corresponding data points after the transformation. The significance of the association is obtained by a Monte Carlo procedure in which the point labels are randomized, M^2 is recomputed, and the M^2 value of the actual pair of datasets is compared to the empirical distribution of M^2 values observed for the permuted datasets. Because M^2 depends on the sample size and the structure of the data, M^2 values typically cannot be directly compared between datasets, and the statistical significance must be computed for each pair of datasets separately. In these studies, we used 1,000 replicates for calculating P-values.

We applied the Bray-Curtis distance metric to the functional data to obtain PCoA coordinates for comparison with the 16S rRNA data. We also tested additional qualitative (Jaccard) and quantitative (Canberra, Gower) distance metrics on the KO data. All of these distance metrics led to the same essential conclusion; the agreement between the weighted UniFrac distances and the KO distance matrix was significant over the first three dimensions of the Procrustes plot ($p<0.05$).

Assessing clustering using the Monte Carlo procedure

Monte Carlo simulations allow researchers to directly determine the probability of obtaining a result more extreme than the observed metric with random sampling of the data. Importantly, this test makes no *a priori* assumptions of the underlying structure of the data (e.g, a parametric distribution) and can thus be powerfully deployed with a range of experimental results. To assess sample clustering by diet, we computed the t-statistic of the UniFrac distances of 16S rRNA data and (separately) the Bray-Curtis distances of KO data, comparing the average distance between the fecal bacterial communities of herbivores to the average distance between carnivore communities. This t-statistic was treated as the "observed" result. Then, the pairwise distance matrix was randomly permuted a set number of times by shuffling the sample labels; in this study, we always performed 1,000 independent permutations. For each permutation, the t-statistic was recalculated using the new permuted labels. The distribution of the t-test statistic for the permutations was compared to the "observed" metric from the real data. The fraction of times a permutation resulted in a metric more extreme than the observed metric is the p-value, the probability that the observed result could have arisen by chance from the underlying data. For example, if 11 of the 1,000 random permutations comparing the average UniFrac distance of herbivore fecal bacterial communities to the average UniFrac distance of carnivore communities had a t-statistic more extreme than the actual observed statistic, the reported p-value would be 11/1000, or 0.011. All Monte Carlo simulations were implemented using QIIME scripts. The same procedure was employed to assess the clustering of conspecific samples, here comparing unweighted and weighted UniFrac distances between animals of the same species to distances between animals of different species.

Regression analysis of data obtained from human subjects

Linear regression was carried out using the R statistical software package with a simple linear model (*33*). Principal Coordinates were generated using QIIME as described previously for the 16S rRNA OTU data and functional KO data using the Bray-Curtis distance metric, and for the 16S rRNA data using the weighted and unweighted UniFrac distances. Principal Coordinates for

the other types of functional annotation were calculated but not included in linear regression model because they had a high Pearson correlation coefficient with the KO Principal Coordinates and thus did not represent independent response variables (E.C correlation=0.97, CAZyme correlation=0.87, peptidase correlation=0.73).

For each individual, the average daily intake of carbohydrates, proteins, and insoluble fiber was calculated based on the dietary records from the four days proceeding fecal sample donation. The position for each calorie-restricted individual's gut bacterial community along Principal Coordinate 1 was regressed against each of the three dietary components using a simple linear model. The p-value resulting from the analysis was multiplied by three to adjust for the multiple hypotheses tested (Bonferroni adjustment). We also implemented a multiple linear regression using all three dietary components and their interactions as explanatory variables, with the Principal Coordinate positions as the response variable. We used backwards stepwise selection to remove non-significant terms from the model. In no case did a mixed model generate significant interactions beyond the simple linear models tested previously.

The average daily calories consumed by our cohort ranged from 1207-2551 kCal/day (mean 1673, standard deviation 395 kCal/day). We regressed total kilocalories consumed against the PC1 coordinate of our samples for all data categories (16S rRNA and functional data). After correcting for multiple-hypothesis testing, total calories were not significantly associated with any of these categories ($p > 0.05$).

Results

Prediction of community functional profiles from species assemblage data using a nearest-neighbor model

As noted in the main text, the strong correlation between bacterial 16S rRNA and functional profiles made us wonder if the functional configuration of a microbiome could be predicted from its 16S rRNA sequences. To test this idea, we developed a nearest-neighbor model. For a given

sample, we predicted its functional composition to be the same as that sample's nearest neighbor (using the weighted UniFrac distance comparison of 16S rRNA data). To assess the quality and significance of these predictions, we compared the average root mean squared error (RMSE) of our model to the average RMSE for one million Monte Carlo trials where each sample's nearest neighbor was chosen at random from the remaining samples. The UniFrac nearest neighbor generated a significantly better functional prediction than a random neighbor for all four types of functional; for KOs, E.C.s, peptidases, and CAZymes, no permutation in the one million trials had a lower RMSE than the UniFrac prediction ($p=0$). Using the unweighted UniFrac distances also led to predicted functional profiles that were significantly better than would be expected by chance (KOs, $p=0$; E.C.s, $p=0$; proteases, $p=0.000252$; CAZymes, $p=0$).

Phylogenetic congruence testing

We examined the congruence between host phylogeny and the presence of bacterial taxa or particular enzymatic (or other protein) functions in host microbiomes by comparing host phylogeny to both the OTUs and KOs present in the fecal samples. We first generated a table of each OTU's or KO's presence or absence in each sample, and selected a number of sequences (without replacement) from each sample's sequences to avoid biases associated with uneven sequencing effort among samples (we selected 1,000 bacterial 16S rRNA sequences from each sample for the OTU analysis, and 5,000 assigned shotgun sequences for the KO analysis). We then eliminated OTUs or KOs present in only one sample, as they were not useful for assessing the congruence between fecal microbiomes and host phylogeny. We then searched for OTUs or KOs that were present in all members of any distinct monophylogenetic lineage and absent in all other samples included in this study. We compared this result to the results obtained on a randomized host phylogeny where the samples we obtained were associated with randomly chosen tips of the host phylogeny. To summarize the results, we grouped OTUs with assignments to identical bacterial genera, and looked for bacterial genera enriched for OTUs matching the mammalian phylogeny. Of the 198 different genera assignments detected in the subsampled OTU dataset, three (Prevotella, Barnesiella, and Bacteroides) were found to be significantly enriched in matching OTUs (see main text;

$p<0.05$ relative to a binomial distribution using the overall abundance of matching OTUs). OTUs that lacked a genera assignment were not reported when identifying bacterial genera congruent with mammalian phylogeny. Of the 668 OTUs not assigned to a named genera, 29 (4%) matched a monophyletic group in the mammalian tree, a similar proportion found in the genus-assigned OTUs.

To account for OTUs and functions whose presence/absence pattern did not exactly match a monophyletic group of mammals, we performed a test for increased presence within a monophyletic group relative to the expected value given the overall abundance of the OTU or function over all samples, using a binomial distribution. Categorizing these OTUs into genera as before, Prevotella, Barnesiella, and Bacteroides were significantly enriched for OTUs whose presence was congruent with the mammalian phylogeny, as well as Ureaplasma, Paludibacter and Pedobacter. Employing this method, we found 90 KOs, 2 CAZymes (of 119 tested), and 1 (of 274 tested) proteases congruent with host phylogeny.

Additionally, we employed the method of Ochman *et. al.* (*34*) as another test for co-phylogeny. Their study examined four wild primates species and their fecal microbiota, and created a maximum parsimony tree of the gut communities using a character matrix of bacterial species normalized abundances that was compared to the primate phylogeny (as determined by mitochondrial DNA sequence). To implement their approach, we used MESQUITE rather than PAUP for parsimony inference and eliminated bacterial taxa that appeared in only a single sample (non-parsimony-informative sites). We also tried a variety of parsimony variants (ordered states based on discretized z-scores, with SPR and NNI heuristics for tree search). We were able to reproduce their results for the four primate species using their source data. However, when applied to our larger mammalian tree of 33 species, we did not see an overall match between the host tree and the parsimony tree inferred from fecal bacterial communities, providing further evidence supporting our conclusion that the distribution of bacterial species does not mirror host phylogeny over the whole of the mammalian tree. We did not sample any one closely related clade with sufficient den-

sity to perform extensive tests of whether the bacterial community signal Ochman and coworkers identified dissipates once a particular evolutionary depth is reached.

Supplemental References

19. http://gordonlab.wustl.edu/Mammals_2011/

20. J. G. Caporaso *et al*., QIIME allows analysis of high-throughput community sequencing data. *Nat. Meth.* **7**, 335 (2010).

21. J. Reeder, R. Knight, Rapidly denoising pyrosequencing reads by exploiting rank-abundance distributions. *Nat. Meth.* **7**, 668 (2010).

22. B. J. Haas *et al*., Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* **21**, 494 (2011).

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

24. N. Youssef *et al*., Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl. Environ. Microbiol.* **75**, 5227 (2009).

25. V. Gomez-Alvarez, T. K. Teal, T. M. Schmidt, Systematic artifacts in metagenomes from complex microbial communities. *ISME J* **3**, 1314 (2009).

26. P. J. Turnbaugh *et al*., A core gut microbiome in obese and lean twins. *Nature* **457**,480 (2009).

27. R. Caspi *et al*., The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathways/genome databases. *Nucleic Acids Res.* **38**, (database issue) D473 (2010).

28. http://www.metacyc.org/.

29. S. Dray, D. Chessel, J. Thioulouse. Co-inertia analysis and the linking of ecological data tables. *Ecology*, **84**, 3078 (2003).

30. K.E. Francl, Schnell G.D. Relationships of human disturbance, bird communities, and plant communities along the land-water interface of a large reservoir. *Environ. Monit. Assess.* **73**, 67 (2002).

31. J. G. Caporaso *et al.*, Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U.S.A.* 108 (suppl. 1), 4516 (2010).

32. R. Knight *et al.*, PyCogent: a toolkit for making sense from sequence. *Genome Biol.* **8**, R171 (2007).

33. R Foundation for Statistical Computing. http://www.R-project.org .

34. H. Ochman *et al.*, Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biol* **8**, e1000546 (2010).

Supplemental Figure Legends

Figure S1. Procrustes analysis is robust to a variety of computational approaches. Procrustes analysis of 16S rRNA sequences (weighted UniFrac, unweighted UniFrac, OTU counts) against KO annotation of shotgun pyrosequencing reads. Every sphere represents a single mammalian fecal community and is colored by host diet. The black end of each line connects to the 16S data for the sample, while the orange end is connected to the functional annotation data. The fit of each Procrustes transformation over the first three dimensions,is reported as the M^2 value. (A) Procrustes analysis of 16S rRNA data (unweighted UniFrac) against KEGG Orthology (KO) groups. (B) Procrustes analysis of OTU counts (Bray Curtis metric) against KOs. (C) Procrustes analysis of 16S rRNA data (weighted UniFrac) against KOs, using only one animal sample from each of the 33 mammalian species.

Figure S2. Bipartite network analysis. (A) Close-up of animal node labels for KO bipartite graph from **Fig. 2B** in main text. **(B-C)** Bipartite network diagrams of evenly sampled CAZymes [glyoside hydrolases] (**B**) ot peptidases (**C**) Labeled circles (nodes) denote animal hosts, and are colored by host diet. Lines (edges) radiating from the host nodes connect to microbiome gene nodes representing a single glycoside hydrolase or peptidase found in the host fecal microbiome.

Figure S3. Procrustes analysis shows that the bacterial lineages and microbiome gene content from humans who practice caloric restriction with adequate nutrition give similar clustering patterns. (A-D) Procrustes analysis of bacterial 16S rRNA sequences (weighted UniFrac) against KOs, CAZymes (glycoside hydrolases), MEROPS (peptidases), and Enzyme Commission numbers (E.C.s). Every sphere represents a single mammalian fecal community and is colored by host diet. The black end of each line connects to the 16S data for the sample, while the orange end is connected to the functional annotation data. The fit of each Procrustes transformation over the first three dimensions is reported as the M2 value. Spheres are colored differently for each human host.

Supplemental Figures

Figure S1.

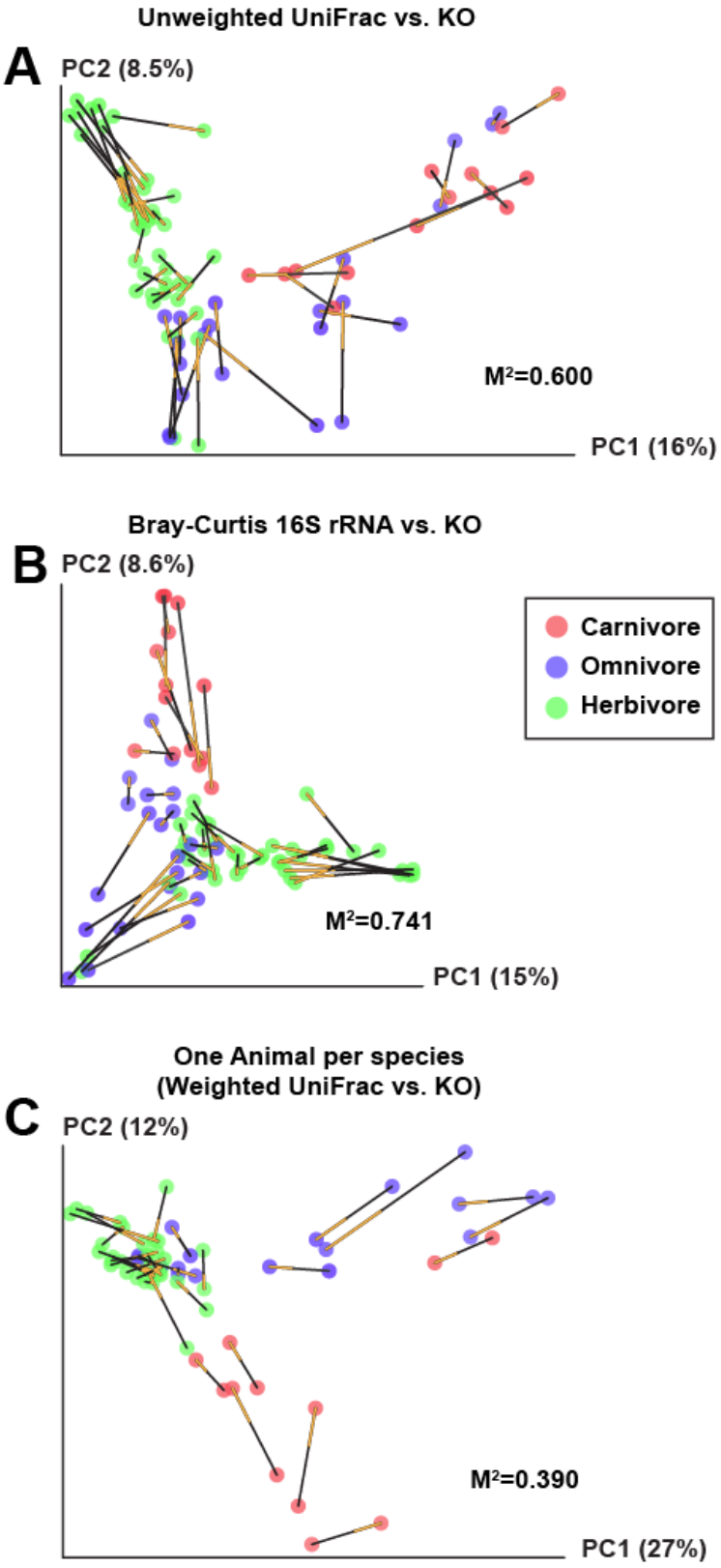


Figure S2.

