Characterization of Human Gut Microbiota Dynamics Using Model Communities in Gnotobiotic Mice

Nathaniel Patrick McNulty

Follow this and additional works at: http://openscholarship.wustl.edu/etd

Part of the Microbiology Commons

Recommended Citation

http://openscholarship.wustl.edu/etd/1063

This Dissertation is brought to you for free and open access by Washington University Open Scholarship. It has been accepted for inclusion in All Theses and Dissertations (ETDs) by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
Characterization of Human Gut Microbiota Dynamics Using Model Communities in Gnotobiotic Mice

by

Nathaniel Patrick McNulty

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

List of Figures ................................................................. vi
List of Tables ......................................................................... viii
Acknowledgements ............................................................. x
Abstract of the Dissertation .................................................. xiii
Epigraph ................................................................................ xv

Chapter 1

Introduction

The human gut microbiota: a metabolic organ of substantial complexity ...................... 2
  Host-microbial mutualism in the human gastrointestinal tract .................................... 2
  Phylogenetic diversity and interpersonal variation .................................................... 3
  Microbial contributions to human physiology and metabolism .............................. 5
  The need for representative models of the human gut microbiota ......................... 7
Creating a defined system: model communities of sequenced human gut microbes .......... 8
  Colonization of germ-free mice with complex, undefined communities ................. 8
  Assembling defined, representative model communities in vivo ......................... 11
Compiling a molecular tool kit: methods for characterizing model community structure and function ......................................................................................................................... 12
  Profiling community structure and the development of COPRO-Seq .................... 12
  Profiling community function using improved methods ...................................... 14
  Genetic manipulation of the community metagenome ....................................... 16
Studying the system: characterization of communities and specific taxa in contexts relevant to human health .............................................................................................................. 17
  The impact of probiotic interventions on the microbiota .................................. 17
  The impact of diet change over time ................................................................ 18
References ................................................................................ 20
Chapter 2

The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins

Abstract ...........................................................................................................................................27
Results .............................................................................................................................................30
  Human studies ...............................................................................................................................30
  Studies in gnotobiotic mice ...........................................................................................................35
Discussion .......................................................................................................................................45
Materials and methods ..................................................................................................................49
  Human studies ...............................................................................................................................49
  Studies in gnotobiotic mice ...........................................................................................................49
Acknowledgements .......................................................................................................................50
Footnotes ..........................................................................................................................................51
References .......................................................................................................................................52
Figure legends .................................................................................................................................57
Figures ............................................................................................................................................61
Supplementary material ..................................................................................................................68
  Supplementary materials and methods ..........................................................................................68
  Supplementary results ....................................................................................................................82
  Supplementary references .............................................................................................................86
  Supplementary figure legends .......................................................................................................91
  Supplementary figures ..................................................................................................................95
  Supplementary table legends .......................................................................................................104

Chapter 3

Effects of diet on resource utilization by a model human gut microbiota containing *Bacteroides cellulosilyticus* WH2, a symbiont with an extensive glycobiome

Abstract ..........................................................................................................................................109
Introduction .....................................................................................................................................110
Results and discussion ....................................................................................................................112
Chapter 4

Future directions

Introduction......................................................................................................................176
Creating more representative models of the human microbiota..................................176
   Culturing the ‘unculturable’ ..................................................................................176
   Personal culture collections ..............................................................................178
Developing new and improved tools and techniques for characterizing model communities ...179
   Absolute quantitation of microbial biomass and taxon abundance ......................179
   An expanded role for INSeq in model community studies ..................................180
   New imaging techniques for evaluating gut biogeography ...................................181
Further study of specific taxa and ecological guilds ...................................................182
Concluding statements .................................................................................................................. 184
References .................................................................................................................................... 185

Appendices

Appendix A
Appendix B
List of Figures

Chapter 2

The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins

Figure 1. Experimental design for human and mouse studies .......................................................... 61
Figure 2. Metagenomic studies of human fecal microbiomes sampled over time ..................... 62
Figure 3. Correspondence analysis of B. animalis subsp. lactis CAZyme gene expression ....... 63
Figure 4. ‘Top-down’ analysis of the effects of the FMP strain consortium on the model 15-member community’s metatranscriptome ................................................................. 64
Figure 5. Mouse and human communities share transcriptional responses to the FMP strain consortium involving ECs related to carbohydrate metabolism .......................... 65
Figure 6. Select urinary metabolites whose levels are altered after the introduction of the FMP strain consortium into mice harboring a defined model human gut microbiota .......................................................................................................................... 66
Figure 7. Shared transcriptional responses to FMP strain exposure in mice and humans ...... 67
Figure S1. Levels of B. animalis subsp. lactis (CNCM I-2494) in human fecal samples collected prior to, during, and after consumption of a FMP ........................................ 95
Figure S2. KEGG pathway coverage ratios suggest that the model human gut microbiome encodes many of the functions present in more complex human fecal communities .................................................................................................................. 96
Figure S3. CAZyme profiles of the 20 bacterial strains introduced into gnotobiotic mice .... 97
Figure S4. Summary of analysis pipelines utilized in this study .................................................. 98
Figure S5. COPRO-Seq-based time series analysis of the abundance of members of the model human microbiota and of the FMP strain consortium in the feces of gnotobiotic mice .................................................................................................................. 99
Figure S6. Top-down analysis of the model community’s transcriptional response to the FMP strain consortium reveals upregulation of genes involved in interconversion of propionate and succinate .............................................................................................................. 100
Figure S7. A species’ contribution to the metatranscriptome is not necessarily proportional to its abundance in the 15-member community ............................................................................. 101
Figure S8. Bottom-up analysis of genes whose expression changes significantly following introduction of the FMP strain consortium ............................................................................. 102
Figure S9. The number of RNA-Seq reads obtained from human fecal samples that map to
genomes in the FMP strain consortium peaks shortly after FMP consumption begins.

Chapter 3

Effects of diet on resource utilization by a model human gut microbiota containing \textit{Bacteroides cellulosilyticus} WH2, a symbiont with an extensive glycobiome

- Figure 1. COPRO-Seq analysis of the structure of a 12-member model human microbiota as a function of diet and time. ................................................................. 140
- Figure 2. \textit{B. cellulosilyticus} WH2 CAZyme expression in mice fed different diets. ....... 141
- Figure 3. Top-down analysis of fecal microbiome gene expression in mice receiving oscillating diets. ...................................................................................... 142
- Figure 4. Two xylanase-containing \textit{B. cellulosilyticus} WH2 PULs demonstrating strong diet-specific expression patterns \textit{in vivo}. ................................................................. 143
- Figure 5. \textit{In vitro} microbial RNA-Seq profiling of \textit{B. cellulosilyticus} WH2 during growth on different carbohydrates. ................................................................. 144
- Figure S1. Phylogenetic relatedness of \textit{B. cellulosilyticus} WH2 to other \textit{Bacteroides} species. .................................................................................................................. 161
- Figure S2. Representation of all putative glycoside hydrolase (GH) families identified in the \textit{B. cellulosilyticus} WH2 genome compared to their representation in other sequenced \textit{Bacteroidetes} species. ...................................................................................... 162
- Figure S3. Design and sampling schedule for experiments E\textsubscript{1} and E\textsubscript{2}. .................. 163
- Figure S4. COPRO-Seq analysis of the proportional representation of component taxa in the 12-member model human gut microbiota as a function of time after colonization of gnotobiotic mice and the diet they were consuming. .............................................. 164
- Figure S5. Further COPRO-Seq analysis of the relative abundance of components of the 12-member model human microbiota as a function of diet and time. .............. 165
- Figure S6. GeneChip profiling of the cecal metatranscriptome in mice fed different diets. 166
- Figure S7. Dissecting the model human gut microbiota’s \textit{in vivo} expression of EC 3.2.1.8 (endo-1,4-β-xylanase). ...................................................................................... 167
- Figure S8. Shotgun metaproteomic analysis of cecal samples from gnotobiotic mice colonized with the model 12-member human gut microbiota. .................. 168
List of Tables

Chapter 2

The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins

Table S1. Characteristics of adult female monozygotic (MZ) twins enrolled in study.
Table S2. Summary of human fecal metagenomic datasets.
Table S3. Features of the microbial genomes in the 5-member FMP strain consortium and the 15-member model human gut microbiota.
Table S4. Carbohydrate active enzyme (CAZy) annotation data.
Table S5. COPRO-Seq analysis of bacterial species abundance in mouse fecal samples.
Table S6. INSeq analysis.
Table S7. Differentially expressed \textit{B. animalis} subsp. \textit{lactis} (CNCM I-2494) genes.
Table S8. Top-down function-level analysis of the impact of the FMP strain consortium on the model human gut microbiota’s metatranscriptome.
Table S9. Model human gut microbiota membrane transport genes demonstrating \(\geq4\)-fold increases or decreases in their expression following introduction of the FMP strain consortium.
Table S10. Bottom-up (gene-level) analysis of the impact of the FMP strain consortium on the model community’s metatranscriptome.
Table S11. Results of Random Forests supervised classification analysis.
Table S12. Urine metabolites whose levels change significantly in transitions between colonization states.
Table S13. ShotgunFunctionalizeR analysis of EC-level changes in the metatranscriptome as a function of FMP strain introduction into mice and humans.
Table S14. Primers and amplification conditions used for quantitative PCR assays of FMP consortium strains in fecal DNA.
Table S15. List of 136 microbial genomes used to analyze human fecal RNA-Seq data.

Chapter 3

Effects of diet on resource utilization by a model human gut microbiota containing \textit{Bacteroides cellulosilyticus} WH2, a symbiont with an extensive glycobiome

Table S1. Sequencing statistics for \textit{B. cellulosilyticus} WH2.
Table S2. *B. cellulosilyticus* WH2 genome features with relevance to carbohydrate metabolism.

Table S3. Bacterial strains included in this study.

Table S4. Composition of the 12-member consortium inoculated by oral gavage into germ-free animals.

Table S5. COPRO-Seq quantitation of the relative abundance of community members over time.

Table S6. GeneChip measurements of cecal gene expression for the 12 bacterial species comprising the model human gut microbiota studied in experiment E₁.

Table S7. List of EC numbers whose representation within the fecal metatranscriptome is significantly impacted by diet.

Table S8. Summary of theoretical peptidome statistics.

Table S9. Number of proteins detected within each cecal sample for each species in our custom SEQUEST database.

Table S10. Raw and normalized MS/MS spectral counts for detectable proteins in E₁ cecal samples.

Table S11. Growth of *B. cellulosilyticus* WH2 and *B. caccae* on a panel of structurally diverse carbohydrates.

Table S12. Preparation of minimal medium for *in vitro* gene expression profiling of *B. cellulosilyticus* WH2.

Table S13. Carbohydrate substrates tested during *in vitro* gene expression profiling of *B. cellulosilyticus* WH2.

Table S14. RNA-Seq gene expression values for *B. cellulosilyticus* WH2 grown *in vitro* on 31 simple and complex saccharides.
Acknowledgements

The renowned Nobel laureate Linus Pauling was once asked how he managed to come up with so many good ideas. His response (to paraphrase) was that he simply had a lot of ideas and threw the bad ones away. Over the years, I have had the great fortune of being surrounded by a number of people who have made it possible for me to think a great many thoughts and who have taught me how to recognize those that are insightful from those that are not. As a result, I would like to believe that this dissertation contains at least a few ‘good’ ideas. These I attribute entirely to the mentors, colleagues, friends and family who have supported and influenced me over the years. I would be completely remiss if I began this dissertation with anything other than an acknowledgement of their invaluable contributions to this work and to my development as a scientist.

First, I would like to thank my thesis advisor, Jeffrey Gordon, for allowing me the privilege of conducting research in his laboratory. The vibrant, collaborative atmosphere in which I have had the pleasure of working and learning is entirely a product of his vision and passion for scientific discovery, and I doubt there is any other place in the world where I would have received comparable support. I will always be grateful for his generosity, enthusiastic encouragement, and willingness to grant me more intellectual autonomy than I was ever entitled to.

I would also like to recognize the remainder of my thesis committee, who generously donated significant amounts of their time and energy over the years. Daniel Goldberg, in particular, deserves special mention for serving as my chair. I know that to the academic, time is one of the most valuable resources there is, and my committee members’ willingness to volunteer theirs for my benefit is deeply appreciated.

Many scientists can pinpoint an event or person in their formative years that contributed in an outsized way to their decision to pursue a career in research. For me, that person was my high school biology teacher, Mr. William Schwindt, whose lectures provided a foundation for my understanding of living things and the natural world. It is with great sincerity that I thank him for his excellent instruction and encouragement in the earliest days of my education as a scientist.
I must also thank a number of previous and current lab members whose mentorship and intellectual influence have particularly impacted me. Daniel Peterson oversaw my initial rotation, provided me with my first hands-on exposure to the world of gut microbes (in particular, \textit{B. thetaiotaomicron}) and convinced me that I was properly equipped to tackle difficult scientific questions. After joining the lab, I was surrounded by a host of brilliant post-docs and graduate students, each with a unique complement of technical skills, research interests and ways of thinking. Though I have learned important things from every one of them, I would like to single out Janaki Lelwala-Guruge, Justin Sonnenburg, Eric Martens, Andrew Goodman, Federico Rey, and Jeremiah Faith for being particularly influential in shaping the way I think about the biological questions addressed in this dissertation. I also want to acknowledge the many graduate students (both my senior and my junior, and too numerous to list individually) who have provided me with direct guidance or advice from time to time or have served as exceptional role models.

Occasionally, my research has created opportunities to meet and work with scientists outside of the lab. I learned a great deal from the individuals at Danone Research with whom we partnered in our studies of the impact of fermented milk strains on the gut microbiota. It was also a true pleasure to work on this same study with experts in mass spectrometry and metabolomics from the Newgard Lab at Duke University. Additionally, our efforts to understand how diet change impacts model gut communities led to a fruitful collaboration with scientists in the Hettich Lab at Oak Ridge National Laboratories, and Alison Erickson deserves special recognition for her significant contributions to this partnership.

I would also like to thank every staff member in the Gordon Lab and the Center for Genome Sciences and Systems Biology whose hard work allows students to keep their energy focused on performing research and thinking about science. Sabrina Wagoner, Jill Manchester, David O’Donnell, and Maria Karlsson are especially deserving of praise in this regard, as are Eric Martin, Brian Koebbe, and Jessica Hoisington-Lopez.
While I am deeply appreciative of the mentorship and support I have received over the years, I am even more grateful for the many, many friendships I’ve formed during this time, both inside and outside of the lab. It is these friends who have made doing research here such a pleasure, regardless of how tough science (or life) has gotten at times. It would take many pages to thank each of these individuals for the numerous positive ways they have affected me, so I will restrain myself from doing so here.

When I reflect on the many influences in my life that have led to this moment in time, all pale in comparison with that of my family. I thank my parents for a lifetime of love, unwavering support, and pride in my intellectual pursuits. No child could ask for more devoted fans. I also thank my grandmother, whose love and encouragement I will always cherish. Above all, I am grateful for the two women who sustain me on a daily basis. The first, my beautiful and brilliant wife Samantha, has been an endless source of support, encouragement, and love. This work is as much her achievement as it is mine. The second, my precious daughter Eva, is and will continue to be my single greatest joy in life. It is to them that I dedicate this work.
ABSTRACT OF THE DISSERTATION

Characterization of Human Gut Microbiota Dynamics Using Model Communities in Gnotobiotic Mice

by

Nathaniel Patrick McNulty

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 human gut is colonized by a diverse array of microbes, collectively referred to as the microbiota. The microbiota’s complexity poses significant challenges in characterizing the rules dictating its assembly, inferring the functional roles of its component species, and understanding how communities sense and respond to changes in their habitat. We developed defined, representative model communities comprised of sequenced human gut bacteria that could be characterized in a highly controlled manner in gnotobiotic mice, plus a suite of scalable molecular tools for assaying community properties. These tools were first used to evaluate how the microbiota is impacted by probiotic bacterial strains found in fermented milk products (FMP). Introduction of a consortium of five FMP strains resulted in only minimal changes in the structural configuration of a 15-member model microbiota. However, RNA-Seq and follow-up mass spectrometry revealed numerous functional responses, many related to carbohydrate metabolism. Results from a study performed in monozygotic twin pairs confirmed many of our observations in the model microbiota, showing that lessons learned from preclinical models can inform the design and interpretation of human studies. In a second set of experiments, we evaluated the impact of food on both a model community and its constituent taxa by feeding gnotobiotic mice oscillating diets of disparate compositi-
tion. In addition to prompt and reversible structural reconfigurations suggesting rules-based diet effects, we noted consistent, staggered changes in the representation of many functions within the metatranscriptome related to carbohydrate and amino acid metabolism. One prominent community member, *Bacteroides cellulosilyticus* WH2, was identified as an adaptive forager that tailors its versatile carbohydrate utilization strategy to the dietary polysaccharides available. The specific carbohydrates that trigger expression of many of this organism’s 113 predicted polysaccharide utilization loci were identified by RNA-Seq analysis during *in vitro* growth on 31 distinct carbohydrate substrates, aiding our interpretation of *in vivo* RNA-Seq and high resolution proteomics data. These results offer insight into how gut microbes adapt to dietary perturbations, both at a community level and from the perspective of a well-adapted symbiont with exceptional saccharolytic capabilities, and illustrate the value of studying defined models of the human gut microbiota.
“In the vernacular of the times, we must unite and confront these microbes with a non-negotiable demand to increase our time of usefulness to them. Mankind should no longer die of infectious diseases; hospitals can be made infection-free; we should live in peaceful coexistence with ‘our microbes,’ and they should contribute positively to our continued health and well-being.”

–T. D. Luckey (1972)
Chapter 1

Introduction
The human gut microbiota: a metabolic organ of substantial complexity

Host-microbial mutualism in the human gastrointestinal tract

While nearly every surface of the human body is colonized with microbes, the greatest density of microorganisms is found within the distal colon \((10^{11}-10^{12} \text{ CFUs/ml})\) [1]. Other, more proximal sites in the gut, including the stomach and duodenum, are also inhabited by microbes, but in smaller numbers \((<10^3 \text{ to } 10^4 \text{ CFUs/ml})\). The prevailing thought among ecologists today is that this population of microbes, collectively referred to as the human gut ‘microbiota’, represents one of the earth’s most dynamic and densely-populated ecosystems [2]. If one assumes that the human body consists of approximately \(10^{13}\) cells (as postulated by Theodosius Dobzhansky in 1970, [3]) and that the human gut harbors some \(10^{14}\) microbial cells (as calculated by T.D. Luckey in 1972, [1]), it becomes clear that the sum of microbial cells inhabiting each of us exceeds our ‘own’ number of \textit{H. sapiens} cells by an order of magnitude. Going one step further, the number of distinct microbial genes within the collective genomes of these microbes (the human gut ‘microbiome’) is estimated to exceed our human genome’s gene content by perhaps as much as 100-fold. Thus, we are living amalgamations of both microbial and non-microbial cells, whose biological operations are inextricably linked with one another in a finely tuned mutualism.

Humans and microbes have co-evolved [4]. In exchange for the shelter we provide them, our gut microbes carry out important functions that the proteins encoded by our own human genes cannot perform. Recent comparative studies of humans, non-human primates, and other mammals have found that each species’ microbiota is tailored to include symbionts with traits well-suited to the host’s anatomy, diet, and other biological parameters [5,6]. And, the results of microbiota transplantation studies in mice and zebrafish suggest that there is selection within a host for those microbes well-suited to a particular habitat [7]. While the human genome is probably not changing very quickly over modern timescales, the conditions in which people live and operate, and the environmental factors shaping our gut communities, are changing at a rapid pace. Thus, if we are to understand the complexities of our biology and how we are evolving as a species, we must
understand who our microbial symbionts are and how our relationship with them is affected by the changing ways we live, eat and interact.

**Phylogenetic diversity and interpersonal variation**

Substantial advances in sequencing technology over the last ten years have allowed microbial ecologists to greatly expand the breadth and depth of culture-independent (metagenomic) surveys of the human gut [8]. As the per nucleotide cost of DNA sequencing has dropped precipitously, at rates even exceeding those predicted by Moore’s Law, researchers have seized on the opportunity to enumerate the microbial diversity in a number of human populations differentiated by many metrics (e.g., geography, health status, cultural tradition). Collectively, the results of these studies provide us with our most comprehensive understanding yet of the incredible diversity that exists in the gut, highlighting both similarities and differences that exist between individuals [9,10].

The intestine is home to species from all three domains of life, though it is dominated by the Bacteria. Representatives of at least 14 bacterial phyla have been identified in human fecal samples to date, including *Bacteroidetes, Firmicutes, Proteobacteria, Actinobacteria, Verrucomicrobia, Fusobacteria, Tenericutes, Lentisphaerae, TM7, Cyanobacteria, Synergistetes, Spirochaetes, SR1*, and *Deinococcus-Thermus* [11]. When more resolved levels of taxonomy are considered, the number of bacterial phylotypes that colonizes human beings increases dramatically. One helpful, previously proposed analogy for visualizing the phylogenetic structure of the gut microbiota suggests that a community can be represented as a grove of palm trees with long trunks (one tree per bacterial phyla), each trunk topped with fans representative of divergent lineages of related bacteria, and each fan comprised of leaflets representative of individuals strains [4]. This illustration emphasizes that a great deal of bacterial diversity in the gut exists at the strain level. Although bacteria are without question the dominant constituents of the gut ecosystem, this microbial menagerie also includes representatives from two other domains (Archaea and Eukaryota), as well as viruses (bacteriophage). However, our understanding of non-bacterial diversity in the gut is in its infancy, and only a small number of studies addressing this area have been reported so far [12-14].
Results from some of the deepest surveys suggest that a typical healthy human is colonized by approximately 100-150 bacterial phylotypes [10]. Comparisons of these many bacterial phylotypes across individuals suggest that each of us is also colonized with a unique assemblage of species [15]. Efforts in recent years to identify a subset of species that reside in all human beings that might constitute a ‘core microbiota’ at the phylogenetic level have not been successful, suggesting it is likely that many different configurations of gut phylotypes can achieve stable or quasi-stable equilibria, presumably because of functional redundancy between different strains/subspecies. These observations suggest that the human gut microbiota’s ‘alpha diversity’ (i.e., the number of distinct microbial organisms observable within an individual) and ‘beta diversity’ (i.e., the degree of differentiation between individuals) are both quite high. The total species diversity within human guts, typically represented by ‘gamma diversity’ (the product of alpha and beta diversity measures), is therefore quite significant. This staggering complexity presents those studying the gut microbiota in human subjects with significant challenges as they attempt to compile generalizable principles about the system and its constituents.

Despite the exceptional diversity described above, there are common themes in community structure and composition when comparing the microbiota of different individuals. For example, although at least 14 bacterial phyla have been observed in human fecal samples, the overwhelming majority of bacteria hail from just two: the Bacteroidetes and the Firmicutes (collectively, taxa from these two phyla constitute approximately 95% of adult gut communities based on recent surveys) [11]. Representatives of the Proteobacteria and Actinobacteria are also common, but are typically observed in low numbers (i.e., <3%) in healthy adults. Although microbes common to all human communities have proven elusive, some individual taxa and related groups of taxa are observed on a regular basis in a large proportion of individuals. One example is Clostridium cluster XIVa of the Firmicutes, which has been suggested to account for as much as 60% of the total mucosa-associated microbiota [16]. This cluster includes a number of species (e.g., Ruminococcus torques, Ruminococcus obeum, Clostridium symbiosum) that have been detected consistently in different subjects from large-scale surveys of the human fecal microbiota [10]. The incidence of
some particular species from the genus *Bacteroides* is also quite high in gut surveys. Some of the most common representatives include *Bacteroides ovatus*, *Bacteroides vulgatus*, and *Bacteroides caccae*, though there are many others as well. Of those bacteria hailing from the phylum *Actinobacteria*, *Collinsella aerofaciens* appears to be most prevalent. A confounding factor in accurately deciphering the representation of different bacterial taxa among humans representing different ages, physiological states, and geographic regions/cultural traditions is the lack of methodological standards in microbiota research. Studies may rely on different methods for extracting microbial DNA from samples, different PCR primers and cycling conditions to amplify different regions of the 16S rRNA gene used as a phylogenetic marker, different sequencing platforms, and different bioinformatics tools to deal with artifacts (sequencing errors, chimeras) arising from amplicon sequencing. Nevertheless, consistent similarities (and differences) between healthy gut communities such as those described above are important for the researcher to consider when attempting to build experimental models of the healthy human gut microbiota.

**Microbial contributions to human physiology and metabolism**

The previously described diversity of our gut microbes (and their genes) affords them a central role in our physiology and metabolism [17]. These microbes encounter innumerable substrates (e.g., dietary compounds, drugs, host molecules, and microbial metabolites, to name a few), and these compounds enter the intestinal ecosystem in a number of ways (orally, as components of biliary and mucosal secretions, or through the circulatory system). This multiplicity of compounds and routes of introduction makes characterizing the biotransformation of different molecules by gut microbes especially challenging.

One way the gut microbiota benefits its human host is by salvaging energy from compounds that we are incapable of efficiently processing with our own metabolic machinery [18]. *Bacteroides* spp., for example, are known for their substantial capacity to degrade and ferment a broad range of plant-derived polysaccharides and oligosaccharides found within our diets, as well as host-derived glycans (e.g., mucin). Gut microbes are thought to liberate energy from some-
where between 20 and 70 g of carbohydrates per day, suggesting that their impact on the extraction of calories from food could be significant [19]. As gut microbes disassemble poly- and oligosaccharides, they ferment their constituent subunits, generating a number of short-chain fatty acids (SCFA; e.g., acetate, propionate, and butyrate) and other end-products (e.g., lactate, succinate, and formate) [20]. In addition to serving as important sources of energy for a subset of host cells (e.g., colonocytes), these SCFA also stimulate epithelial renewal and facilitate the absorption of salt and water in the distal gut [17]. Unfortunately, little is currently known about which microbes within these complex communities are responsible for degrading particular dietary substrates.

In addition to strongly influencing macronutrient processing in the distal gut, the microbiota also impacts how we process a host of smaller molecules that are exogenously as well as endogenously derived. Several examples of gut microbial biotransformation of orally administered drugs have been documented, including the inactivation of digoxin by *Eggerthella lenta* [21] and the deglucuronidation of the chemotherapeutic irinotecan by microbial beta-glucuronidases [22]. Gut microbes are also known to play an important role in the deconjugation of bile acids, ensuring efficient reabsorption of free bile acids from the ileum and colon and the loss of only around 5% of all bile acids in feces [23]. Finally, the microbiota plays an important role in the synthesis of some vitamins. Previous work has shown that while adults maintained on a diet low in vitamin K do not develop deficiencies even after several weeks, the administration of antibiotics that have activity against gut microbes significantly impacts the host’s ability to carry out processes requiring vitamin K [24]. Experiments in conventionally-raised and germ-free rats have demonstrated that gut microbial synthesis of biotin is sufficient to meet the animals’ requirements [25]. As with macronutrient processing, a detailed understanding of how individual components of the microbiota cooperate and compete to process and synthesize micronutrients and other small molecules is currently lacking. New approaches to studying the microbiota that reduce the complexity of the system while preserving the way in which it operates should allow us to monitor the activities of individual members so that we can better understand the roles played by key taxa and the metabolic guilds of which they are members.
Though beyond the scope of this dissertation, it is important to note that gut microbes are also known to play a key role in shaping features of the innate and adaptive immune systems [26]. Some gut microbes prevent colonization of the intestine by pathogens or other allochthonous (transient) microbes by colonizing mucosal surfaces and/or secreting antimicrobial peptides. Others have been shown to exert strong influences on specific subsets of immune cells. For example, researchers have reported that members of the genus *Clostridium* induce the regulatory lineage of CD4+ T cells (Tregs) in mice [27]. And still other work has implicated reduced diversity of the microbiota, especially during infancy, with increased rates of atopy [28]. Although there have been a small number of cases in which specific human gut microbial lineages have been tied to immune phenotypes, further dissection of these important and complex interactions is needed.

**The need for representative models of the human gut microbiota**

Although the importance of studying the gut microbiota in human subjects is undeniable and has resulted in many important insights, such studies are limited in several ways and present challenges that can be overcome by studying animal models of the human microbiota. First, it is logistically difficult, expensive and time-consuming to study the microbiota in recruited human subjects. Animal studies (e.g., in mice) are by comparison relatively inexpensive, require shorter amounts of time, and are easier to plan and oversee. The expense and logistical challenges of human studies also typically restricts researchers from carrying out experiments in large cohorts, restricting the amount of biological replication that can be built into a study, as well as the number of times large-scale human studies can be replicated. Small, unrepeatable human studies may not provide the statistical power required to work out complex relationships in a system whose biological signals may be faint and noisy. Studies in animal models, by comparison, are much more amenable to replication. In contrast with human subjects, animals can also be maintained in habitats where environmental variables such as diet can be tightly controlled, along with important host parameters such as genotype, age, and health status. Additionally, longitudinal studies of the human gut
microbiota have in nearly every case been limited to analyses of feces, whereas studies of model human microbiota in animals allow for sampling along the length of the gut at the time of sacrifice.

The complexity of the human microbiota makes it extremely difficult to define and characterize relationships between different taxa and between taxa and ecosystem parameters. For example, if we assume that in a microbiota comprised of 500 distinct taxonomic groups, all components interact with one another, there would be 124,750 distinct pairwise combinations of relationships to characterize. Experimental models of the microbiota, on the other hand, can intentionally reduce this complexity while preserving the key features of the microbiota. A calculation similar to the one above in which a model microbiota of 15 taxa are considered yields a more manageable 105 distinct pairwise interactions between community members. Additionally, our ability to precisely manipulate the composition of the microbiota in humans to test particular hypotheses is extremely limited; such manipulations are usually restricted to the administration of antibiotics whose microbial targets constitute a broad range of organisms. Taken together, the many advantages of animal studies argue strongly for the development of representative models of the human microbiota in mice whose characterization can complement and inform research in human subjects (e.g., by generating/testing hypotheses, by helping to define effect sizes, by identifying surrogate and mechanism-based biomarkers, etc.).

Creating a defined system: model communities of sequenced human gut microbes

Colonization of germ-free mice with complex, undefined communities\cite{29}

Germ-free (GF) mice can be colonized with microbial communities of different origins and degrees of complexity at defined stages of host development. When such colonizations are performed using one or more known microbial species, the resulting animals are often described as ‘gnotobiotic’ (from the Greek roots gnostos for ‘known’ and bios for ‘life’), though technically the term applies to GF animals as well. The microbial communities used as inocula in gnotobiotic experiments can be harvested from any number of habitats, including various body sites of con-
ventionally-raised mice with defined genotypes and physiological phenotypes of interest. These communities can be introduced into GF recipients (possessing a desired genotype) to determine how much of a donor’s phenotype is transferrable to the resulting ‘conventionalized’ mice via their microbiota. If partial or complete phenotypic transfer occurs, follow-up studies can be performed in order to define the composition of the donor community, the mechanisms by which the donor community impacts host physiology, and how the recipient affects the transplanted microbiota/microbiome. These types of studies have typically been performed using gut contents [30], but in principle could be extended to communities harvested from any body habitat.

Procedures for transplanting human fecal microbial communities into GF mice to generate ‘humanized’ mouse models have been developed by members of our laboratory [31]. A remarkable portion of human fecal microbial community diversity can be transferred in this fashion even when a donor specimen has been frozen at -80°C for several years. Past studies suggest that all bacterial phyla, up to 90% of class-level and genus-level taxa, and 60-90% of species-level phylotypes in a given human donor’s sample can be identified in recipient gnotobiotic mice when the 16S rRNA gene is sequenced from fecal DNA. While some human gut taxa may be refractory to transplantation, these results seem to suggest that a large number of the most prevalent human gut microbes may be capable of colonizing Mus musculus. Once engrafted, the transplanted human microbial communities are remarkably stable, can be reliably transmitted across generations of animals and exhibit well defined and reproducible biogeographical features along the length of the mouse gut [31]. Efficient intergenerational transfer of transplanted human fecal microbiota also allows the microbiota and the host’s innate/adaptive immune system to co-evolve beginning at birth in ‘second generation’ mice.

‘Humanized’ gnotobiotic mice can be used for proof-of-mechanism studies that cannot be readily conducted in humans where potentially confounding variables, including variations in host genotype, diet, and antibiotic consumption are notoriously difficult to control. As such, they are a powerful model for conducting first-generation experiments whose goal is to determine if microbes influence a host phenotype or physiological parameter of interest.
Like any experimental model, humanized gnotobiotic mice are subject to limitations. When studying a biological system, it is obviously important to have as complete an understanding as possible of its components and parameters. Unfortunately, a researcher colonizing germ-free mice using a human fecal suspension is typically unable to discern the exact identities of all the microbes within the donor community, even using the most sophisticated metagenomic and 16S rRNA gene sequencing techniques available today. The level of taxonomic resolution that can be achieved when sequencing these donor communities is in fact the same as for the researcher studying the microbial composition of human feces, and the same is true for the sequencing of fecal communities from humanized mice. In some cases, 16S rRNA gene-based identification of taxa can achieve species-level resolution, but often only genus (or higher) levels of classification are possible when microbes hail from taxonomic clades where few representatives have been sequenced. This ‘fuzziness’ when assigning identities to microbes also negatively impacts our ability to accurately estimate the alpha, beta, and gamma diversity present in gnotobiotic recipients of human gut microbiota transplants. Additionally, genome information is often lacking for microbes transplanted from complete (uncultured) human communities. This lack of genomic data limits our ability to make predictions about the functional capacity of any given constituent of the system, as well as the ways in which it might interact with other microbes (and the host). While metagenomic sequencing of a donor community or the humanized mouse fecal community can provide some information about a community’s genetic potential, it is not yet a substitute for a completely defined metagenome whose genes can be attributed to specific taxa. Finally, the use of intact human communities to colonize GF mice currently does not allow for targeted manipulation of the donor community’s composition. This inability to intentionally manipulate the membership of communities before or after their transplantation to GF mice can make testing hypotheses about the operations of the microbiota and its individual components either incredibly challenging or impossible. Thus,
while the humanized mouse model provides several important advantages, there is a substantial need for more defined systems that are more amenable to mechanistic dissection.

**Assembling defined, representative model communities in vivo\(^{29}\)**

Thanks to previous and ongoing efforts by many to culture and sequence human gut microbes [32-35], an opportunity exists to create representative model human gut communities in gnotobiotic mice in which the identities of all community members and the sequences of all of their genomes/genes are completely defined. Members present in these model communities can be selected from both public and private culture collections based on various criteria relevant to the researcher’s experimental focus. Reasons for a microbe’s inclusion in a model community might include its consistent association with specific human physiological or pathophysiological states, its representation in a fecal microbiota that confers a phenotype to recipient germ-free mice when transferred *en masse*, its representation in healthy individuals, its evolutionary ancestry, its functional potential as predicted by *in silico* inspection of its genome, or any number of other criteria. These defined communities can be used to address a number of basic questions in the field of microbiota research, among them: (i) to what extent do established species who presumably sequester limiting space and/or resources exclude potential colonizers and influence community composition; (ii) to what extent is the dynamic behavior of the microbiota determined by inter- and intra-specific interactions [36]; (iii) what are the genetic predictors of niche overlap; (iv) how robust are assembled communities to various environmental perturbations; and (v) how do different gut microbes prioritize available resources in a highly competitive environment. The defined nature of model gut communities comprised exclusively of sequenced microbes allows these questions and others to be addressed at levels of resolution that are simply not possible in undefined systems.

It is important to emphasize here that the introduction of defined assemblages of microbes into gnotobiotic mice is by no means a new experimental approach. In the 1960’s, Dr. Russell W. Schaedler (then at Rockefeller University) was the first to colonize GF mice with panels of bacte-
ria isolated from healthy conventionally-raised mice [37]. Today, the 8-species ‘altered Schaedler flora’ (the result of several rounds of improvement to the original ‘Schaedler flora’ [38]) is internationally recognized as the standard microbiota for colonizing neonatal mice kept in the barrier rooms of all major animal suppliers. In this case, however, the model community’s main utility is as a means of standardization and ensuring that animals remain pathogen-free.

More recently, simple model communities comprised of 2-3 species have been studied in germ-free mice in order to explore some of the metabolic interactions that take place between different microbes (microbe-microbe interactions) and between particular microbes and the host (host-microbe interactions) [39-42]. The studies described in this thesis have utilized communities comprised of as many as 15 different sequenced human gut microbes in an effort to more faithfully reproduce many of the key taxonomic and functional features of the human gut microbiota. Though these communities are significantly more complex and representative of the human microbiota than past systems, further increases in the complexity of model microbiota introduced into mice and other gnotobiotic hosts are anticipated as we strive to improve our understanding of this dynamic ecosystem (discussed in Chapter 4). Methodologies for characterizing the structure and function of these defined models of the gut microbiota, many of which were developed or improved during the course of this thesis, are described below.

**Compiling a molecular tool kit: methods for characterizing model community structure and function**

**Profiling community structure and the development of COPRO-Seq**

While many parameters of a microbial community can be measured and characterized in order to better understand its properties, one of the most important is its ‘structure’. Species richness, species diversity and patterns of species abundance within a microbiota each constitute different elements of this structure. Measuring changes in gut microbial community structure over time provides a researcher with insights about how the different populations with the system are re-
sponding to changes in their environment, how they are responding to one another, and how they are responding to other elements of the larger ecosystem (e.g., changes in host physiology). One of the earliest aims of this dissertation was to develop a robust, rapid, and cost-effective method for evaluating the structure of a defined model microbiota comprised of sequenced human gut-derived bacterial species.

In principle, the structure of a defined microbiota can be assessed in a number of different ways. Because the genome sequence of each species in a defined community is known, primers specific to a particular organism can be designed for quantitative PCR (qPCR)-based assays of species abundance, yielding data for one member of the community at a time. However, this approach is not tractable if the goal is to longitudinally study many microbes in many animals subjected to multiple experimental treatments. An alternative approach is based on sequencing of the 16S rRNA gene. However, this type of approach is hampered by known PCR primer biases [43] and the need for fairly long DNA sequences (which might significantly inflate the cost of evaluating community structure across large numbers of samples).

At the time this dissertation work began, ‘massively-parallel’ or ‘next-generation’ sequencing was a relatively new technology, and the first commercially-available Illumina sequencers were entering the market. The first-generation Illumina platform was capable of generating many millions of short (35 nt) reads – a number that was many orders of magnitude greater than previous technologies. This significant increase ultimately paved the way for a new approach to defining the structure of model communities in gnotobiotic mice, a method I dubbed COmmunity PROfil- ing by Sequencing (‘COPRO-Seq’). COPRO-Seq takes advantage of the fully-defined nature of the model microbiome to precisely measure the relative abundance of every community member in a given sample simultaneously. After generating large numbers of short reads from community DNA using the Illumina platform, each read is quantitatively mapped back to its genome of origin using a short read aligner. In some cases, because of genetic overlap between microbial species, reads cannot be assigned unambiguously to a single genome. These reads of unknown origin are excluded from the analysis. To compensate for their removal, the ‘informative genome size’ (IGS)
of each genome is calculated and used to normalize the sum of the read counts attributed to each species in the community. Each genome’s IGS value is calculated by multiplying two values: (i) the fraction of all k-mers (e.g., 35-mers) that can be derived from a genome’s sequence and occur only once when considering all k-mers within the defined community’s microbiome (i.e., the fraction of a genome that does not overlap with any other genome in the defined community), and (ii) the genome’s total length. Dividing the normalized read counts for each species present in the community by the total normalized read counts in a sample (e.g., feces) results in a comprehensive profile of community structure that provides the relative abundance of every species detected in that sample. All calculations are performed in an automated fashion by a workflow implemented in Perl to ensure reproducibility and to avoid human error. High-throughput sample processing is achieved by generating barcoded Illumina libraries for each sample and pooling multiple barcoded libraries on each lane of an Illumina flow cell. By taking advantage of the high yields of the Illumina platform and the known genome sequences of a defined community microbiome, COPRO-Seq allows a researcher to quantitatively assess the structure of a defined microbiota in many hundreds or even thousands of samples with exceptional sensitivity (species comprising as little as 0.001% of a community are typically detectable) in rapid fashion.

Additionally, because COPRO-Seq can quantify the abundance of a taxon based on a relatively small number of species-specific reads (compared to the total number of sequences generated from each sample characterized), it is well-equipped to characterize defined communities much larger than those studied in this dissertation. Provided all strains in the model community have a sequenced genome available, it should be possible in the future to enumerate the proportional abundance of hundreds or even thousands of sequenced community members simultaneously. Further details regarding the COPRO-Seq methodology are provided in the study described in Chapter 2.

**Profiling community function using improved methods**[^29]

Understanding how different gut communities modulate their activities in response to changes in
diet, host physiological status, or the introduction of new microbial species is a key step in understanding the operations of the gut microbiota. Microbial RNA-Seq can be used to quantify changes in community gene expression at high resolution and with substantial dynamic range [44]. In the case of model communities of sequenced gut microbes, where all species and genes are known, this high-resolution data can be used to verify gene structure/operons, generate in silico reconstructions of expressed metabolic pathways for each member in the community, and make predictions concerning the metabolic niches of taxa of interest. Functional predictions about community operations can also be informed by RNA-Seq analyses of individual members grown in monoculture under highly defined conditions (e.g., minimal medium supplemented with systematically varied carbon sources; see Chapter 3 for an example of this approach involving a prominent member of the gut Bacteroides). In addition, gnotobiotic mice harboring defined collections of sequenced organisms provide an opportunity to further develop and test methods for extracting and characterizing, by high-performance liquid chromatography-mass spectrometry (LC-MS), the proteins expressed by their model microbiota. Given accurate gene predictions based on the sequenced genomes of each community member, theoretical peptidomes/proteomes can be calculated in silico, and peptides detected in community-extracted protein can be mapped back to their proteins of origin. In collaboration with researchers at Oak Ridge National Laboratories, we have picked up where early work in two-member model microbiota left off [40] and significantly improved our methods for characterizing the metaproteomes of model human gut communities comprised of up to 12 species (see Chapter 3). Finally, targeted and untargeted quantitative mass spectrometry (MS)-based measurement of metabolites in the gut (as well as other biocompartments) can provide an even more direct assessment of how community operations change as the result of experimental perturbations. These MS-based metabolomics methods allow the researcher to characterize the effects of systematically manipulating the species composition of a model microbiota or perturb-
ing the activities of key community members [42]. Collectively, these improved systems-level approaches in metatranscriptomics, metaproteomics, and metabolomics allow for unprecedented time-series assessments of expressed functions in model communities.

**Genetic manipulation of the community metagenome**

Community genetics represents another powerful means by which the researcher may dissect the operations of microbial communities. The addition or removal of particular species prior to gavaging the model microbiota into GF mice constitutes the simplest genetic perturbation that can be imposed on the system to identify taxa that confer a benefit or detriment to other community members or the host. Another method for manipulating the gene content of the metagenome is insertion-sequencing (INSeq), which combines genome-wide transposon mutagenesis with massively parallel sequencing [45,46]. In this approach, complex populations of tens of thousands of transposon mutants of a given sequenced community member are generated and introduced into wild-type or genetically manipulated GF mice in the presence or absence of other (sequenced) microbes. The representation of each mutant in both the input community and each experimental sample (output community) is determined by targeted, quantitative sequencing of transposon-adjacent chromosomal DNA. Any differences in mutant representation in input versus output communities provide a quantitative measure of the degree to which loss of a gene’s function imposes a fitness cost on the species of interest within a given experimental framework (i.e., set of selective pressures). Additionally, the inclusion of a species-specific barcode within the transposon construct used to mutagenize individual community members allows multiple mutagenized species to be studied simultaneously, permitting assessment of the importance of fitness determinants within and between multiple species. Thus, by including or excluding species of interest when constructing model communities and by mutagenizing individual members of those communities using a randomly inserting transposon, a researcher studying model communities can directly determine
which genes and which species are most important for specific functional features of a microbiota. This approach proved particularly informative in two studies included here (see Chapter 3, Appendix A).

**Studying the system: characterization of communities and specific taxa in contexts relevant to human health**

If one accepts the premise that the microbiota plays an important role in human health, it is reasonable to think that methods for preserving a ‘healthy’ microbiota configuration or driving the microbiota away from ‘pathological’ states could serve as the basis for new therapeutic modalities. Developing safe, efficacious microbiome-directed therapies requires a much better understanding of how human gut communities assemble, operate, and respond to environmental changes over time. As discussed, characterization of model communities of human gut microbes using molecular tools that leverage the completely defined nature of these systems represents an approach to achieving this goal of greater understanding. Over the course of this dissertation, I have used these systems and tools to study the microbiota in two contexts with significant relevance to human health.

**The impact of probiotic interventions on the microbiota**

Though the term ‘probiotic’ is one whose definition has evolved substantially over time [47], probiotics are generally described as live microorganisms that when ingested confer a health benefit to their host. In 1907, Élie Metchnikoff (often referred to as the ‘father of probiotics’) proposed that ingesting lactic acid bacteria could replace or diminish ‘putrefactive’ bacteria in the human intestine, normalizing gut function and prolonging human life [48]. Since Metchnikoff’s time, the market for probiotic foods and supplements has become a booming business as more and more individuals turn to these products in hopes of improving intestinal health and general well being. Yet, while a small number of probiotic strains of bacteria have been carefully studied [49], in most cases we have a poor understanding of the mechanisms by which these bacteria influence
human health and how their introduction into the gut microbiota influences its operations. Chapter 2 describes a staged ecological ‘invasion’ of a consortium of five strains reputed to have probiotic properties into the guts of monozygotic twin pairs as well as gnotobiotic mice colonized with a defined model human gut microbiota. This study of the model community in parallel with human subjects allowed us to characterize the ways these probiotic microbes influence gut communities at the level of individual genes and taxa, and resulted in the identification of numerous functional responses that are shared between humans and mice. Additionally, the results show that model gut communities can play an important role in identifying biomarkers and mediators of the direct and indirect effects of probiotics on gut microbiota and their hosts.

The impact of diet change over time

Many studies in recent years have suggested that the well-known idiom ‘you are what you eat’ is apt not only in terms of general human health, but also applies to the composition of the human microbiota [50-52]. While these studies have contributed significantly to our appreciation of diet as a paramount environmental factor in shaping the microbiota’s structure and function, further work is needed to determine how the many interactions between individual microbes and food lead to diet-induced effects at the community level, how microbes prioritize utilization of different components of our diets, and how microbes carve out metabolic niches that allow them to persist over time in individuals eating diverse diets. In Chapter 3, we use the model community experimental framework to evaluate how diet change influences the structure of the microbiota over time, the dynamics of community’s functional responses to diet change, and metabolic strategies prominent taxa rely on in a changing nutrient landscape. We then go on to further characterize the carbohydrate utilization of a well-adapted xylanolytic *Bacteroides* species, the genetic factors that allow it to utilize diverse simple and complex saccharides, and which of those factors are most important for *in vivo* fitness in a host harboring a model human gut microbiota and consuming different diets.

Footnote

The text within some sections of this Introduction includes both direct and revised excerpts from
a previously published review (see reference [29]). These sections are designated by a reference number (in superscript) at the end of their titles.
References


paradigm shift in the human gut microbiome study. Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases 18: 1185-1193.


Chapter 2

The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins
Chapter 2

The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins

Nathan P. McNulty1*, Tanya Yatsunenko1*, Ansel Hsiao1*, Jeremiah J. Faith1, Brian D. Muegge1, Andrew L. Goodman1, Bernard Henrissat2, Raish Oozeer3, Stéphanie Cools-Portier4, Guillaume Gobert3, Christian Chervaux3, Dan Knights4, Catherine A. Lozupone5, Rob Knight5,6, Alexis E. Duncan7,8, James R. Bain9,10, Michael J. Muehlbauer9, Christopher B. Newgard9,10,11, Andrew C. Heath7, and Jeffrey I. Gordon1

1Center for Genome Sciences and Systems Biology and 7Department of Psychiatry, Washington University School of Medicine, St. Louis, MO 63108, 2Architecture et Fonction des Macromolécules Biologiques, CNRS, Marseille, France, 3Danone Research, 91 767 Palaiseau Cedex, France, 4Department of Computer Science and 5Department of Chemistry and Biochemistry and 6Howard Hughes Medical Institute, Univ. of Colorado, Boulder, CO 80309, 8George Warren Brown School of Social Work, Washington University, St. Louis 63130, 9Sarah W. Stedman Nutrition and Metabolism Center and 10Department of Medicine and 11Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, NC 27710

*Contributed equally

4Current address: Section of Microbial Pathogenesis and Microbial Diversity Institute, Yale School of Medicine, New Haven, CT 06536

This chapter corresponds to the complete and accepted version of a manuscript published as Sci Transl Med. 2011 Oct 26;3(106):106ra106. doi: 10.1126/scitranslmed.3002701.
Abstract

Understanding how the human gut microbiota and host are affected by probiotic bacterial strains requires carefully controlled studies in humans and in mouse models of the gut ecosystem where potentially confounding variables that are difficult to control in humans can be constrained. Therefore, we characterized the fecal microbiomes and metatranscriptomes of adult female monozygotic twin pairs through repeated sampling 4 weeks before, 7 weeks during, and 4 weeks after consumption of a commercially available fermented milk product (FMP) containing a consortium of *Bifidobacterium animalis* subsp. *lactis*, two strains of *Lactobacillus delbrueckii* subsp. *bulgaricus*, *Lactococcus lactis* subsp. *cremoris*, and *Streptococcus thermophilus*. In addition, gnotobiotic mice harboring a 15-species model human gut microbiota whose genomes contain 58,399 known or predicted protein-coding genes were studied before and after gavage with all five sequenced FMP strains. No significant changes in bacterial species composition or in the proportional representation of genes encoding known enzymes were observed in the feces of humans consuming the FMP. Only minimal changes in microbiota configuration were noted in mice after single or repeated gavage with the FMP consortium. However, RNA-Seq analysis of fecal samples and follow-up mass spectrometry of urinary metabolites disclosed that introducing the FMP strains into mice results in significant changes in expression of microbiome-encoded enzymes involved in numerous metabolic pathways, most prominently those related to carbohydrate metabolism. *B. animalis* subsp. *lactis*, the dominant persistent member of the FMP consortium in gnotobiotic mice, up-regulates a locus *in vivo* that is involved in the catabolism of xylooligosaccharides, a class of glycans widely distributed in fruits, vegetables, and other foods, underscoring the importance of these sugars to this bacterial species. The human fecal metatranscriptome exhibited significant changes, confined to the period of FMP consumption, that mirror changes in gnotobiotic mice, including those related to plant polysaccharide metabolism. These experiments illustrate a translational research pipeline for characterizing the effects of FMPs on the human gut microbiome.
Introduction

Our physiology and physiological differences are not only manifestations of our human genes and epigenomes but also a reflection of the genes and genetic variations that exist in our resident microbial communities (microbiomes). Our microbiomes contain at least 100 times more genes than our human genomes (1). Marked increases in DNA sequencing capacity have led to an explosive increase in the number and types of culture-independent metagenomic studies of intra- and inter-personal variations in human microbial ecology—as a function of our human life cycle, cultural traditions, and health status (2–7). Long-term goals of this quest to understand the genomic and metabolic underpinnings of our mutually beneficial relationships with microbes include using our symbionts as a new class of biosensors and biomarkers of wellness, and devising safe and effective ways to deliberately manipulate the structure and functions of our microbiome to optimize our health, as well as to treat various diseases.

A necessary starting point for assessing how the structure and functions of the human microbiome are related to our biology is to characterize the normal variations that occur in these communities, their gene pools, and their gene expression profiles both within and between individuals. This requires carefully designed studies where potentially confounding variables such as host genotype, diet, and various environmental exposures can be controlled and systematically manipulated. Monozygotic (MZ) twins represent one way to constrain some of these variables, given that they have more similar genotypes and have experienced more similar dietary and other early environmental exposures than any other combination of individuals. A complementary approach is to use germ-free mice colonized at various points in their life with defined collections of microbes, with sequenced genomes, that represent major phylogenetic lineages encountered in the body habitats of human populations of interest. Gnotobiotic mice harboring “synthetic” model human microbiomes, where all component organisms and microbial genes are known, can be reared under conditions where a number of the variables that confound human studies are extremely well controlled. Insights gleaned from these gnotobiotic animals can be applied back to humans (8).
Common intended or unintentional disturbances to our microbiomes include changes in our diets, consumption of antibiotics, and ingestion of live microbial strains posited to improve health. The latter include commercially available probiotics that are incorporated into fermented milk products (FMPs). With increasing regulatory pressure to validate the composition and health claims of probiotics and “functional” foods, it is particularly important to develop informative translational medicine pipelines so that proof-of-concept clinical trials can be performed with validated biomarkers for quantitative phenotyping of subjects and of their responses. The present study demonstrates one type of approach. It uses adult MZ twin pairs and metagenomic methods to first define temporal fluctuations in the organismal and gene content and gene expression profiles of their fecal microbial communities as a function of administration of a widely used commercial FMP. It then takes the five sequenced strains present in the FMP and introduces them as a consortium, at a dose analogous to that experienced by humans, into gnotobiotic mice containing a model human microbiome composed of 15 sequenced human gut symbionts. Quantitative analyses of temporal changes in the proportional representation of microbial species and genes, and of microbiome gene expression and metabolism before and after an ecological “invasion” with the five-member FMP microbial consortium, have provided insights into the ways that FMP strains and the indigenous model gut community respond to one another. The transcriptional responses were used as biomarkers to interrogate metatranscriptome data sets obtained from the MZ twins’ fecal specimens.
**Results**

**Human studies**

**Study design and assessment of intra- and interpersonal variations in the fecal microbiota of MZ twin pairs over a 4-month period.** Details concerning the seven adult female MZ twin pairs recruited for this study are provided in table S1. All had been vaginally delivered; none consumed antibiotics in the 6 months before and during their participation in the present study; none had a history of gastrointestinal diseases (including irritable bowel syndrome) or any other acute or chronic medical conditions; none were consuming dietary supplements or probiotics at the time of enrollment; and none had a history of gluten sensitivity or other food allergies, nor were any vegans or lacto-vegetarians.

Fresh lots of an FMP were shipped every 2 weeks to the participants’ homes from the same pilot production facility; strain-specific quantitative polymerase chain reaction (qPCR)–based assays indicated that, at the time of shipment, each gram of the FMP contained on average $3.2 \times 10^7$ genome equivalents (GEs) of *Bifidobacterium animalis* subsp. *lactis* (strain CNCM I-2494) and $6.3 \times 10^7$ GEs of *Lactobacillus delbrueckii* subsp. *bulgaricus* (CNCM I-1632, CNCM I-1519). These results were consistent with previous measurements of the number of colony-forming units (CFU) in a typical cup of the FMP [4.9 × 10^7 CFU/g (*B. animalis* subsp. *lactis*), 8.4 × 10^7 CFU/g (*L. delbrueckii* subsp. *bulgaricus*)].

Three fecal samples were obtained over the course of a 4-week period before initiation of FMP consumption (“pretreatment phase”; see Fig. 1A). Each co-twin then consumed two servings of the FMP per day for 7 weeks (breakfast and dinner). Four fecal samples were collected at defined intervals during this treatment period, whereas two additional samples were collected during the 4 weeks after cessation of FMP consumption (“posttreatment phase”; Fig. 1A). Participants kept a daily log of their FMP consumption and stool parameters including frequency. Statistical analyses of this log indicated that, in this population, FMP consumption was associated with significantly softer stools but no significant changes in stool frequency (see the Supplemen-
tary Material). However, based on existing regulatory criteria, our study of this small cohort was insufficiently powered to draw clinical conclusions about these stool parameters. Moreover, the MZ twin population recruited was composed entirely of healthy individuals, so these data cannot be used to make statements about the impact of FMP consumption on stool softness in unhealthy patient populations.

All fecal samples collected during the three phases of this study were frozen at −20°C within 30 min of their passage and maintained at this temperature during overnight shipment to a biospecimen repository where they were subsequently stored at −80°C before metagenomic analyses. To assess intra- and interpersonal variations in microbial community structure, we performed multiplex 454 FLX pyrosequencing of amplicons generated from variable region 2 of bacterial 16S rRNA (ribosomal RNA) genes present in fecal DNA. A total of 431,700 sequencing reads were obtained from 126 fecal samples (3426 ± 2665 sequences per sample; table S2A). Noise due to PCR and pyrosequencing artifacts was removed from this data set using software incorporated into the QIIME suite of 16S rRNA analysis tools (9). Denoised reads were binned into species-level phylogenetic types (phylotypes), with a species defined as isolates that share ≥97% identity in their 16S rRNA gene sequence. To ensure even coverage across samples, we subsampled each of the 126 data sets to 1640 reads per fecal microbiota. A phylogenetic tree was built from one representative sequence from each phylotype using FastTree’s approximately maximum-likelihood implementation (10) and communities were compared using unweighted UniFrac (11): The UniFrac metric measures community similarity based on the degree to which their members share branch length on a reference phylogenetic tree of Bacteria.

To quantify temporal variation in community composition within and between MZ twins, we generated a matrix of unweighted UniFrac distances for all pairwise comparisons of all 126 fecal samples obtained from the twins in our study. This matrix allowed us to compare any two fecal communities separated by all possible time intervals between sampling for each individual in each of the seven twin pairs (Fig. 2A). The results indicated that no matter how far apart in time sample collection occurred (1 week to 4 months), the phylogenetic distance between communi-
ties from the same individual was less than the distance between communities between co-twins or unrelated individuals. UniFrac distances between samples harvested from a given individual increased modestly as a function of time during the 4-month period, although the changes were not statistically significant (Fig. 2A).

Each sample contained 163 ± 3 (mean ± SEM) observed species-level phylotypes. Four of the total 1673 phylotypes identified in our data set were found in all 126 samples; all belonged to the family Lachnospiraceae (order Clostridiales; phylum Firmicutes) and represented 2.5 ± 0.04% of the 16S rRNA sequences in each sample. We determined that 26.4 ± 0.4% of species-level phylotypes observed in a given sample were consistently represented in all nine samples from that individual (Fig. 2B): The family-level taxa to which these species belong consist principally of Lachnospiraceae, Ruminococcaceae, and Veillonellaceae (phylum Firmicutes); the Bacteroidaceae and Rikenellaceae (phylum Bacteroidetes); and Coriobacteriaceae (phylum Actinobacteria). Moreover, 13.7 ± 0.2% of the observed phylotypes were represented in all samples from both co-twins (Fig. 2B).

**Impact of FMP consumption on fecal bacterial community composition.** A qPCR assay disclosed that 1 week after initiation of FMP consumption, the level of representation of *B. animalis* subsp. *lactis* (CNCM I-2494) was $10^7$ cell equivalents per gram of feces; this level was sustained in all 14 individuals throughout the ensuing 7 weeks of FMP consumption (that is, there were no statistically significant differences between the 1-, 2-, 4-, and 7-week time points as determined by Friedman test with post hoc correction). The Spearman correlation test revealed no significant effect of human family membership on the levels of *B. animalis* subsp. *lactis during FMP consumption. Levels fell to below the limits of detection of the assay in all but four participants within 2 weeks of cessation of FMP consumption (fig. S1); two of these individuals represented a twin pair, whereas the other two individuals were unrelated to each other.

Co-occurrence analysis (see the Supplementary Material) indicated that with the FMP dosing schedule used, no species-level phylotypes present in the pretreatment microbiota exhibited a
statistically significant change in their proportional representation in feces in any individual, during or after the period of FMP consumption. In addition, no species-level taxa that were undetected in the pretreatment period appeared and persisted during and/or after treatment in any individual [paired t test, analysis of variance (ANOVA)]. It is possible that with even deeper sampling, differences might be revealed in feces or may exist in more proximal regions of the gut. Further details of this co-occurrence analysis, including the results of tests at the genus and family level, plus deeper sequencing of a subset of twin pairs are provided in the Supplementary Material.

**Effects of FMP consumption on the functional gene repertoire of the fecal microbiome.** To determine the effects of FMP consumption on the representation of gene functions in the microbiome, we performed shotgun sequencing on 48 of the fecal DNA preparations generated from four of the MZ twin pairs (n = 6 samples per individual; two fecal samples obtained before, two during, and two after cessation of FMP consumption; see Fig. 1A). Two of these twin pairs lived together, whereas two pairs lived 3 and 932 miles apart. A 634-Mb data set was generated (60,863 ± 28,775 sequences per sample; average length, 238 nucleotides; table S2B). A BLASTX search against version 54 of the Kyoto Encyclopedia of Genes and Genomes (KEGG) GENES database (12–14) yielded a total of 2205 ± 26 (mean ± SEM) KEGG Orthology identifiers (KOs) per microbiome sample: 64% of the KOs in a given sample (1417 ± 46) were consistently represented in all 6 samples from that individual; 55% were consistently represented in all samples from both co-twins; 892 KOs (41% of the total KOs in a given sample) were identified in all 48 samples forming a core set of shared fecal microbiome-associated functions. Figure 2C provides a visual representation of this conserved set of 892 KOs: 38% of the 892 belong to six KEGG categories: “membrane transport,” “carbohydrate metabolism,” “DNA replication and repair,” “amino acid metabolism,” “translation,” and “metabolism of cofactors and vitamins.”

The proportional representation of KEGG pathways and their component KOs was subsequently calculated for each of the 48 microbiomes. The microbiomes were then subjected to all possible pairwise comparisons on the basis of these two classification schemes. The results, quantified using the Hellinger distance metric, disclosed that over time, unlike the UniFrac-based
16S rRNA comparisons of community bacterial species composition, there was no significant difference in the degree of similarity of microbiome functional profiles for a given co-twin compared to the degree of similarity that existed between co-twins (that is, intrapersonal variation was not significantly different from interpersonal variation between co-twins). However, as with the UniFrac results, individual and twin pair microbiomes were significantly more similar to one another than those from unrelated individuals (Fig. 2D). No KEGG pathways or KOs exhibited a statistically significant change in their relative abundance in response to FMP consumption in any of the subjects at any of the time points (Student’s paired t test and two-way ANOVA with Bonferroni post hoc testing).

At this point in our analysis, the human studies indicated that exposure of a healthy individual’s resident gut microbiota to the FMP strains did not produce a detectable perturbation in fecal bacterial species composition, nor did it have a broad effect on the functional profile of fecal microbiome genes. To help guide further analysis of the human data sets, we turned to a simplified in vivo model of the human gut microbiota. We based our selection of model community members on several criteria. All members of this model community, or their close relatives, would be represented in the fecal microbiota of the MZ twins and other sampled human populations. They would encompass the three major bacterial phyla present in this host habitat (Firmicutes, Bacteroidetes, and Actinobacteria) and would have deep draft or finished genome sequences available. Gnotobiotic mice harboring such a model human microbiome would be used to characterize the impact of FMP strain introduction on the community’s species and microbial gene abundances, as well as the microbiome’s transcriptional profile, and to ascertain the impact of the model community on the abundance and gene expression profiles of the FMP strains whose genome sequences were also known. The knowledge gleaned would then be used to help guide further analysis of the human fecal microbiome data sets, including microbial RNA-Seq data sets generated from a subset of the human fecal samples.
Studies in gnotobiotic mice

Study design. A community of 15 sequenced human gut-derived microbes containing a total of 58,399 known or predicted protein-coding genes was constructed (Fig. 1B and table S3). Figure S2 uses assigned KOs to provide evidence of the functional similarity of this model human microbiome to a collection of 127 genomes generated from cultured members of the human gut microbiota, a deeply sampled set of fecal microbiomes obtained from 124 unrelated Europeans (1), and deeply sampled microbiomes from a pair of obese MZ co-twins (15).

Figure 1B presents the study protocol. Two groups of adult 6- to 8-week-old germ-free C57BL/6J male mice were colonized with a single gavage of the 15-member community (6 × 10^6 CFU per member, total of 9 × 10^7 CFU). Each group (n = 5 animals) was maintained on a low-fat, plant polysaccharide–rich diet. Fourteen and 15 days after gavage with the 15-member community, both groups of mice were inoculated with a mixture of the five FMP strains. One group received a second pair of gavages of the five strains 7 and 8 days later, and a third pair 21 and 22 days after the first inoculation of the FMP consortium (multiple-treatment group). Each gavage consisted of a community composed of 2 × 10^7 CFU: 25% (5 × 10^6 CFU) *Streptococcus thermophilus*, 25% *B. animalis* subsp. *lactis*, and 25% *Lactococcus lactis* subsp. *cremoris*, with the remaining 25% split between the two *L. delbrueckii* subsp. *bulgaricus* strains (12.5% each; 2.5 × 10^6 CFU per strain). Dosing was based on the following considerations: (i) a daily dose of two cups of the FMP contains ~10^{10} CFU of *B. animalis* subsp. *lactis*; (ii) assuming ~10^{14} bacteria in the human gut, the ratio of the number of input *B. animalis* subsp. *lactis* CFU to the human gut symbiont population is about 10^{-4}; (iii) to maintain this ratio of 10^{-4} in mice, and assuming 10^{11} to 10^{12} organisms in the mouse gut, we administered a total of 10^7 *B. animalis* subsp. *lactis* CFU per gavage period (one period equals two gavages within 24 hours); (iv) the difference in CFU between the least and the most abundant microbial species in the FMP product remains less than or equal to twofold during manufacture and storage; therefore, each species in the gavage was represented at equivalent levels.
FMPs contain defined collections of microbes, a dairy matrix, and products of microbial metabolism of this matrix generated during manufacturing and storage. In principle, the dairy matrix and microbial metabolic products could be responsible for some of the effects observed when humans consume an FMP. Nonetheless, we chose to administer the strain consortium to mice directly by gavage, rather than by giving them the corresponding commercial FMP. This allowed us to more precisely control dosing. It also allayed concerns about possible unintended colonization of the gnotobiotic mice with microbial species (other than strains deliberately introduced during manufacturing) that could be introduced during handling of an FMP before gavage.