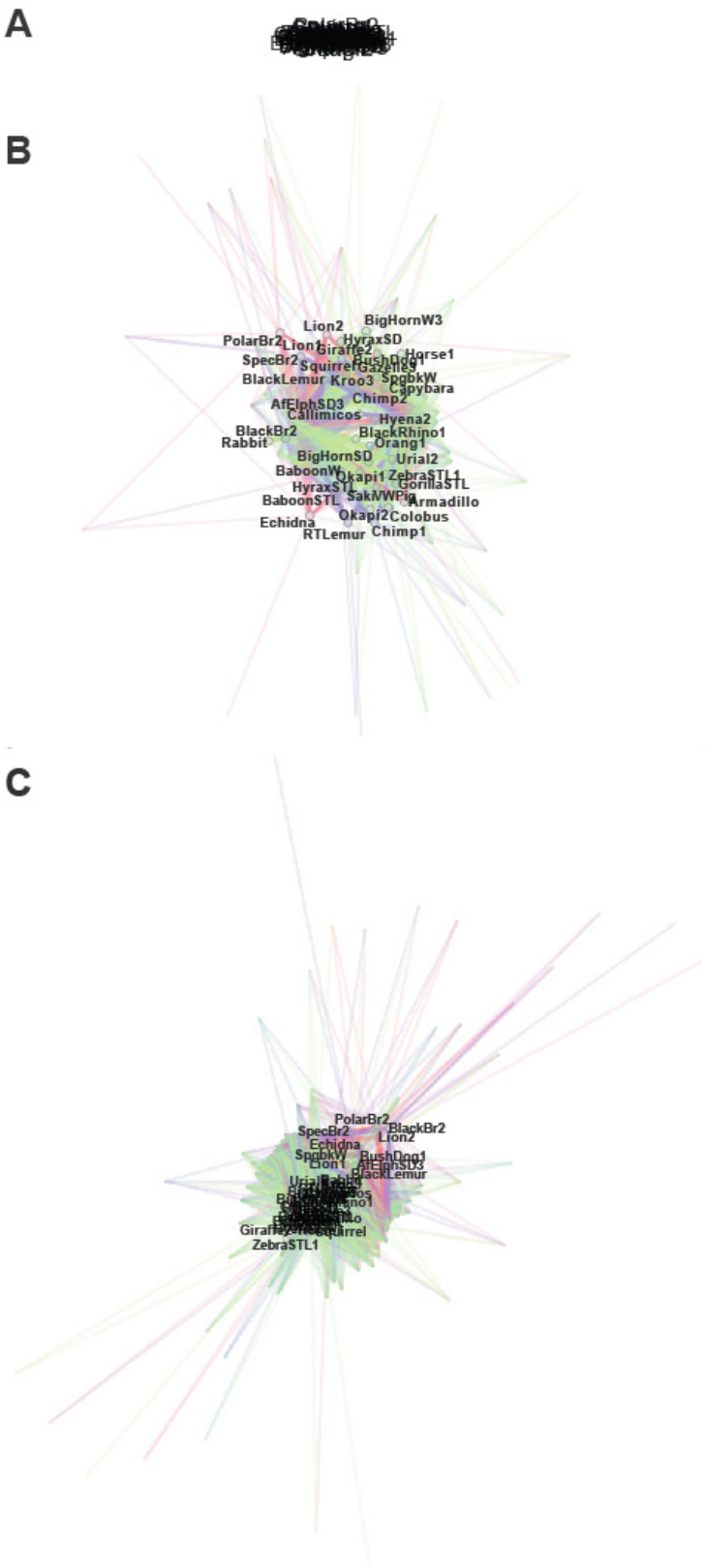
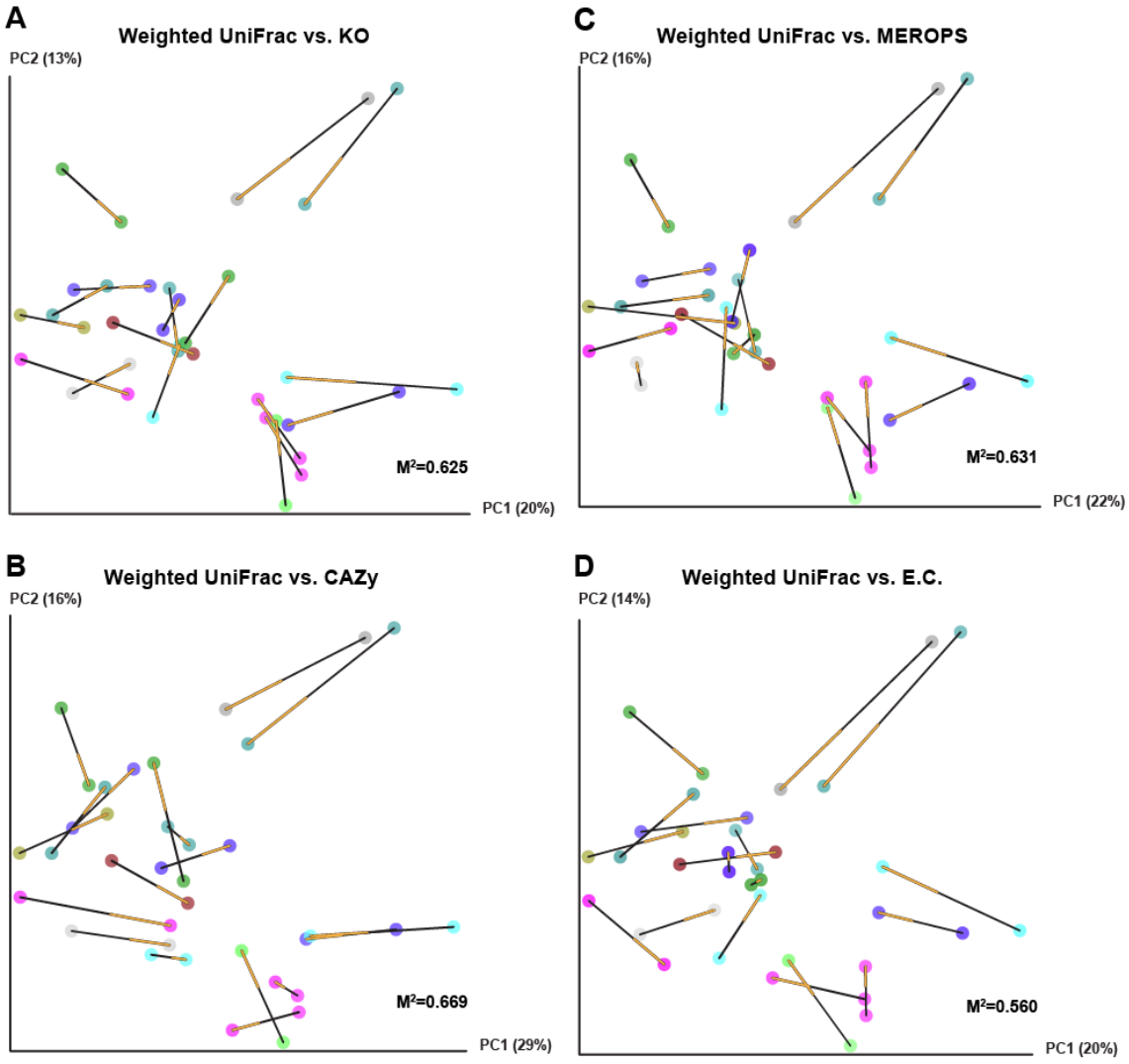

A
B
Lion2
BigHornW3
PolarBr2
HyraxSD
Lion1
Giraffe2
Horse1
SpecBr2
BushDog1
BlackLemur
Kroo3
SpgbkW
Capybara
Chimp2
AfElphSD3
Callimicos
Hyena2
BlackBr2
BlackRhino1
Rabbit
Orang1
BigHornSD
Urial2
BaboonW
Okapi1
ZebraSTL1
HyraxSTL
GorillaSTL
SakiWPig
BaboonSTL
Armadillo
Okapi2
Colobus
Echidna
RTLemur
Chimp1
C
PolarBr2
SpecBr2
BlackBr2
Echidna
Lion2
SpgbkW
BushDog1
Lion1
AfElphSD3
BlackLemur
ZebraSTL1

Figure S3.


A
Weighted UniFrac vs. KO
PC2 (13%)
M²=0.625
PC1 (20%)
C
Weighted UniFrac vs. MEROPS
PC2 (16%)
M²=0.631
PC1 (22%)
B
Weighted UniFrac vs. CAZy
PC2 (16%)
M²=0.669
PC1 (29%)
D
Weighted UniFrac vs. E.C.
PC2 (14%)
M²=0.560
PC1 (20%)

Supplemental Tables

Table S1. Metadata on 39 non-human mammals included in this study, including provenance, diet, gut physiology, and phylogenetic order.

Table S2. Mammal 16S rRNA sequencing statistics.

Table S3. Mammal fecal community DNA shotgun pyrosequencing datasets: statistics.

Table S4. Mammal fecal community DNA shotgun pyrosequencing datasets: phylogenetic assignments. (A) Summary of total hits against the NCBI non-redundant (nr) database. (B) Percentage of all reads assigned to the major phylogenetic divisions (normalized to total number of reads with hit in NCBI nr database).

Table S5. E.C.s encoded by genes whose representation is significantly different between herbivorous and carnivorous microbiomes. *= Poisson model coefficient. †=Akaike Information Criterion, smaller values denote better agreement between model and data. ‡= Benjamini-Hochberg corrected. §="NA" E.C. with no specific annotation.

Table S6. Summary of differences in amino acid metabolism between herbivore and carnivore microbiomes.

Table S7. Metadata on 18 calorie restricted humans included in this study, including host BMI and intake of major food categories.

Table S8. 16S rRNA sequencing statistics from calorie restricted humans.

Table S9. Shotgun pyrosequencing datasets obtained from fecal DNA prepared from calorie restricted humans: statistics.

Table S10. Shotgun pyrosequencing datasets obtained from fecal DNA prepared from calorie restricted humans: phylogenetic assignments. (A) Summary of total hits against the NCBI non-redundant (nr) database. (B) Percentage of all reads assigned to the major phylogenetic divisions (normalized to total number of reads with hit in NCBI nr database).

Table S11. Results of regression analysis comparing position on Principal Coordinate 1 with host dietary intake.

Chapter 3

Assembly of the functional gut microbiome in healthy and malnourished children

Chapter 3

Assembly of the functional gut microbiome in healthy and malnourished children

"Tell me what you eat, and I will tell you what you are."

Jean Anthelme Brillat-Savarin, *The Physiology of Taste*, 1825

Introduction and Background

The previous chapter showed that variations in diet between adult humans are associated with differences in the organismal and gene content of their gut microbial communities. Enormous dietary changes occur during infancy. It is well established that the gut microbiota changes over this same period; it is less clear how the functional properties of the bacterial community evolves during the period of exclusive breast feeding, during the period of supplemental feeding and as the child is fully weaned. A better understanding of assembly dynamics in the infant gut microbiome is required to develop a basic model of factors controlling microbiome operation. Understanding this period of assembly also promises to inform us about the origins of normal variations in supraorganismal (host plus microbial) metabolic and physiologic phenotypes within and between human populations, as well as the origins of pathologic states, including those associated with microbial community level disturbances (e.g., enteropathogen infection or antibiotic associated diarrhea).

In many parts of the world, children in early infancy are afflicted by malnutrition and diarrheal disease. One fifth of the world's population is undernourished, a broad term that includes states of stunting and wasting along with any deficiency of micro- and macronutrients (Black et al., 2008). The impact of undernutrition is particularly severe in children under the age of five; one third of the deaths in this group are caused at least in part by undernutrition (Horton, 2008). Nutritional status affects health directly through altered metabolism, and also by causing immunodeficiency and predisposition to infection (Katona and Katona-Apte, 2008; Kau et al., 2011). In turn, the infectious process often exacerbates the nutritional imbalance. For instance, children who are malnourished are more likely to suffer enteropathogen infection and diarrhea, and diarrhea in turn can increase the sequelae of malnutrition.

With this background, we were motivated to study the development of the gut microbial communities in healthy and malnourished infants. To our knowledge, there has been only one culture-independent characterization of infant gut communities focusing on malnutrition (Gupta et al., 2011). That study was limited because it included a single time point from one malnourished

child and an aged matched control. Although the recent literature is sparse, there is a significant body of work using culture methods to study patterns of gut microbial community assembly in the developing world, with some emphasis on alterations in malnutrition. The most notable series of studies was led by Leonardo Mata, working in rural Guatemala and Central America. He and his colleagues were also interested in the interaction of malnutrition and infection as a cause of morbidity in children (Mata, 1975), and thus studied the successional pattern in these "unsanitary conditions" to understand how the normal bacterial community in the gut might influence disease (Mata et al., 1969; Mata and Urrutia, 1971). Like infants in developed nations, *E. coli* was found in all children in the first few days of life, and Bifidobacteria reached dominance by the end of the first week and persisted for the duration of breast feeding. Bacteroides were variably found in babies who were exclusively breast fed, but when they were present, they were present in high levels. Bacteroides became dominant once solid foods became the sole dietary source. A separate group compared microbial community profiles from infants, children, and adults living in Nigeria and the United Kingdom (Tomkins et al., 1981). Babies from both countries had similar communities, but stronger geographical differences were apparent in weaned children and adults. They also demonstrated that the levels of Bacteroidetes increased in the Nigerian infants when the diet was supplemented with cereal.