The repertoire of carbohydrate-active enzymes in members of the FMP consortium and model human gut microbial community. The genomes of the five FMP strains in this study were sequenced, either completely (B. animalis subsp. lactis) or at a deep draft level (other four strains) for subsequent analyses of their representation in the model community after gavage of gnotobiotic mice, and so we could define their in vivo patterns of gene expression (table S3).

Analysis of the 48 carbohydrate-active enzyme (CAZyme) families (16) identified in the five FMP strains and the 126 CAZyme families identified in the 15-member model human microbiota disclosed that 23 of the 24 CAZyme glycoside hydrolase (GH) families, 11 of the 12 glycosyltransferase (GT) families, 4 of the 4 carbohydrate esterase (CE) families, and 4 of the 8 carbohydrate-binding modules (CBMs) represented in the former were also represented in the latter. The FMP consortium contains only six CAZyme families that were not represented in the model human gut community. Three of these are associated with L. lactis subsp. cremoris: of these, two are predicted to play roles in the binding and metabolism of chitin (fig. S3 and table S4). The other three are from B. animalis subsp. lactis: BALAC2494_01193 encodes a GT39 family mannose transferase involved in O-glycosylation of proteins; BALAC2494_01288 specifies a predicted β-mannanase carrying a C-terminal CBM10 module predicted to bind cellulose; BALAC2494_01971 gives rise to a protein with a CBM23 module predicted to bind mannan.
Minimal changes in the species representation of the 15-member model human gut microbiota after introduction of the FMP strain consortium. We used COmmunity PROfil-ing by Sequencing (COPRO-Seq), a generally applicable method based on highly parallel DNA sequencing (17), to quantify the proportional representation of each component of the 15-member microbiota and of the FMP consortium in our gnotobiotic mice. Sequencing reads generated from fecal DNA samples collected before, during, and after introduction of the FMP strains were analyzed as described in fig. S4A. Briefly, “informative” tags (that is, reads that can be mapped uniquely to a single genome) were first identified. Informative tags were then summed by species to generate digital “counts” of abundance. In cases where a read could not be assigned with certainty during COPRO-Seq analysis, it was ignored. To account for this fact, we normalized species-specific counts using their “informative genome size” (defined as the percentage of all possible k-mers a genome can produce that are unique multiplied by the total genome length). Multiplex sequencing using the Illumina GA-IIx instrument yielded sufficient numbers of reads per sample so that an organism comprising ≥0.003% of the community could be detected: For a mouse colonized at $10^{11}$ to $10^{12}$ CFU/ml cecal contents or feces, this represents $10^6$ CFU/ml.

COPRO-Seq produced several notable findings. First, community assembly before introduction of the FMP strains occurred in a highly reproducible manner, both within and between the two groups of animals (fig. S5A and table S5A). This reproducibility ensured that animals in both treatment groups harbored communities with structures comparable to one another at the time of administration of the five-member FMP strain consortium. Second, within 1 week of introducing the FMP strains, either in a single treatment or in multiple treatments, *B. animalis* subsp. *lactis* and *L. lactis* were detectable in the fecal microbiota (fig. S5B and table S5A). These two species persisted in the gut throughout the study. Importantly, *B. animalis* subsp. *lactis* was the most prominently represented member of the FMP consortium in the model human gut microbiota, exhibiting a progressive increase in its representation during the 28 days after initial introduction, and reaching comparable levels in both the single- and the multiple-treatment groups (up to 1.1%; see fig. S5B). In contrast, *S. thermophilus* and the two strains of *L. delbrueckii* subsp. *bulgari-
*cus* were undetectable or intermittently just over the limit of detection in the single- or the multiple-treatment groups. Third, as with the MZ twin pairs, introduction of the consortium led to minimal rearrangements in overall community structure, whether or not the consortium was administered twice in a 2-day period or on two subsequent occasions (see table S5B for the results of Mann-Whitney tests of significance for each species at each time point surveyed relative to the time point just before initial introduction). *Collinsella aerofaciens*, the lone Actinobacteria in the 15-member community, showed a significant reduction in its abundance in both treatment groups immediately after FMP strain introduction (fig. S5C) that persisted through later time points, raising the possibility of a competitive relationship between this organism and *B. animalis* subsp. *lactis*, the only Actinobacteria in the FMP strain consortium.

The *Bacteroides thetaiotaomicron* component of the 15-member human gut microbiota was composed of a library of 34,544 randomly inserted transposon mutant strains covering 3435 of the organism’s 4779 genes (72%). As noted in the Supplementary Material and table S6, by comparing the representation of mutants in fecal samples before and after introducing the FMP strains, we were able to determine that their presence did not affect the profile of *B. thetaiotaomicron*’s genetic determinants of fitness in the distal gut.

**Microbial RNA-Seq analysis of the response of *B. animalis* subsp. *lactis* to the gut environment and members of the 15-member community to the FMP strain consortium.** Moving beyond COPRO-Seq–based structural analysis, we performed microbial RNA-Seq analysis to determine the functional impact of exposing the established model human community to the FMP strains and to ascertain which FMP consortium genes are most highly expressed in the intestines of these animals. *B. animalis* subsp. *lactis* attained sufficient abundance in gnotobiotic mice to allow profiling of its transcriptome at late time points (days 35, 36, and 42). When its in vivo patterns of gene expression were compared with those documented during mid-log and stationary phase in MRS medium and in the commercial FMP (see the Supplementary Material and table S7), we noted that the *BALAC2494_00604-BALAC2494_00614* locus, encoding enzymes involved in the catabolism of xylooligosaccharides (18), was strongly up-regulated in vivo (average across the
locus; 27-fold at the day 42 time point compared to mid-log phase in MRS monoculture; 128-fold compared to the FMP; table S7). Xylose is the main building block of dietary hemicelluloses. Addition of this pentose sugar is also one of the first steps in O-glycosylation of host mucins. These results support previous observations suggesting that xylooligosaccharides may serve as potent “bifidogenic factors,” whose consumption increases the densities of Bifidobacteria in the gut (19, 20).

Ordination of samples and B. animalis subsp. lactis CAZyme gene expression patterns by correspondence analysis identified additional CAZymes that correlate strongly with the in vivo state (Fig. 3), including members of families expected to play roles in the degradation of dietary plant polysaccharides [GH43 (xylan β-xylosidases), GH77 (4-α-glucanotransferases)]. The analysis revealed sets of B. animalis subsp. lactis CAZymes that corresponded well with each growth condition (that is, MRS medium, commercial dairy matrix, and mice). Within each growth condition, the expressed groups of CAZymes often had related functions (Fig. 3).

We next examined the impact of the FMP strain consortium on expression of genes in the 15-member community. In a “top-down” analysis, genes were binned by function and the community’s metatranscriptome was evaluated in aggregate, ignoring the species from which each transcript arose. A complementary “bottom-up” analysis allowed us to determine how each species in the community responded to the introduction of the FMP consortium.

Top-down analysis of the impact of the FMP strains on the community metatranscriptome revealed significant increases in expression of genes falling within the KEGG categories carbohydrate metabolism and “nucleotide metabolism,” whereas decreases were observed in amino acid metabolism and “lipid metabolism” (Fig. 4A and table S8). Peak responses in both treatment groups occurred 3 weeks after the first gavage of the FMP strains, corresponding to the time of highest representation of B. animalis subsp. lactis in the community.

The genes that exhibited the highest fold change in expression were heavily skewed toward the KEGG categories carbohydrate metabolism and membrane transport. The latter includes a
number of adenosine triphosphate–binding cassette (ABC)– and PTS (phosphotransferase system)–type carbohydrate transporters (table S9). When these KEGG category–level responses were subsequently broken down into KEGG pathways (Fig. 4B), it was apparent that the most significant responses in the carbohydrate metabolism category involved increases in “starch and sucrose metabolism,” “fructose and mannose metabolism,” and “pentose and glucuronate interconversions.”

Transcript data were subsequently binned by enzyme commission (EC) number. The levels of mRNAs encoding these ECs at each time point were compared using ShotgunFunctionalizeR, an R-based statistical and visualization tool originally designed to identify genes significantly enriched or depleted in environmental microbiomes (21, 22). Using this approach, we were able to determine that the starch and sucrose metabolism pathway response to the FMP strains was driven by significant up-regulation of genes encoding three enzymes involved in metabolism of dietary plant polysaccharides: (i) EC 3.2.1.65 (levanase; Fig. 4C), which cleaves 2,6-β-D-fructofuranosidic linkages in 2,6-β-D-fructans (levans); (ii) EC 3.1.1.11 (pectinesterase), which de-esterifies pectin to pectate and methanol; and (iii) EC 2.4.1.20 (cellobiose phosphorylase), which uses cellobiose formed from partial hydrolysis of cellulose as its substrate to generate α-D-glucose-1-phosphate and D-glucose. The genes encoding these ECs, which catalyze early steps in three entry points of the starch and sucrose metabolism KEGG pathway, underwent significant increases in their expression within 24 hours after introduction of the FMP consortium (Fig. 5A). The levels of expression of these genes either increased further (levanase) or were sustained (the other two ECs) in both the single- and the multiple-treatment groups through the remaining 4 weeks of the experiment (Fig. 5A). The levanase response showed remarkable species specificity: This gene is represented in 8 members of the 15-member community, yet the community’s transcriptional response is driven almost exclusively by the levanase in Bacteroides vulgatus (BVU_1663; Fig. 4C). In contrast, the pectinesterase response was distributed across six members of the community (Bacteroides caccae, Bacteroides ovatus, B. thetaiotaomicron, B. vulgatus, Bacteroides WH2, C. aerofaciens), with changes in transcription largely due to pectinesterase genes found in B. ovatus (BA-COVA_03576, BACOVA_03581, BACOVA_04902), B. thetaiotaomicron (BT_4109, BT_4110), B.
vulgatus (BVU_1116), and B. WH2 (BACWH2_3569, BACWH2_3615). Increases in the proportional abundance of cellobiose phosphorylase transcripts reflected the contributions of three community members: Bacteroides uniformis, Eubacterium rectale, and Ruminococcus obeum (table S8).

The KEGG starch and sucrose metabolism, pentose and glucuronate interconversions, and “pentose phosphate” pathways process products generated by these three enzymes. Figure 5B shows that many of the other components of these pathways are up-regulated in the 15-member community when the FMP strain consortium is introduced. ShotgunFunctionalizeR also identified significant increases in the expression of genes encoding five ECs that participate in the generation of propionate and succinate: The induction occurred within 24 hours after the FMP strains were introduced and involved acetate kinase (EC 2.7.2.1; catalyzes a bidirectional reaction between propanoyl phosphate and propionate), phosphate acetyltransferase (EC 2.3.1.8), methylmalonyl-CoA (coenzyme A) decarboxylase (EC 4.1.1.41), propionyl-CoA carboxylase (EC 6.4.1.3), and methylmalonyl-CoA mutase (EC 5.4.99.2, yields succinyl-CoA as its product) (fig. S6). Only a single treatment with the FMP consortium was required to produce a sustained response involving the enzymes that can yield propionate (fig. S6).

A breakdown of RNA-Seq reads by the community member genome to which they mapped revealed that the abundance of a species in the 15-member community did not necessarily correlate with its contribution to the community transcript pool. At the time point sampled immediately before invasion (day 14), two of the most extreme outliers were B. WH2 [comprised 39.6 ± 1.6% (mean ± SD) of the community but only contributed 15.4 ± 2% of the raw reads to the total RNA-Seq read pool] and R. obeum (2.1 ± 0.4% of the community; 18.2 ± 4.4% of the transcript pool) (fig. S7). These observations indicate that community-level transcriptional responses can be driven by species representing small fractions of the microbiota.

Our bottom-up analysis is summarized in fig. S8 and table S10, and disclosed early- and later-responding species. Specifically, there were more significantly highly regulated R. obeum
transcripts within the community metatranscriptome 1 day after gavage than would be expected based on its community representation, and more highly regulated \( R. \) obeum genes in the comparison between day 14 (just before gavage) and day 15 metatranscriptomes than between day 14 and day 42 metatranscriptomes. In contrast, \( B. \) WH2, \( Clostridium \) scindens, and \( B. \) uniformis were defined as late responders to the FMP consortium.

**Identifying predictive features from the model community metatranscriptome data using a random forests classifier.** Machine learning techniques using the random forests classifier can be applied to metagenomic data (23) to learn a function that maps a set of input values or predictors (in this case, relative abundance of KEGG categories, KEGG pathways, or ECs in a community) to a discrete output value (here, the presence/absence of the FMP strains). KEGG categories, KEGG pathways, and ECs were all able to predict pretreatment/posttreatment status with low estimated generalization error (KEGG categories, 6.7%; ECs, 13.3%; KEGG pathways, 10.0%). In all cases, these generalization error rates were less than half of the baseline error rate of 33% (that is, that achieved by always predicting the largest category). There were 11 predictive and 5 highly predictive KEGG categories, 35 moderately predictive ECs, and 27 predictive and 4 highly predictive KEGG pathways (table S11). The predictive ECs identified using our supervised classification approach include a number of carbohydrate metabolism–related functions that were also identified using ShotgunFunctionalizeR in our top-down analysis.

**Metabolomic analyses.** To evaluate the impact of invasion with the five-member FMP consortium on microbial-host co-metabolism, we performed untargeted gas chromatography–mass spectrometry (GC-MS) on urine samples collected at multiple time points (days 0, 14, and 42) from members of the single- and multiple-treatment groups (Fig. 1B). A metabolite profile was constructed for each urine sample (\( n = 19 \)) using the spectral abundances of all identifiable metabolites. A total of 198 metabolites met our reverse match score cutoff of 65% and were present in at least 50% of samples at one or more time points [for an explanation of the reverse match score, see (24) and table S12]. Comparing day 0 and 14 samples revealed 39 metabolites whose levels were significantly higher or lower after colonization with the defined 15-member community (see
The changes included decreases in the levels of oligosaccharides we would expect to be consumed by members of the microbiota [melibiose (87% decline); raffinose/maltotriose (98%); note that oligosaccharides are by their nature difficult to identify with certainty with the present, nontargeted GC-MS technique, and our annotations of these metabolites as melibiose and raffinose/maltotriose are provisional]. The observed 3.4-fold increase in pyrogallol, a polyphenol, is consistent with the known ability of many gut microbes to cleave these molecules from polyphenols present in dietary plant material. A 4.4-fold increase in taurine after the initial colonization of mice was also noted, probably a result of microbial deconjugation of taurine from bile compounds.

Table S12B lists urinary metabolites that change significantly after introducing the five FMP strains (compare day 14 versus day 42 in table S12B). Fructose and xylose were not significantly affected by the introduction of the defined 15-member community but increased significantly after the introduction of the FMP strain consortium (2.3- and 2.9-fold, respectively; Fig. 6, A and B). Increases in fructose may reflect an enhanced capacity of the community to liberate this monosaccharide from levan and other polyfructans via levanase-catalyzed reactions. Increases in xylose might be explained by the additional xylanase activity introduced by B. animalis subsp. lactis (Fig. 3) or, alternatively, by the induction of microbiome genes encoding xylan-degrading enzymes (for example, BACOVA_04387 and BACOVA_04390, which were up-regulated 5.2- and 11.0-fold, respectively, after introduction of the FMP strains; table S10). Changes in other metabolites such as xanthosine (Fig. 6C), a purine metabolite, suggest that the metabolic consequences of FMP strain introduction extend beyond the processing of carbohydrates.

Collectively, our transcriptional and metabolite analyses indicated that introducing FMP strains that constitute a small fraction of a defined model human gut microbiota signals the microbiota to change its metabolic activities, including activities related to carbohydrate metabolism. With this information in hand, we returned to the human fecal samples to determine the extent to which observations made in our gnotobiotic mouse model were applicable to humans.
Microbiome transcriptional responses to FMP strains that are shared by gnotobiotic mice and humans. Microbial RNA-Seq analysis was performed on human fecal samples obtained 1 week before FMP consumption, 1 and 4 weeks into the consumption period, and 4 weeks after cessation (both co-twins from family 1; one co-twin from family 3; see table S1). Using an analysis pipeline comparable to the one used for the mouse data, we first aligned all RNA-Seq reads against a reference set of 131 human gut microbial genomes plus the FMP strain genomes, binned the aligned transcripts on the basis of their EC annotations, and used ShotgunFunctionalizeR to identify ECs whose abundances were significantly changed as a function of FMP exposure (Benjamini-Hochberg adjusted $P < 0.01$).

Categorical analysis of the responses of the human fecal community to FMP consumption revealed that significantly up-regulated ECs were principally distributed among the KEGG categories carbohydrate metabolism, amino acid metabolism, and metabolism of cofactors and vitamins (see table S13 for a complete list of ECs identified from the various pairwise comparisons of time points).

Figure 7 highlights the 86 ECs that were significantly changed ($P < 0.01$) in the same direction in all humans and in all sampled mice as a function of exposure to the FMP strain consortium. Similar to our findings in mice, the most prominently represented KEGG category among up-regulated gene functions in all comparisons of human metatranscriptomes was carbohydrate metabolism (Fig. 7). The three ECs involved in entry points in the KEGG starch and sucrose metabolism pathway [levanase (EC 3.2.1.65), pectinesterase (EC 3.1.1.11), and cellobiose phosphorylase (EC 2.4.1.20)] were significantly up-regulated within 1 week after FMP consumption was initiated in the humans surveyed. This transcriptional response was sustained in the case of levanase and pectinesterase and ceased (fell to below the limits of detection) within 4 weeks after FMP administration was stopped (Fig. 5A).

ECs involved in succinate and propionate metabolism (EC 2.7.2.1 and EC 6.4.1.3) were also up-regulated in the human fecal metatranscriptome within 1 week of the initiation of FMP
consumption (FMP1 versus Pre1, Fig. 7). As with levanase, pectinesterase, and cellobiose phosphorylase, this response was sustained during and subsided after the period of FMP consumption (see “FMP4 versus Pre1” and “FMP1 versus Pre1” in Fig. 7 and table S13).

Human fecal transcripts were detected that mapped to the *B. animalis* subsp. *lactis* genome (see the Supplementary Material). The presence of these transcripts was limited to the period of FMP consumption, supporting the notion that they emanated from the FMP strain rather than from a related species present within the microbiota (fig. S9). This clear linkage to FMP consumption was not evident in the case of other members of the consortium, so we could not confidently analyze their patterns of gene expression in vivo. The highest number of mapped reads to the *B. animalis* subsp. *lactis* genome was obtained 1 week after FMP administration began: Among the 4000 reads, we were able to detect transcripts from all but 1 of the 10 genes in the *BALAC2494_00604-BALAC2494_00614* locus that encodes enzymes involved in the catabolism of xylooligosaccharides, leading us to conclude that this locus is highly expressed in the distal human gut, just as it is in our mouse model.

**Discussion**

Repeated sampling of seven healthy MZ adult twin pairs over a 4-month period emphasized that intrapersonal variation in bacterial community structure was less than interpersonal variation, with co-twins having significantly more similar phylogenetic and taxonomic structure in their fecal microbiota compared to those from unrelated individuals (9, 25, 26). The results also showed that (i) consumption of an FMP containing five bacterial strains was not associated with a statistically significant change in the proportional representation of resident community members within and between individuals; (ii) the appearance and disappearance of strains comprising the FMP consortium did not exhibit familial patterns in the fecal microbiota; and (iii) *B. animalis* subsp. *lactis* CNCM I-2494 was the most prominent assayed member of the consortium represented in the microbiota during the 7-week period of FMP consumption. Analyses of the fecal gene repertoire over the course of the 16 weeks of the experiment indicated that (i) variations in the functional fea-
tures of the (fecal) microbiome were less than the variations in bacterial species composition; (ii) there was no significant difference in the degree of similarity in representation of KEGG orthology group functions for a given co-twin at each time point compared to the degree of similarity that existed between co-twins, whereas individual and twin pair microbiomes were significantly more similar to one another than those from unrelated individuals; and (iii) there were no statistically significant changes in the representation of these functions when the FMP strain consortium was being consumed. With these findings in mind, and with each individual as well as each genetically identical co-twin serving as a control, we concluded that at least at the depth and frequency of sampling used for this small healthy cohort, the bacterial species and gene content of their fecal microbiota/microbiome were not an informative biomarker for understanding whether or how this commercial FMP affected microbial community properties.

Gnotobiotic mice harboring a model 15-member gut microbial community that represented the three principal bacterial phyla present in the human gut microbiota, and whose 58,399 known or predicted protein-coding genes encompassed many of the prominent functions present in the normal adult human fecal microbiome, provided a means for characterizing the impact of the five-member FMP strain consortium on expressed gut microbial community functions and then applying the results to the human fecal specimens collected for this study. As with the MZ twins, introduction of the five-member strain consortium did not significantly affect the representation of the 15 species that make up the model human microbiota. As with the MZ twins, B. animalis subsp. lactis exhibited the greatest fitness of the five FMP strains in the gut, as judged by its prominence and persistence. Unlike the human arm of the study, where all subjects consumed the FMP twice daily, the design of the mouse study, with its single- versus multiple-treatment regimens, allowed us to directly compare the persistence of FMP consortium members. Only B. animalis subsp. lactis and L. lactis subsp. cremoris were able to maintain a foothold in the gut ecosystem at detectable levels for the entire 4-week monitoring period after a single dose. In addition, colonization levels were not affected by the number of times the FMP strains were administered to mice.
An advantage of constructing the model human gut microbiome was that its entire predicted gene repertoire was known. This allowed us to define the impact of introducing the FMP strain consortium on the functions expressed by the overall community as well as by its individual components. A major theme emanating from our analysis was the effect of introducing the FMP consortium on carbohydrate metabolism by the community, as well as the effect of the community on a feature of carbohydrate metabolism by *B. animalis* subsp. *lactis*. The model 15-member community responded to the FMP consortium by inducing genes encoding enzymes involved in catalyzing reactions that represent the three entry points into the KEGG “starch and sucrose metabolic pathway,” as well as enzymes that catalyze fermentation of carbohydrates to propionate. The mechanism by which the FMP strains elicit this response is unclear at present, but the effect is rapid (occurring within the first 24 hours after invasion) and was persistent whether the consortium was introduced in a single set of gavages during a 1-day period or with subsequent repeated gavage over a several-week period. The persistence of both the carbohydrate pathway response and of *B. animalis* subsp. *lactis* suggests but does not prove that the latter may be instrumental in instigating and maintaining the former.

Intriguingly, the carbohydrate response showed features of “differentiation.” As noted in the Results section, the levanase response was driven almost entirely by changes in transcription in just a single species (*B. vulgatus*), the pectinesterase response by six community members (*B. caccae, B. ovatus, B. thetaiotaomicron, B. vulgatus, B. WH2, and C. aerofaciens*), and the cellobiose phosphorylase response by three components of the defined model human gut microbiota (*B. uniformis, E. rectale, and R. obeum*). Of the 50 genes with predicted xylan-degrading capacity in the model microbiome (that is, those encoding enzymes in ECs 3.2.1.37 and 3.2.1.8), only *BACOVA_04387* and *BACOVA_04390* (both from *B. ovatus*) were significantly up-regulated after FMP strain introduction (this is ignoring xylanase genes encoded by FMP strains like *B. animalis* subsp. *lactis*). This up-regulation in a limited subset of the model community coincides with an increase in urinary xylose.
The ability to attribute EC-level changes to individual genes in specific bacterial species was not possible with our RNA-Seq analysis of the human fecal samples. The differentiation of carbohydrate responses among bacterial species documented in gnotobiotic mice emphasizes a challenge and opportunity that can be addressed in these models, namely, to further delineate the niches, interactions, and adaptive resource switching behaviors of community members by intentional addition, removal or substitution of taxa, and/or their modification through genetic manipulation. Although requiring significantly more animals and loss of the ability to use an animal as its own control, future studies could be expanded to include sampling of community gene expression in different segments of the small intestine.

The increased expression of genes encoding enzymes involved in the interconversion of propionate and succinate is intriguing given the fact that this short-chain fatty acid has been linked in some reports to effects on gastrointestinal transit time. However, work in this area has yielded varying results and conclusions, perhaps because of the diversity of models and methodological approaches used (27–30). Propionate may also link the gut microbiota and human physiology through its effects on hepatic and adipose tissue metabolism (31). Notably, another group has reported that in the T-bet−/−Rag2−/− mouse model of colitis, consumption of an FMP containing a dairy matrix plus the same strains used in this study led to increased cecal propionate levels and a reduction in intestinal inflammation (32).

The extent of translatability of data from gnotobiotic mouse models harboring collections of sequenced representatives of the human gut microbiota to humans themselves needs to be tested further, not only at the transcriptional level but also at the level of community-host co-metabolism. Although current models can and should be evolved to embrace more of the diversity present in our gut communities, even with current limitations, they can serve as part of a preclinical discovery pipeline designed to identify candidate biomarkers and mediators of the effects of existing or new probiotic strains on the properties of microbial communities and their hosts. They also represent an analytical tool for characterizing the effects of specified dietary components on the indigenous gut community and on probiotic species that are deliberately consumed. The results could yield
new candidate prebiotics that may affect the representation and metabolic properties of probiotic species or entrenched members of our gut microbiota and provide the proof-of-mechanism and proof-of-principle observations needed to justify, direct, and interpret human studies.

**Materials and methods**

**Human studies**

**Subject recruitment.** Seven MZ female twin pairs aged 21 to 32 years with body mass indices ranging from 20 to 25 kg/m$^2$ were recruited for this study. These twins were long-standing participants in the Missouri Adolescent Female Twin Study (MOAFTS) (26, 33). Procedures for obtaining consent, for providing fecal samples, and for maintaining diaries of FMP consumption, and stool frequency and consistency were approved by the Washington University Human Studies Committee.

**Other procedures.** Methods used for the production and distribution of the FMP to study participants, analysis of the effects of FMP consumption on stool parameters, qPCR analysis of fecal levels of FMP strains (table S14), multiplex pyrosequencing of 16$S$ rRNA genes in fecal samples and the FMP, co-occurrence analysis, and shotgun sequencing of human fecal microbiomes are described in the Materials and Methods section of the Supplementary Material. Sequenced human gut–associated microbial genomes used to analyze the human fecal RNA-Seq data are listed in table S15.

**Studies in gnotobiotic mice**

**Colonization of germ-free mice.** The justification for using mice and the protocols used for treating them were approved by the Washington University Animal Studies Committee. Animals belonging to the C57BL/6J inbred strain were maintained in plastic flexible film gnotobiotic isolators and fed a standard autoclaved chow diet (B&K rat and mouse autoclavable chow #7378000, Zeigler Bros Inc.) ad libitum. Two groups of 6- to 8-week-old germ-free male animals ($n = 5$ per group) were
colonized with a single gavage of 500 μl of supplemented TYG medium (TYG₄) (34) containing 15 sequenced human gut-derived bacterial symbionts (6 × 10⁶ CFU per strain; total of 9 × 10⁷ CFU for the community). The B. thetaiotaomicron component of this community was composed of a library of 34,544 transposon mutants prepared as described (34). Fourteen and 15 days later, both groups of mice were gavaged with a mixture of the five FMP strains (each species at 5 × 10⁶ CFU) in 300 μl of TYG₄. One group of mice received a second pair of gavages 7 and 8 days later, and a third pair of gavages 21 and 22 days after the initial FMP strain introduction.

**Other procedures.** Methods used for sampling animals, COPRO-Seq, INSeq, microbial RNA-Seq, and nontargeted metabolomics via GC-MS, are described in the Methods section of the Supplementary Material, as are methods for sequencing and annotating FMP strain genomes.

**Acknowledgements**

We thank Jill Manchester, Jessica Hoisington-López for assistance with DNA sequencing, Maria Karlsson, David O’Donnell and Sabrina Wagoner for help with gnotobiotic mouse husbandry, Su Deng for assistance in preparing Illumina DNA libraries, Stacy Marion and Deborah Hooper for their contributions to the human study, Deanna Carlsen for coordination of FMP production and logistics, Stephan Baumann and Steven Fischer (Agilent Corp) for kindly providing the Fiehn GC/MS Metabolomics RTL library used for metabolomics analyses, members of the Gordon lab for valuable suggestions during the course of this work, and Gerard Denariaz for his continued support. We are also grateful to Integrated Genomics for generating the draft genome sequences of B. animalis subsp. lactis (CNCM I-2494) and S. thermophilus (CNCM I-1630).

**Funding:** Supported by grants from the NIH (DK30292, DK70977) and Danone Research. Maintenance of the MOAFTS twin cohort is supported by NIH grants AA09022, AA11998, AA17915 and HD49024.
Accession numbers: The complete genome sequence of *Bifidobacterium animalis* subsp. *lactis* (CNCM I-2494) has been deposited at GenBank under accession number CP002915. The Whole Genome Shotgun projects for *Lactobacillus delbrueckii* subsp. *bulgaricus* (CNCM I-1632, CNCM I-1519), *Streptococcus thermophilus* CNCM I-1630, and *Lactococcus lactis* subsp. *cremoris* CNCM I-1631 have been deposited at DDBJ/EMBL/GenBank under accession numbers AGFO00000000, AGHW00000000, AGFN00000000, and AGHX00000000, respectively. The version of each of these genomes described in this paper is the first version (CP002915.1, AGFO01000000, AGHW01000000, AGFN01000000, and AGHX01000000, respectively). COPRO-Seq and RNA-Seq data are deposited in GEO (accession numbers GSE31943 and GSE31670, respectively), while 16S rRNA pyrosequencing reads and shotgun pyrosequencing reads of human fecal community DNA are deposited in MG-RAST (accession numbers qiime:803 and 4473933-4473980, respectively).


Competing interests: The authors declare that they have no competing interests.
References


Figure legends

Figure 1. Experimental design for human and mouse studies. (A) Human study. Seven healthy lean MZ female twin pairs were sampled before, during, and after FMP consumption. (B) Gnotobiotic mouse study. Two groups of five germ-free mice were colonized by oral gavage at 6-8 weeks of age with a 15-member microbial consortium constituting a model human gut microbiota (day of gavage denoted by black arrows). Two weeks later, the five species FMP strain consortium was administered by oral gavage to each group of mice twice over two days (denoted by green arrows). Mice in the single treatment group underwent no further manipulations, whereas animals in the multiple treatment group received additional two-day gavages one and three weeks following the first gavage. Samples were collected at the indicated time points for profiling bacterial community membership (shotgun and 16S rRNA gene sequencing for human fecal samples, COPRO-Seq for mouse fecal and cecal samples), gene expression profiling (microbial RNA-Seq) and metabolite analysis (urines, GC/MS). The species comprising the model 15-member human community and the 5-member FMP consortium are listed in the gray and green boxes, respectively.

Figure 2. Metagenomic studies of human fecal microbiomes sampled over time. (A) 16S rRNA-based time course study of intra- and interpersonal variations in fecal bacterial community structure during the course of the 4-month study. Unweighted UniFrac measurements of community distances, from pairwise comparisons of all samples obtained from a given individual, from co-twins, and from unrelated individuals are plotted as mean values ±SEM. (B) Colored boxes represent the proportion of bacterial phylotypes that were consistently present within an individual over time (gray), between co-twins over time (orange), and in all 126 fecal samples (red). The white box represents the average number of species-level phylotypes found in a given sample. All measures of spread provided in parentheses represent ±SEM. (C) KEGG Orthology groups (KOs) consistently present within the fecal microbiome of an individual over time (gray), between co-twins over time (orange), and in all 48 microbiomes analyzed from the 4 sets of MZ twins during the 4 month study (red). The white box indicates the average number of unique KOs (±SEM) identified in a particular sample. All measures of spread provided in parentheses represent ±SEM.
(D) Hellinger distance measurements of fecal microbiomes based on their KO content. Tests of statistical significance are based on 1000 permutations of a Hellinger distance matrix. Mean values (±SEM) are shown for the three types of comparisons (self-self; co-twin-co-twin; unrelated-unrelated individual).

**Figure 3. Correspondence analysis of B. animalis subsp. lactis CAZyme gene expression.** RNA-Seq data for all *B. animalis* subsp. *lactis* genes encoding known or predicted CAZymes were subjected to unconstrained correspondence analysis using the ‘vegan’ package in R. Correspondence analysis (CA) allows for the generation of biplots in which samples and genes can be plotted in the same ordinate space to reveal associations/anti-associations between the two. Circles represent individual CAZymes (genes). The genes ordinating furthest from the origin in the direction of one of the sample clusters (treatment groups) are labeled according to their locus number and are colored based on CAZyme family assignment (see Table to the right of the Figure for details; the abbreviation NA refers to no designation). Red triangles represent samples and are labeled according to the following nomenclature: LX, logarithmic phase cells in MRS with X being the technical replicate number (e.g. L1 refers to the first technical replicate harvested in log phase); SX, stationary phase cells in MRS with accompanying replicate number; MX, feces from designated gnotobiotic animals obtained 4 weeks after the initial invasion with the FMP strain consortium; PX, samples obtained after fermentation in the FMP dairy matrix. Each cluster of samples from a particular treatment is associated with a functionally related set of expressed CAZymes.

**Figure 4. ‘Top-down’ analysis of the effects of the FMP strain consortium on the model 15-member community’s metatranscriptome.** RNA-Seq reads were mapped to the sequenced genomes of the 15 community members. Transcript counts were normalized [reads per kb of gene length per million reads (RPKM), see Supplementary Material] and binned using the hierarchical levels of functional annotation employed by KEGG. For each KEGG category (A) or pathway (B) shown, boxplots depict the proportion of normalized read counts assignable to that annotation out of all reads which could be assigned annotations for that hierarchical level. Data shown correspond to the ‘multiple’ treatment group of mice (the group for which the most time points were collected),
however, data for all mice are provided in Table S8. (C) Illustration of how a model community’s functional response (e.g., the increased expression of levanase-encoding genes) can be dissected to identify the subset of genes/species driving the response. Boxes denote top quartile, median, and bottom quartile. Whisker length represents 1.5x inter-quartile range (IQR), except where there are no outliers; in these situations, whiskers span the range from minimum to maximum values. Box color denotes the day fecal samples were obtained (day 14 is the pre-treatment timepoint immediately preceding gavage of the FMP strain consortium). When an asterisk is centered over a box, it indicates that there was a statistically significant change following administration of the FMP consortium relative to the pre-treatment timepoint (p<0.05 by paired, two-tailed Student’s t test). The positioning of asterisks above versus below a box emphasizes the direction of change (above, upregulation; below, downregulation).

**Figure 5. Mouse and human communities share transcriptional responses to the FMP strain consortium involving ECs related to carbohydrate metabolism.** (A) Box plots of the proportion of all RPKM-normalized reads in mouse and human fecal metatranscriptomes represented by three ECs involved in plant biomass degradation. Individual samples are shown as black dots (n=2-10). Boxes are also colored by fold-change, as determined by comparing mean values at a given time point to the value at the pre-treatment time point [for gnotobiotic mice pre-treatment refers to day 14; in the case of humans, pre-treatment refers to the fecal sample collected 1 week prior to initiation of FMP consumption (sample ‘Pre1’ in Fig. 1A)]. Statistical significance was determined using the ShotgunFunctionalizeR package in R and an adjusted p-value cutoff of <0.01. Pre-treatment time points, and subsequent time points where expression levels were not significantly different from the pre-treatment mean are colored white. (B) Components of KEGG ‘starch and sucrose metabolism’, ‘pentose and glucuronate interconversions’ and ‘pentose phosphate’ pathways whose expression in the 15-member model community changed compared to pre-treatment values when the 5-member FMP strain consortium was introduced. Gray indicates that the fold-change was statistically significant (adjusted p-value <0.01). Ovals highlight the three enzymes shown in panel A. Dashed arrows indicate that multiple enzymatic reactions lead from these ECs and their indicated
substrates to the products shown. These intermediate reactions have been omitted for clarity or because the omitted ECs did not manifest significant changes in their expression.

**Figure 6. Select urinary metabolites whose levels are altered after the introduction of the FMP strain consortium into mice harboring a defined model human gut microbiota.** The statistical significance in pairwise comparisons shown in panels A-C was evaluated using a two-tailed Student’s *t* test on the log-transformed spectral abundance of the metabolite in each sample. Values for the statistical significance of differences between time points as evaluated by one-way ANOVA, followed by FDR-correction and a post-hoc Tukey HSD test are also provided in Table S12. Horizontal bars represent group means, vertical bars represent ± SEM. n/s; not significant.

**Figure 7. Shared transcriptional responses to FMP strain exposure in mice and humans.** The heatmap shows ECs that exhibit a statistically significant change in their expression (Shotgun-FunctionalizeR, adjusted p < 0.01) and manifest a consistent direction of change in their expression in all four comparisons shown. Comparisons include those where the pre-treatment timepoint was compared with a timepoint shortly after FMP strains were introduced (mouse: ‘d15 vs d14’, human: ‘FMP1 vs Pre1’) and those where the pre-treatment period was compared to a timepoint several weeks after strain introduction (mouse: ‘d42multi vs d14’, human: ‘FMP4 vs Pre1’). ‘d42multi’ indicates the multiple-treatment group at day 42 of the mouse experiment. The colored boxes correspond to the KEGG categories that contain the ECs shown to the right of the heatmap. The scale refers to fold-difference in the mean of relative abundance of each EC between treatment and pre-treatment groups based on the mean number of normalized reads (RPKM) of transcripts assigned to a given EC. The 18 ECs shown at the bottom of the Figure are not associated with the five prominent KEGG categories listed. Their assigned categories are provided in Table S13.
Figures

Figure 1.

A (Human)

B (Mouse)

Synthetic Community Species
- Bacteroides caccae
- Bacteroides ovatus
- Bacteroides thetaiotaomicron
- Bacteroides uniformis
- Bacteroides vulgatus

Fermented Milk Product (FMP) Species
- Bifidobacterium animalis subsp. lactis CNCM I-2494
- Lactobacillus delbrueckii subsp. bulgaricus CNCM I-1632
- Lactococcus lactis subsp. cremoris CNCM I-1631
- Streptococcus thermophilus CNCM I-1630

Legend:
- Feces (16S sequencing)
- Feces (shotgun sequencing)
- Feces (RNA-Seq)
- Cecal contents (COPRO-Seq)
- Feces (INSeq)
- Urine (GC/MS)
Figure 2.

A) UniFrac distance versus number of weeks between samples.

B) Bar graph showing average OTUs per timepoint, average individual core, average twin pair core, and completely conserved core. Details include:
- Average OTUs per timepoint: 163±3 OTUs/sample
- Average individual core: 25% of OTUs; 40±2 OTUs/sample
- Average twin pair core: 14% of OTUs; 22±2 OTUs/sample
- Completely conserved core: 2.5% of OTUs found in all 126 samples; 4 OTUs, Firmicutes (family Lachnospiraceae)

C) Histogram showing average KOs/sample: 2205±26 KOs/sample.
- 64% of KOs in a given sample (1417±46 KOs/sample) consistently represented in all samples of an individual and her cotwin.
- 55% of KOs in a given sample (1212±33 KOs/sample) consistently represented in all samples of an individual and her cotwin.
- 41% of KOs (882) found in all 48 samples.

D) Hellinger Distance comparison:
- Self: p<0.001
- Twin-Twin: p<0.001
- Unrelated: NS
Figure 3.

<table>
<thead>
<tr>
<th>Gene</th>
<th>CAZy family</th>
<th>EC</th>
<th>Activity predicted by EC</th>
</tr>
</thead>
<tbody>
<tr>
<td>BALAC2494_00644</td>
<td>GH2</td>
<td>3.2.1.23</td>
<td>β-galactosidase</td>
</tr>
<tr>
<td>BALAC2494_00604</td>
<td>GT2</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>BALAC2494_01233</td>
<td>GT2</td>
<td>2.4.1.-</td>
<td>Hexosyltransferase</td>
</tr>
<tr>
<td>BALAC2494_01233</td>
<td>GT2</td>
<td>2.4.1.-</td>
<td>Hexosyltransferase</td>
</tr>
<tr>
<td>BALAC2494_01349</td>
<td>GT2</td>
<td>2.4.1.-</td>
<td>Hexosyltransferase</td>
</tr>
<tr>
<td>BALAC2494_01359</td>
<td>GT2</td>
<td>2.4.1.83</td>
<td>β-D-mannosytransferase</td>
</tr>
<tr>
<td>BALAC2494_00644</td>
<td>GH2</td>
<td>3.2.1.23</td>
<td>β-galactosidase</td>
</tr>
<tr>
<td>BALAC2494_00604</td>
<td>GT2</td>
<td>3.2.1.23</td>
<td>β-galactosidase</td>
</tr>
<tr>
<td>BALAC2494_00605</td>
<td>GH43</td>
<td>3.2.1.23</td>
<td>Glycosidase</td>
</tr>
<tr>
<td>BALAC2494_00606</td>
<td>GH43</td>
<td>3.2.1.37</td>
<td>xylan 1,4-β-xylosidase</td>
</tr>
<tr>
<td>BALAC2494_00612</td>
<td>GH43</td>
<td>3.2.1.37</td>
<td>xylan 1,4-β-xylosidase</td>
</tr>
<tr>
<td>BALAC2494_01173</td>
<td>GH77</td>
<td>2.4.1.25</td>
<td>4-β-glucanotransferase</td>
</tr>
</tbody>
</table>
Figure 5.

A

<table>
<thead>
<tr>
<th>EC 2.4.1.20</th>
<th>EC 3.1.1.11</th>
<th>EC 3.2.1.65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cellobiosephosphorylase</td>
<td>Pectinesterase</td>
<td>Levanase</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>% of normalized counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-5}$</td>
</tr>
<tr>
<td>$10^{-4.5}$</td>
</tr>
<tr>
<td>$10^{-4}$</td>
</tr>
<tr>
<td>$10^{-3.5}$</td>
</tr>
<tr>
<td>$10^{-3}$</td>
</tr>
<tr>
<td>$10^{-2.5}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relative fold-change in expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>-10</td>
</tr>
<tr>
<td>-2</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

B

1 day after FMP strain introduction vs pre-FMP
4 weeks after FMP strain introduction (multiple treatment) vs pre-FMP
1 day after FMP strain introduction vs pre-FMP
4 weeks after FMP strain introduction (multiple treatment) vs pre-FMP
Figure 6.

A  Fructose (urine)  

B  Xylose (urine)  

C  Xanthosine (urine)  

Fructose (urine)  

Xylose (urine)  

Xanthosine (urine)  

p = 0.048  

p = 0.047  

p = 0.0033  

p = 0.016
Supplementary material

Supplementary materials and methods

Microbial genome sequencing. The following bacterial strains are incorporated into the commercially available FMP: *Bifidobacterium animalis* subsp. *lactis* (strain CNCM I-2494), *Lactobacillus delbrueckii* subsp. *bulgaricus* (strains CNCM I-1632, CNCM I-1519), *Lactococcus lactis* subsp. *cremoris* (strain CNCM I-1631), and *Streptococcus thermophilus* (strain CNCM I-1630). We performed shotgun 454 FLX pyrosequencing of both *L. delbrueckii* subsp. *bulgaricus* strains (CNCM I-1632, CNCM I-1519), plus the *L. lactis* subsp. *cremoris* strain (33-, 32- and 51-fold coverage, respectively). Using the Newbler assembler (v1.1, 454 Life Sciences) and already sequenced strains of these species, we obtained draft genome assemblies with N50 contig sizes of 69,466 and 55,626 and 55,851 bp, respectively. The total sizes of the assembled *L. delbrueckii* genomes were 1,767,897 bp (CNCM I-1632) and 1,795,730 bp (CNCM I-1519), while the *L. lactis* assembly had an aggregate genome size of 2,511,332 bp. A finished genome sequence of the *B. animalis* subsp. *lactis* genome and a deep draft assembly of the *S. thermophilus* genome were previously generated by Integrated Genomics (see Table S3 for a summary of genome metrics).

Annotation and comparative genomic analysis. The predicted proteins of all the sequenced bacterial species used in this study were annotated using BLASTP searches (*E*-value $<10^{-5}$) against version 54 (v54) of the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (*SI-3*) and the Carbohydrate Active Enzyme (CAZy) database (*S4*). BLAST results were parsed into a lookup table for each genome, and each fecal microbiome, using a Perl script (lookup_KEGG_for_genomes_BLAST.pl) that assigned to each gene the KEGG Orthology (KO) identifier associated with its top BLAST hit (*E*-value $<10^{-5}$). In cases where the query protein’s top match to the reference database was ambiguous (i.e., when there were multiple top matches with equally good scores), the gene was annotated with the KOs of all the top-matching database entries.

These lookup tables were then used to calculate the ‘coverage ratio’ of each KEGG metabolic pathway for each bacterial genome in this study using the Perl script kegg_key.pl. For each
edge (enzymatic reaction) of each KEGG metabolic pathway, this script checks for the presence of a gene whose assigned KO corresponds with that reaction in (i) a given microbial genome, (ii) defined collections of microbial genomes, or (iii) larger, incompletely sequenced microbiomes. KEGG pathway coverage ratios for several combinations of genomes, as well as several published gut microbiomes, were calculated and used to perform average linkage hierarchical clustering in Cluster 3.0 (S5) using an un-centered correlation similarity metric. A heatmap visualization of this clustered data was generated using the Java Treeview application (jtreeview.sourceforge.net).

**Culturing of B. animalis subsp. lactis.** A frozen stock of *B. animalis* subsp. *lactis* (strain CNCM I-2494) was streaked out on MRS-agar plates (BD/Difco) and transferred to a Coy chamber for overnight growth at 37°C under an atmosphere of 5% H₂, 20% CO₂ and 75% N₂. Single colonies were picked and inoculated into 10 ml of pre-reduced MRS broth (BD/Difco) that had been stored for 24 h in an anaerobic chamber. The medium was not supplemented with cysteine. Each culture was passaged four times to stationary phase, and growth curves were recorded to discern growth kinetics under these conditions. 100 μl of an overnight culture was used to inoculate 10 ml of fresh anaerobic MRS in 27 ml Balch tubes with rubber stoppers and aluminum crimp tops. The initial headspace of the tubes was composed of 5% H₂, 20% CO₂, and 75% N₂ at ambient pressure. Tubes were incubated at 37°C, and 4 ml aliquots were collected at mid-log phase (9 h post-inoculation, OD₆₀₀ =0.2) and during late stationary phase (27 h post-inoculation, OD₆₀₀ =2.6). Each aliquot from each culture (n=2) was immediately combined with 8 ml of RNAProtect Bacteria Reagent (Qiagen), incubated for 5 min at room temperature, then centrifuged (3,220 x g; 15 min at 25°C). The pellets were snap-frozen in liquid nitrogen and stored at -80°C, and total cellular RNA was subsequently isolated as described previously (S6).

**Human studies**

**Production and distribution of the FMP to study participants.** The FMP used for this study was produced in Danone’s pilot plant located in Fort Worth, TX. Batches were shipped directly to subjects by an independent delivery service so that the names of study participants would remain
unknown to all but those in the MOAFTS study group. Each subject received 4 shipments of FMP, spaced at 2-week intervals. Each shipment was composed of sufficient numbers of cups (pots) so that study participants could consume one 4 oz serving twice a day (each serving consisted of a single pot). Each co-twin chose her flavors (strawberry, vanilla, and/or peach). The same flavor selection was shipped each time. Each subject was allowed to vary the sequence of selected flavors according to her wishes.

**Analysis of the effects of FMP consumption on stool parameters.** Stool consistency, difficulty of passage, and frequency were assessed using a daily stool diary in which participants recorded the time of day for each bowel movement. Participants rated the stool consistency using the seven point Bristol Stool Form Scale (S7) and the difficulty of passage using a five-point scale (no difficulty to extreme difficulty).

**Quantitative PCR (qPCR) analysis of fecal levels of FMP strains.** qPCR was used to define the levels of selected FMP strains in fecal samples obtained from MZ co-twins and gnotobiotic mice (S8-12). The PCR primer sets targeting each strain’s 16S rRNA gene or CRISPR locus are described in Table S14, as are the amplification conditions. Samples were analyzed on an Applied Biosystems 7900HT instrument using SYBR Green chemistry. Standard curves were constructed using genomic DNA prepared from a known number of bacterial cells harvested from monocultures grown to stationary phase (cells were DAPI stained and counted by microscopy); 

To calculate the concentration of a given bacterial strain in each fecal sample, three serial dilutions of extracted fecal DNA (10 ng, 1 ng, 0.1 ng) were assayed in at least two independent qPCR reactions. 

For human samples, data were log-transformed and normalized to fecal mass ($\log_{10}$ CE/g feces). For mouse samples, data were log-transformed and normalized to mass of template DNA ($\log_{10}$ CE/μg DNA).
**Multiplex pyrosequencing of 16S rRNA genes in fecal samples and the FMP.** A total of 126 fecal samples (9 samples per individual) were collected over the course of 4 months according to the schedule shown in Fig. 1. All fecal samples were frozen at -20°C within 30 min after they were produced. The samples were maintained at this temperature for the period of time required to ship them to a biospecimen repository (less than 24 h). As soon as samples were received, they were de-identified and stored at -80°C. DNA was extracted from frozen, pulverized fecal samples by bead beating followed by phenol-chloroform extraction, as described previously (S13). Methods for generating and performing multiplex pyrosequencing of amplicons from variable region 2 (V2) of bacterial 16S rRNA genes are described in (S13). Bacterial V2 16S rRNA gene sequencing data were pre-processed to remove sequences with low quality scores, sequences with ambiguous characters, and sequences outside the length bounds (200-300 nt). All subsequent data processing and analyses were done using QIIME software (S14). Pyrosequencing ‘noise’ was removed with an algorithm implemented in QIIME. 16S rRNA reads were binned according to their sample-specific, error-correcting barcode incorporated into the reverse primer. Similar sequences were binned into phylotypes using CD-HIT with minimum pairwise identity of 97% (S15).

Aliquots of freshly produced as well as 30 day-old FMP from 6 batches of each flavor were sent directly from the pilot production plant to one of our labs using the same shipping protocol that was used to deliver the FMP to study participants. DNA was extracted and amplicons from the V2 region of bacterial 16S rRNA were generated and sequenced using the protocols described above. 49,959 high quality reads were obtained from a total of 33 FMP samples (1,332±187 reads/sample (mean ± S.D)): 43,729 reads of these were classified using the GreenGenes database (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi) and belonged to the genera Streptococcus, Lactococcus, Lactobacillus or Bifidobacterium.

**Co-occurrence analysis.** To determine whether there were statistically significant associations between the presence of *B. animalis* subsp. *lactis* and the occurrence of resident gut bacterial species-level phylotypes in human fecal samples, a co-occurrence analysis was performed using software tools present in QIIME under the script `otu_category_significance.py`. We used this
script to employ an ANOVA test to search our fecal 16S rRNA datasets for phylotypes whose relative abundances were higher in samples in which *B. animalis* subsp. *lactis* was present versus samples in which this strain was absent, as determined using qPCR. To avoid biases that might be introduced by differences in sample sequencing depth, we randomly selected an even number of sequences/sample (1,644 sequences) prior to performing the analysis. The raw *p*-values were corrected for multiple tests using the false discovery rate (fdr) correction (*S16*). We also performed the analysis at the genus and family levels by binning all operational taxonomic units (OTUs) that mapped to the same family or genus based on classification with the RDP classifier, using the `summarize_taxa.py` script in QIIME.

**Shotgun sequencing of fecal microbiomes.** Forty-eight fecal samples from four twin pairs were selected for multiplex shotgun pyrosequencing of total community DNA (454 FLX chemistry). For each individual, two fecal samples were analyzed before initiation of FMP consumption, two samples during the period when FMP was being consumed, and two samples after consumption had ceased.

Each fecal community DNA sample was randomly fragmented to an average length of 500 bp by nebulization, and then labeled with a distinct multiplex identifier (MID; Roche) using the manufacturer’s protocol. Equivalent amounts of 12 MID-labeled samples from each family were pooled prior to each pyrosequencing run. Shotgun reads were subsequently filtered using publicly available software (*S17*) to remove (i) all reads less than 60 nt, (ii) LR70 reads with at least one degenerate base (N) or reads with two continuous and/or three total degenerate bases, (iii) all duplicates, defined as sequences whose initial 20 nt were identical and shared an overall identity of >97% throughout the length of the shortest read, and (iv) all sequences with significant similarity to human reference genomes (i.e., having a BLASTN match with *E*-value <10^-5, bitscore ≥50, and percent identity ≥75%) to ensure continued de-identification of samples.

Reads obtained from shotgun sequencing of the twins’ fecal microbiomes were used to query v54 of the KEGG GENES database using BLASTX (*E*-value <10^-5, bitscore >50, and
A comparable annotation was performed for published fecal microbiomes that had been generated from 124 deeply sampled unrelated adult Europeans (S18), and from a pair of obese adult MOAFTS twins (S19).

**Microbial RNA-Seq analysis of human fecal metatranscriptomes.** Methods used to prepare fecal RNA, generate RNA-Seq data, and perform computational analyses of these datasets are described in detail below and in one of our previously published studies (S6, Fig. S4B). The only difference in the analysis pipeline used for human versus mouse RNA-Seq datasets was the choice of reference genomes. To map RNA-Seq reads from human fecal samples to genes, we used a collection of 131 bacterial genomes encompassing much of the diversity documented in human gut communities, plus the genomes of the five FMP strains; these genomes are listed in Table S15.

**Studies in gnotobiotic mice**

**Using INSeq to assay the determinants of fitness in a saccharolytic member of the 15-species model human microbiota.** INsertion Sequencing (INSeq) is a method based on a mutagenic transposon modified so that discrete fragments of adjacent chromosomal DNA can be captured when the transposon is excised from bacterial genomes by the restriction enzyme MmeI (S20). Sequencing fragments excised from a mixed population of tens of thousands of transposon mutants provides information about the location of each transposon in the genome. The number of occurrences of the transposon insertion site sequence mirrors the relative abundance of that mutant in the mixed population. By identifying mutants that significantly decrease in relative abundance after passage through a selective condition, INSeq allows a genome-wide map of in vivo fitness determinants to be created.

To determine whether introduction of the FMP strain consortium results in differences in the in vivo fitness requirements of a human gut symbiont present in the 15-member community, *B. thetaiotaomicron* strain VPI-5482 was mutagenized with the INSeq transposon (S20). A library of 34,544 randomly inserted transposon mutant strains covering 3,435 of the organism’s 4,779 genes was introduced, by gavage, together with the other 14 non-mutagenized members of the
community into germ-free mice. Fecal samples were subsequently collected from each mouse (n = 10) before (d7, d14), immediately after (d16), and 7 d after (d21) initial introduction of the FMP strains. INSeq libraries were prepared as described (S20) and sequenced using an Illumina GA IIx instrument (~1,000,000 36 nt reads/sample; Table S6). Resulting sequences were mapped to the *B. thetaiotaomicron* reference genome and quantified as described (S20). We found that insertions in 626 genes showed a significantly decreased relative abundance in the day 14 fecal microbiota (multiple hypothesis testing-corrected q<0.001), reflecting a fitness requirement for these genes in the colonization process (Table S6). Analysis of fecal samples collected just prior to, one day after, and 7 days after introduction of the FMP consortium established that exposure to the FMP strains did not impose significant new fitness pressures on specific genes present in this saccharolytic bacterial species.

**Animal sampling.** Fecal samples were obtained from each animal at the time points indicated in Fig. 1. Each fecal sample was collected directly as it emerged from the anus into a 1.7 ml screw cap Eppendorf tube, which was immediately deposited in a stainless steel dewer containing liquid N$_2$ (the dewer was introduced into the gnotobiotic isolator on the day of collection after it had been sterilized in the isolator’s entry port with chlorine dioxide spray (Clidox-S; PRL Pharmaceutical)). Various subsets of samples were subjected to COPRO-Seq, INSeq and microbial RNA-Seq analyses. Blood samples were collected into lithium heparin tubes (Becton Dickinson), placed immediately on ice, and then centrifuged (2,700 x g; 3 min at 4°C). The resulting plasma supernatant was stored at -80°C until assay. Urine was collected directly into Eppendorf tubes and immediately frozen in liquid N$_2$. Upon sacrifice, ceca were dissected and cecal contents were frozen immediately at -80°C.

**Isolation of DNA from cecal contents and feces.** Microbial community DNA was prepared in a two-step process consisting of a crude extraction step followed by additional purification and RNAse treatment.
Crude extraction. The sample (typically 25-100 mg of frozen feces or 50-125 mg of frozen cecal contents) was combined with 250 μl of 0.1 mm zirconium beads (BioSpec Products), 500 μl Buffer A (200 mM NaCl, 200 mM Tris, 20 mM EDTA), 210 μl SDS (20% v/v, filter-sterilized), and 500 μl phenol:chloroform:isoamyl alcohol (25:24:1, pH 7.9; Ambion), and the mixture was briefly chilled on ice. Samples were then disrupted using a Mini-BeadBeater-8 (BioSpec) set to 'homogenize' (bead-beating for 2 min at room temperature, followed by placement on ice for 1-2 min, followed by bead-beating for an additional 2 min). The aqueous phase (~600 μl) was collected after centrifugation (6,800 x g; 3 min at 4°C) and combined with an equal volume of phenol:chloroform in 2 ml ‘light’ phase-lock gel tubes (5Prime) per the manufacturer’s protocol. After spinning, the aqueous phase was combined with 1 volume of chilled 100% isopropanol (-20°C) and 1/10 volume sodium acetate (3 M, pH 5.5). Following incubation at -20°C for 1 h, the precipitate was pelleted (20,800 x g; 20 min at 4°C), washed in 100% EtOH, dried, and resuspended in 5 μl TE (pH 7.0) per milligram of original sample material.

RNase treatment and further purification. Aliquots of crude DNA were transferred to a 96-well plate. Buffer PM (Qiagen) was mixed with RNase A (Qiagen) to a final concentration of 1.3 mg/ml. Three volumes of this mixture were added to each well and the reactions were allowed to incubate at room temperature for 2 min. Following RNase digestion, samples were applied to a QIAquick 96 PCR purification plate (Qiagen) and processed according to the manufacturer’s instructions using a QIAvac 96 vacuum manifold. DNA was eluted in 100 μl of Buffer EB (Qiagen). DNA quality and purity were verified using a NanoDrop spectrophotometer (model ND-1000).

Preparing DNA libraries for Illumina sequencing and COPRO-Seq analysis in a 96-well format. DNA libraries were prepared for sequencing using a modified version of Illumina’s sample preparation protocol for generating libraries from genomic DNA. The six steps of this modified preparation are as follows:

(i) Fragmentation. Two micrograms of each purified DNA sample was suspended in 100 μl Buffer EB and fragmented by sonication in 1.7 ml Eppendorfs using the BioruptorXL multi-sam-
ple sonicator (Diagenode) set on ‘high’. Samples were sonicated over the course of 20 min using successive cycles of 30 sec ‘on’ followed by 30 sec ‘off’. Sonicated samples were subsequently cleaned up using the MinElute 96 UF PCR Purification Kit (Qiagen) per the manufacturer's instructions. Each sonicated DNA sample in each well of the 96-well plate was eluted with 22 μl Buffer EB.

(ii) ‘Add-only’ enzymatic modification. Ten microliters of each eluate from step (i) was transferred to a 96-well plate where it was subjected to enzymatic blunting. Each 20 μl reaction contained: 10 μl DNA, 2 μl T4 DNA ligase buffer [10X; New England Biolabs (NEB), catalogue number B0202S], 1 μl dNTPs (1 mM; NEB, N1201AA), 0.5 μl Klenow DNA polymerase (5 U/μl; NEB, M0210S), T4 PNK (10 U/μl; NEB, M0201S), and 6 μl molecular grade water. Blunting reactions were incubated (25°C, 30 min) then heat-inactivated (75°C, 20 min). Residual dNTPs were dephosphorylated by adding 1 μl of shrimp alkaline phosphatase (1 U/μl; Promega, M820A) to each reaction. Reactions were incubated (37°C, 30 min) and heat-inactivated (75°C, 30 min). Adenine tailing reactions were set up in 30 μl reaction volumes that contained 21 μl of the inactivated phosphatase reaction, 6.4 μl T4 DNA ligase buffer (diluted to 1X; NEB, B0202S), 0.6 μl dATP (5 mM), and 2 μl Klenow 3’->5’ exo’ (5 U/μl; NEB, M0212L). Reactions were incubated (37°C, 30 min) and heat-inactivated (75°C, 20 min).

(iii) Ligation. Customized Illumina adapters containing maximally distant 4 bp barcodes described elsewhere (S6) were ligated to the polyA-tailed DNA in 50 μl reactions as follows. Thirty microliters of the inactivated A-tailing reaction described in the preceding paragraph was added to 5 μl T4 DNA ligase buffer (10X), 5 μl adapter mix (1 μM final concentration per adapter), and 9 μl water at 4°C. One microliter of T4 DNA ligase (2,000,000 U/μl; NEB M0202M) was subsequently added and reactions were incubated (16°C, 1 h), followed by heat-inactivation (65°C, 10 min). Ligation reactions were cleaned up using the MinElute 96 UF PCR Purification Kit (Qiagen) according to the manufacturer’s recommended protocol. DNA was eluted in 22 μl Buffer EB.
(iv) **Gel purification.** 10 μl of each elution was separated by gel electrophoresis on 2% agarose. DNA migrating at 200 bp was excised and gel slices were purified using a QIAquick 96 PCR Purification Kit (Qiagen).

(v) **PCR amplification.** Each library was PCR amplified for 19 cycles using Illumina’s standard amplification primers with modifications to impart barcode-specificity (S6) and Illumina’s recommended amplification conditions/reagents. Products were purified using a QIAquick 96 PCR Purification Kit (Qiagen). An aliquot of this purification was also subjected to 2% agarose gel electrophoresis to confirm the absence of any significant adapter-dimer contamination.

(vi) **Library pooling and sequencing.** The concentration of each purified library was quantified using the Qubit dsDNA HS Assay Kit (Invitrogen). Barcoded libraries were subsequently pooled (typically in groups of 16) at equivalent final concentrations. Sequencing was performed using the standard Illumina GA IIx sequencing protocol, with libraries loaded on the flow cell at a concentration of 2.0-2.5 pM).

A custom software pipeline was written in Perl for performing COPRO-Seq data processing in a computer cluster environment running Sun Grid Engine. These data processing steps are schematized in Fig. S4A. Briefly, raw Illumina GA IIx reads from a sequencing pool were first de-plexed by barcode and trimmed to 34 bp (30 bp genome sequence + 4 bp barcode). Trimmed reads were aligned to the genomes of the 20 microbial strains used in this study using Illumina’s ELAND aligner. Perfect, unique alignments to the reference genomes were retained, while those mapping less than perfectly or having multiple possible alignments to the reference genomes were filtered out, ensuring that only high-quality, unambiguous reads were used. Hits to each genome were then tallied, after which the summed counts for each genome were normalized by that genome’s ‘informative genome size’ (term defined in Results) to adjust for both genome size and uniqueness relative to all other genomes in the experiment. The Perl scripts supporting the COPRO-Seq analytic pipeline can be downloaded from: http://gordonlab.wustl.edu/projects/2011-McNulty_etal.
**Characterizing gene expression with microbial RNA-Seq.** Following extraction of total nucleic acid with phenol-chloroform, and precipitation with isopropanol, fecal samples were subjected to DNAse digestion (S6). Total RNA was then: (i) passed through a MEGAClear column (Ambion) to deplete RNAs <200 nt (removing most 5S rRNA and tRNA species; (S6)); (ii) subjected to another round of DNAse digestion; (iii) passed through another MEGAClear column; and (iv) subjected to a hybridization-based pulldown of 16S and 23S rRNA using custom-designed biotinylated oligonucleotides that contain short rDNA sequences conserved across a set of 37 human gut-derived sequenced microbial genomes (S6). The depletion protocol, which has been adapted to 96-well format, is described elsewhere (S6). PCR (30 cycles) targeting a short (<100 bp) region from the genome of an abundant community member (typically *Bacteroides WH2*) was used to verify the absence of detectable gDNA in each RNA preparation.

Doubled stranded cDNA (dscDNA) was synthesized using random hexanucleotide primers. At the conclusion of the reaction, Illumina adapters containing sample-specific 4 nt barcodes were ligated to the dscDNA. Multiplex sequencing was performed using the Illumina GA IIx instrument. We typically sequenced two barcoded *in vitro* samples/lane of the 8-lane flow cell; *in vivo* samples were not multiplexed (i.e. 1 sample was analyzed per sequencing lane). This allowed us to identify mRNA present at levels representing ≥0.001% of all reads.

The pipeline for processing microbial RNA-Seq data is presented in **Fig. S4B**. The 8-20 million 36 nt cDNA reads from each sequencing lane were separated by barcode, and mapped against the relevant set of genome sequences using the SSAHA2 algorithm (S21) to determine the raw unique-match ‘counts’ (reads) for each gene present in the relevant microbial genome or microbiome. Reads that mapped non-uniquely were added to each gene in proportion to each gene’s fraction of unique-match counts (e.g., a non-unique read that maps equally well to gene A with 18 unique reads and gene B with 2 unique reads would be scored as 0.9 of a count to gene A, and 0.1 of a count to gene B; the influence of ties is negligible for RNA-Seq given the small numbers of distinct genomes, but would become more important with more complex communities). Raw counts were then normalized to reads/kb gene length/million mapped reads (RPKM) using one or more gene position file(s) in conjunction with custom Perl scripts.
Data normalization was carried out at two different levels in this study. For our ‘top-down’ analysis, data were normalized at the level of the entire community metatranscriptome (i.e., raw counts from all species were normalized simultaneously using a single gene position file that included the positions of all genes in the model community metatranscriptome). Data normalized in top-down fashion allowed us to determine, after binning gene expression values by function, how the collective operations of the model community were changing as the result of experimental perturbations. In our ‘bottom-up’ analysis, data were normalized at the level of individual species (i.e., raw counts from each individual species were normalized separately from one another, in each case using a species-specific gene position file describing the positions of only that species’ genes). Data normalized in bottom-up fashion allowed us to interrogate what statistically significant gene expression changes were occurring within a given species of interest.