Mata also looked at differences in the bacterial communities of children with malnutrition (Mata et al., 1972). He studied a cohort of 13 children presenting with acute malnutrition, ranging in age from 1 to 6 years, along with a control group of 4 children who were not malnourished, ranging in age from 10 to 22 months. In addition to the age difference between the groups, the cohort had the confounding variable that 10 of 13 cases presented with diarrhea. Nonetheless, the fecal microbiota of the two groups was found to be different, with lower abundance of Bifidobacteria and more frequent detection of atypical organisms like Proteus and Pseudomonas species in the malnourished children. The malnourished children also had a greater number of facultative anaerobes in their stools. Jejunal aspirates from the malnourished children showed higher bacterial loads in these children than in the controls (the jejunum is thought to harbor a very small

community under normal conditions). Heyworth and Brown also studied the jejunal microbiota in 25 malnourished children from Gambia (Heyworth and Brown, 1975). They too observed higher bacterial loads in children with chronic diarrhea and malnutrition. Furthermore, the jejunal communities of children suffering from chronic disease tended to dominated by only 1 or 2 culturable organisms, while controls had a relatively more diverse bacterial collection.

This chapter describes our metagenomic study of the functional development of the gut microbiome in children living in Dhaka, Bangladesh. Our work was done as part of a larger study of childhood malnutrition and infection sponsored by the Bill and Melinda Gates Foundation. The site in Dhaka was chosen because of the remarkable scientific and clinical infrastructure in the International Center for Diarrhoeal Disease Research, Bangladesh. Researchers there enrolled the cohort, performed regular surveillance and clinical phenotyping of the children, and did initial DNA extraction from the collected stools.

Monthly stool samples were collected from 103 children from birth through the first two years of life using a uniform protocol for prompt freezing at -20^{o}C followed by storage at -80^{o}C. For all samples, we generated and sequenced amplicons from the V2 region of bacterial 16S rRNA genes present in the fecal microbiota. For a subset of the children, we performed shotgun gene sequencing to characterize the functional features represented in the microbiome. Several questions motivated our analysis. Studies in western adults have demonstrated a high level of stability of a person's microbiota over time: at what point during the first two years of life, if any, does the gut community reach a stable state with a 'personal' community profile? Are there environmental factors, such as diet or malnutrition that cause reproducible changes in microbial communities across many children? Are there any signatures in the microbiota or microbiome that reflect host nutritional status? Conversely, are there states of the bacterial community that predict the future development of undernutrition? Is there a functional core of genes found in infant microbiomes, and if so, what are the predicted properties of this community? Are there any consistent functional differences in the microbiomes of children who are stunted versus those who are healthy?

Cohort Characteristics and Sequencing Results

Samples from 103 children were used for this study; 98 of these children remained in the cohort for two full years. Nutritional status for each child was classified using height-for-age or weight-for-age Z scores (HAZ and WAZ, respectively). For both HAZ and WAZ, a score greater than or equal to -2 is considered normal, while scores less than -2 denote "stunting" (HAZ) or "underweight" (WAZ) status. Children born with a low score at birth were classified as "Born Stunted" or "Born Underweight." Of the children born with a normal score, those who had a score $\leq$ -2 at 12 months of age were classified as "Stunted at 12 months" (HAZ) or "Underweight at 12 months" (WAZ). The remaining children were classified as "Healthy." Two children lost from the cohort before 12 months of age were excluded from the classification (Supplementary Table 1). In Bangladesh, the stunting classification has more significant clinical implications than their underweight classification, and thus most of our analysis related to malnutrition focuses on HAZ and stunting.

By 12 months of age, the average child not in the "Healthy" bin was malnourished by both measures and remained so for the duration of the study period. Even the "Healthy" children become significantly more stunted as they aged (paired U-test of HAZ at birth vs. HAZ at 24 months; P value < 0.001) (Figure 1). Healthy versus malnourished infants and children in our cohort did not differ significantly in terms of the number of days they had documented diarrhea, the number of diarrheal episodes, or the age at which food other than breast milk was first introduced (U-test; P value>0.05) (Table 1).

Children were enrolled in the study shortly after birth and the first fecal sample was collected at the time of enrollment (mean=6.8 ± 3.2 days old at first sample). A single fecal sample was collected every month thereafter for a total of 24 samples. Bacterial 16S rRNA datasets were generated by multiplex 454 FLX Titanium pyrosequencing from 1,892 stool samples (mean=18.4 ± 3.5 samples per child; 2,657 ± 856 high-quality 16S reads/sample; average processed read length 241 nt). A total of 242 shotgun pyrosequencing datasets were generated using the same fecal DNA from 14 children born at a normal height-for-age Z score; 8 of them had normal HAZ scores at 12

months of age and 6 had become stunted (17.3 ± 3.5 samples/child; 86,209 ± 40,360 high-quality reads/microbiome; total size of the dataset, 7.57 Gbp).

Results and Discussion

Bacterial taxonomic succession with aging

The average relative abundance of bacterial class-level phylotypes was calculated from samples at each month of life (Figure 2). Consistent with the literature, the class Actinobacteria (predominantly composed of bifidobacteria species) was most abundant during the earliest months of infancy and declined as the children aged. Actinobacteria were replaced almost entirely by rising levels of Bacteroidia. The total abundance of bacteria in the Firmicute phylum also increased over time; while the class Bacilli declined approximately 50% from birth to the second year of life, Clostridia expanded 10-fold over the same interval.

Although the aggregate pattern of succession is relatively smooth, the successional changes within a single child are much less predictable (Supplementary Figure 1). The composition of each individual's microbiota can fluctuate dramatically from monthly sample to monthly sample. To quantify this change, pairwise distances between all samples were calculated using the UniFrac metric, which compares each pair of samples in terms of their evolutionary distance (Lozupone and Knight, 2005). Considering only the pairwise comparisons of a child with him- or herself, we asked what fraction of times the nearest neighbor for an individual timepoint came from the samples immediately before or after that time point in the child's life. On average, only 28% (weighted UniFrac) to 43% (unweighted UniFrac) of the samples had an adjacent month time point as the nearest neighbor. Even more dramatically, when considering pairwise comparisons between all children, only 7% (weighted UniFrac) to 14% (unweighted UniFrac) of samples had a nearest neighbor from the same child (Supplementary Figure 2). This contrasts with a previous study of 14 USA infants that demonstrated a relatively high level of nearest neighbor matching within samples from a single child (Palmer et al., 2007). It is possible that the difference arises because

most samples in that study were collected over a period of several weeks, and almost all samples were collected in the first 3 months of the child's life. The lower level of similarity in our monthly sampling suggests that whatever intrapersonal stability exists within an infant may be limited to periods of days or weeks.

Bacterial community changes with aging and diet switches

We next asked if there were any events in the life of the children that explained the changes in the bacterial community over time. Given the high level of variation within and between individuals, we chose to analyze each child separately. We reasoned that a combined analysis would be complicated by the differences between children, while analyzing each child separately allowed us to ask if there were common responses to external factors across many independent replicates. Both diet and age of the child had profound effects on community composition. As a case study, we highlight the data from Child 7114, one of the children classified as "Healthy" and selected for shotgun gene sequencing (Figure 3). The position of each sample on Principal Coordinate 1, the axis of maximal variation between samples, was significantly correlated with the child's age (Figure 3a and Figure 3c).

The principal coordinate values were also significantly different between diet categories (Figure 3b and Figure 3d). We tested the impact of a single host characteristic, presence of family food in the diet, as a discriminator of microbial community differences in each child separately. In 85% of the children, the average unweighted UniFrac distance between samples with the same exposure to family food was significantly lower than the average distance between samples with different exposures (Monte Carlo permutation test with 1000 iterations, p-value ≤ 0.05).

Of course the progression of an infant's diet is highly conflated with aging, so it is difficult to tease apart the relative contributions of both factors. If the introduction of family food to the diet is causing a reproducible shift in fecal microbial community configuration, we reasoned that children who had family food introduced at early ages should have an altered community configuration compared to children who experienced a later introduction. We selected the ten children with

the earliest introduction of family food and compared them to the ten children with the latest age of introduction. For both groups of children, the major variation along PC1 was associated with host age rather than diet (Figure 4a). Children who had family food introduced at a very early age followed a nearly identical trajectory along PC1 compared to children who did not receive family food. Similarly, we wondered if the withdrawal of breast milk at weaning was responsible for major changes in community structure. In Bangladesh, children are often breast fed for several years; only 11 of the children in our cohort were totally weaned within the study period (Supplementary Table 1). Of those children who were weaned, we selected the 5 children weaned at the earliest age and plotted change in Principal Coordinate 1 compared to host age (Figure 4b). The withdrawal of breast milk from these infants was not reliably associated with any changes in community structure along PC1.

Taken together, these results demonstrate that the developing infant gut microbiota is highly dynamic, and in general highly associated with age. Diet could be the cause of these changes; however, comparing children with different diet patterns did not reveal any diet-associated changes that were reproducible across the members of the cohort. The major limitation to our analysis of diet is that we treat the food groups as monotonous bins (e.g, "breast milk"). We have no data on the quantity or quality (composition) of these foods and are thus unable to make any conclusions about dose-dependent affects. Additionally, the composition of milk changes over the course of lactation (e.g., alterations in Human Milk Oligosaccharides), as do immunoactive compounds. These changes may play an important role in shaping the microbial community of the gut (Sela and Mills, 2010).

Impact of HAZ on community differences

To test for an association of malnutrition status with microbial community structure, we generated a series of age-matched comparisons. Anthropometric measurements were recorded every three months by the field workers; for each quarterly time point, we collected bacterial 16S rRNA data from all stool samples provided within 30 days of the child's anthropometric measurement.

We categorized the samples in two ways for comparison. The first method was the 'nutrition bins' described earlier, where each child was assigned a status ("Healthy," "Born," or "Became" stunted or underweight at 12 months of life). The second method was to transform all of the Z scores to quintiles based on the distribution of values recorded over the first two years in all children. A Z score in the lowest 20% of all measurements was scored as "1", while a Z score in the highest 20% was scored as "5". The first method allowed us to ask if children with different nutritional fates had any age-matched differences in their fecal microbiota. The second method allowed us to ask if the nutritional status at each timepoint was associated with microbial community differences, regardless of past or future occurrences.

Our test for association was Monte Carlo permutation analysis. We calculated the average pairwise unweighted UniFrac distance of samples within an age-matched nutrition category compared to the average pairwise distance between samples in that category and samples outside of the category (e.g., the average pairwise distance between samples assigned to HAZ quintile 1, compared to the average pairwise distance between quintile 1 and quintile 5 samples). The statistical significance of the difference was determined by comparing the observed difference to the distribution of differences measured with 1000 random permutations of category assignments. The results of this analysis require further validation, but there is some evidence of statistical differences between children who are healthy at 12 months of life and children who are stunted. At several different time points, the average distance within the "Healthy" category is significantly different than the distance between "Healthy" and "Born Stunted" or "Stunted at 12 months" (p-value ≤ 0.05). However, the magnitude of the average distance differences is small, and there is no obvious separation of the samples by category in Principal Coordinate space over the first 10 dimensions. Furthermore, the differences are not consistent over time. Healthy children are significantly different than children born stunted at the birth, 6 month, and 12 month recordings, but not at 3 months or 9 months. The significance of these findings needs to be tested in at least two additional ways. First, we will randomly assign samples to one of the three groups, and repeat the Monte Carlo analysis. The distribution of p-values obtained from these analyses will help us to

establish the False Discovery Rate for the method. Second, we will use machine learning methods (Knights et al., 2011) to look for bacterial species in the dataset that separate the age-matched nutritional categories. The validity of the modeling can be tested by using some of the samples for training and withholding other samples for testing.

Change in community functional profile with age

We used the metadata from the cohort to select 14 children for shotgun gene sequencing. We first selected 8 children who were classified as "Healthy" by HAZ scores at 12 months of life, with preference given to children with diarrheal incidence and history of diet switches that were fairly representative of population averages. The goal was to create a dataset of fairly similar children as a reference for "normal" development. We then selected 6 children who were born healthy but who were classified as "Stunted" at the 12th month of life. Where possible, these children were selected to match our healthy subset in terms of diarrheal burden and age at diet switches so that the major distinguishing characteristic between the children was nutritional status (Supplementary Table 1).

Shotgun gene sequencing datasets were generated for all 14 children using fecal samples from birth through the 24th month of life. Only 5 of the samples provided at the time of enrollment yielded enough DNA for shotgun library preparation; for the remaining 23 months, we generated an average of 10.3 ± 1.6 shotgun libraries per time point. Datasets were annotated using the KEGG database. Principal coordinate analysis shows that the major axis of community variation is correlated with the children's age and diet (Figure 5a and Figure 5b). It is interesting to note that the samples from the first 14 days of life ("Life Phase 1" in the Cooperstock and Zedd model) span the entire range of PC1 (Figure 5c) (Cooperstock and Zedd, 1983). By the second month of life, samples are much more tightly clustered by Life Phase. This argues that although the infant gut can initially be colonized by a range of organisms with widely varied functional capabilities, there is rapid equilibration and selection for organisms with functional similarity.

Functional maturation of the gut microbiome during postnatal development

To understand the metabolic functions that change in the developing gut microbiome, we aggregated the counts of KO's into KEGG metabolic pathways. Using data from the 8 healthy children, we correlated the abundance of each pathway in the samples with the age of the host using the Spearman non-parametric correlation coefficient. Of the 129 pathways included in the analysis, 83 were significantly correlated with age (Supplementary Table 2). We visualized the pathways and their correlation with age using iPath (Yamada et al., 2011). Consistent with a previous time series study of a single healthy USA infant (Koenig et al., 2011), younger microbiomes were functionally enriched in genes encoding proteins involved in fatty acid metabolism and ABC transporters. Older microbiomes were enriched in a variety of metabolic pathways, notably for amino acid and carbohydrate utilization. One of the enriched 'superfamilies' of pathways was for glycan metabolism (Figure 6). The gut epithelium is covered by a glycan-rich mucus layer, and it is known that many bacteria in the microbiota can degrade these mucins (e.g., members of Bacteroides). While no glycan biosynthetic pathways were enriched in the microbiome, a variety of degradative pathways were significantly and positively correlated with host age.