Bottom-up normalized transcript data were analyzed by Cyber-T (S22) to identify mRNAs that exhibited significant differences in their levels of expression between samples. For each comparison, transcripts were then binned into a list where the magnitude of the difference in their expression was greater than or equal to 4-fold. Binned transcripts were subsequently annotated using the kegg_counting.pl Perl script described above. Each resulting annotated dataset was used to determine the representation of KEGG pathways and categories within these lists.

Further functional comparisons were carried out using ShotgunFunctionalizeR, an R package designed to analyze differences between metagenomic datasets using a Poisson statistical model (S23). The RPKM-normalized counts for each gene in the model microbiome were then binned by EC number using each gene’s previously assigned annotation. Summed reads in each EC bin were rounded to the nearest integer, and the data were imported into ShotgunFunctionalizeR, which was used to identify ECs that were differentially expressed.

Identification of predictive KEGG categories, pathways and ECs using a Random Forests classifier. To identify KEGG categories, ECs, or pathways that were significantly differentiated across treatment states, we used the Random Forests classifier (S24) described in (S25). Mouse
samples were divided into 10 pre-treatment samples (experimental day 14) and 20 post-treatment samples (experimental days 15 and 42). To estimate the generalization error of the classifier we used leave-one-out cross-validation, in which each sample’s group was predicted by a classifier trained on the other 29 samples. Training was done using default settings for the randomForest package in R (S24). Each feature’s predictiveness was estimated by calculating the mean increase in estimated generalization error when the values of that feature were permuted at random. Features whose removal caused an average error increase of at least 0.1% were labeled as ‘predictive’; those with an increase of at least 1% were labeled as ‘highly predictive’.

**Non-targeted metabolomics via gas chromatography/mass spectrometry (GC/MS).** Urines were first assayed for creatinine content as measured by a modified Jaffe method using the three microliter “random-urine” routine and CR-S 3000 reagent on the UniCel DxC 600 Synchron Clinical System (Beckman Instruments, Brea, CA). A urine volume equivalent to 0.2 micromoles of creatinine was then aliquoted and spiked with perdeuterated myristic acid (D$_{27}$-C14:0) as an internal standard for retention-time locking (RTL IS). Following treatment with 7.5 volumes of methanol, the mixture was centrifuged and the supernatant was decanted and dried.

Derivatization of all dried supernatants for GC/MS followed a method adapted with modifications from that of (S26). Reagents were from Sigma-Aldrich (St. Louis, MO), unless otherwise noted. Briefly, certain reactive carbonyls were first methoximated at 50ºC with a saturated solution of methoxyamine hydrochloride in dry pyridine, followed by replacement of exchangeable protons with trimethylsilyl (TMS) groups using N-methyl-N-(trimethylsilyl) trifluoroacetamide with a 1% v/v catalytic admixture of trimethylchlorosilane (Thermo-Fisher Scientific, Rockford, IL) at 50ºC.

GC/MS methods generally followed those of Fiehn (S27) and Kind (S28), and used a 6890N GC connected to a 5975 Inert single-quadrupole MS (Agilent, Santa Clara, CA). A large-volume, ProSep inlet enabled programmed-temperature vaporization and diversion of heavy contaminants away from the GC and MS, as described below, greatly reducing maintenance time (Apex Technologies, Inc., Independence, KY). The two wall-coated, open-tubular GC columns connected in
series were both from J&W/Agilent (part 122-5512), DB5-MS, 15 m in length, 0.25 mm in diameter, with a 0.25 µm luminal film. Prior to each run, initial inlet pressures were empirically adjusted such that the resulting retention time (RT) of the TMS-D27-C14:0 standard was set at ~16.727 min. Under these conditions, derivatized metabolites eluted from the column and reached the electron-ionization (EI) source in the MS at known times (e.g., bis-TMS-lactic acid at ~6.85 min, and TMS-cholesterol at ~27.38 min). A mid-column, microfluidic splitter (Agilent) provided a means for hot back-flushing of the upstream GC column at the end of each run while the oven was held at 325ºC for a terminal “bake-out” (another antifouling and anti-carryover measure analogous to that described in (S29)). During this terminal “bake-out,” the inlet was also held at 325ºC, and it was purged with a large flow of the carrier gas, helium. Positive ions generated with conventional EI at 70 eV were scanned broadly from 600 to 50 m/z in the detector throughout the run.

Raw data from Agilent’s ChemStation software environment were imported into the freeware, Automatic Mass Spectral Deconvolution and Identification Software (AMDIS) (developed by Drs. Steve Stein, W. Gary Mallard, and their coworkers at National Institute of Standards and Technology (S30-32); http://chemdata.nist.gov/mass-spc/amdis/). Deconvoluted spectra were identified, to the extent possible, using several commercial and public spectral libraries. Our primary source was the Fiehn GC/MS Metabolomics RTL Library (a gift from Agilent Technologies, Santa Clara, CA, part number G1676-90000). Additional spectra for comparison were gleaned from the Golm Metabolome Library (courtesy of Dr. Joachim Kopka and coworkers at the Max Planck Institute of Molecular Plant Physiology, Golm, Germany (S33); http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html), the commercial NIST/EPA/NIH Mass Spectral Library and our own purpose-built spectral library. Where indicated, peak alignment was performed with SpectConnect freeware (courtesy of Dr. Gregory Stephanopoulos, Massachusetts Institute of Technology (S34); http://www.spectconnect.mit.edu). Chemometrics were performed with Mass Profiler Professional (a recent descendant of GeneSpring MS, purchased from Agilent), along with our own custom macros, written in Visual Basic for use in the Excel software environment.
The statistical significance of differences in the log$_2$ spectral abundances of each metabolite in samples obtained at different time points was tested using two approaches. A first-pass, highly-permissive set of pairwise comparisons was calculated between each combination of samples (d0 versus d14, d0 versus d42, d14 versus d42) using a simple two-tailed Student’s $t$-test. The resulting $p$-values, which were not corrected for multiple hypothesis testing, are listed in Table S12. Given the high ratio of hypotheses tested to samples per group (198 metabolites; 5-8 samples per group), we also guarded against false discovery by performing a more stringent set of calculations that produced a shorter list of metabolites with significant differences in abundance. This latter procedure consisted of first taking metabolite data from all three time points and subjecting them to a one-way ANOVA. The resulting $p$-values were then adjusted using Benjamini-Hochberg correction, generating $q$-values. The log$_2$ spectral abundances of all metabolites whose $q$-values were below 0.05 were then subjected to Tukey’s HDS (Honestly Significantly Different) post-hoc test to determine which time points were significantly different from one another. All Tukey’s HDS $p$-values that were calculated are provided in Table S12.

Supplementary results

**Human studies**

*Analysis of the effects of FMP consumption on stool consistency, difficulty of passage, and frequency.* To determine whether there were differences in stool consistency and difficulty of passage of stools between pre-treatment, treatment, and post-treatment study periods, we first constructed a dataset in which the unit of analysis was ‘bowel movement’. Using ordinal logistic regression, we analyzed separate models predicting stool consistency and difficulty of stool passage using ‘treatment period’ as the reference group. We adjusted for clustering of observations using a Huber-White robust variance estimator (STATA 2004). When data from the entire study period were included in the analyses, no significant differences were observed between study periods for either stool consistency or difficulty passing stool. Next, we conducted an alternate
analysis in which data from the first two weeks of the treatment and of the post-treatment phases were omitted. We found that women had lower stool consistency scores during the last two weeks of the post-treatment phase compared to the last two weeks of the treatment phase: i.e., stools were softer during the treatment period (OR=0.69; \( p=0.04 \)). The difference in stool consistency between the pre-treatment phase and the last two weeks of the treatment phase was not significant; however, there was a significant difference between the pre-treatment versus treatment compared to the post-treatment versus treatment odds ratios (\( p=0.005 \)).

Analyses for stool frequency were conducted similarly to those above with the exception that the unit of analysis was the ‘person-day’ (i.e., one observation per person per day) and the dependent variable was number of bowel movements per day. We did not find stool frequency to be associated with study period regardless of which study days were included in the analysis.

One participant had a diarrheal illness on three of the FMP treatment days, with a dramatic increase in stool frequency and decrease in stool firmness on these days. She reported taking four 2 mg tablets of loperamide [4-(p-chlorophenyl)-4-hydroxy-N, N-dimethyl-α, α-diphenyl-1-piperidinebutyramide monohydrochloride] to relieve her symptoms during this period. Therefore, data from these bowel movements were excluded from the analyses.

**Co-occurrence analysis.** As noted above, to identify species-level phylotypes that consistently increase or decrease in abundance when *B. animalis* subsp. *lactis* is present in human fecal samples, we performed a co-occurrence analysis using QIIME (see Supp. Methods). This analysis indicated that no OTUs present in the pre-treatment microbiota exhibited a statistically significant change in their proportional representation in feces during the period of FMP consumption or during the post-treatment period in any individual after correction for multiple tests. The OTU that most nearly achieved significance was closely related to *Lactococcus lactis* (raw \( p=0.00067 \); ANOVA: \( p>0.05 \) after FDR correction). A follow-up co-occurrence analysis for all genera also identified the genus Lactococcus as being significantly more abundant when *B. animalis* subsp. *lactis* was present (the latter determined by qPCR). It is reasonable that *L. lactis* would co-occur
with *B. animalis* subsp. *lactis* given the presence of both strains in the FMP. A co-occurrence analysis performed at the family level of taxonomy failed to identify any significant differences.

Our ability to identify *L. lactis* in our co-occurrence analysis was encouraging, but raised the question of why an OTU representing *B. animalis* subsp. *lactis* did not achieve significance, given that nearly every sample collected during the period of FMP consumption was positive for this strain by qPCR. Of the 58 samples deemed positive for *B. animalis* subsp. *lactis* by qPCR, only 7 yielded an OTU in our 16S rRNA dataset with a 100% identity match to *B. animalis* subsp. *lactis*. This OTU was not detected in any of the samples deemed negative by qPCR. Therefore, we concluded that the discrepancy between *B. animalis* subsp. *lactis* being called ‘present’ by qPCR and by 16S rRNA sequencing was due to inadequate depth of sequencing. Extrapolating, there could be other rare species whose changes in abundance were not detected. To explore this latter possibility, the fecal microbiota of two healthy MZ twin pairs, similar in age and body mass index, but with marked differences in their degree of geographical proximity (*Table S1*), were subjected to deeper sequencing (n=36 samples, yielding an additional 411,177 16S rRNA sequences, resulting in 14,241±2,144 (mean±SD) reads/sample from these individuals). No significant changes at any level of bacterial taxonomy were observed in this small sample dataset.

**Studies in gnotobiotic mice**

**Measurement of adiposity.** The body weights and epididymal fat pad weights of mice from both treatment groups were measured at the time of sacrifice. We observed no significant differences between the single and multiple treatment groups in either measurement (*p*=0.6865, *p*=0.3516, respectively; two-tailed Student’s *t*-test). Furthermore, all measurements of adiposity and weight were in line with those of mice from other studies that had involved animals from the same inbred strain, who were similarly aged, the same gender, on the same diet, and who harbored comparable defined model human gut microbiota (without FMP strains).
In vitro studies

RNA-Seq profiling of B. animalis subsp. lactis during growth in vitro. Sequencing of transcripts expressed by B. animalis subsp. lactis during mid-log phase growth in MRS medium (1.5-2.9 million reads per technical replicate; n=2 independent cultures) detected products from 1,618 of the organism’s 1,660 predicted genes, while profiling during late stationary phase indicated that 1,609 of its genes were expressed. The transition from logarithmic to stationary phase was accompanied by significant up- or down-regulation of 98 and 194 genes, respectively, including those involved in various aspects of carbohydrate, amino acid and nucleotide metabolism (see Table S7A).
Supplementary references


Supplementary figure legends

Figure S1. Levels of *B. animalis* subsp. *lactis* (CNCM I-2494) in human fecal samples collected prior to, during, and after consumption of a FMP. (A) qPCR assays; each dot represents a sample from a given individual. The green bar denotes the period of FMP consumption. (B) Comparison of qPCR results to the number of shotgun reads mapped to the genomes of three *B. animalis* subsp. *lactis* strains. qPCR results are plotted on the X-axis, while the proportional representation of reads that mapped to the *B. animalis* subsp. *lactis* genomes is presented on the Y-axis.

Figure S2. KEGG pathway coverage ratios suggest that the model human gut microbiome encodes many of the functions present in more complex human fecal communities. Genes in (i) the genomes of the five-member FMP strain consortium, (ii) the 15-member model human gut microbiota, (iii) a highly simplified two-member human gut microbiota composed of *B. thetairotaomicron* and a prominent gut Firmicute (*Eubacterium rectale*) (S35), (iv) a reference set of 127 sequenced human gut microbial isolates, (v) the deeply sampled fecal microbiomes of 124 unrelated adult Europeans from a recent METAHIT study (S18), (vi) the fecal microbiomes of the 7 twin pairs characterized in the present study, and (vii) the deeply sequenced fecal microbiomes of an obese adult MZ twin pair (S19) were annotated using v54 of the KEGG GENES database. The presence/absence of each KO in each KEGG pathway was determined for every set of genes and the pathway coverage ratio (i.e., % of a pathway’s components called ‘present’; BLASTP *E*-value <10⁻⁵) was calculated and depicted as the heatmap shown using Cluster 3.0/Treeview.

Figure S3. CAZyme profiles of the 20 bacterial strains introduced into gnotobiotic mice. The indicated genomes were annotated for all glycoside hydrolases (GH), glycosyltransferases (GT), carbohydrate binding modules (CBM), and polysaccharide lyases (PL) using the CAZy classification scheme. The Bacteroides possess a larger and more diverse arsenal of CAZymes relative to the Firmicutes/Actinobacteria, though most CAZyme families encoded in the genomes of the FMP strains were also present in defined community members. The small number of FMP strain-specific CAZyme families (CBM5, CBM10, CBM23, CBM33, GH85, and GT39) are highlighted in red. The scale refers to the number of genes in a given CAZy family in a given genome.
Figure S4. Summary of analysis pipelines utilized in this study. (A) COPRO-Seq. (B) RNA-Seq.

Figure S5. COPRO-Seq-based time series analysis of the abundance of members of the model human microbiota and of the FMP strain consortium in the feces of gnotobiotic mice. Relative abundance, expressed as the log_{10} of percent representation of all detected community members was calculated over time (d0, time of colonization with the model 15-member community; d14, time of first gavage with the FMP consortium for the single and multiple treatment groups; d21 and d35, times of subsequent gavage with the FMP consortium for the multiple treatment group). For each treatment, animals were gavaged twice over a 24 h period. Mean values ± SEM are plotted (n=5 animals/treatment group; 1 fecal sample/animal/time point; limit of detection = 0.003%). (A) COPRO-Seq data for 13 members of the 15 member community (F. prausnitzii and C. spiroforme were at levels below the limit of detection throughout the study and are not shown). (B) Data obtained from the two members of the FMP consortium that persisted at levels above the limit of detection following their introduction into mice. (C) Data from panel A representing the response of C aerofaciens to introduction of the FMP strain consortium (see text for details).

Figure S6. Top-down analysis of the model community’s transcriptional response to the FMP strain consortium reveals upregulation of genes involved in interconversion of propionate and succinate. RPKM-normalized RNA-Seq data were binned at the level of EC, and comparisons were made between early responses (d14 versus d15, representing time points just before and 1 day after gavage with the strain consortium) and late responses (d14 versus d42). Boxes and lines are colored according to the key shown above the pathway map and in the legend to panel B of Fig. 5.

Figure S7. A species’ contribution to the metatranscriptome is not necessarily proportional to its abundance in the 15-member community. Microbial RNA-Seq data from d14 of the mouse study were parsed by species to determine the total number of reads that each community member contributed to the total sequenced transcript pool (‘metatranscriptome’) (both raw and normalized
reads as defined in Fig. S4B). Data were further broken down into reads that could be mapped to genes with known functions (as defined by KEGG) and those with unknown functions (lacking any KO in v54 of the KEGG GENES database). Mean values ± SD are plotted for each of the four types of data presented. Note, for example, the large number of raw reads attributed to *R. obeum* despite its low proportional abundance in the community. Conversely, *Bacteroides WH2* contributes a far smaller proportion of total raw RNA-Seq reads to the pool than its relative abundance in the community might have suggested.

**Figure S8. Bottom-up analysis of genes whose expression changes significantly following introduction of the FMP strain consortium.** (A) Volcano plots of the >48,000 expressed genes detected in at least one fecal RNA sample. Colored points represent genes whose difference in expression following introduction of the FMP strain consortium was (i) ≥4-fold (increase or decrease) relative to the d14 pre-treatment time point and (ii) statistically significant (*p*<0.05; two-tailed Student’s *t*-test). Dots are colored according to each gene’s species of origin (color key shown to the right of the panel). Black dots represent genes whose change in expression is less than 4-fold at the time points indicated and/or not statistically significant. See Table S10 for a complete list of all the genes shown as colored dots. (B) Breakdown by species of differentially expressed genes. (C) Breakdown by KEGG category of differentially expressed genes, showing that late responses are more numerous than immediate ones, and that there is a noticeable bias towards genes involved in carbohydrate and glycan metabolism, particularly in the late response to introduction of the FMP strain consortium.

**Figure S9. The number of RNA-Seq reads obtained from human fecal samples that map to genomes in the FMP strain consortium peaks shortly after FMP consumption begins.** RNA-Seq reads derived from selected human samples were mapped back to the five genomes in the FMP consortium to determine which species’ transcripts could be detected over time. *B. animalis* transcripts were detected only during periods of FMP consumption. Reads attributed to *L. lactis*, *S. thermophilus*, and *L. delbrueckii* at time points before consumption began may reflect ‘spurious’ mapping to related endogenous strains in the gut community. Facet labels located at the top
of each bar chart correspond to the sample labels shown in Fig. 1A of the main text. Numbers are presented below each bar chart correspond to the code number assigned to each de-identified co-twin (see Table S1).
Supplementary figures

Figure S1.

A

Abundance of *B. animalis* subsp. *lactis* log₁₀(cell equivalents/g feces)

Detection Limit

B

*B. animalis* subsp. *lactis* abundance by direct sequencing log₁₀(% of reads mapping to *B. animalis* subsp. *lactis* genomes)

B. *animalis* subsp. *lactis* abundance by qPCR log₁₀(cell equivalents/g feces)

\[ R^2 = 0.62 \]
Figure S2.
Figure S3.
Figure S4.

A COPRO-Seq library preparation and sequencing

- DNA isolation
- Illumina library preparation with barcoded adapters
- Illumina GA IIx sequencing

Mapping and normalization

- Split reads by barcode
- Map reads to reference genomes (ELAND)
- Perfect, unique hits to references
- Non-mapping reads
- Map to other possible sources (adapters, mouse genome)
- Generate mapping stats report

Informative genome size (IGS) calculations

- Create FASTA of all k-mer sequences a reference could generate
- Map k-mers to all reference genomes used in COPRO-Seq library mapping
- Count k-mers that hit uniquely (informative k-mers)
- Multiply unique fraction by genome length to determine IGS
- Repeat for each reference genome

Normalizing

- Normalize each species sum by its IGS

Selective plating

- Selective plating of one community member
- Proportional abundance
- Absolute abundance

B Library preparation and sequencing

- RNA isolation
- 5S/16S/23S rRNA, tRNA depletion
- cDNA synthesis
- Illumina library preparation
- Illumina GA IIx sequencing

Mapping and normalization

- Map reads to reference genomes (SSAHA2)
- Sum hit counts by gene

RPKM normalization by community (metatranscriptome)

- List of metatranscriptome-normalized transcript counts

List of species-normalized transcript counts

Annotation

- BLASTP predicted proteins from reference genomes against v34 of KEGG database

Bin genes into KEGG Hierarchies and sum counts

Data analysis

- ShotgunFunctionalizerR analysis of EC level changes
- Analysis of changes in KEGG functional hierarchies

A 'Bottom-up' analysis

B 'Top-down' analysis
Figure S5.

A

B

C

B. animalis

L. lactis subsp. cremoris

B. caccae

C. aero-
faciens

B. ovatus

B. thetaiota-
omicron

B. uniformis

B. vulgatus

P. distasonis

R. obeum

R. torques

C. aero-
faciens

C. scindens

D. longi-
catena

E. rectale

99
Figure S6.

1 day after FMP strain introduction vs pre-FMP
4 weeks after FMP strain introduction (multiple treatment) vs pre-FMP
1 day after FMP strain introduction vs pre-FMP
4 weeks after FMP strain introduction (multiple treatment) vs pre-FMP

Propanoyl phosphate → Propanoate → Propanoyl-CoA

(S)-2-Methyl-malonyl-CoA → EC 5.1.99.1
(R)-Methyl-malonyl-CoA → Succinyl-CoA → Succinate

Propionyl-adenylate → Propionyl-CoA → Acryloyl-CoA

EC 2.7.2.1
EC 2.3.1.8
EC 4.1.1.41
EC 2.8.3.1
EC 6.2.1.1
EC 6.4.1.3
EC 1.3.1.-
EC 6.2.1.5
EC 6.4.1.3
EC 5.4.99.2
Figure S7.
Figure S8.

A

'Early' (d14-v-d15) response

'Late' (d14-v-d42) response

B

C

Species

-log_{10}(p-value)

-log_{10}(p-value)

Species

-log_{10}(p-value)

Species

-log_{10}(p-value)

Species

-log_{10}(p-value)

Species

# genes differentially expressed

Species

# genes differentially expressed

Species

# genes differentially expressed

Species

# genes differentially expressed

Species

# genes differentially expressed

Species

# genes differentially expressed

Species

# genes differentially expressed

Species

# genes differentially expressed

Species

# genes differentially expressed

Species

# genes differentially expressed

Species

# genes differentially expressed

Species

# genes differentially expressed
Figure S9.
Supplementary table legends

Table S1. Characteristics of adult female monozygotic (MZ) twins enrolled in study.

Table S2. Summary of human fecal metagenomic datasets. (A) Multiplex pyrosequencing of fecal bacterial 16S rRNA V2 amplicons. (B) Multiplex shotgun pyrosequencing of total fecal community DNA.

Table S3. Features of the microbial genomes in the 5-member FMP strain consortium and the 15-member model human gut microbiota.

Table S4. Carbohydrate active enzyme (CAZy) annotation data. (A) CAZy summaries by genome. (B) CAZy annotations for the 20 bacterial species in this study.

Table S5. COPRO-Seq analysis of bacterial species abundance in mouse fecal samples. (A) Proportional representation of the 20 bacterial species in this study in mouse fecal samples as measured by COPRO-Seq. (B) Statistical significance and fold-change of differences in pairwise comparisons of abundance calculated from data in panel A. The group mean for each day/treatment/species combination at time points after the introduction of the FMP strain consortium was compared to the mean for the same treatment/species at d14 (the last time point collected prior to introduction of the FMP strains) using Welch’s t-test. Values have not been corrected for multiple hypothesis testing. p-values <0.05 are highlighted in pink. Fold-changes greater than 2 or less than -2 are highlighted in pink and green, respectively.

Table S6. INSeq analysis. (A) INSeq analysis sequencing statistics. Scale factor corresponds to counts per million normalization; underrepresented samples were re-sequenced and combined with original data so that all samples were represented by ~1 million reads. (B) Genes required by B. thetaiotaomicron for survival in the intestines of mice harboring the 15-member model human gut microbiota. The table describes the relative abundance of transposon insertions in each gene (rows) in the input community (average of two independent technical replicates) and in the output communities (fecal samples collected from 10 mice two weeks after introduction of the synthetic
model community, immediately prior to introduction of the FMP strain consortium). A z-test was used to identify genes whose log-transformed output to input ratios were significantly different from the overall distribution (a uniform value of 1 was added to all counts, and genes with no insertions were removed to allow ratios to be calculated). Resultant p-values were corrected for multiple hypothesis testing by q-test (S36). Genes assigned a q-value <0.001 are highlighted in red.

Data filtering, normalization, mapping, and statistical analysis were conducted in Perl and Matlab.

**Table S7. Differentially expressed B. animalis subsp. lactis (CNCM I-2494) genes.** (A) Log versus stationary growth in MRS medium. (B) *In vivo* (mouse) versus *in vitro* (logarithmic phase in MRS) growth.

**Table S8. Top-down function-level analysis of the impact of the FMP strain consortium on the model human gut microbiota’s metatranscriptome.** (A) Proportional representation of assignable normalized RNA-Seq counts binned by KEGG category in fecal samples collected over time from singly- and multiply-treated animals. (B) Proportional representation of assignable normalized RNA-Seq counts binned by KEGG pathway in fecal samples collected over time from singly- and multiply-treated animals.

**Table S9. Model human gut microbiota membrane transport genes demonstrating ≥4-fold increases or decreases in their expression following introduction of the FMP strain consortium.**

**Table S10. Bottom-up (gene-level) analysis of the impact of the FMP strain consortium on the model community’s metatranscriptome.** (A) Breakdown by microbial species of significantly up/down-regulated genes. (B) Breakdown by KEGG category of significantly up/down-regulated genes. (C) Model community microbiome genes demonstrating significant increases/decreases in their expression following introduction of the FMP strain consortium.

**Table S11. Results of Random Forests supervised classification analysis.**
**Table S12.** Urine metabolites whose levels change significantly in transitions between colonization states. The ‘Reverse match score’ column contains the AMDIS, dot-product, reverse-match scores (S32) which in this case evaluate not only mass-spectral concordance, but also the goodness of fit of chromatographic retention-time index made by comparison to (i) commercial and public target-compound libraries of small metabolites (S28, 33), and (ii) our own in-house reference library. Metabolites with match scores less than an arbitrary threshold of 65% were excluded from these results. (A) Day 0 (germ-free) versus day 14 (colonized with 15-member model community). (B) Day 14 (colonized with 15-member model community) versus day 42 (colonized with 15-member model community plus five-member FMP consortium).

**Table S13.** ShotgunFunctionalizeR analysis of EC-level changes in the metatranscriptome as a function of FMP strain introduction into mice and humans. The table shows fold-change in mean proportional representation of each EC between groups for each comparison. Values for non-significant EC changes in a comparison (adjusted $p \geq 0.01$, ShotgunFunctionalizeR) are reported as “NS”. Note that ECs can have multiple KEGG pathway and category assignments.

**Table S14.** Primers and amplification conditions used for quantitative PCR assays of FMP consortium strains in fecal DNA. (A) Primers used to assay human fecal samples. (B) Primers used to assay mouse fecal samples. (C) Amplification conditions.

**Table S15.** List of 136 microbial genomes used to analyze human fecal RNA-Seq data.
Chapter 3

Effects of diet on resource utilization by a model human gut microbiota containing *Bacteroides cellulosilyticus* WH2, a symbiont with an extensive glycobiome
Chapter 3

Effects of diet on resource utilization by a model human gut microbiota containing \textit{Bacteroides cellulosilyticus WH2}, a symbiont with an extensive glycobiome

Nathan P. McNulty\textsuperscript{a}, Meng Wu\textsuperscript{a}, Alison R. Erickson\textsuperscript{b,c}, Chongle Pan\textsuperscript{a}, Brian K. Erickson\textsuperscript{e}, Eric C. Martens\textsuperscript{d}, Nicholas A. Pudlo\textsuperscript{d}, Brian D. Muegge\textsuperscript{a}, Bernard Henrissat\textsuperscript{e}, Robert L. Hettich\textsuperscript{b,c}, Jeffrey I. Gordon\textsuperscript{a}

\textsuperscript{a}Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108, USA; \textsuperscript{b}Graduate School of Genome Science and Technology, University of Tennessee-Oak Ridge National Laboratory, Knoxville, TN 37996, USA; \textsuperscript{c}Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA; \textsuperscript{d}Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI 48109, USA; and \textsuperscript{e}Architecture et Fonction des Macromolecules Biologiques, CNRS and Aix-Marseille University, Marseille, France
Abstract

The human gut microbiota is an important metabolic organ. However, little is known about how its individual species interact, establish dominant positions, and respond to changes in environmental factors such as diet. In the current study, gnotobiotic mice colonized with a simplified model microbiota composed of 12 sequenced human gut bacterial species were fed oscillating diets of disparate composition. Rapid, reproducible and reversible changes in community structure were observed. Time series microbial RNA-Seq analyses revealed staggered functional responses to diet shifts throughout the community that were heavily focused on carbohydrate and amino acid metabolism. High-resolution shotgun metaproteomics analysis confirmed many of these responses at a protein level. One member, *Bacteroides cellulosilyticus* WH2, proved exceptionally fit regardless of diet. Its genome has the largest collection of carbohydrate active enzymes of any *Bacteroides* species reported to date. Whole genome transcriptional profiling indicated that *B. cellulosilyticus* WH2 is an adaptive forager that tailors its versatile carbohydrate utilization strategy to the dietary polysaccharides available, with a strong emphasis on highly branched plant cell wall xylans that are abundant in dietary staples like cereal grains. Two highly expressed, diet-specific polysaccharide utilization loci (PULs) in *B. cellulosilyticus* WH2 were identified, one with characteristics similar to xylan utilization systems. Introduction of a *B. cellulosilyticus* WH2 library of 26,750 isogenic transposon mutants into gnotobiotic mice along with other model community members confirmed that these loci represent some of its most critical diet-specific fitness determinants. The specific carbohydrates that trigger overexpression of these two loci and many of the organism’s 111 other predicted PULs were identified by RNA-Seq analysis during in vitro growth on 31 distinct carbohydrate substrates, allowing us to better interpret the in vivo RNA-Seq and proteomics data. These results offer insight into how gut microbes adapt to dietary perturbations, both at a community level and from the perspective of a well-adapted symbiont with exceptional saccharolytic capabilities, and illustrate the value of studying defined models of the human gut microbiota.
Introduction

A growing body of evidence indicates that the tens of trillions of microbial cells that inhabit our gastrointestinal tracts (microbiota) extend our biological capabilities in important ways [1,2]. Microbial enzymes process many compounds that would otherwise pass through our intestines unaltered [3], and cases of particular nutrient substrates favoring the growth of particular taxa are being reported [4,5]. Changes in diet are therefore expected to lead to changes in the composition and function of the microbiota [6]. However, our understanding of diet-microbiota interactions at a mechanistic level is still in its infancy.

The absence of a complete catalog of the microbial strains and associated genome sequences that comprise a given microbiota complicates efforts to describe how particular dietary substrates influence individual taxa, how taxa cooperate/compete to utilize nutrients, and how these many interactions in aggregate lead to emergent host phenotypes. Gnotobiotic mice colonized with defined consortia of sequenced human gut microbes, on the other hand, provide an in vivo model of the microbiota in which the identity of all taxa and genes comprising the system are known [7]. In these systems, expressed mRNAs and proteins can be attributed to their genome, gene and species of origin, and findings of interest can be pursued in follow-up in vitro or in vivo experiments. These models afford an opportunity to tightly control experimental variables to a degree not possible in human studies and have proven useful in studying microbial invasion [8], inter-phylum microbe-microbe interactions [9], and the metabolic roles of key ecological guilds [10]. Studies aiming to better understand community-level assembly, resilience, and adaptation are also likely to benefit from a focus on such fully defined systems.

Culture-independent surveys of the healthy adult gut microbiota consistently conclude that it is composed primarily of Bacteroidetes and Firmicutes species [11-13]. The dominance of these two bacterial phyla suggests that their representatives in the human gut are exquisitely adapted to its dynamic conditions, which include a constantly evolving nutrient environment. Members of the genus Bacteroides are known to be adept at utilizing both plant- and host-derived polysaccharides
Comparisons of available *Bacteroides* genomes with those from other gut species indicate that the former are enriched in genes involved in the acquisition and metabolism of various glycans, including glycoside hydrolases and polysaccharide lyases, as well as linked environmental sensors that control their expression (e.g., hybrid two-component systems, extracytoplasmic function (ECF) sigma factors and anti-sigma factors). Many of these genes are organized into polysaccharide utilization loci (PULs) that are distributed throughout the genome [17]. Recent studies have focused on better understanding the evolution, specificity and regulation of PULs in the genomes of species like *Bacteroides thetaiotaomicron* and *Bacteroides ovatus* [18,19]. Little is known, however, about the metabolic strategies adopted by multiple competing species in more complex communities, how dietary changes lead to shifts in the *Bacteroides* fraction of communities [20] through changes in individual species, or how dietary context influences the repertoires of genes the most dominant species rely on to remain competitive with other microbes, including those genes that are components of PULs.

Here, we adopt a multi-faceted approach to the study of model communities in gnotobiotic mice fed changing diets in order to better understand: (i) the process by which communities reconfigure themselves structurally in response to changes in host diet; (ii) how aggregate community function, as judged by the metatranscriptome and the metaproteome, is impacted when host diet is altered; (iii) how the metabolic strategies of individual microbes change, if at all, when the nutrient milieu is dramatically altered, with an emphasis on one prominent but understudied member of the human gut *Bacteroides*; and (iv) whether a microbe’s metabolic versatility/flexibility correlates with competitive advantage in a microbiota comprising numerous related and unrelated species.
Results and discussion

Sequencing the *Bacteroides cellulosilyticus* WH2 genome

Though at least 8 complete and 68 draft genomes of *Bacteroides* spp. are currently available [21], there are numerous examples of distinct clades within this genus where little genomic information exists. To further explore the genome space of one such clade, we obtained a human fecal isolate whose four 16S rRNA gene sequences indicate a close relationship to *Bacteroides cellulosilyticus* (Figs. S1A, B). The genome of this isolate, which we have designated *B. cellulosilyticus* WH2, was sequenced deeply, yielding a high quality draft assembly (23 contigs with an N50 value of 798,728 bp; total length of all contigs in the assembly, 7.1 Mb; Table S1). Annotation of its 5,244 predicted protein-coding genes using the carbohydrate active enzyme (CAZy) database [22] revealed an extraordinary complement of 503 CAZymes comprising 373 glycoside hydrolases (GHs), 23 polysaccharide lyases (PLs), 28 carbohydrate esterases (CEs), and 84 glycosyltransferases (GTs) (see Table S2 for all annotated genes in the *B. cellulosilyticus* WH2 genome predicted to have relevance to carbohydrate metabolism). One distinguishing feature of gut *Bacteroides* genomes is the substantial number of CAZymes they encode relative to those of other intestinal bacteria [23]. The *B. cellulosilyticus* WH2 CAZome is enriched in a number of GH families even when compared with prominent representatives of the gut *Bacteroidetes* (Fig. S2A). When we expanded this comparison to include all 86 *Bacteroidetes* in the CAZy database, we found that the *B. cellulosilyticus* WH2 genome had the greatest number of genes for 19 different GH families, as well as genes from two GH families that had not previously been observed within a *Bacteroidetes* genome (Fig. S2B). Altogether, *B. cellulosilyticus* WH2 has more GH genes at its disposal than any other *Bacteroidetes* species analyzed to date.

In *Bacteroides* spp., CAZymes are often located within PULs [24]. At a minimum, a typical PUL harbors a pair of genes with significant homology to the *susC* and *susD* genes of the starch utilization system (Sus) in *B. thetaiotaomicron* [24]. Other genes encoding enzymes capable of liberating oligo- and monosaccharides from a larger polysaccharide are also frequently present.
The *susC*- and *susD*-like genes of these loci encode the proteins that comprise the main outer membrane binding and transport apparatus and thus represent key elements of these systems. A search of the *B. cellulosilyticus* WH2 genome for genes with strong homology to the *susC*- and *susD*-like genes in *B. thetaiotaomicron* VPI-5482 (see Methods) revealed an unprecedented number of *susC/D* pairs (a total of 118). Studies of other prominent *Bacteroides* spp. have found that the evolutionary expansion of these genes has played an important role in endowing the *Bacteroides* with the ability to degrade a wide range of host- and plant-derived polysaccharides [18,25]. Deep sampling of the gut microbiota of adult humans indicates that *B. cellulosilyticus* strains are common in humans, colonizing approximately 77% of 124 adult Europeans characterized in one study [11] and 62% of 139 individuals living in the USA examined in another survey [12]. We hypothesized that the apparent success of *B. cellulosilyticus* in the gut is derived in part from its substantial arsenal of genes involved in carbohydrate utilization.

**Measuring changes in the structural configuration of a 12-member model microbiota in response to a dietary perturbation**

To test the fitness of *B. cellulosilyticus* WH2 in relation to other prominent gut symbionts, and the importance of diet on its fitness, we carried out an experiment in gnotobiotic mice (experiment 1, “E1”, Fig. S3). Two groups of 10-12 week-old male germ-free C57BL/6J animals were moved to individual cages within gnotobiotic isolators (n=7 animals/group). At day zero, each animal was colonized by oral gavage with a synthetic community comprising 12 prominent, sequenced human gut bacterial species (Fig. 1A) whose taxonomic and functional attributes (as judged by their annotated gene content) are representative of a more complex human microbiota (Tables S3, S4). For two weeks, each treatment group was fed a standard low-fat, high-plant polysaccharide (LF/HPP) mouse chow, or a ‘Western’-like diet where calories are largely derived from fat, starch, and simple sugars [high-fat/high-sugar (HF/HS)] [9]. Over the course of six weeks, diets were changed twice at two-week intervals, such that each group began and ended on the same diet, with an intervening two-week period during which the other diet was administered (Fig. S3).
Using fecal DNA as a proxy for microbial biomass, the plant polysaccharide-rich LF/HPP diet supported 2- to 3-fold more total bacterial growth (primary productivity) despite its lower caloric density (3.7 kcal/g versus 4.5 kcal/g for the HF/HS diet; Fig. S4A). The HF/HS diet contains carbohydrates that are easily metabolized and absorbed in the proximal intestine (sucrose, corn starch and maltodextrin), with cellulose being the one exception (4% of the diet by weight versus 46.3% for the other carbohydrate sources). Thus, in mice fed the HF/HS diet, diet-derived simple sugars are likely to be rare in the distal gut where the vast majority of gut microbes reside. This may provide an advantage to those bacteria capable of utilizing other carbon sources (e.g., proteins/oligopeptides, host glycans). In mice fed the LF/HPP diet, on the other hand, plant polysaccharides that are indigestible by the host should provide a plentiful source of carbon for saccharolytic members of the community.

To evaluate the impact of each starting diet and the diet switch on the structural configuration of the model human microbiota, we subjected DNA isolated from fecal samples collected throughout the course of the experiment and from cecal DNA at sacrifice to shotgun sequencing (COmmunity PROfiling by Sequencing; COPRO-Seq) [8]. The resulting relative abundances of each species in each sample (defined by the number of sequencing reads that could be unambiguously assigned to community members after adjusting for genome uniqueness) were subjected to ordination by principal coordinates analysis (PCoA) (Fig. S5A). As expected, diet was found to be the predominant explanatory variable for observed variance (see separation along principal coordinate 1, ‘PC1’, which accounts for 52% of variance). Overall community structure achieved quasi-equilibrium before the midpoint of the first diet phase, as evidenced by the lack of any significant movement along PC1 after day five. Reconfiguration of overall community structure following a transition to the second diet phase also took place over the course of ~5 days. Notably, communities showed a near-perfect inversion in their positions along PC1 after the first diet switch, with animals transitioning from LF/HPP to HF/HS diets adopting a structure like that of the communities that were initially assembled while animals were consuming a starting HF/HS diet, and vice versa. The second diet switch from phase two to three resulted in a similar movement along PC1.
in the opposite direction, indicating a reversion of each community to its originally assembled configuration. These results, in addition to demonstrating the significant impact of these two diets on the structure of this 12-member model human gut microbiota, also suggest that communities of this size are capable of demonstrating resilience in the face of substantial diet perturbations.

The assembly process and observed diet-induced reconfigurations also proved to be highly reproducible as evidenced by COPRO-Seq results from a replication of E<sub>1</sub> (experiment 2, “E<sub>2</sub>”). In this follow-up experiment, fecal samples were collected more frequently than in E<sub>1</sub>, providing a dataset with improved temporal resolution. Ordination of E<sub>2</sub> COPRO-Seq data by PCoA (Fig. 1B) showed that the model microbiota in each E<sub>2</sub> treatment group assemble in a manner similar to their counterparts in E<sub>1</sub>, that structural reconfigurations in response to diet occur with the same timing as in E<sub>1</sub>, and that the quasi-equilibrium community configurations achieved during each diet phase are highly similar between experiments for each treatment group. As in E<sub>1</sub>, cecal data for each E<sub>2</sub> treatment group overlap with their corresponding fecal samples, and DNA yields from E<sub>2</sub> fecal samples vary substantially as a function of host diet (Fig. S4B).

COPRO-Seq provides precise measurements of the proportional abundance of each species present in a defined microbiota. Data collected in both E<sub>1</sub> and E<sub>2</sub> (Table S5) revealed significant differences between microbiota members in terms of the maximum abundance levels they achieved, the rates at which their abundance levels were impacted by diet shifts, and the degree to which each species demonstrated a preference for one diet over another (Fig. S4C). Changes in each species’ abundance over time replicated well across animals in each treatment group, suggesting the assembly process and diet-induced reconfigurations occur in an orderly and rules-based fashion, with minimal stochasticity. A species’ relative abundance immediately after colonization (i.e., 24 hours after gavage / day 1) was, in general, a poor predictor of its abundance at the end of the first diet phase (i.e. day 13) (E<sub>1</sub> R<sup>2</sup>=0.23; E<sub>2</sub> R<sup>2</sup>=0.27), suggesting early dominance of the founder population was not strongly tied to relative success in the assembly process.
In communities that assembled in mice initially fed a HF/HS diet, four Bacteroides spp. (*B. caccae*, *B. cellulosyiticus* WH2, *B. thetaiotaomicron*, and *B. vulgatus*) each achieved relative abundances of ≥10% by the end of the first diet phase (day 13 post-gavage), with *B. caccae* attaining the highest levels (37.1±4.9% and 34.2±5.5%; group mean ± SD in E₁ and E₂, respectively). In animals fed the plant polysaccharide-rich LF/HPP chow during the first diet phase, *B. cellulosi-lyticus* WH2 was dominant, achieving levels of 37.1±2.0% (E₁) and 41.6±3.9% (E₂) by day 13. *B. thetaiotaomicron* and *B. vulgatus* also attained relative abundances of >10%.

Changes in diet often resulted in rapid, dramatic changes in a species’ proportional representation. Because the dynamic range of abundance values observed when comparing multiple species was substantial [lowest, *D. longicatena* (<0.003%); highest *B. caccae* (55.0%)], comparing diet responses on a common scale using raw abundance values was challenging. To represent these changes in a way that scaled absolute increases/decreases in community proportion to the abundance range of each strain, we also normalized community representation at each time-point to the maximum level each microbe achieved across all time-points within each mouse. Plotting the resulting measure of abundance (percentage of maximum achieved, ‘PoMA’) over time demonstrates which microbes are strongly responsive to diet (experience significant swings in PoMA value following a diet switch) and which are relatively diet-insensitive (experience only modest or no significant change in PoMA value following a diet switch). Heatmap visualization of E₁ PoMA values (Fig. S5B) indicated that those microbes with a preference for a particular diet in one animal treatment group also tended to demonstrate the same diet preference in the other. Likewise, diet insensitivity was also consistent across treatment groups; diet-insensitive microbes were insensitive regardless of the order in which diets were introduced.

Of the diet-sensitive components of the model microbiota, those showing the most striking responses were *B. caccae* and *B. ovatus*, which strongly preferred the ‘Western’-like HF/HS diet and the polysaccharide-rich LF/HPP diet, respectively (Figs. 1C, S4C). Of the diet-insensitive components of the microbiota, *B. thetaiotaomicron* showed the most stability in its representation (Figs. 1C, S4C), consistent with its reputation as a versatile forager. Paradoxically, *B. cellulosi-
lyticus WH2 was found to be both diet-sensitive and highly fit on its less-preferred diet. Although this strain clearly achieved higher levels of community representation in animals fed the LF/HPP diet, it maintained strong levels of representation in animals fed HF/HS as well (Figs. 1C, S4C).

When taking into account the abundance data for all 12 community members, proportional representation at the end of the first diet phase (i.e., day 13) was a good predictor of representation at the end of the third diet phase (i.e., day 42) ($E_1 R^2=0.77$; $E_2 R^2=0.84$), suggesting that the intervening dietary perturbation had little effect on the ultimate outcomes for most species within the community. However, one very low-abundance strain (*Dorea longicatena*) achieved significantly different maximum percentages across the two treatment groups in each experiment, suggesting that steady-state levels of this strain may have been impacted by diet history. In communities that assembled in mice fed LF/HPP chow, *D. longicatena* was found to persist throughout the experiment at low levels on both diet regimens. In communities that assembled in mice fed the HF/HS diet, *D. longicatena* dropped below the limit of detection before the end of the first diet phase, was undetectable by the end of the second diet phase, and remained undetectable throughout the rest of the time course. This interesting example raises the possibility that for some species, irreversible hysteresis effects may play a significant role in determining the likelihood that they will remain members of a microbiota over long periods of time. For most strains, however, these data suggest that diet-driven increases or decreases in a species’ abundance are likely to be reversible over short time spans.

**The cecal community metatranscriptome sampled at the time of sacrifice**

These diet-induced reconfigurations in community structure led us to examine the degree to which members of the community were modifying their metabolic strategies. To establish an initial baseline static view of expression data for each microbe on each diet, we developed a custom GeneChip whose probe sets were designed to target 46,851 of the 48,023 known or predicted protein-coding genes within our model human gut microbiome (see Methods). Total RNA was collected from the cecal contents of each animal in $E_1$ at the time of sacrifice and hybridized to this GeneChip. The to-
tal number of genes whose expression was detectable on each diet was remarkably similar (14,929 and 14,594 detected in the LF/HPP→HF/HS→LF/HPP and HF/HS→LF/HPP→HF/HS treatment groups, respectively). A total of 11,373 genes were expressed on both diets (Fig. S6A) while 2,003 (4.3%) were differentially expressed to a statistically significant degree, including 161 (6%) of the 2,640 genes in the model human gut microbiome encoding proteins with CAZy-recognized domains. Fig. S6B illustrates the fraction of the community CAZome and several species-level CAZomes expressed on each diet (see Table S6 for a comprehensive list of all genes, organized by species and fold-change in expression, whose cecal expression was detectable on each diet and all genes whose expression was significantly different when comparing data from each treatment group).

Of the community members demonstrating obvious diet preferences (as judged by relative abundance data), Bacteroides caccae and Bacteroides cellulosilyticus WH2 served as a particularly interesting pair of organisms whose CAZy-level responses to diet change are different in several respects. Our ability to evaluate shifts in B. caccae’s metabolic strategy in the gut was limited by the very low abundance of this species in animals fed LF/HPP chow (i.e., at these low levels of abundance, our mRNA and subsequent protein assays were often not sensitive enough to exhaustively sample B. caccae’s metatranscriptome and metaproteome). Our observations regarding this species’ carbohydrate utilization capabilities and preferences are summarized in Supplementary Results. In contrast, the abundance of Bacteroides cellulosilyticus WH2, which favored the LF/HPP diet, remained high enough on both diets to allow for a comprehensive analysis of its expressed genes and proteins. This advantage, along with the exceptional carbohydrate utilization machinery encoded within the genome of this organism, encouraged us to focus on further dissecting the responses of B. cellulosilyticus WH2 to diet changes.

Detailed inspection of the expressed B. cellulosilyticus WH2 CAZome (503 CAZymes in total) provided an initial view of this microbe’s sophisticated carbohydrate utilization strategy. A comparison of the top decile of expressed CAZymes on each diet disclosed many shared elements between the two lists, spanning many different CAZy families, with just over half of the 50 most
expressed enzymes on the plant polysaccharide-rich LF/HPP chow also occurring in the list of most highly expressed enzymes on the sucrose-, corn starch- and maltodextrin-rich HF/HS diet (Fig. 2A). 25 of the 50 most expressed CAZymes on the LF/HPP diet were significantly upregulated compared to the HF/HS diet; of these, seven were members of the GH43 family (Fig. 2B). The GH43 family consists of enzymes with activities required for the breakdown of plant-derived polysaccharides such as hemicellulose and pectin. Inspection of the enzyme commission (EC) annotations for the most upregulated GH43 genes shows that they encode xylan 1-4-β-xylosidases (EC 3.2.1.37), arabinan endo-1,5-α-L-arabinosidase (EC 3.2.1.99), and α-L-arabinofuranosidases (EC 3.2.1.55). The GH10 family, which is currently comprised exclusively of endo-xylanases (EC 3.2.1.8, EC 3.2.1.32), was also well represented among this set of 25 genes, with four of the seven putative GH10 genes in the \( B. \) cellulosilyticus WH2 genome making the list. Strikingly, of the 45 predicted genes with putative GH43 domains in the \( B. \) cellulosilyticus WH2 genome, none were upregulated on the ‘Western’-style HF/HS chow.

The most highly expressed \( B. \) cellulosilyticus WH2 CAZyme on the plant polysaccharide-rich chow (which was also highly-expressed on the HF/HS chow) was BWH2_1228, a putative α-galactosidase from the GH36 family. These enzymes, which are not expressed by humans in the stomach or intestine, cleave terminal galactose residues from the non-reducing ends of raffinose family oligosaccharides (RFOs, including raffinose, stachyose, and verbascose), galacto(gluco)mannans, galactolipids, and glycoproteins. RFOs, which are well represented in cereal grains consumed by humans, are expected to be abundant in the LF/HPP diet given its ingredients (e.g., soybean meal), but potential substrates in the HF/HS diet are less obvious, possibly implicating a host glycolipid or glycoprotein target. Surface glycans in the intestinal epithelium of rodents are decorated with terminal fucose residues [26] as well as terminal sialic acid residues and sulfate [27]. Hydrolysis of the α-2 linkage connecting terminal fucose residues to the galactose-rich extended core is thought to be catalyzed in large part by GH95 and GH29 enzymes [28]. The \( B. \) cellulosilyticus WH2 genome is replete with putative GH95 and GH29 genes (total of 12 and 9, respectively), but only a few (BWH2_1350/2142/3154/3818) were expressed in vivo on at least
one diet, and their expression was low relative to many other CAZymes (see Table S6). Cleavage of terminal sialic acids present in host mucins by bacteria is usually carried out by members of the GH33 family. *B. cellulosilyticus* WH2 has two GH33 genes that are expressed on either one diet (BWH2_3822, HF/HS) or both diets (BWH2_4650), but neither is highly expressed relative to other *B. cellulosilyticus* WH2 CAZymes. Therefore, utilization of host glycans by *B. cellulosilyticus* WH2, if it occurs, likely requires partnerships with other members of the community that express GH29/95/33 enzymes (see Table S6 for a list of members that express these enzymes in a diet-independent and/or diet-specific fashion).

Among the 50 most highly expressed *B. cellulosilyticus* WH2 CAZymes, 12 were significantly upregulated on the HF/HS diet compared to the LF/HPP diet, with members of family GH13 being most prevalent. While the enzymatic activities and substrate specificities of GH13 family members are varied, most relate to the hydrolysis of substrates comprising chains of glucose subunits, including amylose (one of the two components of starch) and maltodextrin, both prominent ingredients in the HF/HS diet.

GeneChip-based profiling of the E₁ cecal communities provided a snapshot of the metatranscriptome on the final day of the final diet phase in each treatment group. The analysis of *B. cellulosilyticus* WH2 CAZyme expression suggested that this strain achieves a ‘generalist’ lifestyle not by relying on substrates that are present at all times (e.g., host mucins), but rather by modifying its resource utilization strategy to effectively compete with other microbes for diet-derived polysaccharides not metabolized by the host.

**Community-wide analysis of diet-induced changes in microbial gene expression**

To develop a more complete understanding of the dynamic changes that occur in gene expression over time and throughout the community following diet perturbations, we performed microbial RNA-Seq analyses using feces obtained at select time-points from mice in the LF/HPP→HF/HS→LF/HPP treatment group of E₂ (Fig. S3).
We began with a ‘top-down’ analysis in which every RNA-Seq read count from every gene in the model microbiome was binned based on the functional annotation of the gene from which it was derived, regardless of its species of origin. In this case, the functional annotation used as the binning variable was the predicted enzyme commission (EC) number for a gene’s encoded protein product. Expecting that some changes might occur rapidly, while others might require days or weeks, we searched for significant differences between the terminal time-points of the first two diet phases (i.e., points at which the model human gut microbiota had been allowed 13 days to acclimate to each diet). The 157 significant changes we identified were subjected to hierarchical clustering by EC number to determine which functional responses occurred with similar kinetics. The results revealed that unlike the rapid diet-induced structural reconfigurations in community composition, community-level changes in gene expression occurred with highly variable timing that differed from function to function. These changes were dominated by EC numbers associated with enzymatic reactions relevant to carbohydrate and amino acid metabolism (see Table S7 for a summary of all significant changes observed, including aggregate expression values for each functional bin (EC number) at each time-point). Significant responses could be divided into one of three groups: ‘rapid’ responses denoting EC numbers whose representation in the transcriptome increased/decreased dramatically within 1-2 days of a diet switch; ‘gradual’ responses where the representation of EC numbers changed notably, but slowly, between the two diet transition points; and ‘delayed’ responses where no significant change occurred immediately but did occur by the end of a diet phase (Fig. 3, Table S7). EC numbers associated with reactions important in carbohydrate metabolism and transport were distributed across all three response types for each of the two diets. Nearly all genes encoding proteins with EC numbers related to amino acid metabolism that were significantly upregulated on HF/HS chow binned into the ‘rapid’ or ‘gradual’ groups, suggesting this diet put immediate pressure on the community to increase its repertoire of expressed amino acid biosynthesis and degradation genes. Genes with assigned EC numbers involved in amino acid metabolism that were significantly upregulated on the other, polysaccharide-rich, LF/HPP diet were spread more evenly across the three groups (Fig. 3).
Careful inspection of our top-down analysis results and a complementary ‘bottom-up’ analysis in which normalization was performed at the level of individual species, rather than the community, allowed us to identify other important responses that would have gone undetected were it not for the fact that we were dealing with a defined community where all genes in its component members’ genomes (microbiome) were known. For example, an assessment of the representation of EC 3.2.1.8 (endo-1,4-β-xylanase) within the metatranscriptome before and after the first diet switch (LF/HPP→HF/HS) initially suggested that this activity was reduced to a statistically non-significant degree as a result of the diet switch (day 13 versus day 27, \(p=0.25\); Fig. S7A). Aggregation by species of all sequencing read counts assignable to mRNAs encoding proteins with this EC number revealed that over 99% of the contributions to this functional bin originated from *B. cellulosilyticus* WH2 (Fig. S7B), implying that the community-level response and the response of this *Bacteroides* species were virtually one and the same. A tally of all sequencing reads assignable to *B. cellulosilyticus* WH2 at each time-point disclosed that although this strain maintains high proportional representation in the community throughout each diet oscillation period (range: 10.3-42.5% and 11.6-43.3% for \(E_1\) and \(E_2\), respectively), its contribution to the community metatranscriptome is substantially decreased during the HF/HS diet phase (Fig. S7C). This dramatic reduction in the extent to which *B. cellulosilyticus* WH2 contributes to the community metatranscriptome in HF/HS-fed mice ‘masks’ the highly significant upregulation of EC 3.2.1.8 (endo-1,4-β-xylanase) that occurs within the *B. cellulosilyticus* WH2 transcriptome following the first diet shift (Fig. S7D). A further breakdown of endo-1,4-β-xylanase upregulation in *B. cellulosilyticus* WH2 when mice are switched to the HF/HS diet reveals that most of this response is driven by two genes, BWH2_4068 and BWH2_4072 (Fig. S7E). Our realization that one of the most significant diet-induced gene expression changes in the second most abundant strain in the model community was undetectable when inspecting functional responses at the community level provides a strong argument for expanding the use of model microbiota comprised exclusively of sequenced species, whose individual contributions to community activity can be evaluated in a rigorous way that is not possible with consortia of unknown or poorly defined gene composition.
High-resolution profiling of the cecal metaproteome sampled at the time of sacrifice

In principle, protein measurements can provide a more direct readout of expressed community functions than an RNA-level analysis, and thus a deeper understanding of community members’ interactions with one another and with their habitat [29,30]. For these reasons and others, much work has been dedicated to applying shotgun proteomics techniques to microbial ecosystems in various environments [31,32]. Though these efforts have provided illustrations of significant methodological advances, they have been limited by the complexity of the metaproteomes studied and by the difficulties this complexity creates when attempting to assign peptide identities uniquely to proteins of specific taxa. Recognizing that a metaproteomics analysis of our model community would not be subject to such uncertainty given its fully-defined microbiome and thus fully-defined theoretical proteome, we subjected cecal samples from two mice from each diet treatment group in E1 (n=4 total) to high-performance liquid chromatography-tandem mass spectrometry (LC-MS/MS; see Methods). We had three goals: (i) to evaluate how our ability to assign peptide-spectrum matches (PSMs) to particular proteins within a theoretical metaproteome is affected by the presence of close homologs within the same species and within other, closely-related species; (ii) to test the limits of our ability to characterize protein expression across different species given the substantial dynamic range we documented in microbial species abundance; and (iii) to collect semi-quantitative peptide/protein data that might validate and enrich our understanding of functional responses identified at the mRNA level, particularly with respect to the niche (profession) of CAZyme-rich \textit{B. cellulosilyticus} WH2.

Given the evolutionary relatedness of the strains involved, we expected that some fraction of observed PSMs from each sample would be of ambiguous origin due to non-unique peptides shared between species’ proteomes. To assess which species might be most affected by this phenomenon when characterizing the metaproteome on different diets, we catalogued each strain’s theoretical peptidome using an \textit{in silico} tryptic digest. This simulated digest took into account both the potential for missed trypic cleavages and the peptide mass range that could be detected using our methods. The results (\textbf{Fig. S8A}, Table S8) demonstrated that for a community of modest com-
plexity, the proportion of peptides within each strain’s theoretical peptidome that are ‘unique’ (i.e., assignable to a single protein within the theoretical community metaproteome) varies substantially from species to species, even among those that are closely related. We found the lone representative of the *Actinobacteria* in the community, *C. aerofaciens*, to have the highest proportion of unique peptides (94.2%), while *B. caccae* had the lowest (63.0%). Interestingly, there was not a strong correlation between the fraction of a species’ peptides that were unique and the total number of unique peptides that species contributed to the theoretical peptidome. For example, *C. aerofaciens* (2,367 predicted protein-coding genes) contributed only 81,894 (1.5%) unique peptides, the lowest of any community member, despite having a proteome composed of mostly unique peptides. *B. cellulosilyticus* WH2, on the other hand (5,244 predicted protein-coding genes), contributed 241,473 (4.5%) unique peptides, the highest of any community member, despite a high fraction of non-unique peptides (18.4%) within its theoretical peptidome. The evolutionary relatedness of the *Bacteroides* components of the model gut microbiota appeared to negatively affect our ability to assign their peptides to specific proteins; their six theoretical peptidomes had the six lowest uniqueness levels. However, their greater number of proteins and peptides relative to the *Firmicutes* and *Actinobacteria* within the community more than compensated for this deficiency; over 60% of unique peptides within the unique theoretical metaproteome were contributed by the *Bacteroides*.

We also found that the proportion of PSMs uniquely assignable to a single protein within the metaproteome varied significantly by function, suggesting that some classes of proteins can be traced back to specific microbes more readily than others. For example, when considering all theoretical peptides that could be derived from the proteome of a particular bacterial species, those from proteins with roles in categories with high expected levels of functional conservation (e.g., translation and nucleotide metabolism) were on average deemed unique more often than those from proteins with roles in functions we might expect to be less conserved (e.g., glycan biosynthesis and metabolism) (See Table S8 for a summary of how peptide uniqueness varied across different KEGG categories and pathways, and across different species in the experiment). However,
even in KEGG categories and pathways with high expected levels of functional conservation, the vast majority of peptides were found to be unique when a particular species was not closely related to other members of the model community.

Next, we determined the average number of proteins that could be experimentally identified in our samples for each microbial species within each treatment group in E1. The results (Fig. S8B, Table S9) illustrate two important conclusions. First, although equal concentrations of total protein were evaluated for each sample, slightly less than twice as many total microbial proteins were identified in samples from the LF/HPP-fed mice as those from mice fed the HF/HS diet (4,659 versus 2,777, respectively). While there are a number of possible explanations, this finding and the higher number of mouse proteins detected in samples from this treatment group is consistent with our fecal DNA analysis that had indicated that the HF/HS diet supports lower levels of gut microbial biomass compared to the LF/HPP diet (Figs. S4A, B). Second, a breakdown of all microbial proteins detected by species of origin (Fig. S8B) revealed that the degree to which we could inspect protein expression for a given species was dictated to a large degree by its relative abundance and the diet to which it was exposed.

Next, our ability to detect a substantial number of B. cellulosilyticus WH2’s expressed transcripts and proteins in samples from mice on either diet allowed us to determine the degree to which RNA and protein data for an abundant, active member of the community might correlate, and whether or not the type of genes considered might influence the degree of correlation between these two datasets. We first performed a spectral count-based correlation analysis on the diet-induced, log-transformed, average fold-differences in expression for all B. cellulosilyticus WH2 genes that were detectable at both the RNA and protein level for both diets. The results revealed a moderate degree of linear correlation between RNA and protein observations (Fig. S8C, black plot; r=0.53). However, subsequent division of these genes into functionally related subsets, which were each subjected to their own correlation analysis, revealed striking differences in the degree to which RNA level and protein-level expression changes agreed with one another. For example, diet-induced changes in mRNA expression for genes involved in translation showed virtually no
correlation with changes measured at the protein level (Fig. S8C, red plot; \( r=0.03 \)). Correlations for other categories of \( B. \) cellulosilyticus WH2 genes, such as those involved in energy metabolism (Fig. S8C, green plot; \( r=0.36 \)) and amino acid metabolism (Fig. S8C, orange plot; \( r=0.48 \)) were also poorer than the correlation for all genes collectively. In contrast, the correlation for the 110 genes with predicted involvement in carbohydrate metabolism was quite strong (Fig. S8C, blue plot; \( r=0.69 \)), and was in fact the best correlation identified for any functional category of genes considered. The significant range of correlations observed in different categories of genes suggests that the degree to which RNA-based analyses provide an accurate picture of microbial adaptation to environmental perturbation may be strongly impacted by the functional nature of the genes involved. Additionally, these data further emphasize the need for enhanced dynamic range metaproteome measurements and better bioinformatic methods to assign/bin unique and non-unique peptides so that deeper and more thorough surveys of the microbial protein landscape may be performed and evaluated alongside more robust transcriptional datasets. Below we focus on two highly expressed, strongly diet-specific loci relevant to \textit{in vivo} carbohydrate utilization by \( B. \) cellulosilyticus WH2.

**Identifying two diet-inducible, xylanase-containing PULs whose genetic disruption results in diet-specific loss of fitness**

Several of the most highly expressed and diet-sensitive \( B. \) cellulosilyticus WH2 genes in this study fell within two putative PULs. One PUL (BWH2_4044-55) contains 12 ORFs that include a dual \( \text{susC/D} \) cassette, three putative xylanases assigned to CAZy families GH8 and GH10, a putative multifunctional acetyl xylan esterase/\( \alpha \)-L-fucosidase, and a putative hybrid two-component system regulator (Fig. 4A). Gene expression within this PUL was markedly higher in mice consuming the plant polysaccharide-rich LF/HPP diet at both the mRNA and protein level. mRNA analysis disclosed that BWH2_4047 was the most highly expressed \( B. \) cellulosilyticus WH2 \( \text{susD} \) homolog on this diet. Likewise, BWH2_4046/4, the two \( \text{susC} \)-like genes within this PUL, were the 2nd and 4th most highly expressed \( B. \) cellulosilyticus WH2 \( \text{susC} \)-like genes in LF/HPP-fed animals, and
exhibited reductions in expression levels of 99.5% and 93% in animals consuming the HF/HS diet. The same LF/HPP diet bias was observed for other genes within this PUL (Figs. 2A, 4B) but not for neighboring genes. The same trends were obvious and amplified when we quantified protein expression (Fig. 4C). In mice fed LF/HPP chow, only three *B. cellulosilyticus* WH2 SusC-like proteins had higher protein levels than BWH2_4044/6, and only two SusD-like proteins had higher levels than BWH2_4045/7. Strikingly, we were unable to detect a single peptide from 9 of the 12 proteins in this PUL in samples obtained from mice fed the HF/HS diet, emphasizing the strong diet specificity of this locus.

A second PUL in the *B. cellulosilyticus* WH2 genome composed of a *susC/D*-like pair (BWH2_4074/5), a putative hybrid two-component system regulator (BWH2_4076), and a xylanase (GH10) with dual carbohydrate binding module domains (CBM22) (BWH2_4072) (Fig. 4A) demonstrated a strong but opposite diet bias, in this case exhibiting significantly higher expression in animals consuming the HF/HS ‘Western’-like diet. mRNA-level analysis showed that this xylanase was the second most highly-expressed *B. cellulosilyticus* WH2 CAZyme in animals consuming this diet (Fig. 2A). As with the previously described PUL, shotgun metaproteomics validated the transcriptional analysis (Figs. 4B, C); except the gene encoding the PUL’s presumed transcriptional regulator (BWH2_4076), the diet specificity was substantial with protein-level fold changes ranging from 10.0-33.3 across the locus (Table S10).