The Spearman analysis is particularly well suited to detect components of the data that increase monotonically with age. Alternatively, functions in the microbiome may follow a discrete progression, where they are found with a certain level of abundance in early life and then increase or decrease after an environmental trigger. We looked for genes encoding enzymes that were significantly different in early life (0-3 months) compared to later life (6-9) months using a Poisson model (Kristiansson et al., 2009). We analyzed the healthy and stunted infants separately. In both groups, the most significantly different enzyme function was sialate-o-acetylesterase (Enzyme Commission number 3.1.1.53). This function is not detected in any microbiome before the 130th day of life, but is present in most microbiomes by the first year of life and in all children by 2 years (Figure 7). It is likely that this enzyme is used by bacteria to metabolize sialic acid decorations on secreted host mucins. The expression of glycans, and the carbohydrate modifications of these glycans, changes over the course of infant development (Biol et al., 1992). It will be fasci-

nating to relate changes in an individual child's gut mucosal glycan profile to the capacity of their gut microbiome to degrade those glycans. This also raises the intriguing possibility that structural variation in the host mucosa is one of the non-diet based factors driving the changes seen in the microbiota as infants age.

Conclusions and Future Directions

Our study of microbiota succession in 103 human infants demonstrates some common patterns of community assembly in these children. Dietary change is associated with changes in a single individual's community; however, when children are compared to each other, the age of the infant is a better predictor of community similarity than shared diet. The relative importance of age and the limited ability of diet switches to differentiate communities from different children suggests that stochastic processes are likely important during infant assembly, and that non-nutritive environmental or host events also act to shape the community. At the same time that the assemblage of species in the microbiota is evolving, the functional capacity of that community is maturing. Numerous pathways, including those needed for amino acid, carbohydrate, and glycan metabolism, are positively associated with the age of the child. We are in the process of identifying specific enzymatic functions that change in these communities over time, and relating those functions to the occurrence of life events such as the introduction of family food.

Several areas of this project require additional work. First, preliminary analysis suggests that there are differences between the bacterial phylogenetic composition of malnourished and healthy children. These results need to be confirmed by comparison to null models, and the responsible OTUs identified with machine learning methods. Second, we are continuing to classify the patterns of assembly for bacterial functions in these infants. Our focus will be on identifying functions that change in the same way in all or most members of the cohort, as these most likely represent the functional core requirements of the gut community. We are also working to develop appropriate paired statistical models that will allow us to look at the changes from one timepoint to the next; these methods will be used to classify the functional and taxonomic changes associated

with discrete life events. Lastly, the work will be enhanced by establishing quantitative models that describe bacterial species and gene abundance as a function of host age and diet. It is clear that diet is associated with the change of the community within a single individual, but it is not clear if similar diet transitions have the same affect in different host. If diet is truly associated with reproducible changes across diverse infants, then discriminant analysis should be able to accurately classify microbiota from hosts with the same diet. If diet is not associated with reproducible changes, it will strongly argue that there are additional factors changing in infancy that may be more proximal in driving community evolution.

Figure Legends

Figure 1. Nutritional status of cohort infants during the first two years of life. Height-for-age (HAZ) and weight-for-age (WAZ) Z scores were measured every three months, beginning at birth. Children with normal Z scores (≥-2) at birth and also at 12 months of life were classified as "Healthy." Children with normal scores at birth but low scores (<-2) at 12 months of life were classified as "Stunted at 12 months" (HAZ) or "Underweight at 12 months" (WAZ). Children with a low score at birth were classified as "Born Stunted" or "Born Underweight." Error bars represent 95% confidence intervals. (a) Mean HAZ scores over time for each stunting category. (b) Mean WAZ scores over time for each underweight category.

Figure 2. Change in relative abundance of bacterial classes over first two years of life. Operational Taxonomic Units (OTUs) were determined using a 97% nucleotide sequence similarity threshold for the V2 region of the 16S rRNA gene, and taxonomy was assigned using the RDP database. The OTU taxonomy was cut at the class level, and the abundance of each bacterial class was determined in every sample. This plot charts the average abundance of the major bacterial classes from all of the samples collected in the same month of life. The bacterial phyla is noted in parentheses after the class name: A=Actinobacteria, B=Bacteroides, F=Firmicutes, P=Proteobacteria, T=Tenericutes, TM=TM7.

Figure 3. Community changes in child 7114 are associated with age and diet. Principal coordinates analysis (PCoA) was performed using the unweighted UniFrac distances between all samples from child 7114. (a) Principal coordinates plot, colored by age. (b) Principal coordinates plot, colored by diet category. Diet categories: "BM" = breast milk only, "BM/CM" = breast milk and cow's milk, "BM/CM/RP"= breast milk and cow's milk and rice powder, "BM/FF" = breast milk and family food, "FF' = family food only. (c) Linear regression of child age against microbiota position on principal coordinate 1. (d) Spread of data on principal coordinate 1, grouped by diet.

Figure 4. Impacts of diet and age in children with different exposure to family food. (a) Data from the 10 children with the earliest first exposure to family food (mean age 167 days) were plot-

ted together with data from the 10 children with the latest first exposure to family food (mean age 375 days). Samples are colored according to diet category as in Figure 3. (b) The 5 children with the earliest age of complete weaning (mean age 421 days) are plotted together. Each sample is colored by "Life Phase": Phase 1= first 14 days of life, Phase 2= exclusive breast feeding, Phase 3= breast feeding with food supplement, Phase 4 = completely weaned.

Figure 5. Changes in fecal microbiome functional profiles are strongly associated with host age and diet. The Bray-Curtis metric was used to calculate the pairwise distances between all shotgun datasets. (a) Principal coordinates plot of all microbiomes. Samples are colored by host age. (b) Linear regression of host age versus position of the microbiome along principal coordinate 1. (c) Variation of samples along PC1, grouped by Life Phase.

Figure 6. Glycan degradation pathways increase in aging microbiomes. Shotgun reads from the healthy children were annotated with the KEGG database and assigned to pathways. The abundance of each pathway in each sample was correlated with the age of the host at the time of sampling using the non-parametric Spearman method. Those pathways that were statistically significantly correlated with age were visualized in iPath; green pathways are negatively correlated with age, while red pathways are positively correlated.

Figure 7. Presence of sialate-o-acetylesterase in infant microbiomes. Microbiomes were grouped by host age into bins for every 75 days of life. The average relative abundance of EC.3.1.1.53 in the samples in the bin was calculated (red bars, left axis). The prevalence of the E.C. in samples from the 8 healthy children and 6 stunted children is also plotted for each bin (green and blue lines, right axis).

Figures

Figure 1.

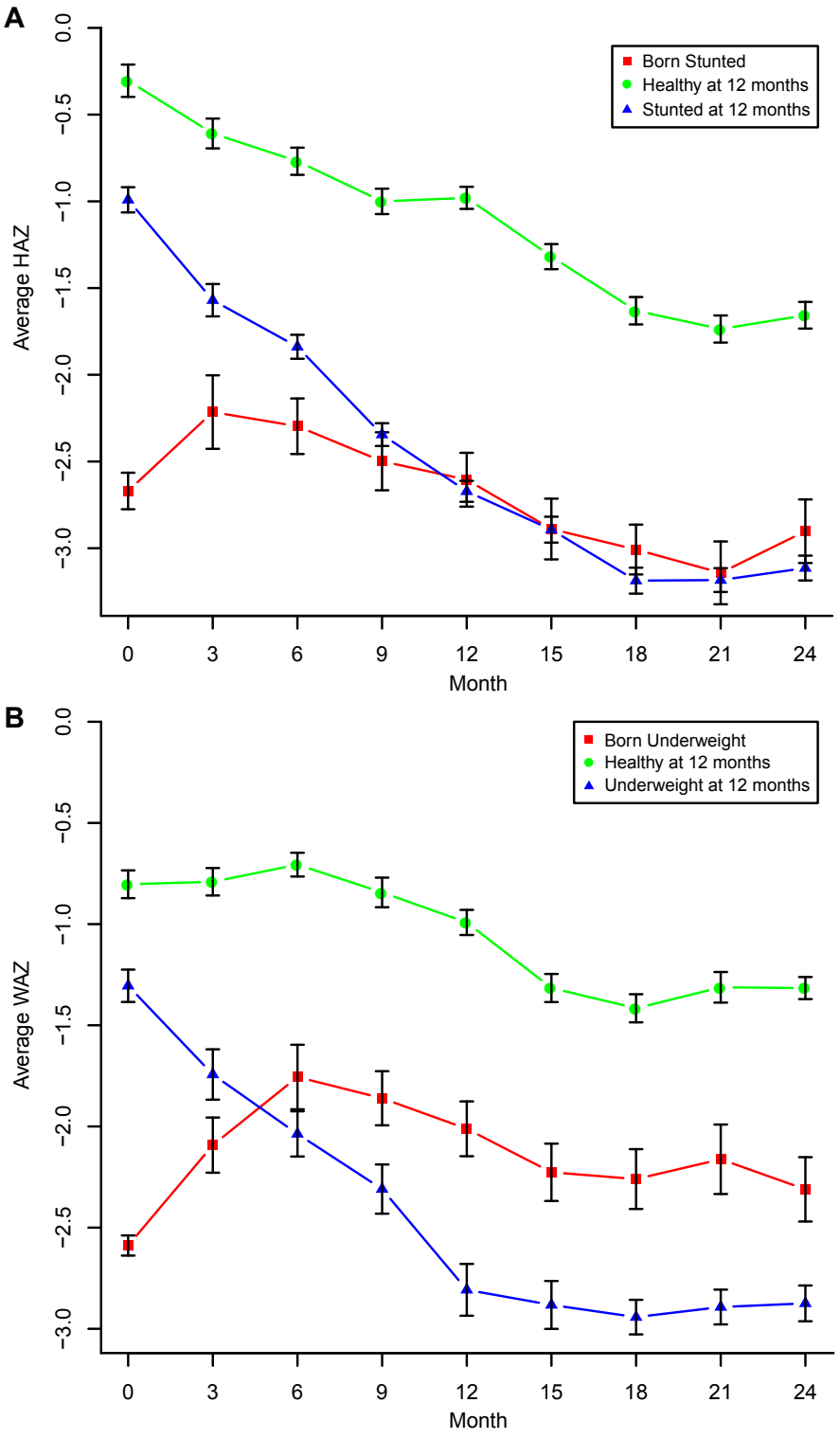


Figure 2.

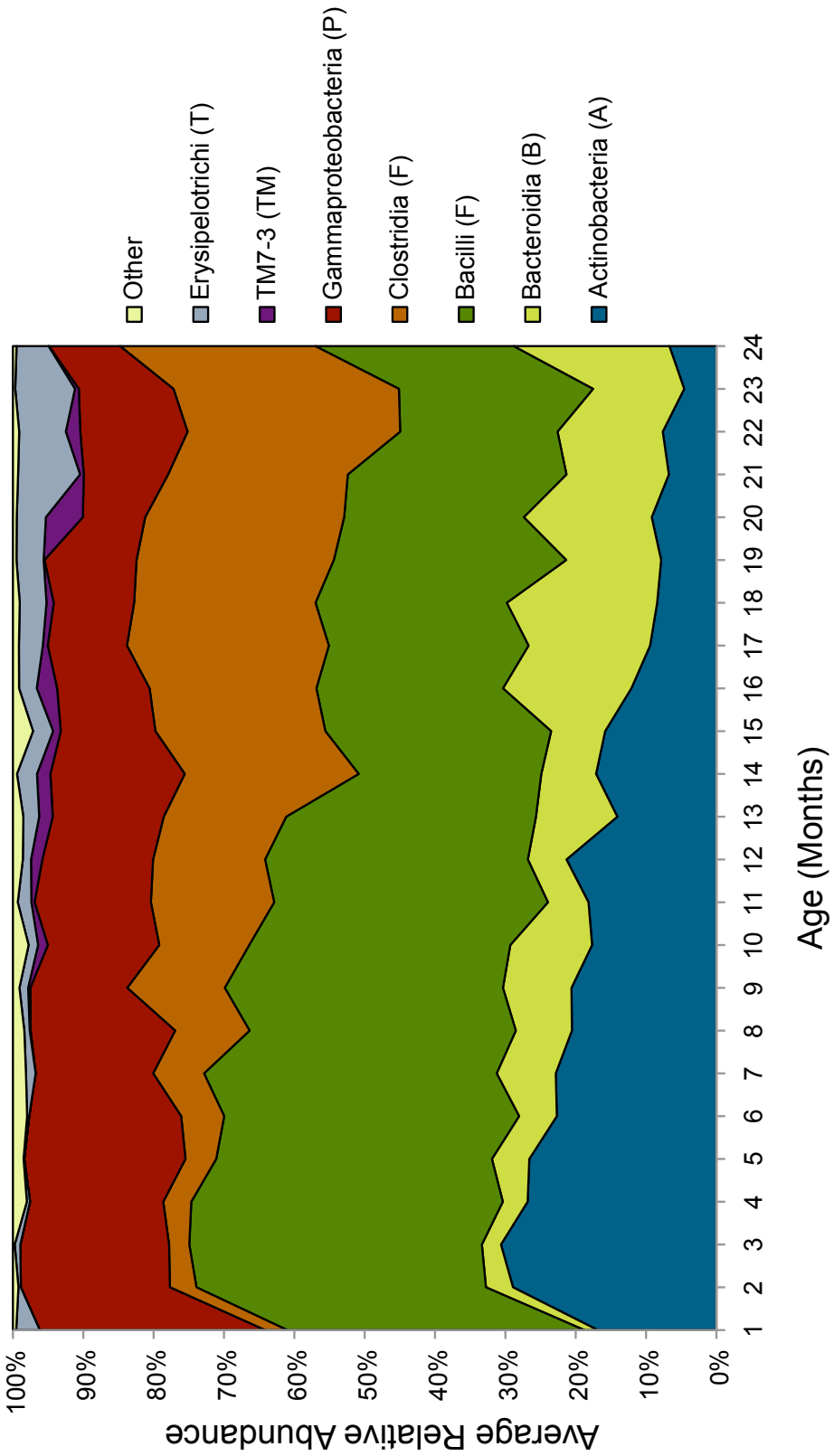

Other
Erysipelotrichi (T)
TM7-3 (TM)
Gammaproteobacteria (P)
Clostridia (F)
Bacilli (F)
Bacteroidia (B)
Actinobacteria (A)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Average Relative Abundance
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Age (Months)

Figure 3.