Recent work by Dodd and colleagues has done much to advance our understanding of the regulation and metabolic role of xylan utilization system gene clusters in xylanolytic members of the *Bacteroidetes*, particularly within the genus *Prevotella* [33]. The ‘core’ gene cluster of the prototypical xylan utilization system they described consists of two tandem repeats of *susC/susD* homologs (*xusA/B/C/D*), a downstream hypothetical gene (*xusE*) and a GH10 endoxylanase (*xyn10C*). The 12-gene PUL identified in our study (BWH2_4044-55) appears to contain the only instance of this core gene cluster within the *B. cellulosilyticus* WH2 genome, suggesting that this PUL, induced during consumption of a plant polysaccharide-rich diet, is likely to be the primary xylan utilization system within this organism. A recent study characterizing the carbohydrate utili-
zation capabilities of *B. ovatus* ATCC 8483 also identified two PULs involved in xylan utilization whose gene configurations differ from the paradigm described by Dodd [18]. The smaller of the two xylanase-containing PULs described above appears to be evolutionarily related to a subset of the genes found within *BACOVA_04385-94*, highlighting the fact that *Bacteroidetes* have likely evolved multiple strategies for the utilization of this important hemicellulose.

Having established that expression of BWH2_4044-55 and BWH2_4072-6 is strongly dictated by diet, we next sought to determine if these PULs are required by *B. cellulosilyticus* WH2 for fitness *in vivo*. A follow-up study was performed in which mice were fed either a LF/HPP or HF/HS diet after being colonized with a model human gut microbiota similar to the one utilized in E1 and E2 (see *Methods*). In this community, the wild-type *B. cellulosilyticus* WH2 strain used in our previous experiments was replaced with a transposon mutant library consisting of 26,750 distinct transposon insertion mutants in over 78% of all predicted ORFs (average 5.3 distinct insertion mutants / ORF). The library was constructed using methods similar to those reported by Goodman *et al.* (see *Methods*) so that the relative proportion of each insertion mutant in both the input (orally gavaged) and output (fecal) populations could be determined using insertion sequencing (INSeq) [34]. The results of this analysis revealed clear, diet-specific losses of fitness when components of these loci were disrupted (*Fig. 4D*). Additionally, as observed in E1 and E2, expression of each PUL was strongly biased by diet, with genes BWH2_4072-5 demonstrating upregulation on the HF/HS diet and BWH2_4044-55 on the LF/HPP diet. The extent to which a gene’s disruption impacted the fitness of *B. cellulosilyticus* WH2 on one diet or the other correlated well with whether or not that gene was highly expressed on a given diet. For example, four of the five most highly expressed genes in the BWH2_4044-55 locus were the four genes shown to be most crucial for fitness on the LF/HPP diet. Of these four genes, three were homologs of either the *susC* or *susD* gene in *B. thetaiotaomicron* (the fourth was the putative endo-1,4-β-xylanase thought to constitute the last element of the xylan utilization system core). Though the fitness cost of disrupting genes within BWH2_4044-55 varied from gene to gene, disruption of any one component of the BWH2_4072-6 PUL had serious consequences for *B. cellulosilyticus* WH2 in animals fed the
HF/HS diet. This could suggest that while disruption of some components of the BWH2_4044-55 locus can be rescued by similar or redundant functions elsewhere in the genome, the same is not true for BWH2_4072-5. Notably, disruption of the BWH2_4076 gene (which is predicted to encode a hybrid two-component regulatory system) had negative consequences on either diet tested, suggesting that the regulation of this locus is crucial even when the PUL is not actively expressed. While many genes outside of these two PULs were also found to be important for the in vivo fitness of *B. cellulosilyticus* WH2, those within these PULs were among the most essential to diet-specific fitness, suggesting these loci are central to the metabolic lifestyle of *B. cellulosilyticus* WH2 in the gut.

**Characterizing the carbohydrate utilization capabilities of *B. cellulosilyticus* WH2 and *B. caccae***

The results described in the preceding section indicate that *B. cellulosilyticus* WH2 prioritizes xylan as a nutrient source in the gut and that it tightly regulates the expression of its xylan utilization machinery. Moreover, the extraordinary number of putative CAZymes and PULs within the *B. cellulosilyticus* WH2 genome suggests that it is capable of growing on carbohydrates with diverse structures and varying degrees of polymerization. To characterize its carbohydrate utilization capabilities, we defined its growth in minimal medium supplemented with one of 46 different carbohydrates [18]. Three independent growths, each consisting of two technical replications, yielded a total of six growth curves for each substrate. Of the 46 substrates tested, *B. cellulosilyticus* WH2 grew on 39 (Table S11); they encompassed numerous pectins (6 of 6), hemicelluloses/β-glucans (8 of 8), starches/fructans/α-glucans (6 of 6) and simple sugars (14 of 15), as well as host-derived glycans (4 of 7) and one cellooligosaccharide (cellobiose). The seven substrates that did not support growth included three esoteric carbohydrates (carrageenan, porphyran, and alginic acid), the simple sugar N-acetylneuraminic acid, two host glyans (keratan sulfate and mucin O-glycans) and fungal cell wall-derived α-mannan. *B. cellulosilyticus* WH2 clearly grew more robustly on some carbohydrates compared to others. Excluding simple sugars, fastest growth was achieved
on dextran (0.099±0.048), laminarin (0.095±0.014 OD$_{600}$ units/h), pectic galactan (0.088±0.018), pullulan (0.088±0.026), and amylopectin (0.085±0.003). Although one study has reported that the type strain of $B. \text{cellulosilyticus}$ degrades cellulose [35], the $B. \text{cellulosilyticus}$ WH2 strain of $B. \text{cellulosilyticus}$ failed to demonstrate any growth on minimal medium plus cellulose (specifically, Solka-Floc 200 FCC, International Fiber Corp.) after five days. Maximum cell density was achieved with amylopectin (1.17±0.02 OD$_{600}$ units), dextran (1.12±0.20), cellobiose (1.09±0.08), laminarin (1.08±0.08), and xyloglucan (0.99±0.04). Total $B. \text{cellulosilyticus}$ WH2 growth (i.e., maximum cell density achieved) on host-derived glycans was typically very poor, with only two substrates achieving total growth above 0.2 OD$_{600}$ units (chondroitin sulfate, 0.50±0.04; glyco- gen, 0.99±0.02). The disparity between total growth on plant polysaccharides versus host-derived glycans, including O-glycans which are prevalent in host mucin, indicates a preference for diet-derived saccharides, consistent with our in vivo mRNA and protein expression data.

We determined how the growth rate of $B. \text{cellulosilyticus}$ WH2 on these substrates compared to the growth rates for other prominent gut $Bacteroides$ spp.. After subjecting $B. \text{caccae}$ to the same phenotypic characterization as $B. \text{cellulosilyticus}$ WH2, we combined our measurements for these two strains with previously published measurements for $B. \text{thetaiotaomicron}$ and $B. \text{ovatus}$ [18]. The results underscored the competitive growth advantage $B. \text{cellulosilyticus}$ WH2 likely enjoys when foraging for polysaccharides in the intestinal lumen. For example, of the eight hemicelluloses and β-glucans tested in our carbohydrate panel, $B. \text{cellulosilyticus}$ WH2 grew fastest on six while $B. \text{ovatus}$ grew fastest on two (Table S11). $B. \text{caccae}$ and $B. \text{thetaiotaomicron}$, on the other hand, failed to grow on any of these substrates. Across all the carbohydrates for which data are available for all four species, $B. \text{cellulosilyticus}$ WH2 grew fastest on the greatest number of polysaccharides (11 of 26) and tied with $B. \text{caccae}$ for the greatest number of monosaccharides (6 of 15). $B. \text{thetaiotaomicron}$ and $B. \text{caccae}$ did, however, outperform the other two $Bacteroides$ tested with respect to utilization of host glycans in vitro, demonstrating superior growth rates on four of five substrates tested (Table S11).
B. cellulosilyticus WH2’s rapid growth to high densities on xylan, arabinoxylan, and xylol- glucan, as well as xylose, arabinose, and galactose, is noteworthy given our prediction that two of its most tightly-regulated, highly-expressed PULs appear to be involved in the utilization of xylan, arabinoxylan, or some closely-related polysaccharide. To identify specific mono- and/or polysaccharides capable of triggering the activation of these two PULs, as well as the 111 other putative PULs within the B. cellulosilyticus WH2 genome, we used RNA-Seq to characterize its transcriptional profile at mid-log phase in minimal medium (Table S12) plus one of 16 simple sugars or one of 15 complex sugars (Table S13) (see Methods; n= 2-3 cultures/substrate; 5.2-14.0 million raw Illumina HiSeq reads generated for each of the 90 transcriptomes). After mapping each read to the B. cellulosilyticus WH2 reference gene set, counts were normalized using DESeq to allow for direct comparisons across samples and conditions. Hierarchical clustering of the normalized dataset resulted in a well-ordered dendrogram in which samples clustered almost perfectly by the carbohydrate on which B. cellulosilyticus WH2 was grown (Fig. 5A). The consistency of this clustering illustrates how (i) technical replicates within each condition exhibit strong correlation with one another, suggesting any differences between cultures in a treatment group (e.g., small differences in density or growth phase) had at best minor effects on aggregate gene expression, and (ii) growth on different carbohydrates results in distinct, substrate-specific gene expression signals capable of driving highly discriminatory differences between treatment groups. The application of rigorous bootstrapping to our hierarchical clustering results also revealed several cases of higher-level clusters in which strong confidence was achieved. These dendrogram nodes (illustrated as white circles), indicate sets of growth conditions that yield gene expression patterns more like each other than like the patterns observed for other substrates. Two notable examples were xylan/arabinoxylan (which are structurally related and share the same xylan backbone) and L-fucose/L-rhamnose (which are known to be metabolized via parallel pathways in E. coli).

Importantly, these findings suggested that when combined with in vivo expression data from the model community, it is possible to identify the particular carbohydrates to which B. cellulosi- lyticus WH2 is exposed and accessing within its gut environment. To explore this concept further,
we compared expression of each gene in each condition to its expression on our control treatment, minimal medium plus glucose (MM-G). The results revealed a dynamic PUL activation network in which some PULs were activated by a single substrate, some by multiple substrates, and some that were transcriptionally silent across all conditions tested. Of the 118 putative susC/D pairs in *B. cellulosilyticus* WH2 that we have used as markers of PULs, 30 were dramatically activated on one or more of the substrates tested; in these cases, both the *susC*- and *susD*-like genes in the cassette were upregulated an average of >100-fold relative to MM-G across all technical replicates (Fig. 5B). At least one *susC/D* activation signature was identified for every one of the 17 oligosaccharides and polysaccharides and for six of the 13 monosaccharides tested (Table S14). The lack of carbohydrate-specific PUL activation events for some monosaccharides (fructose, galactose, glucuronic acid, sucrose, and xylose) was expected, given that these loci are primarily dedicated to polysaccharide acquisition. Further inspection of gene expression outside of PULs disclosed that *B. cellulosilyticus* WH2 prioritizes use of its non-PUL-associated carbohydrate machinery, such as putative phosphotransferase system (PTS) components and monosaccharide permeases, when grown on these monosaccharides (Table S14).

Several carbohydrates activated the expression of multiple PULs. Growth on water-soluble xylan and wheat arabinoxylan produced significant upregulation of five *susC/D*-like pairs (BWH2_0865/0866, 0867/0868, 4044/4045, 4046/4047, and 4074/4075). No other substrate tested activated as many loci within the genome, again hinting at the importance of xylan and arabinoxylan to this strain’s metabolic strategy *in vivo*. Cecal expression data from E1 showed that 15 of these activated PULs were expressed *in vivo* on one or both of the diets tested (see circles to the right of heatmap in Fig. 5B). In mice fed the polysaccharide-rich LF/HPP chow, *B. cellulosilyticus* WH2 upregulates three *susC/D* pairs (BWH2_2717/8, 4044/5, 4046/7) whose expression is activated *in vitro* by arabinan and xylan/arabinoxylan. The three most significantly upregulated *susC/D* pairs (BWH2_1736/7, 2514/5, 4074/5) in mice fed the HF/HS diet rich in sugar, corn starch and maltodextrin are activated *in vitro* by amylopectin, ribose, and xylan/arabinoxylan, respectively. All three PULs identified as being upregulated at the RNA level in LF/HPP-fed mice
were also found to be upregulated at the protein level (Fig. 5B). Two of the three PULs induced at the mRNA level in HF/HS-fed mice were shown to be upregulated at the protein level as well, along with four other PULs whose upregulation was only evident in our LC-MS/MS data. The presence of a dextran (starch)-activated PUL among the latter four loci is noteworthy, given the significant amount of starch present in the HF/HS diet. Such examples of loci whose induction is expected but evident only in one data type or the other reinforce the notion that protein data both complement and supplement mRNA data when profiling microbes of interest.

Two of the five susC/D pairs activated by xylan/arabinoxylan form the four-gene cassette in the previously discussed PUL comprising BWH2_4044-55 that is activated in mice fed the plant polysaccharide-rich chow (see Fig. 4A). Another one of these five is the susC/D pair found in the PUL comprising BWH2_4072-6 that is activated in mice fed the HF/HS ‘Western’-like chow (see Fig. 4A). Thus, we have identified a pair of putative PULs in close proximity to one another on the B. cellulosilyticus WH2 genome that encode CAZymes with similar predicted functions, are subject to near-identical levels of specific activation by the same two polysaccharides (i.e., xylan, arabinoxylan) in vitro, but are discordantly regulated in vivo in a diet-specific manner. The highly expressed nature of these PULs in the diet environment where they are active, their shared emphasis on xylan/arabinoxylan utilization, and their tight regulation indicate that they are likely to be important for the in vivo success of this organism in the two nutrient environments tested. However, the reasons for their discordant regulation are unclear. One possibility is that in addition to being activated by xylan/arabinoxylan and related polysaccharides, these loci are also subject to repression by other substrates present in the lumen of the gut, and this repression is sufficient to block activation. Alternatively, the specific activators of each PUL may be molecular moieties shared by both xylan and arabinoxylan that do not co-occur in the luminal environment when mice are fed the diets tested in this study.
Prospectus

Elucidating generalizable ‘rules’ for how microbiota operate under different environmental conditions is a substantial challenge. As our appreciation for the importance of the gut microbiota in human health and well-being grows, so too does our need to develop such rules using tractable experimental models of the gut ecosystem that allow us to move back and forth between in vivo and ex vivo analyses, using one to inform the other. Here, we have demonstrated the extent to which high-resolution DNA-, mRNA- and protein-level analyses can be applied (and integrated) to study a model community of sequenced human gut microbes colonizing gnotobiotic mice. Our efforts have focused on characterizing community-level and species-level adaptation to dietary change over time and ‘leverage’ results obtained from in vitro expression profiling of individual community member responses to a panel of purified carbohydrates to deduce glycan exposures and consumption strategies in vivo. This experimental paradigm could be applied to any number of questions related to microbe-microbe, environment-microbe and host-microbe interactions, including for example the metabolic fate of particular nutrients of interest (metabolic flux experiments), microbial succession, and biotransformations of xenobiotics. Studying representative model communities in gnotobiotic mice also allows us to evaluate the technical limitations of current molecular approaches for characterizing native gut communities. For example, the structure of a defined community can be evaluated over time at low cost using short read shotgun DNA sequencing data mapped to all microbial genomes within the community (COPRO-Seq). This allows for a much greater depth of sequencing coverage (i.e., more sensitivity) and much less ambiguity in the assignment of reads to particular taxa than traditional 16S rRNA gene-based sequencing allows. Short read cDNA sequences transcribed from total community RNA can also often be assigned to the exact microbial gene from which they were derived, and the same is also often true for peptides derived from particular microbial proteins. However, substantial dynamic range in species/transcript/protein abundance within any microbiota, defined or otherwise, imposes limits on our ability to characterize the least abundant elements of these systems.
The effort to obtain a more complete understanding of the operations and behaviors of minor components of the microbiota is an area deserving of significant attention, given known examples of low-abundance taxa that play key roles within their larger communities and in host physiology [5,36]. Developing such an understanding requires methods and assays that are collectively capable of assessing the structure and function of a microbiota at multiple levels of resolution. The need for high sensitivity and specificity in these approaches will become increasingly relevant as we transition towards experiments involving defined communities of even greater complexity, including bacterial culture collections prepared from the fecal microbiota of humans [37]. We anticipate that the study of culture collections transplanted to gnotobiotic mice will be instrumental in determining the degree to which physiologic or pathologic host phenotypes can be ascribed to the microbiota as well as specific constituent taxa.

One of the greatest advantages of studying defined model microbiota is the ability they afford to interrogate the biology of key bacterial species of interest in a focused manner. The synthetic model human gut microbiota we studied included *Bacteroides cellulosilyticus* WH2, a species that warrants further study as a model gut symbiont given its exceptional carbohydrate utilization capabilities, its apparent fitness advantage over many other previously characterized gut symbionts, and its genetic tractability. This genetic tractability should facilitate future experiments in which transposon mutant libraries are screened *in vivo* as one component of a larger model community in order to identify this strain’s most important fitness determinants under a wide variety of dietary conditions. Identifying the genetic elements that allow *B. cellulosilyticus* to persist at the relatively high levels observed, regardless of diet, should provide microbiologists and synthetic biologists with new ‘standard biological parts’ that will aid them in developing the next generation of prebiotics, probiotics, and synbiotics.
Figure legends

Figure 1. COPRO-Seq analysis of the structure of a 12-member model human microbiota as a function of diet and time. (A) The 12 bacterial species comprising the model microbiota in this study. (B) Principal coordinates analysis (PCoA) was applied to relative abundance data generated by COPRO-Seq from two experiments (E\textsubscript{1}, E\textsubscript{2}), each spanning six weeks. Following colonization (day 0), mice were switched between two different diets at two-week intervals as described in Fig. S3. COPRO-Seq data from E\textsubscript{1} and E\textsubscript{2} were ordinated in the same multi-dimensional space. For clarity, only data from E\textsubscript{2} are shown here (for the E\textsubscript{1} PCoA plot, see Fig. S5A). Red/blue, feces; magenta/cyan, cecal contents. (C) Proportional abundance data from E\textsubscript{1} illustrating the impact of diet on fecal levels of a diet-sensitive strain with higher representation on HF/HS chow (B. caccae), a diet-sensitive strain with higher representation on LF/HPP chow (B. ovatus), a diet-insensitive strain with no obvious diet preference (B. thetaiotaomicron), and a diet-sensitive strain with a preference for the LF/HPP diet that also achieves a high level of community representation on the HF/HS diet (B. cellulosilyticus WH2). Mean values ± SEM are shown. Plots illustrating changes in abundance over time for all species in both experiments are provided in Fig. S4C.

Figure 2. B. cellulosilyticus WH2 CAZyme expression in mice fed different diets.

(A) Overview of the 50 most highly expressed B. cellulosilyticus WH2 CAZymes (i.e., glycoside hydrolases (GH), glycosyltransferases (GT), polysaccharide lyases (PL) and carbohydrate esterases (CE)) for samples from each diet treatment group. List position denotes the rank order of gene expression for each treatment group, with higher expression levels situated at the top of each list. Genes common to both lists are identified by a connecting line, with the slope of the line indicating the degree to which a CAZyme’s prioritized expression is increased/decreased from one diet to the other. CAZy families in bold, colored letters highlight those list entries found to be significantly upregulated relative to the alternative diet (i.e., a CAZyme with a bold green family designation was upregulated on the LF/HPP diet; a bold orange family name implies a gene was upregulated significantly on HF/HS diet). Statistically significant fold-changes between diets are denoted in
the “F.C.” column (non-significant fold-changes are omitted for clarity). (B) Breakdown by CAZy family of the top 10% most expressed CAZymes on each diet whose expression was also found to be significantly higher on one diet than the other. Note that for each diet, the family with the greatest number of upregulated genes was also exclusively upregulated on that diet (LF/HPP, GH43; HF/HS, GH13). In total, 25 genes representative of 27 families and 12 genes representative of 13 families are shown for the LF/HPP and HF/HS diets, respectively.

**Figure 3. Top-down analysis of fecal microbiome gene expression in mice receiving oscillating diets.** The fecal metatranscriptomes of four animals in the LF/HPP➔HF/HS➔LF/HPP treatment group of E₂ were analyzed using microbial RNA-Seq at seven time-points to evaluate the temporal progression of changes in expressed community functions triggered by a change in diet. After aligning reads to genes in the defined model human gut microbiome, raw counts were collapsed by the functional annotation (EC number) of the gene from which the corresponding reads originated. Total counts for each EC number in each sample were normalized, and any EC numbers demonstrating a statistically significant difference in their representation in the metatranscriptome between the final days of the first two diet phases were identified using a model based on the negative binomial distribution [38]. Normalized expression values for 157 significant EC numbers (out of 1,021 total tested) were log-transformed, mean-centered, and subjected to hierarchical clustering, followed by heatmap visualization. ‘Rapid’ responses are those where expression increased/decreased dramatically within 1-2 days of a diet switch; ‘gradual’ responses are those where expression changed notably, but slowly, between the two diet transition points; and ‘delayed’ responses denote changes that did not achieve statistical significance until the end of a diet phase. EC numbers specifying enzymatic reactions relevant to carbohydrate metabolism and/or transport are denoted by purple markers, while those with relevance to amino acid metabolism are indicated using orange markers. A full breakdown of all significant responses over time and the outputs of the statistical tests performed are provided in Table S7.

**Figure 4. Two xylanase-containing *B. cellulosilyticus WH2* PULs demonstrating strong diet-specific expression patterns *in vivo.* (A) The PUL spanning genes BWH2_4044-55 includes a
four-gene cassette comprising two consecutive susC/D pairs, multiple genes encoding glycoside hydrolases (GHs) and carbohydrate esterases (CEs), and a gene encoding a putative hybrid two-component system (HTCS) presumed to play a role in the regulation of this locus. GH10 enzymes are endo-xylanases (most often endo-β-1,4-xylanases), while some GH5 and GH8 enzymes are also known to have endo- or exo-xylanase activity. CE6 enzymes are acetyl xylan esterases, as are some members of the CE1 family. A second PUL spanning genes BWH2_4072-6 contains a susC/D cassette, an endo-xylanase with dual GH10 modules as well as dual carbohydrate (xylan) binding modules (CBM22), a hypothetical protein of unknown function, and a putative HTCS. (B) Heat map visualization of GeneChip expression data for BWH2_4044-55 and BWH2_4072-6 showing marked upregulation of these putative PULs when mice are fed either a plant polysaccharide-rich LF/HPP diet or a diet high in fat and simple sugar (HF/HS), respectively. Data are from cecal contents harvested from mice at the endpoint of experiment E1. (C) Mass spectrometry-based quantitation of the abundance of all cecal proteins from the BWH2_4044-55 and BWH2_4072-6 PULs that were detectable in the same material used for GeneChip quantitation in panel B. Bars represent results (mean ± SEM) from two technical runs per sample. For each MS run, the spectral counts for each protein were normalized against the total number B. cellulosilyticus WH2 spectra acquired. (D) Comparison of in vivo PUL gene expression as measured by RNA-Seq (top) and the degree to which disruption of each gene within each PUL by a transposon impacts the fitness of B. cellulosilyticus WH2 on each diet, as measured by insertion sequencing (INSeq, bottom). For the lower plots, fitness measurements are calculated by dividing a mutant’s representation (normalized sequencing counts) within the fecal output community by the mutant’s representation within an input population that was introduced into germ-free animals by oral gavage along with other model community members. For cases in which no instances of a particular mutant could be measured in the fecal output community (resulting in a fitness calculation denominator of zero), the data are plotted as “<0.01” and are drawn without error bars.

Figure 5. In vitro microbial RNA-Seq profiling of B. cellulosilyticus WH2 during growth on different carbohydrates. (A) Hierarchical clustering of the gene expression profiles of 90 cul-
tures grown in minimal medium supplemented with one of 31 simple or complex sugars (n=2-3 replicates per condition). Circles at dendrogram branch points identify clusters with strong bootstrapping support (>95%; 10,000 repetitions). Solid circles denote clusters comprising only replicates from a single treatment group/carbohydrate, while open circles denote higher-level clusters comprising samples from multiple treatment groups. Colored rectangles indicate the type of carbohydrate on which the samples within each cluster were grown. (B) Unclustered heat map representation of fold-changes in gene expression relative to growth on minimal medium plus glucose (MM-Glc) for 60 of the 236 paired susC- and susD-like genes identified within the B. cellulosilyticus WH2 genome (for a full list of all paired and unpaired susC and susD homologs, see Table S2). Data shown are limited to those genes whose expression on at least one of the 31 carbohydrates tested demonstrated a >100-fold increase relative to growth on MM-Glc for at least one of the replicates within the treatment group. Yellow boxes denote areas of the map where both genes in a susC/D pair were upregulated >100-fold for at least two of the replicates in a treatment group and where the average upregulation for each gene in the pair was >100-fold across all replicates of the treatment group. Two sets of columns to the right of the heatmap indicate PULs that were detectably expressed at the mRNA level (left set of columns) and/or protein level (right set of columns) in experiment 1 (E1). Red and black circles indicate that both genes in a susC/D pair were consistently expressed on a particular diet, as determined by GeneChip analysis of cecal RNA (≥5 of 7 animals assayed) or LC-MS/MS analysis of cecal protein (2 of 2 animals assayed). In both cases, a red circle denotes significantly higher expression on one diet compared to the other.
Figures

Figure 1.

A) Bacteroides caccae
Bacteroides cellulosilyticus WH2
Bacteroides ovatus
Bacteroides thetaiotaomicron
Bacteroides uniformis
Bacteroides vulgatus
Clostridium scindens
Clostridium spiroforme
Collinsella aerofaciens
Dorea longicaudata
Parabacteroides distasonis
Ruminococcus obeum

B) LF/HPP  |  HF/HS  |  LF/HPP  |  HF/HS
0  |  14  |  28  |  42

C) Relative abundance (%)

B. caccae
B. thetaiotaomicron
B. ovatus
B. cellulosilyticus WH2

Day
0  |  14  |  28  |  42

Relative abundance (%)
0  |  10  |  20  |  30  |  40  |  50

Day
0  |  14  |  28  |  42

Relative abundance (%)
0  |  10  |  20  |  30  |  40  |  50
Figure 2.

A

<table>
<thead>
<tr>
<th>CAZy family</th>
<th>F.C. Gene</th>
<th>Gene</th>
<th>F.C. CAZy family</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH36</td>
<td>BWH2_1228</td>
<td>BWH2_4167</td>
<td>4.1 GH13</td>
</tr>
<tr>
<td>GH3</td>
<td>BWH2_2714</td>
<td>BWH2_4072</td>
<td>9.7 GH10-GH10</td>
</tr>
<tr>
<td>GH26</td>
<td>BWH2_1876</td>
<td>BWH2_1495</td>
<td>2.7 GH2</td>
</tr>
<tr>
<td>GH10</td>
<td>BWH2_4049</td>
<td>BWH2_1228</td>
<td>-3.5 GH36</td>
</tr>
<tr>
<td>CE1</td>
<td>BWH2_2721</td>
<td>BWH2_4107</td>
<td>5.3 GH97</td>
</tr>
<tr>
<td>GH43-GH10</td>
<td>BWH2_1895</td>
<td>BWH2_3997</td>
<td>GH17</td>
</tr>
<tr>
<td>GT4</td>
<td>BWH2_0888</td>
<td>BWH2_1876</td>
<td>GH26</td>
</tr>
<tr>
<td>GH8</td>
<td>BWH2_2829</td>
<td>BWH2_1876</td>
<td>GH26</td>
</tr>
<tr>
<td>GH43</td>
<td>BWH2_2719</td>
<td>BWH2_4066</td>
<td>GH43</td>
</tr>
<tr>
<td>GH10</td>
<td>BWH2_4050</td>
<td>BWH2_4016</td>
<td>6.2 GH97</td>
</tr>
<tr>
<td>GH57</td>
<td>BWH2_3997</td>
<td>BWH2_1184</td>
<td>3.5 GH13</td>
</tr>
<tr>
<td>GH43</td>
<td>BWH2_2713</td>
<td>BWH2_1164</td>
<td>4.7 GH13</td>
</tr>
<tr>
<td>GH77</td>
<td>BWH2_1495</td>
<td>BWH2_3676</td>
<td>CE11</td>
</tr>
<tr>
<td>GT2</td>
<td>BWH2_2294</td>
<td>BWH2_4068</td>
<td>GH10-GH43</td>
</tr>
<tr>
<td>GH10-CE1</td>
<td>BWH2_1708</td>
<td>BWH2_2294</td>
<td>GT2</td>
</tr>
<tr>
<td>GH51</td>
<td>BWH2_0846</td>
<td>BWH2_0480</td>
<td>GT2</td>
</tr>
<tr>
<td>GH51</td>
<td>BWH2_2781</td>
<td>BWH2_4003</td>
<td>2.1 GT35</td>
</tr>
<tr>
<td>GH5</td>
<td>BWH2_3132</td>
<td>BWH2_5229</td>
<td>GH13</td>
</tr>
<tr>
<td>GH8</td>
<td>BWH2_4054</td>
<td>BWH2_1532</td>
<td>2.5 GH51</td>
</tr>
<tr>
<td>GH10-GH43</td>
<td>BWH2_4088</td>
<td>BWH2_4017</td>
<td>2.7 GH92</td>
</tr>
<tr>
<td>GH10</td>
<td>BWH2_4088</td>
<td>BWH2_3999</td>
<td>GT5</td>
</tr>
<tr>
<td>GH13</td>
<td>BWH2_4167</td>
<td>BWH2_0846</td>
<td>GH10</td>
</tr>
<tr>
<td>GH43</td>
<td>BWH2_1896</td>
<td>BWH2_4064</td>
<td>GH67</td>
</tr>
<tr>
<td>GH26</td>
<td>BWH2_1875</td>
<td>BWH2_4294</td>
<td>GH23</td>
</tr>
<tr>
<td>GH9</td>
<td>BWH2_2716</td>
<td>BWH2_1875</td>
<td>GH26</td>
</tr>
<tr>
<td>GH43</td>
<td>BWH2_2654</td>
<td>BWH2_1705</td>
<td>GH26</td>
</tr>
<tr>
<td>GT2</td>
<td>BWH2_1498</td>
<td>BWH2_2300</td>
<td>GT9</td>
</tr>
<tr>
<td>GH130</td>
<td>BWH2_1874</td>
<td>BWH2_2295</td>
<td>GT2</td>
</tr>
<tr>
<td>GT32</td>
<td>BWH2_4925</td>
<td>BWH2_4113</td>
<td>GT4</td>
</tr>
<tr>
<td>CE1</td>
<td>BWH2_4052</td>
<td>BWH2_0345</td>
<td>GH2</td>
</tr>
<tr>
<td>GH43</td>
<td>BWH2_2715</td>
<td>BWH2_1630</td>
<td>2.8 GH2</td>
</tr>
<tr>
<td>GH5</td>
<td>BWH2_3131</td>
<td>BWH2_0846</td>
<td>GH10</td>
</tr>
<tr>
<td>GH3</td>
<td>BWH2_2830</td>
<td>BWH2_0811</td>
<td>GH10</td>
</tr>
<tr>
<td>GT2</td>
<td>BWH2_2295</td>
<td>BWH2_2143</td>
<td>GH3</td>
</tr>
<tr>
<td>CE1</td>
<td>BWH2_3876</td>
<td>BWH2_1498</td>
<td>GT2</td>
</tr>
<tr>
<td>GH3</td>
<td>BWH2_1885</td>
<td>BWH2_4294</td>
<td>GH23</td>
</tr>
<tr>
<td>GH32</td>
<td>BWH2_1235</td>
<td>BWH2_1034</td>
<td>GH23</td>
</tr>
<tr>
<td>GH5</td>
<td>BWH2_4051</td>
<td>BWH2_1674</td>
<td>GH130</td>
</tr>
<tr>
<td>GH13</td>
<td>BWH2_5229</td>
<td>BWH2_1874</td>
<td>GH130</td>
</tr>
<tr>
<td>CE7</td>
<td>BWH2_1883</td>
<td>BWH2_2253</td>
<td>GH3</td>
</tr>
<tr>
<td>GT5</td>
<td>BWH2_3999</td>
<td>BWH2_4010</td>
<td>GH2</td>
</tr>
<tr>
<td>GH2</td>
<td>BWH2_0345</td>
<td>BWH2_1062</td>
<td>CE1</td>
</tr>
<tr>
<td>GH23</td>
<td>BWH2_4294</td>
<td>BWH2_2022</td>
<td>GH9</td>
</tr>
<tr>
<td>CE15</td>
<td>BWH2_2127</td>
<td>BWH2_1515</td>
<td>GH43-GH35</td>
</tr>
<tr>
<td>GH43</td>
<td>BWH2_2646</td>
<td>BWH2_2714</td>
<td>-5.7 GH3</td>
</tr>
<tr>
<td>GH97</td>
<td>BWH2_2055</td>
<td>BWH2_1885</td>
<td>GH3</td>
</tr>
<tr>
<td>GH97</td>
<td>BWH2_2651</td>
<td>BWH2_4071</td>
<td>GH35</td>
</tr>
<tr>
<td>GH43-GH35</td>
<td>BWH2_1515</td>
<td>BWH2_3243</td>
<td>GH20</td>
</tr>
<tr>
<td>GH97</td>
<td>BWH2_4107</td>
<td>BWH2_2924</td>
<td>GH52</td>
</tr>
<tr>
<td>GT9</td>
<td>BWH2_2300</td>
<td>BWH2_4002</td>
<td>GT3</td>
</tr>
</tbody>
</table>

B

Incidence among highly expressed, upregulated CAZymes

- Up on HF/HS
- Up on LF/HPP

CAZy family
Figure 3.