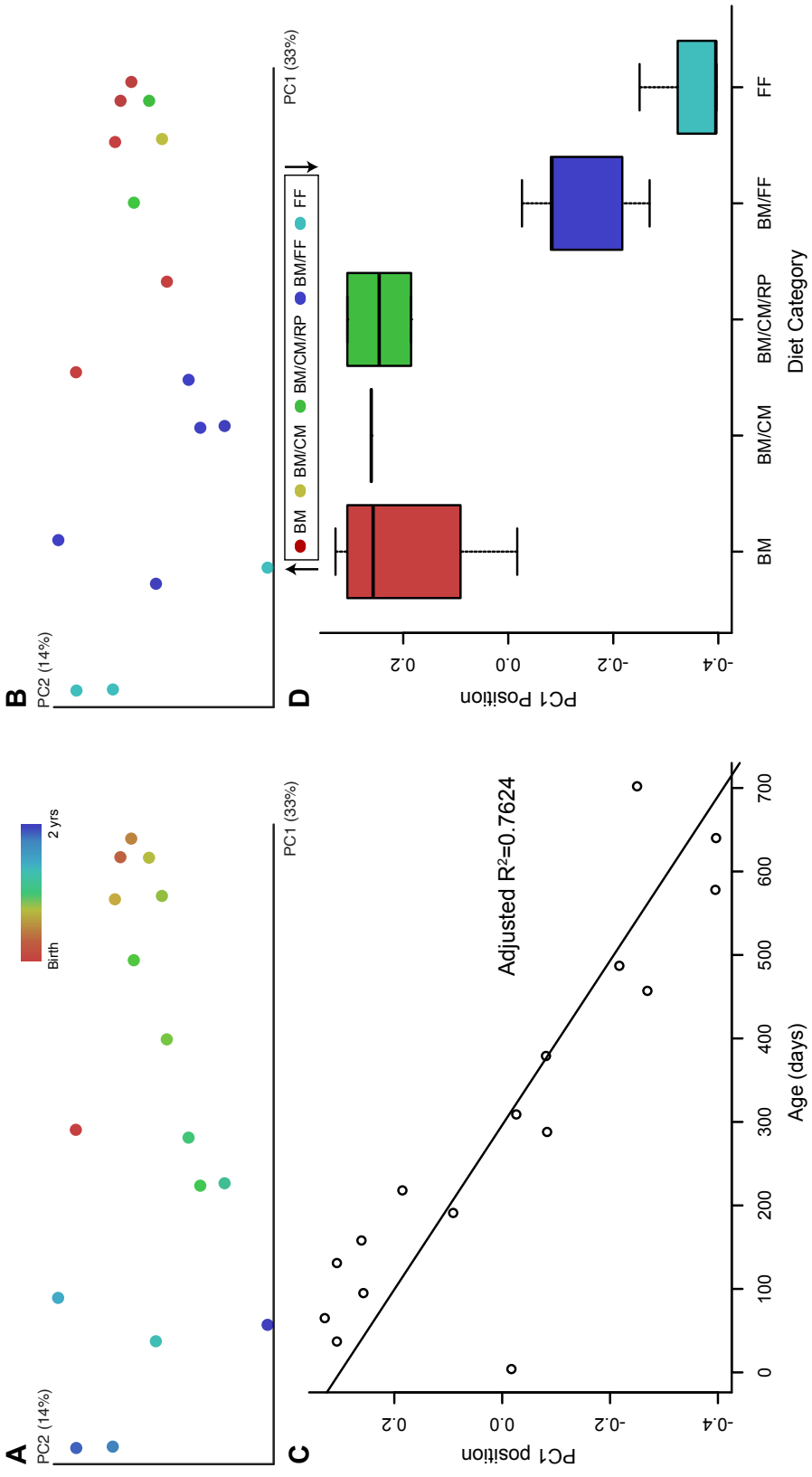

A
PC2 (14%)
PC1 (33%)
Birth
2 yrs
B
PC2 (14%)
PC1 (33%)
BM
BM/CM
BM/CM/RP
BM/FF
FF
C
PC1 position
Age (days)
Adjusted R²=0.7624
D
PC1 Position
Diet Category
BM
BM/CM
BM/CM/RP
BM/FF
FF

Figure 4.

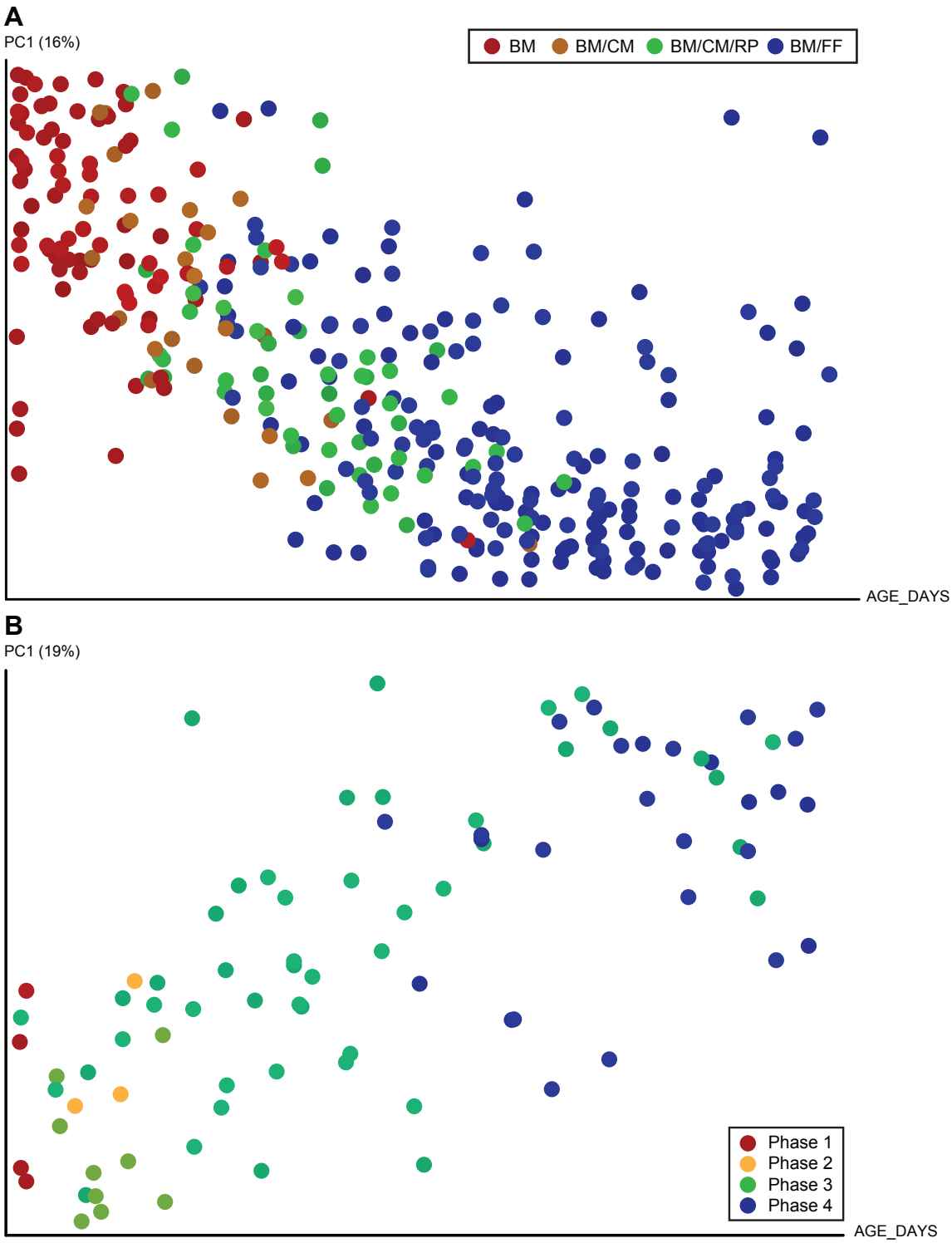

A
PC1 (16%)
BM
BM/CM
BM/CM/RP
BM/FF
AGE_DAYS
B
PC1 (19%)
Phase 1
Phase 2
Phase 3
Phase 4
AGE_DAYS

Figure 5.

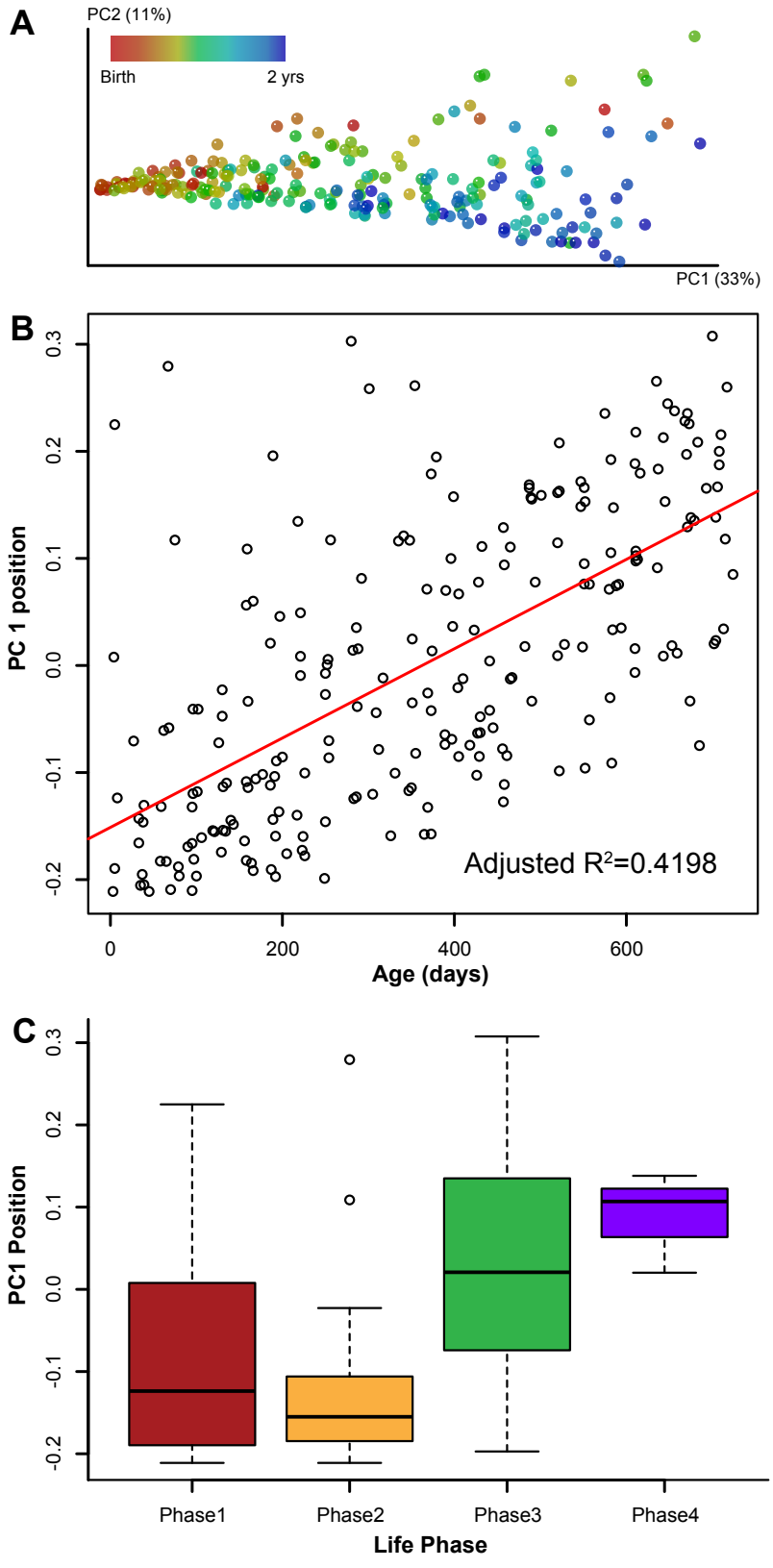

A
PC2 (11%)
Birth
2 yrs
PC1 (33%)
B
PC 1 position
0.3
0.2
0.1
0.0
-0.1
-0.2
Adjusted R²=0.4198
0
200
400
600
Age (days)
C
PC1 Position
0.3
0.2
0.1
0.0
-0.1
-0.2
Phase1
Phase2
Phase3
Phase4
Life Phase

Figure 6.

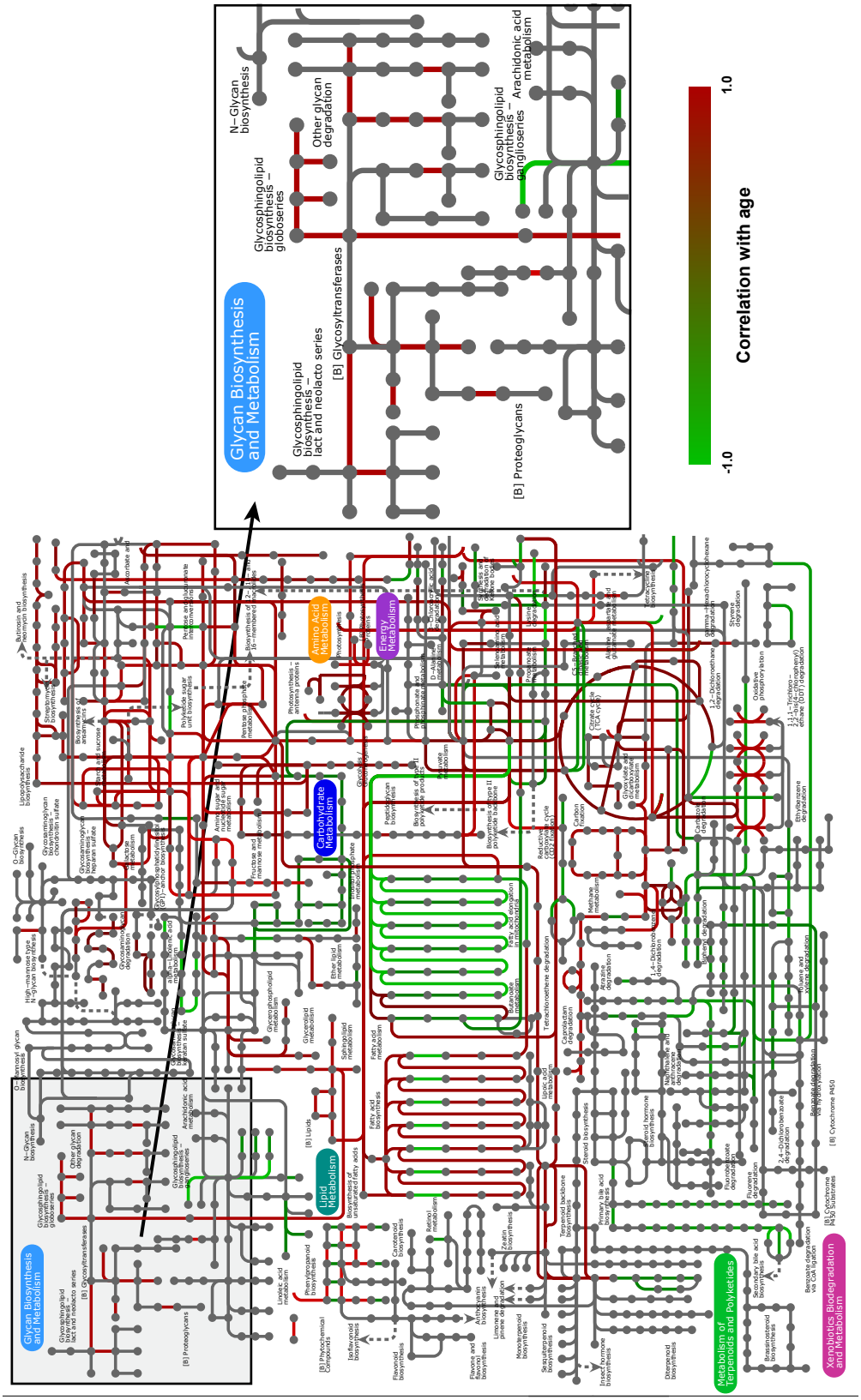
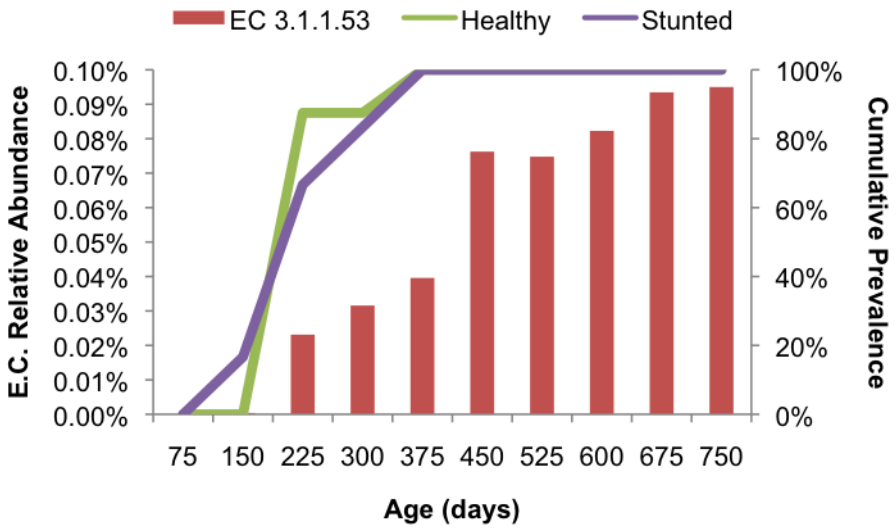

Glycan Biosynthesis and Metabolism
Glycosphingolipid biosynthesis – lacto and neolacto series
[B] Glycosyltransferases
Glycosphingolipid biosynthesis – globoseries
N-Glycan biosynthesis
Other glycan degradation
Glycosphingolipid biosynthesis – ganglioseries
Arachidonic acid metabolism
[B] Proteoglycans
Correlation with age
1.0
-1.0
Amino Acid Metabolism
Energy Metabolism
Carbohydrate Metabolism
Lipid Metabolism
Metabolism of Terpenoids and Polyketides
Xenobiotics Biodegradation and Metabolism

Figure 7.


EC 3.1.1.53
Healthy
Stunted
E.C. Relative Abundance
0.10%
0.09%
0.08%
0.07%
0.06%
0.05%
0.04%
0.03%
0.02%
0.01%
0.00%
Cumulative Prevalence
100%
80%
60%
40%
20%
0%
75
150
225
300
375
450
525
600
675
750
Age (days)

Table

Table 1. Characteristics of cohort children by nutritional category. Nutritional categories were defined using HAZ and WAZ scores as described in the text. Diarrhea was defined as ≥3 loose stools in a 24 hour period. Diarrheal events that occurred at least three days after the most recent diarrhea were considered to be new episodes.

A) Metadata by HAZ status

	Healthy at 12 months	Stunted at 12 months	Born Stunted
# (%Male)	50 (52)	35 (60)	16 (56.3)
Mean #Diarrheal Days (SD)	28.6 (14)	35.1 (21.6)	33.7 (24.7)
Mean # Diarrheal Episodes (SD)	7.8 (3.3)	8.9 (3.7)	8.4 (4.7)
Mean HAZ at Birth (SD)	-0.3 (1)	-1 (0.6)	-2.7 (0.6)
Mean HAZ 12 Months (SD)	-1 (0.7)	-2.7 (0.5)	-2.6 (0.9)
Mean HAZ 24 Months (SD)	-1.7 (0.8)	-3.1 (0.6)	-2.9 (1.1)
Mean Day of first Breast Milk Supplement (SD)	126.3 (71.7)	138.4 (87.9)	122.1 (80.9)

B) Metadata by WAZ status

	Healthy at 12 months	Underweight at 12 months	Born Underweight
# (%Male)	51 (51)	19 (63.2)	31 (58.1)
Mean #Diarrheal Days (SD)	31.2 (17.6)	33.8 (17.6)	31 (21.9)
Mean # Diarrheal Episodes (SD)	8.2 (3.4)	9.1 (3.4)	8 (4.3)
Mean WAZ at Birth (SD)	-0.8 (0.7)	-1.3 (0.5)	-2.6 (0.4)
Mean WAZ 12 Months (SD)	-1 (0.7)	-2.8 (0.8)	-2 (1.1)
Mean WAZ 24 Months (SD)	-1.3 (0.6)	-2.9 (0.6)	-2.3 (1.3)
Mean Day of first Breast Milk Supplement (SD)	133.8 (82.2)	131.1 (79.7)	122.5 (73.4)

Supplementary Information

Material and Methods

Sample recruitment

Full details of the human subject recruitment have been described previously (Mondal *et al*, in submission). All work was conducted with approval and oversight by the Institutional Review Board at the University of Virginia and the Ethical Review Committee of the International Center for Diarrhoeal Disease Research, Bangladesh (ICCDR-B). Briefly, trained Field Research Assistants (FRA) working in Mirpur, an urban slum in Dhaka, Bangladesh, identified pregnant mothers in 2007. Within 72 hours of the baby's birth, a medical officer affiliated with the ICCDR-B evaluated the child and enrolled the baby in the study after obtaining appropriate written consent. The first stool sample was collected at this time. For the next two years, stool samples were collected every month in the child's home by the FRAs. Additionally, FRAs visited each study household every two days to assess the child's health with particular focus on diarrheal disease and feeding history. Height and weight were recorded every quarter and converted to Z scores using WHO standards.

Isolation of fecal DNA

Fecal samples were frozen at -20^{o}C within 4h after they were produced by the child, and then maintained at -80^{o}C prior to processing. Microbial community DNA was prepared from the fecal samples in two steps. First, crude DNA was isolated in Dhaka. Each fecal sample was pulverized at -80^{o}C using a mortar and pestle. An aliquot (typically 500 mg) of frozen pulverized feces was combined with 500µl of 0.1mm zirconium beads (BioSpec Products), 500µl ice cold Buffer A (200mM NaCl, 200mM Tris, 20mM EDTA), 210µl room temperature SDS (20% v/v, filter-sterilized), and 500µl phenol:chloroform:isoamyl alcohol (25:24:1, Applied Biosystems). Samples were disrupted using a bead beater (BioSpec; 2 min on high setting at room temperature). After centrifugation (8000 rpm, 3 min, 4^{o}C), the aqueous phase was removed and mixed by inversion with 500 µl of the phenol:chloroform:isoamyl alcohol solution. The samples were centrifuged again (13000 rpm,

3 min, 4 °C), and then the aqueous layer was extracted and mixed with 600 mL of chilled isopropanol (-20°C). DNA was left to precipitate for 20 min on ice and then pelleted (13000 rpm, 20 min, 4°C). The pellet was washed once with 500 µl cold 100% ethanol. DNA was resuspended in 100 mL H20. At this point, crude DNA was shipped to St. Louis on dry ice; the courier service refilled the dry ice several times during transit to maintain the samples. DNA was cleaned in St. Louis using QIAquick 96 PCR purification plates (Qiagen). Buffer PM (Qiagen) was mixed with RNase A (Qiagen) to a final concentration of 1.33 mg/ml. Three volumes of this mixture were added to a volume of crude DNA and incubated at room temperature for 2 min. Following RNase digestion, samples were applied to the filter plate and processed according to the manufacturer's instructions using a QIAvac 96 manifold. DNA was eluted in Buffer EB (Qiagen).

Multiplex pyrosequencing

An aliquot of the cleaned DNA was used for PCR amplification and sequencing of bacterial 16S rRNA genes. ~330bp amplicons, spanning variable region 2 (V2) of the gene were generated by using (i) modified primer 8F (5′- CCTATCCCCTGTGTGCCTTGGCAGTCTCAG *AGAGTTTGATCCTGGCTCAG*-3') which consists of 454 Titanium Shotgun Primer A (underlined) and the universal bacterial primer 8F (italics) and (ii) modified primer 338R (5' CCATCTCATCCCTGCGTGTCTCCGACTCAGNNNNNNNN*TGCTGCCTCCCGTAGGAGT* 3') which contains 454 Titanium Shotgun Primer B (underlined), a sample specific, error correcting 8-mer Hamming barcode (N's)(Hamady et al., 2008), and the bacterial primer 338R (italics). Three replicate polymerase chain reactions were performed for each fecal DNA sample; each 20-mL reaction contained 25 ng of cleaned DNA, 8 ml 2.5X HotMaster PCR Mix (Eppendorf) and 0.2 µM of each primer. PCR conditions consisted of an initial denaturation step performed at 95 °C for 2 min, followed by 30 cycles of denaturation (95°C, 20 sec), annealing (52°C, 20 sec) and amplification (65°C, 1 min). Amplicons generated from each set of three reactions were pooled and quantified using the Quant-iT double stranded DNA High Sensitivity assay (Invitrogen). Barcoded samples were pooled in equimolar amounts, and the pool was cleaned with Ampure magnetic purification

beads (Agencourt).

For multiplex shotgun 454 FLX pyrosequencing, each fecal community DNA sample was randomly fragmented by nebulization to 500-800 bp (FLX Titanium), and then labeled with a distinct MID (Multiplex IDentifier; Roche) using the manufacturer's protocol (Rapid Library preparation for FLX Titanium). Equivalent amounts of up to 12 MID-labeled samples were pooled for each pyrosequencer run (454 Titanium chemistry).

Sequencing libraries were prepared from the 16S rRNA or shotgun pools using the manufacturer's Shotgun Library protocol. The only exception was reducing the amplification primer volume 75% in the 16S rRNA emulsion to reduce the occurrence of short reads.

For both amplicon and shotgun gene sequencing, PTP images were processed using the default shotgun analysis pipeline.

16S rRNA data processing and analysis

The 16S rRNA sequence data was processed using Qiime, v1.3.0 (Caporaso et al., 2010). The data from each 454 run was demultiplexed using the Hamming barcodes, and reads were removed if the sequence length was less than 150 nucleotides or if the average quality of the read was less than 25 in sliding windows of 50 bases. The reads were processed with the program "otupipe," which performs error correction and chimera removal of 454 amplicon reads, along with clustering into Operational Taxonomic Units (OTU) (http://www.drive5.com/usearch/). Bacterial taxonomy was assigned to each OTU using the Ribosomal Database Project classifier.

Shotgun sequence data processing and functional annotation

Shotgun metagenomic data was processed with publicly available software and custom Perl scripts. First, the sequences associated with each barcoded sample were extracted from the multiplexed run using the 454 utilities sffinfo and sfffile. No mismatches were allowed in the Rapid Library barcode sequences. The data were filtered to remove (i) reads less than 60 bases in length and (ii)

reads with two contiguous and/or three total degenerate bases (N's). Emulsion PCR is known to introduce duplication artifacts into pyrosequencing data; therefore, duplicate reads with exact sequence match over the initial 20 nucleotides and an overall identity of ≥97% throughout the length of the shortest read were identified and removed (Gomez-Alvarez et al., 2009). To insure the anonymity of samples, shotgun gene datasets were scrubbed of all reads with homology to the human genome using BLASTN (e-value ≤ 1e-5, bitscore ≥ 50, percent identity ≥ 75%).

Searches against the KEGG (version 58) and MEROPS (version 9.5) databases were carried out using BLASTX with the bacterial translation table (minimum e-value ≤ 1e-5, bitscore ≥ 50, and percent identity ≥ 50%). For KEGG, the BLAST protein database was constructed using only the sequences that had a KO number assigned to the read. The effective database size of the blast search was adjusted with option '–z 2214788408' to reflect the size of the entire database.

For each database search, an abundance tally was created with all database entries initially assigned a count of 0. Each time that an entry from the database was found to be the single best blast hit of a shotgun read, that entry's count in the abundance tally was incremented by 1. If a read had n equivalent best BLAST hits, the count for each of the entries was incremented by 1/n in the tally. For KEGG, the database entry was annotated with one or more KO numbers, and each KO number was sometimes further associated with KEGG pathways or E.C. numbers. The count increment for the database entry was added to the abundance tally for each of these levels of annotation.

Taxonomic composition of the shotgun sequence data

Taxonomy was inferred with a database of 127 sequenced genomes from bacterial and archaeal isolates recovered from human intestine (Supplementary Table 3 for details). The search against the database was performed with BLASTN using previously published BLAST thresholds (e-value ≤ 1e-20, bitscore ≥ 50, percent identity ≥ 50, percent alignment ≥ 80%) (Arumugam et al., 2011). The effective abundance of each genome in the determined by normalizing the number of reads assigned to the genome by the genome length of that organism.

Other statistical analysis

Spearman analysis and hierarchical clustering were performed using R (R Development Core Team, 2011), and ecologically relevant data transformations and distance calculations were used as implemented in the vegan package (Oksanen et al., 2011). For analysis at the level of KEGG pathways, all samples with at least 30,000 shotgun reads were used. The pathway counts for each sample were evenly sampled without replacement. Pathways that had no variance across the samples were removed. Other pathways were retained if and only if they met one of the following criteria: (i) the pathway was found with at least 0.1% relative abundance in at least one half of the samples, or (ii) the pathway was found with a relative abundance of at least 0.4% in at least one sample. These filtering criteria left 129 pathways for analysis. The samples corresponding to children binned as "Healthy" or "Stunted at 12 months" were processed separately. For each group, the pathway counts were correlated with the age of the child at the time of sample using the Spearman non-parametric method. The Spearman correlation coefficient, rho, was converted to a red-green color gradient for each pathway, where full red means perfect positive correlation of the pathway with increasing age, and full green means perfect negative correlation with increasing age. The pathways were visualized using iPath (Yamada et al., 2011).

References

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011). Enterotypes of the human gut microbiome. Nature *473*, 174–180.

Biol, M. C., Martin, A., and Louisot, P. (1992). Nutritional and developmental regulation of glycosylation processes in digestive organs. Biochimie *74*, 13–24.

Black, R. E., Allen, L. H., Bhutta, Z. A., Caulfield, L. E., de Onis, M., Ezzati, M., Mathers, C., and Rivera, J. (2008). Maternal and child undernutrition: global and regional exposures and health consequences. The Lancet *371*, 243–260.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat Meth *7*, 335–336.

Cooperstock, M. S., and Zedd, A. J. (1983). Intestinal flora of infants. In Human intestinal microflora in health and disease, Hentes, David J., ed. (New York: Academic Press), pp. 79–100.

Gomez-Alvarez, V., Teal, T. K., and Schmidt, T. M. (2009). Systematic artifacts in metagenomes from complex microbial communities. ISME J *3*, 1314–1317.

Gupta, S., Mohammed, M., Ghosh, T., Kanungo, S., Nair, G., and Mande, S. (2011). Metagenome of the gut of a malnourished child. Gut Pathogens *3*, 7.

Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J., and Knight, R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat. Methods *5*, 235–237.

Heyworth, B., and Brown, J. (1975). Jejunal microflora in malnourished Gambian children. Arch. Dis. Child *50*, 27–33.

Horton, R. (2008). Maternal and child undernutrition: an urgent opportunity. The Lancet *371*, 179.

Katona, P., and Katona-Apte, J. (2008). Clinical Practice: The Interaction between Nutrition and

Infection. Clinical Infectious Diseases *46*, 1582–1588.

Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., and Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. Nature *474*, 327–336.

Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., Knight, R., and Kelley, S. T. (2011). Bayesian community-wide culture-independent microbial source tracking. Nat. Methods *8*, 761–763.

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., and Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. Proc. Natl. Acad. Sci. U.S.A *108 Suppl 1*, 4578–4585.

Kristiansson, E., Hugenholtz, P., and Dalevi, D. (2009). ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. Bioinformatics *25*, 2737–2738.

Lozupone, C., and Knight, R. (2005). UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Appl. Environ. Microbiol. *71*, 8228–8235.

Mata, L. J. (1975). Malnutrition-infection interactions in the tropics. Am. J. Trop. Med. Hyg *24*, 564–574.

Mata, L. J., Carrillo, C., and Villatoro, E. (1969). Fecal microflora in health persons in a preindustrial region. Appl Microbiol *17*, 596–602.

Mata, L. J., Jiménez, F., Cordón, M., Rosales, R., Prera, E., Schneider, R. E., and Viteri, F. (1972). Gastrointestinal flora of children with protein--calorie malnutrition. Am. J. Clin. Nutr *25*, 118–126.

Mata, L., and Urrutia, J. (1971). Intestinal colonization of breast-fed children in a rural area of low socioeconomic level. Annals of the New York Academy of Sciences *176*, 93–109.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., O'Hara, R. B., Simpson, G. L., Steven, M. H. H., and Wagner, H. (2011). vegan: Community Ecology Package.

Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., and Brown, P. O. (2007). Development of the Human Infant Intestinal Microbiota. PLoS Biology *5*, e177 EP -.

R Development Core Team (2011). R: A language and environment for statistical computing. Available at: http://www.R-project.org/.

Sela, D. A., and Mills, D. A. (2010). Nursing our microbiota: molecular linkages between bifidobacteria and milk oligosaccharides. Trends in Microbiology *18*, 298–307.

Tomkins, A. M., Bradley, A. K., Oswald, S., and Drasar, B. S. (1981). Diet and the faecal microflora of infants, children and adults in rural Nigeria and urban U.K. J Hyg (Lond) *86*, 285–293.

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., and Bork, P. (2011). iPath2.0: interactive pathway explorer. Nucleic Acids Research. *39*, W412-W415.

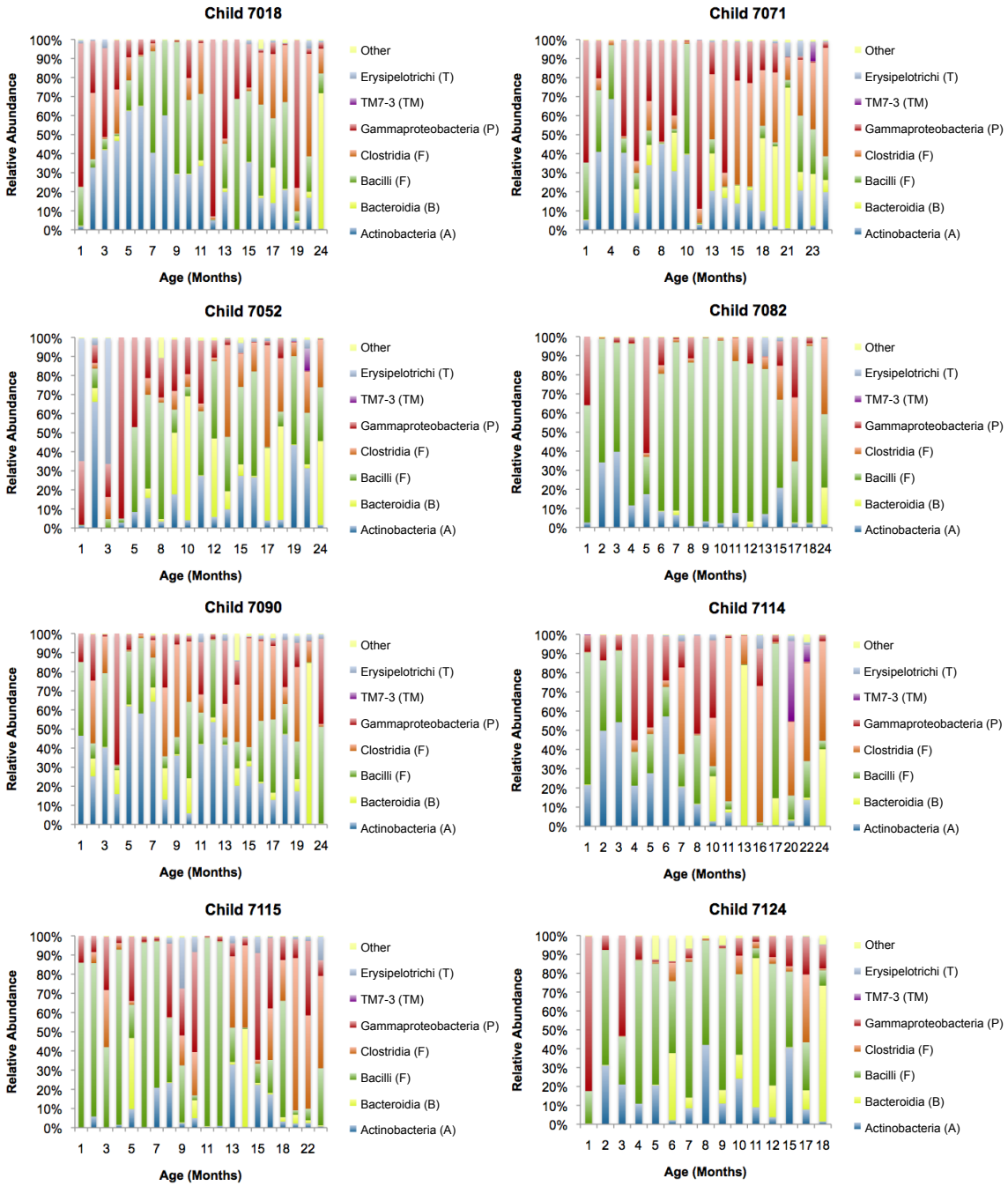
Supplementary Figure Legends

Supplementary Figure 1. Interpersonal variation in bacterial class changes over time in 8 healthy children. The children presented in this figure are the "Healthy" members of the cohort (by HAZ status) who were chosen for shotgun gene sequencing. Plots are constructed as described in Figure 2. The bacterial phyla is noted in parentheses after the class name: A=Actinobacteria, B=Bacteroides, F=Firmicutes, P=Proteobacteria, T=Tenericutes, TM=TM7.

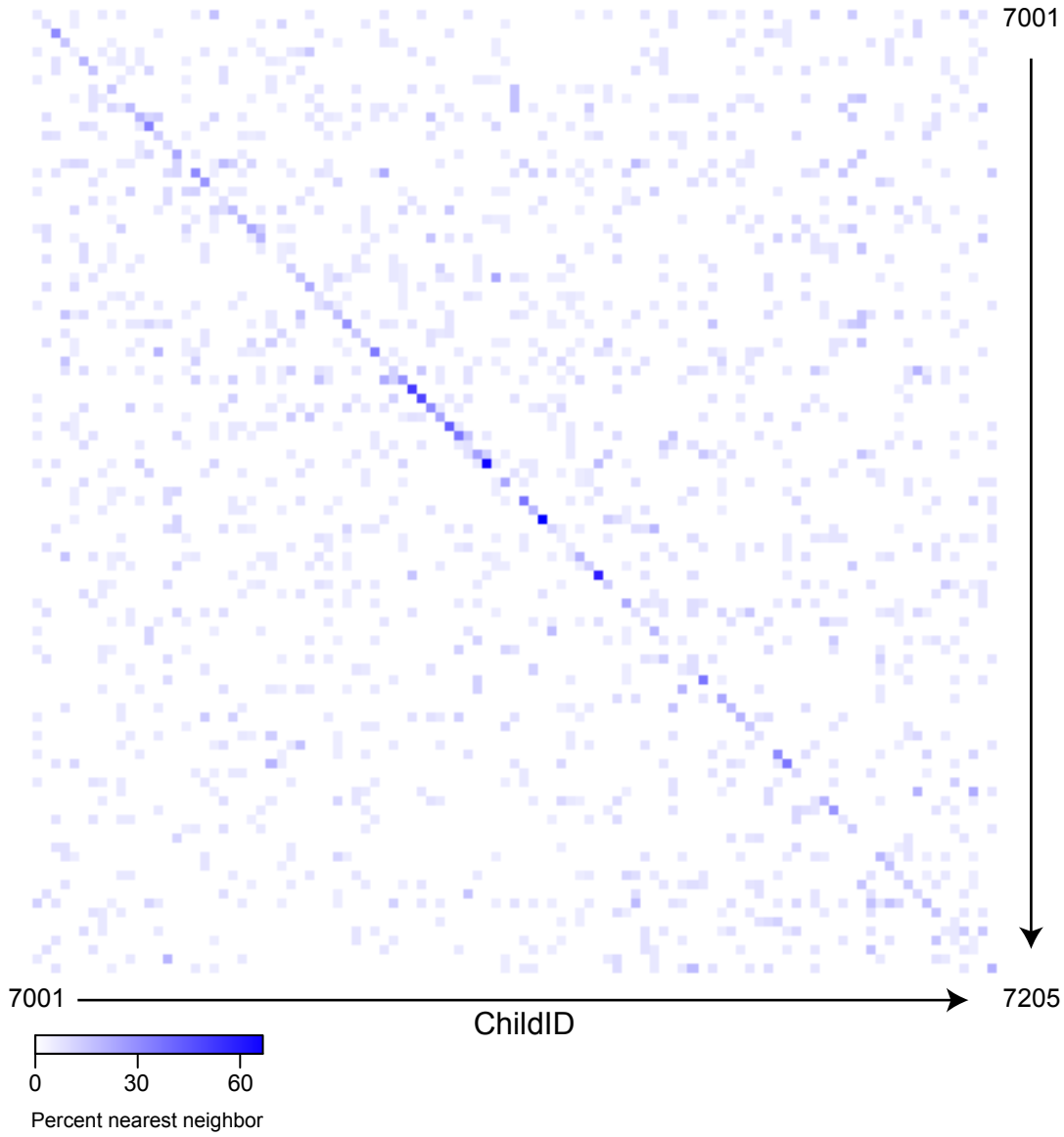
Supplementary Figure 2. Low intrapersonal stability of the gut microbiota over time. Unweighted UniFrac distances were calculated for all pairwise comparisons of samples. The heatmap presents the percentage of samples from a child (row) whose nearest neighbor was found in a sample from the individual listed in the column. Entries on the diagonal represent the self-self comparisons.

Supplementary Figures

Supplementary Figure 1.



Supplementary Figure 2.



Supplementary Tables

Supplementary Table 1. Metadata on all children in the birth cohort.

Supplementary Table 2. Spearman correlation of pathway changes in healthy children over time.

Supplementary Table 3. List of 127 gut-isolated bacterial and archael genomes used for shotgun taxonomy assignment.

Chapter 4

Prospectus

Chapter 4: Prospectus

"It is nice to think that there are so many unsolved puzzles ahead for biology, although I wonder whether we will ever find enough graduate students."

Lewis Thomas (Thomas, 1978)

Collectively, the studies described in this thesis demonstrate that the functional properties of the gut microbiome are influenced by host diet. We showed the power of comparative metagenomic analysis to reveal differences in microbial communities, and our work has created a list of several functional metabolic pathways that seem especially influenced by host diet. Yet as I conclude my graduate work, I am left with the belief that our research has opened more new questions to explore than we have answered; I hope that this means that my time in the lab was well spent. Much work remains before the effect of diet on the microbes in the gut can be predicted with any accuracy, and we are even farther from being able to design diets that might modulate the operation of the community. In the next few pages, I will discuss three different research aims that would springboard from this thesis to study different aspects of the host-diet-microbe interaction.

Testing theories in gnotobiotic mice: focus on weaning and glycans

Broad human surveys such as the work presented in Chapter 3 are often the first step towards understanding new biology. However, human studies alone will never reveal the total picture. In the microbiome of infants, this is especially true because of the enormous interpersonal variation between the members of the cohort and our inability to intervene in their lives with dietary or other manipulation. Biology's solution to these problems is to use animal models, and the stage is now set to move our studies of microbiome assembly and infant diets into gnotobiotic mice.

The highly controlled environmental conditions achievable in gnotobiotic isolators means that a single microbial community can be spread to multiple replicate animals, allowing hypothesis testing with control animals gavaged with the exact same microbial community. When these animals are gavaged with small, defined communities of bacteria, modern sequencing technologies can be used to monitor community abundance and gene expression (Faith et al., 2010). For instance, Jeremiah Faith recently developed quantitative predictive models of the response to defined diets by 10 sequenced human gut bacterial species in gnotobiotic mice (Faith et al., 2011). The bacteria transferred into mice can also represent a complex natural community. Work by our lab and others has shown that human fecal samples can be gavaged into mice and that most of the

microbial diversity present in the input is captured in the animals (Goodman et al., 2011; Turnbaugh et al., 2009). Members of the lab have already transferred stool from the Bangladeshi cohort into gnotobiotic mice and are monitoring the effect of altered diets and the patterns of vertical inheritance at birth. That work and planned future experiments will also start to test the response of these communities to malnourished diets.

The study presented in Chapter 3 can be directly extended in mice to tease apart the contributions of diet and other factors in early community assembly. This study should begin with a very simple question; are the kinetics of community assembly altered if weaning is delayed? Although the connection between weaning and microbial community changes is part of the dogma in the field, there are no studies that actually test this belief. The initial experiment to test this question is a simple one. Pups would be raised in a cage with their mother and another lactating dam. At the 15th day of life (several days before the expected weaning transition), all solid food would be removed from the cage; the dams would be alternately cycled between the cage with the pups and the cage with the food so that the pups always have access to milk. Fecal samples would be collected from pups and mothers every day, and the experiment would proceed until the pups stopped getting adequate nourishment from the available milk. The composition of the microbiota in the fecal DNA would be quantified with 16S rRNA sequencing, along with complementary measurements of microbial gene abundance by shotgun sequencing, mRNA expression by microbial RNA-Seq, and cecal metabolites by mass spectrometry or NMR. The microbial community in the pups exposed to extending suckling would be compared with a control group, born to mothers inoculated with the same bacterial community, but allowed to wean at the normal time.

I hypothesize that at least two groups of microbes would be identified by this analysis. The first will be a group of organisms that never become abundant until solid food is introduced in the diet (or conversely, a group that is abundant as long as milk is present but rarely thereafter). This group would represent the collection of microbes that are “diet” dependent. The second group will be the “age” dependent set of microbes that follow similar kinetics in the control and experimental groups of animals. If such an ensemble were identified, it would strongly suggest that one or more

developmental patterns in the mouse mucosa or immune system are shaping the bacterial species in the gut. Initial candidates for factors known to mature at weaning and interact with gut microbes include glycan modifications on the surface of the epithelium and in secreted mucus, intestinal angiogenesis, and the expression of defensin molecules by members of the innate immune system (Hooper, 2004). The potential impact of glycans is especially intriguing because of the significant changes in glycan metabolism identified in our human studies, and also because previous studies in gnotobiotic mice have demonstrated the intimate interrelationship between the prominent human saccharolytic bacterium, *Bacteroides thetaiotamicron* with production and harvesting of epithelial fucosylated glycoconjugates (Bry et al., 1996; Hooper et al., 1999).

This experimental base could give rise to a number of followup studies. Gnotobiotic animals could be humanized with samples from different infants to look for differences in intestinal mucosal glycan production or responses in the innate immune system due to these different microbial communities. The experiments could be extended by giving the mice diets representative of the human host's diet that are either sufficient or deficient in specified micro- and/or macronutrients. Bacteria that are responsive to different conditions could be isolated from the animals for genome sequencing and in vitro studies, and then reintroduced to mice for hypothesis testing. The role of host factors could be further addressed with genetically engineered mice with specified alterations in host factors postulated to play an important role in initial colonization of the gut or community dynamics. The combined goal of all of these studies should be to work towards a complete ecological model of the assembly process in the gut (Helmus et al., 2007; Robinson et al., 2010). Ecological models have been developed for environmental microbial communities, but none exist at a similar level of detail for complex host-associated communities. The set of experiments outlined above could provide the first quantitative model of the factors that regulate host microbiome operation, at least in a one phase of life. They are also likely to provide new insight to the physiology of weaning.

Microbiome hunters: in search of extreme communities

One enduring lesson from my project is that the discovery is much easier when the effect size is large. If mammalian diet has any role in changing the functional properties of a microbiome, than comparing echidnas to elephants seems like a good place to start the search. Similarly, our human studies in infants focused on the single most dynamic period of human microbial ecology and diet change.

To continue exploring the role of human diet in shaping the microbiome, there is still important work to be done characterizing the microbial response at the extremes of our diet. One approach would be to seek out populations of people with extremely unusual diets. For example, the Inuit people of Greenland and northern Canada depend on locally caught or harvested food, notably animal blubber. The traditional Inuit diet is almost devoid of carbohydrates and contains more than 50% of calories from fat; current research on the cardioprotective role of omega-3 fatty acids began when epidemiologists found that Inuits had lower rates of acute myocardial infarction than matched controls (O'Keefe and Harris, 2000). What are the responses of the human microbiome, which we think is geared towards the fermentation of complex polysaccharides, when the carbohydrates are absent from the diet? Are the microbial communities 'plastic' enough to adapt with the seasonal variation in Inuit diets? A study of this population would be even stronger if another pastoralist culture could be identified and included; similar microbial responses in both populations would likely be more directly linked to diet.

Another paradigm for biology at the extremes is to look at the change within a single host with extreme dietary shifts. Just as the Bangladeshi infant study was empowered by the ability to use each child as their own control, dietary interventions in adults should include a thorough characterization of the microbial community before any sort of manipulation. This is especially important at this point in time because the field does not have a firm understanding of the role that past exposures plays in shaping the operation of the gut microbiome.

These types of analyses should be aided and abetted by 'translational' studies involving gnotobiotic mice who (i) have received transplants of fecal microbiota from human donors (in-

cluding those who normally consume various 'extreme' diets) and (ii) are fed the very extreme diets of the donor and various systematically manipulated derivatives of these diets.

Role of diet in disease: *C. difficile* **colitis**

The work of this thesis established that diet can and does change the structure of the gut microbiome. We based these conclusions on the presence of genomic signatures in fecal DNA, representing the potential functional capacity of the system. However, there is another level of analysis needed: the regulation of gene expression under different conditions as a means to alter community operation. A microbe in the gut with a constant genome can demonstrate very different behaviors depending on the environmental conditions. For example, work from our lab in vitro and in gnotobiotic mice has demonstrated that the model gut symbiont *B. thetaiotaomicron* contains a vast repertoire of polysaccharide utilization loci and that these PULs are differentially regulated by different glycan structures (Bjursell et al., 2006; Martens et al., 2008, 2009; Sonnenburg et al., 2005)

Clostridium difficile is another organism that can behave very differently in the gut at different life stages. In adults, *C. difficile* is generally virulent and is the causative agent of pseudomembranous colitis after antibiotic therapy (Kyne et al., 2001). This bacterial species is undetectable or present at very low levels in healthy adults. By contrast, 40% or more of human infants are asymptomatic carriers of *C. difficile*, with a higher incidence when they are formula- compared to breast milk fed (Cooperstock et al., 1983). The reasons for this are completely unknown, although it is tempting to speculate that host diet may be part of the explanation. First, gnotobiotic animal studies have demonstrated that mice mono-associated with *C. difficile* have lower mortality and lower measurable levels of toxin when raised on high protein or low carbohydrate diets compared to normal chow (Mahe et al., 1987). Second, one of the unreported findings from our mammal study is that the guts of healthy adult carnivores are significantly enriched with OTU's assigned as clostridial species, notably *C. perfringens*.

These facts lead to the hypothesis that the protein-rich, carbohydrate-deficient diet characteristic of infancy may create a favorable environment for the growth of *C. difficile*, but with an

expressional profile that is not pathogenic to the human host. The basis for this may be that the availability of amino acid or lipid substrates as energy sources shifts the organism's metabolism away from toxin production. These questions lend themselves to several microbiological experiments. What are the differences in gene expression and toxin expression when *C. difficile* is grown in media with energy sources typical of the infant intestinal tract compared to the adult tract? Are the same expressional patterns observed when the organism is introduced into gnotobiotic mice fed different diets? Can *C. difficile* pathology in animal models be ameliorated with a switch to an infant-like diet devoid of complex polysaccharides and enriched for proteins and lipids? If this hypothesis is upheld in these experiments, then clinical trials using carbohydrate deficient diets as part of the management of *C. difficile* associated diarrhea may be warranted.

The idea of using food to alter gut microbial metabolism is a very old one. Many studies of the human microbiota at the dawn of the 20th century focused on the "proteolytic flora" and the metabolic products they created that were thought to damage host tissues. Kendall advocated the idea of "bromotherapy" where high carbohydrate diets would be administered to patients with active intestinal infection in an attempt to drive the pathogen's metabolism towards a more benign state (Kendall, 1921, 625-628). Later in the century, bacterial metabolism of diet was again implicated in disease, this time the onset of colon cancer though microbial production of carcinogens from diets rich in fats and red meat (Guarner and Malagelada, 2003). The theme in these studies is the enduring question that also motivated this thesis; how do the foods that we eat influence our gut microbes, and in turn our health? The proposed experiments enumerated in this chapter are simply the vanguard of many studies to come that will identify functional components of food that could be used as therapeutic interventions to enhance the physiologic or metabolic phenotype of the host by modulating the microbiome.

References

Bjursell, M. K., Martens, E. C., and Gordon, J. I. (2006). Functional Genomic and Metabolic Studies of the Adaptations of a Prominent Adult Human Gut Symbiont, Bacteroides thetaiotaomicron, to the Suckling Period. Journal of Biological Chemistry *281*, 36269–36279.

Bry, L., Falk, P. G., Midtvedt, T., and Gordon, J. I. (1996). A model of host-microbial interactions in an open mammalian ecosystem. Science *273*, 1380–1383.

Cooperstock, M., Riegle, L., Woodruff, C. W., and Onderdonk, A. (1983). Influence of age, sex, and diet on asymptomatic colonization of infants with Clostridium difficile. J. Clin. Microbiol *17*, 830–833.

Faith, J. J., Rey, F. E., O'Donnell, D., Karlsson, M., McNulty, N. P., Kallstrom, G., Goodman, A. L., and Gordon, J. I. (2010). Creating and characterizing communities of human gut microbes in gnotobiotic mice. ISME J *4*, 1094–1098.

Faith, J. J., McNulty, N. P., Rey, F. E., and Gordon, J. I. (2011). Predicting a Human Gut Microbiota's Response to Diet in Gnotobiotic Mice. Science *333*, 101–104.

Goodman, A. L., Kallstrom, G., Faith, J. J., Reyes, A., Moore, A., Dantas, G., and Gordon, J. I. (2011). Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. Proceedings of the National Academy of Sciences *108*, 6252–6257.

Guarner, F., and Malagelada, J.-R. (2003). Gut flora in health and disease. Lancet *361*, 512–519.

Helmus, M. R., Savage, K., Diebel, M. W., Maxted, J. T., and Ives, A. R. (2007). Separating the determinants of phylogenetic community structure. Ecol Letters *10*, 917–925.

Hooper, L. V. (2004). Bacterial contributions to mammalian gut development. Trends in Microbiology *12*, 129–134.

Hooper, L. V., Xu, J., Falk, P. G., Midtvedt, T., and Gordon, J. I. (1999). A molecular sensor that allows a gut commensal to control its nutrient foundation in a competitive ecosystem. Proceedings of the National Academy of Sciences of the United States of America *96*, 9833–9838.

Kendall, A. I. (1921). Bacteriology: general, pathological and intestinal (Lea & Febiger).

Kyne, L., Farrell, R. J., and Kelly, C. P. (2001). Clostridium difficile. Gastroenterol. Clin. North Am *30*, 753–777, ix-x.

Mahe, S., Corthier, G., and Dubos, F. (1987). Effect of various diets on toxin production by two strains of Clostridium difficile in gnotobiotic mice. Infect. Immun *55*, 1801–1805.

Martens, E. C., Chiang, H. C., and Gordon, J. I. (2008). Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. Cell Host Microbe *4*, 447–457.

Martens, E. C., Roth, R., Heuser, J. E., and Gordon, J. I. (2009). Coordinate regulation of glycan degradation and polysaccharide capsule biosynthesis by a prominent human gut symbiont. J. Biol. Chem *284*, 18445–18457.

O'Keefe, J. H., and Harris, W. S. (2000). From Inuit to implementation: omega-3 fatty acids come of age. Mayo Clinic Proceedings *75*, 607–614.

Robinson, C. J., Bohannan, B. J. M., and Young, V. B. (2010). From Structure to Function: the Ecology of Host-Associated Microbial Communities. Microbiol. Mol. Biol. Rev. *74*, 453–476.

Sonnenburg, J. L., Xu, J., Leip, D. D., Chen, C.-H., Westover, B. P., Weatherford, J., Buhler, J. D., and Gordon, J. I. (2005). Glycan Foraging in Vivo by an Intestine-Adapted Bacterial Symbiont. Science *307*, 1955–1959.

Thomas, L. (1978). Lives of a Cell: Notes of a Biology Watcher (Penguin (Non-Classics)).

Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., and Gordon, J. I. (2009). The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice. Science Translational Medicine *1*, 6ra14–ra16ra14.

Appendices

Appendix A

Crawford P.A., Crowley J., Sambandam N., **Muegge B.D.**, Costello E., Hamady M., Knight R., Gordon J.I.

"Regulation of myocardial ketone body metabolism by the gut microbiota during nutrient deprivation."

Proceedings of the National Academy of Sciences USA. **2009**. 106(27): 11276-81.

Appendix B

Caporaso J.G., Kuczynski J., Stombaugh J., Bittinger K., Bushman F.D., Costello E.K., Fierer N., Peña A.G., Goodrich J.K., Gordon J.I., Huttley G.A., Kelley S.T., Knights D., Koenig J.E., Ley R.E., Lozupone C.A., McDonald D., **Muegge B.D.**, Pirrung M., Reeder J., Sevinsky J.R., Turnbaugh P.J., Walters W.A., Widmann J., Yatsunenko T., Zaneveld J., Knight R.

"QIIME allows analysis of high-throughput community sequencing data."

Nature Methods. **2010**. 7(5): 335-6.