Rapid responses (33) (higher on HF/HS)
Gradual responses (29) (higher on HF/HS)
Delayed responses (19) (higher on HF/HS)
Rapid responses (24) (higher on LF/HPP)
Gradual responses (22) (higher on LF/HPP)
Delayed responses (30) (higher on LF/HPP)

Mean-centered, log2 expression level

Carbohydrate metabolism and transport
Amino acid metabolism
Figure 4.

A  BWH2_4044-55

<table>
<thead>
<tr>
<th>4044</th>
<th>4045</th>
<th>4046</th>
<th>4047</th>
<th>4048</th>
<th>4049</th>
<th>4050</th>
<th>4051</th>
<th>4052</th>
<th>4053</th>
<th>4054</th>
<th>4055</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH10</td>
<td>GH10</td>
<td>GH5</td>
<td>CE1</td>
<td>CE5-GH95</td>
<td>GH8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BWH2_4072-6

<table>
<thead>
<tr>
<th>4072</th>
<th>4073</th>
<th>4074</th>
<th>4075</th>
<th>4076</th>
</tr>
</thead>
</table>

GH10-(CBM22),GH10

B

LF/HPP  HF/HS

susc-like  susD-like  carbohydrate esterase (CE)  hypothetical  hybrid two-component system

C

Mouse #1 (LF/HPP)  Mouse #2 (LF/HPP)  Mouse #9 (HF/HS)  Mouse #12 (HF/HS)

Normalised spectral counts

D

Expression

Fitness

Mouse #1 (LF/HPP)  Mouse #2 (LF/HPP)  Mouse #9 (HF/HS)  Mouse #12 (HF/HS)
Figure 5

A

B

Fold-change relative to Glucose-MM

<10 10 100 500 >500

Expressed

Upregulated
Supplementary material

Supplementary materials and methods

**B. cellulosilyticus WH2 genome sequencing.** A strain of *B. cellulosilyticus* designated ‘WH2’ (see Figs. S1A, B) was isolated from a human fecal sample during an iteration of the Microbial Diversity Summer Course overseen by A. Salyers (Univ. of Illinois, Urbana-Champaign) at the Marine Biological Laboratory (Woods Hole, MA). The genome of this isolate was sequenced using a combination of long-read and short-read technologies, yielding 51,819 plasmid and fosmid end reads (library insert sizes: 3.9, 4.9, 6.0, 8.0, and 40 kb; ABI 3730 platform), 333,883 unpaired 454 reads (FLX+ and XL+ chemistry), and 10 million unpaired Illumina reads (HiSeq; 42 nt read length). A hybrid assembly was constructed using MIRA v3.4.0 (method: *de novo*; type: genome; quality grade: accurate) with default settings [39,40]. Gene calling was performed using the YACOP metatool [41]. Additionally, the four ribosomal RNA (rRNA) operons within the *B. cellulosilyticus* WH2 genome were sequenced individually to ensure high sequence accuracy in these difficult to assemble regions. Further details for the *B. cellulosilyticus* WH2 assembly are provided in Table S1.

**Bacterial strains.** Details regarding the 12 bacterial strains used in this study are provided in Table S3. Cells were grown in supplemented TYG (TYG$_s$; [34]) at 37°C under anaerobic conditions in a Coy anaerobic chamber (atmosphere; 75% N$_2$, 20% H$_2$, 5% CO$_2$). After reaching stationary phase, cells were pelleted by centrifugation and resuspended in TYG$_s$ medium supplemented with 20% glycerol. Individual aliquots containing 400-800 µL of each cell suspension were stored at -80°C in 1.8 mL borosilicate glass vials with aluminum crimp tops. The identity of each species was verified prior to its use in experiments by extracting DNA from a frozen aliquot of cells, amplifying the 16S rRNA gene by PCR using primers 8F/27F (AGAGTTTGATCCTGGCTCAG; [42]) and 1391R (GACGGGCGGTGWGTRCA; [43]), sequencing the entire amplicon with an ABI 3730 capillary sequencer (Retrogen, Inc.), and comparing the assembled 16S rRNA gene sequence to the known reference sequence.
**Preparation of strains for oral gavage.** Details regarding the construction of each inoculum are provided in Table S4. The inocula used to gavage germ-free mice in each experiment were prepared either directly from frozen stocks (experiment #1, E₁) or from a combination of frozen stocks and overnight cultures (experiment #2, E₂). The recoverable cell density for each batch of frozen stocks used in inoculum preparation was determined prior to pooling, while the same values for overnight cultures were calculated after pooling. An aliquot of cells from each overnight culture or set of frozen stocks was used to prepare a ten-fold dilution series in phosphate-buffered saline (PBS), and each dilution series was plated on brain-heart-infusion (BHI; BD Difco) agar supplemented with 10% (v/v) defibrinated horse blood (Colorado Serum Co.). Plates were grown for up to 3 d at 37°C under anaerobic conditions in a Coy chamber, colonies were counted, and the number of colony-forming units per milliliter (CFUs/mL) was calculated. The volume of each cell suspension included in the final inoculum was normalized by its known or estimated viable cell concentration in an effort to ensure that no species received an early advantage during establishment of the model human gut microbiota in the germ-free animals. Total CFUs per gavage were estimated at 8.0x10⁷ and 4.2x10⁸ for experiments E₁ and E₂, respectively.

**Mice.** Experiments were performed using protocols approved by the animal studies committee of the Washington University School of Medicine (animal studies protocol #20110232). For each experiment, two groups of 10-12 week-old male germ-free C57BL/6J mice were maintained in flexible film gnotobiotic isolators under a strict 12 h light cycle, during which time they received sterilized food and water *ad libitum*. Animals were fasted for 4 h prior to gavage with 500 µL of a cell suspension inoculum containing the 12 sequenced, human-gut derived bacterial symbionts. After gavage, animals were maintained in separate cages throughout the course of the experiment. Fresh fecal pellets were periodically collected directly into screw-cap sample tubes that were immediately frozen in liquid nitrogen. At the time of sacrifice, the contents of each animal’s cecum were divided into thirds and snap-frozen in liquid nitrogen for later use in DNA, RNA, and total protein isolations.
**Diets.** Animals were subjected to dietary oscillations comprising three consecutive phases of two weeks each (see Fig. S3). Prior to inoculation, germ-free mice were maintained on a standard autoclaved chow diet low in fat and rich in plant polysaccharides (‘LF/HPP’, B&K rat and mouse autoclavable chow #73780000, Zeigler Bros, Inc). Three days prior to inoculation, one group of germ-free animals was switched to a ‘Western’-like chow high in fat and simple sugars (‘HF/HS’; Harlan Teklad TD96132), while the other continued to receive LF/HPP chow. After gavage, each group of animals was maintained on its respective diet for two weeks, after which each treatment group was switched to the alternative diet. Two weeks later, the diets were switched back to their original starting diet and were retained on this diet up through the time of sacrifice. The ingredients in plant polysaccharide-rich LF/HPP chow include: wheat, soybean meal, wheat middlings, corn, fish meal, lignin sulfoate, soy oil, dicalcium phosphate, limestone, sodium chloride, calcium propionate, L-lysine, choline chloride, zinc sulfate, vitamin E, and other vitamin and mineral premixes. The ingredients in the ‘Western’-like HF/HS diet include (% by weight): casein (23.60%), sucrose (18.262%), corn starch (16.0%), maltodextrin (12.0%), shortening (10.0%), beef tallow (10.0%), mineral mix [TD 94046] (4.13%), cellulose (4.0%), vitamin mix [Teklad 40060] (1.18%), calcium phosphate [dibasic] (0.472%), DL-methionine (0.354%), and ethoxyquin (0.002%).

**DNA and RNA extraction.** DNA and RNA were extracted from fecal pellets and cecal contents as previously described [8].

**COmmunity PROfiling by Sequencing (COPRO-Seq).** COPRO-Seq measurements of the proportional representation of all species present in each fecal/cecal sample analyzed were performed as previously described [8] using short-read (36 nt) data collected from an Illumina GAIIX sequencer. After deplexing of each barcoded pool, reads were trimmed to 25 bp prior to alignment to the reference genomes. An abundance threshold cutoff of 0.003% was set, based on the proportion of reads from each experiment that were found to spuriously align to distractor reference genomes of bacterial species not included in this study. Normalized counts for each bacterial species in each sample were used to calculate a simple intra-sample percentage (“% of community”).
In order to make changes in abundance over time more easily comparable between species with significantly different relative abundances, these percentages were also normalized by the maximum abundance (%) observed for a given species across all time-points from a given animal. This transformation resulted in a value representing the ratio of a given percentage to the maximum percentage achieved (“% of % max”) that was used to evaluate which species were most/least responsive to dietary interventions.

**Ordination of COPRO-Seq data using QIIME.** COPRO-Seq proportional abundance data were subjected to ordination using scripts found in QIIME v1.5.0-dev [44]. Data from both E₁ and E₂ were combined to generate a single tab-delimited table conforming to QIIME’s early (pre-v1.4.0-dev) OTU table format. This pseudo-OTU table was subsequently converted into a biom-formatted table object that was used as the input for beta_diversity.py to calculate the pairwise distances between all samples using a Hellinger metric. Principle coordinates analysis (PCoA) calculations were performed using principal_coordinates.py. These coordinates and sample metadata were passed to make_3d_plots.py to generate PCoA plots. Plots shown are visualized using v2.21 of the KiNG software package [45].

**Metatranscriptomics (GeneChip).** A custom Affymetrix GeneChip (‘SynComm1’) with perfect match/mismatch (PM/MM) probe sets targeting 97.6% of the predicted protein-coding genes within the genomes of the 12 bacterial species in this study (plus three additional species not included in the model human gut microbiota) was designed and manufactured in collaboration with the Affymetrix chip design team. Control probes targeting intergenic regions from each genome were also tiled onto the array to allow detection of any contaminating gDNA. Hybridizations were carried out with 0.9-5.1 µg cDNA using the manufacturer’s recommended protocols. Details regarding the design of this GeneChip are deposited under Gene Expression Omnibus (GEO) accession GPL9803.

Custom mask files were generated for each species on the GeneChip for the purposes of performing data normalization one species at a time. Normalization of raw intensity values was
carried out in Affymetrix Microarray Suite (MAS) v5.0. MAS output was exported to Excel where advanced filtering was used to identify those probe sets called present in at least 5 of 7 cecal RNA samples in at least one diet tested. Data from probe sets that did not meet these criteria (i.e., genes that were not expressed on either condition) were not included in subsequent analyses. Normalized, filtered data were evaluated using the Cyber-T web server [46] to identify differentially expressed genes. Genes were generally considered significantly differentially expressed in cases meeting the following three criteria: \( p < 0.01 \), \( \text{PPDE}(\< p) \geq 0.99 \), and \( |\text{fold-change}| \geq 2 \).

**Metatranscriptomics (microbial RNA-Seq).** Methods for extracting total microbial RNA from mouse feces and cecal contents, depleting small RNAs (e.g., tRNA) and ribosomal RNA (5S, 16S, and 23S rRNA), and for converting depleted RNA to double-stranded cDNA were described previously [10]. Illumina libraries were prepared [8] from 26 fecal samples obtained from the second diet oscillation experiment (4 animals, 6-7 time-points surveyed per animal), using 500 ng of input double-stranded cDNA/sample/library. RNA-Seq reads were aligned to the reference genomes using the SSAHA2 aligner [47]. Normalization of the resulting raw counts was performed using the DESeq package in R [38]. Raw counts derived from the community metatranscriptome were normalized either at the community level (i.e., counts from all genes were included in the same table during normalization) for purposes of looking at community-level representation of functions (ECs) of interest, or at the species level (i.e., counts from each species were independently normalized) for purposes of looking at gene expression changes within individual species. Data adjustment (logarithmic transformation) and hierarchical clustering were performed using Cluster 3.0 [48] and GENE-E. Heatmap visualizations of expression data were prepared using JavaTreeView [49] and Microsoft Excel. The *B. cellulosilyticus* WH2 *in vitro* gene expression dendrogram presented was prepared using GENE-E. Bootstrap probabilities at each edge of the dendrogram were calculated using the ‘pvclust’ package in R (10,000 replications). Clusters with bootstrap \( p \)-values >0.95 were considered strongly supported and statistically significant.

**Metaproteomics (sample preparation).** Cecal contents were collected from four mice and solubilized in 1 mL SDS lysis buffer (4% w/v SDS, 100mM Tris-HCl (pH 8.0), 10mM dithioth-
reitol (DTT)), lysed mechanically by sonication, incubated at 95°C for 5 min, and centrifuged at 21,000 x g. Crude protein extracts were precipitated using 100% trichloroacetic acid (TCA), pelleted by centrifugation, and washed with ice-cold acetone to remove lipids and excess SDS. Protein precipitates were resolubilized in 500 µL of 8M urea and 100mM Tris·HCl (pH 8.0), reduced by incubation in DTT (final concentration of 10mM) for 1 h at room temperature, and sonicated in an ice water bath (Branson (model SSE-1) sonicator; 20% amp; 2 min total (cycles of 5 sec on, 10 sec off)). An aliquot of each protein extract was quantified using a bicinchoninic acid (BCA)-based protein assay kit (Pierce). Protein samples (1 mg) were subsequently diluted with 100mM Tris·HCl and 10mM CaCl₂ (pH 8.0) to a final urea concentration below 4M. Proteolytic digestions were initiated with sequencing grade trypsin (1/100, w/w; Promega) and incubated overnight at room temperature. A second aliquot of trypsin was added (1/100) after the reactions were diluted with 100mM Tris·HCl (pH 8.0) to a final urea concentration below 2M. After incubation for 4 h at room temperature, samples were reduced by incubation in 10 mM DTT for 1 h at room temperature. Finally, the peptides were acidified (protonated) in 200mM NaCl and 0.1% formic acid, filtered, and concentrated with a 10 kDa molecular weight cutoff spin column (Sartorius).

Metaproteomics (LC-MS/MS data collection). The peptide mixture from each mouse was analyzed in technical duplicate via two-dimensional liquid chromatography (LC)-MS/MS on a hybrid LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific). Peptides (~100 µL per sample) were separated using a split phase 2D (strong-cation exchange (SCX) and C₁₈ reverse phase (RP))-LC column over a 12-step gradient for each run. All MS analyses were performed in positive ion mode. Mass spectral data were acquired using Xcalibur (v2.0.7) in data-dependent acquisition mode for each chromatographic separation (22 h run). One precursor MS scan was acquired in the Orbitrap at 30K resolution followed by ten data-dependent MS/MS scans (m/z 400-1700) at 35% normalized collision energy with dynamic exclusion enabled at a repeat count of 1. MS/MS spectra were searched with SEQUEST (v.27; [50]) using the following settings: enzyme type = trypsin; precursor ion mass tolerance = 3.0 Da; fragment mass tolerance = 0.5 Da; fully tryptic peptides and those resulting from up to 4 missed cleavages only. All datasets were filtered
with DTASelect (v1.9; [51]) using the following parameters: Xcorrs of 1.8, 2.5, and 3.5 for singly, doubly, and triply charged precursor ions; DeltCN ≥ 0.08; ≥ 2 fully tryptic peptides per protein.

A custom-built FASTA target-decoy database [52,53] was generated and searched with SEQUEST at a peptide-level false positive rate (FPR) estimated at ≤ 0.5%. The database contained theoretical proteomes predicted from the genomes of the 12 bacterial species characterized in this study (see Tables S3, S8), some diet components (e.g., rice and yeast) and common contaminants (e.g., keratins). Three additional theoretical bacterial proteomes predicted from the genomes of *Eubacterium rectale, Faecalibacterium prausnitzii* M21/2 and *Ruminococcus torques* were included as distractors (negative controls) that were not expected to be present in any of the samples analyzed. An in silico tryptically digested protein sequence database was also used to generate a theoretical peptidome of unique peptides within a mass range of 600-4,890 Da and ≤ 1 miscleavings.

**Metaproteomics (analysis of proteomic datasets).** Spectral counts for each protein were normalized by either the total number of spectra collected for all species in a sample (normalization by community, ‘NBC’), or by the total number of spectra collected for all proteins from a given species (normalization by species, ‘NBS’). *p* values for each protein were calculated using the Mann–Whitney *U* test. To correct for multiple comparisons, *q* values were calculated using an optimized false discovery rate (FDR) approach with the ‘qvalue’ package in BioConductor. Regardless of the normalization strategy employed, *p* and *q* values were only calculated for proteins with at least 3 valid runs, where a valid run was one with more than 5 spectral counts. In NBC data, *p* and *q* values were calculated for all proteins within the model metaproteome. In NBS data, *p* and *q* values were calculated for each species-specific set of proteins. Differences in spectral counts between treatment groups (diets) were calculated using group medians. A protein was designated as ‘UP’-regulated if both its *p* and *q* values were less than 0.05 and the spectral count difference between treatment groups was greater than 5. The same criteria were applied in the opposite direction for proteins labeled as ‘DOWN’. For proteins labeled ‘NULL’, there was insufficient evidence to report any significant difference between the two treatment groups. Finally, a protein was con-
sidered detected, or ‘present’, in a sample if at least four (raw) spectral counts were assigned to that protein when aggregating the results from the two runs (technical replications) performed on the sample.

**Phenotypic screen of the growth of Bacteroides spp. on various carbohydrates.** The ability of *Bacteroides cellulosilyticus* WH2 and *Bacteroides caccae* (ATCC 43185) to grow on a panel of 47 simple and complex carbohydrates was evaluated using a phenotypic array whose composition has been previously described [18]. Growth measurements were collected in duplicate (2 wells per substrate) over the course of three days at 37°C under anaerobic conditions. A total of three independent experiments were performed for each species tested (i.e., n=6 growth profiles/per substrate/species). Total growth (Δₐₜₒₜ) was calculated from each growth curve as the difference between the maximum and minimum optical densities (OD₆₀₀) observed (i.e., A_max - A_min). Growth rates were calculated as total growth divided by time (Δₐₜₒₜ / (t_max - t_min)), where t_max and t_min correspond to the time-points at which A_max and A_min, respectively, were collected. Consolidated statistics from all six replicates for each of the 47 conditions tested for each species are provided in Table S11.

**Profiling of B. cellulosilyticus WH2 gene expression during growth in defined minimal medium containing various carbohydrates (microbial RNA-Seq).** To characterize the impact of select mono- and polysaccharides on the in vitro gene expression of *B. cellulosilyticus* WH2, cells were cultured in minimal medium (‘MM’) supplemented with one of 31 distinct carbohydrates (for the formulation of MM and a list of the carbohydrates used as growth substrates, see Tables S12, S13). After recovery from a frozen stock on BHI blood agar, a single colony was picked and inoculated into 5 mL of MM containing 5 mg/mL glucose. Anaerobic conditions were generated within each individual culture tube using a previously described method [54] with the following modifications: (i) the cotton plug was lit and extinguished before being pushed below the lip of the culture tube; and, (ii) 200 µL of saturated sodium bicarbonate was combined with 200 µL 35% (w/v) pyrogallate solution on top of the cotton plug before a bare rubber stopper was used to seal the tube. Cultures were grown overnight at 37°C. Twenty microliters of this ‘starter’ culture was
subsequently inoculated into a series of ‘acclimatization’ cultures, each containing 5 mL of MM
plus one of the 31 carbohydrates to be tested (5 mg/mL final concentration), and anaerobic culturing was carried out as above. This second round of culturing served two purposes: (i) it ensured cells were acclimated to growth on their new carbohydrate substrate prior to the inoculation of the final cultures that were harvested for RNA; and (ii) it provided an opportunity to obtain OD_{600}
measurements indicating, for each carbohydrate, the range of optical densities corresponding to \textit{B. cellulosilyticus} WH2’s logarithmic phase of growth. Finally, 50 µL of each ‘acclimatization’ culture was inoculated into triplicate 10 mL volumes of medium of the same composition, and the 93 ‘harvest’ cultures were grown anaerobically at 37°C. At mid-log phase, 5 mL of cells were immediately preserved in Qiagen RNAProtect Bacteria Reagent according to the manufacturer’s instructions. Cells were then pelleted, RNAProtect reagent was poured off, and the bacteria were stored at -80°C.

After thawing, while still cold, each bacterial cell pellet was combined with 500 µL Buffer B (200mM NaCl, 20mM EDTA), 210 µL of 20% SDS, and 500 µL of acid phenol:chloroform:isoamyl alcohol (125:24:1, pH 4.5). The pellet was resuspended by manual manipulation with a pipette tip and transferred to a 2 mL screwcap tube containing acid-washed glass beads (Sigma, 212-300 µm diameter). Tubes were iced, bead-beaten for 2 min at room temperature (BioSpec Mini-Beadbeater-8; set to ‘homogenize’), placed on ice, and bead-beaten for an additional 2 min, after which time the RNA extraction protocol proceeded as for fecal/cecal samples.

\textit{Identification of diet-specific fitness determinants within the B. cellulosilyticus WH2 genome using INsertion Sequencing (INSeq).} Whole genome transposon mutagenesis of \textit{B. cellulosilyticus} WH2 was performed using protocols originally developed for \textit{B. thetaiotaomicron} [Goodman 2009, 2011], with some modification. Initial attempts to transform \textit{B. cellulosilyticus} WH2 with the pSAM\_Bt construct reported by Goodman \textit{et al.} yielded very low numbers of antibiotic resistant clones, which we attributed to poor recognition of one or more promoters in the mutagenesis plasmid. Replacement of the promoter driving expression of the transposon’s erythromycin resistance gene (\textit{ermG}) with the promoter for the gene encoding EF-Tu in \textit{B. cellulosi-
lyticus WH2 (BWH2_3183) dramatically improved the number of resistant clones recovered after transformation. The resulting library consisted of 26,750 distinct transposon insertion mutants in 78.8% of all predicted ORFs (mean = 5.3 distinct insertion mutants / ORF).

At 11-weeks-of-age, male germ-free C57BL/6J mice (individually caged) were fed either a diet low in fat and rich in plant polysaccharides (LF/HPP) or high in fat and simple sugars (HF/HS). After a week on their experimental diet, the animals received a single gavage containing the B. cellulosilyticus WH2 transposon library and 14 other species of bacteria (i.e., the model community consisted of the 12 species listed in Fig. 1A, plus B. thetaiotaomicron 7330, Eubacterium rectale ATCC 33656, and Clostridium symbiosum ATCC 14940). After 16 d, fecal pellets were collected, and total fecal community DNA was extracted.

Five hundred nanograms of each fecal DNA extraction was diluted in 15 µL of TE buffer and digested with MmeI (4 U, New England Biolabs) in a 20 µL reaction supplemented with 10 pmoles of 12 bp DNA containing an MmeI restriction site (to improve the efficiency of restriction enzyme digestion) [Goodman, 2009]. The reaction was incubated for 1 h at 37°C and then terminated (80°C for 20 min). MmeI-digested DNA was subsequently purified using 125 µL of AMPure beads (after washing the beads once with 100 µL of sizing solution (1.2 M NaCl and 8.4% PEG 8000)). The digested DNA was added to the beads and the solution incubated at room temperature for 5 min. Beads were pelleted with a Magnetic Particle Collector (MPC), washed twice (each time using a mixture composed of 20 µL TE (pH 7.0) and 100 µL sizing solution, with bead recovery via MPC after each wash), followed by two ethanol washes (180µL 70% ethanol/wash) and air-drying for 10 min. Samples were resuspended in 18 µL TE (pH 7.0) and the DNA was removed after pelleting beads with the MPC. Ligation of adapters was performed in a 20 µL reaction that contained 16 µL of purified DNA, 1 µL of T4 Ligase (2000 U/µL; NEB), 2 µL 10X ligase buffer and 10 pmoles of barcoded adapter (incubation for 1 h at 16°C). Ligations were subsequently diluted with TE (pH 7.0) buffer to a final volume of 50 µL, mixed with 60 µL of AMPure beads, and incubated at room temperature for 5 min. Beads with bound DNA were pelleted using the MPC and washed twice with 70% ethanol as above. After allowing the ethanol to evaporate for 10
min, 35 µL of nuclease-free water was added and the mixture was incubated at room temperature for 2 min before collecting the beads with the MPC. Enrichment PCR was performed in a final volume of 50 µL using 32µL of the cleaned up sample DNA, 10 µL 10X Pfx amplification buffer (Invitrogen), 2 µL 10mM dNTPs, 0.5 µL 50mM MgSO₄, 2 µL of 5µM amplification primers (Forward primer: 5’CAAGCAGAAGACGGCATACG3’, Reverse primer: 5’AATGATACGGCGAC-CACCGAACACTCTTTCCCTACACGA3’), and 1.5 µL Pfx polymerase (2.5 U/µL; Invitrogen) 22 cycles of denaturation at 94°C for 15 sec, annealing at 65°C for 1 min and extension at 68°C for 30 sec. The 134 bp PCR product from each reaction was purified [4% metaphere gel; MiniElute Gel extraction kit (Qiagen)] in a final volume of 20 µL, and was quantified (Qubit, dsDNA HS Assay Kit; Invitrogen). Reaction products were then combined in equimolar amounts into a pool that was subsequently adjusted to 10nM and sequenced (Illumina Hi-Seq2000 instrument).

**Supplementary results**

*Evaluating the carbohydrate utilization capabilities and preferences of B. caccae, a HF/HS diet-adapted species.* Comparing CAZyme expression between three diet-insensitive *Bacteroides* spp. (*B. thetaiotaomicron*, *B. vulgatus*, and *B. cellulosilyticus* WH2) and HF/HS-favoring *B. caccae* revealed that these two groups have dissimilar profiles. While diet-insensitive strains express many CAZymes on both diets, and roughly equal percentages of their encoded CAZymes in a diet-specific manner, *B. caccae*’s CAZyme utilization is heavily skewed (**Fig. S6B**). While 19% of *B. caccae* CAZymes were expressed in mice regardless of the diet consumed, an additional 28% of this species’ predicted CAZymes were expressed in animals consuming the HF/HS diet. In contrast, *B. caccae* expressed only 1% of its predicted CAZymes in a LF/HPP diet-specific manner.

Phenotypic characterization of *B. caccae* on the same carbohydrate growth array we used to characterize *B. cellulosilyticus* WH2’s substrate utilization capabilities revealed significant deficiencies in *B. caccae*’s ability to utilize many simple and complex sugars (**Table S11**). Of particular note were its complete lack of any starch utilization system (as evidenced by its inability to grow on amylopectin derived from both potato and maize, as well as dextran and pullulan) and
its inability to utilize any type of hemicellulose or β-glucan tested. These deficiencies strongly contrast the strong growth we observed for \textit{B. cellulosilyticus} WH2 when it was grown on several such compounds. A complete comparison of the growth capabilities of \textit{B. cellulosilyticus} WH2 and \textit{B. caccae} reveals the striking fact that with the exception of one monosaccharide (N-acetylneuraminic acid), \textit{B. cellulosilyticus} WH2 growth outperforms that of \textit{B. caccae} on every carbohydrate tested.
Supplementary figure legends

Figure S1. Phylogenetic relatedness of *B. cellulosilyticus* WH2 to other *Bacteroides* species. (A) Near full-length 16S rRNA gene sequences from the *B. cellulosilyticus* WH2 isolate, its three closest relatives (*B. xylanisolvens*, *B. intestinalis*, and *B. cellulosilyticus*), and *Parabacteroides distasonis* (the latter was included as an outgroup) were aligned against the SILVA SEED using the SINA aligner [55]. The 5’ and 3’ ends of the resulting multiple sequence alignment were trimmed to remove ragged edges, and the final alignment was used to construct an approximately-maximum-likelihood phylogenetic tree using FastTree v2.1.4 [56]. Sequences in the trimmed alignment used to generate the tree shown correspond to bases 22-1498 of the *Escherichia coli* 16S rRNA gene [57]. Parenthetical identifiers indicate the locus tag (for *B. cellulosilyticus* WH2, whose genome contains four copies of the 16S rRNA gene) or accession number (for all other strains) of each sequence included in the phylogenetic analysis. (B) Identity matrix summarizing the pairwise similarities (as % nucleotide sequence identity) for all 16S rRNA gene sequences used to construct the tree shown in panel A.

Figure S2. Representation of all putative glycoside hydrolase (GH) families identified in the *B. cellulosilyticus* WH2 genome compared to their representation in other sequenced *Bacteroidetes* species. Enumeration of the GH repertoire of *Bacteroides cellulosilyticus* WH2 relative to (A) the six other *Bacteroidetes* species included in the model community described in Fig. 1A, and (B) the 86 *Bacteroidetes* currently annotated in the CAZy database. GH numbers in red signify CAZy families whose representation is greater in *B. cellulosilyticus* WH2 than any of the other *Bacteroidetes* to which it is being compared. An asterisk following a GH family number indicates genes encoding proteins from that family were found exclusively in the *B. cellulosilyticus* WH2 genome. In (B), GH family numbers are ordered from left to right and from top to bottom by their average representation within the 87 *Bacteroidetes* genomes included in the analysis.

Figure S3. Design and sampling schedule for experiments E₁ and E₂.
Figure S4. COPRO-Seq analysis of the proportional representation of component taxa in the 12-member model human gut microbiota as a function of time after colonization of gnotobiotic mice and the diet they were consuming. (A) Average DNA yields from fecal and cecal samples collected from each treatment group in experiment $E_1$. (B) DNA yields from samples collected in experiment $E_2$. (C) COPRO-Seq quantitation of the 12 bacterial species comprising the synthetic community used to colonize germ-free mice in experiments $E_1$ and $E_2$. Vertical dashed lines at days 14 and 28 denote time-points at which diets were switched. Panels A-C share a common key, provided in the upper right. Circles and triangles denote samples from experiments $E_1$ and $E_2$, respectively. Cecal sample data points (obtained at sacrifice on day 42 of the experiment) are plotted as for fecal sample data, but with inverted colors (i.e., colored outline, solid black fill). For all panels (A-C), data shown are mean ± SEM.

Figure S5. Further COPRO-Seq analysis of the relative abundance of components of the 12-member model human microbiota as a function of diet and time. (A) Plot of the ordination results for experiment 1 ($E_1$) from the principle coordinates analysis (PCoA) described in Fig. 1B. COPRO-Seq data from $E_1$ and $E_2$ were ordinated in the same multi-dimensional space. For clarity, only data from $E_1$ are shown (for the $E_2$ PCoA plot, see Fig. 1B). Color code: red/blue, feces; magenta/cyan, cecal contents. (B) Heatmap representation of the relative abundance data from $E_1$ normalized to each species’ maximum across all time-points within a given animal (“% of % max”). Each heatmap cell denotes the mean for one treatment group (n=7 animals), and each treatment group is shown as its own heatmap.

Figure S6. GeneChip profiling of the cecal metatranscriptome in mice fed different diets. (A) Venn diagram illustrating the number of bacterial genes whose expression was scored as ‘present’ (i.e., detectable in $\geq 5$ of 7 animals), only in mice that were consuming the plant polysaccharide-rich LF/HPP diet, only in mice that were consuming a ‘Western’-like HF/HS diet, or in both groups. (B) Overview of the diet-specificity of CAZyme gene expression in the 12-member model microbiota and in four prominent community members that maintained a proportional representation in the cecal microbiota that was $>5\%$ on each diet.
Figure S7. Dissecting the model human gut microbiota’s *in vivo* expression of EC 3.2.1.8 (endo-1,4-β-xylanase). (A) Gene expression in E₂ fecal samples was evaluated by microbial RNA-Seq. After data from all 12 species in the model microbiome were binned by EC number annotation and normalized (i.e., data were ‘community-normalized’ at the level of ECs), no significant difference in the representation of EC 3.2.1.8 in the metatranscriptome was observed when comparing day 13 (LF/HPP diet) and day 27 (HF/HS diet) data. (B) Transcribed *B. cellulosilyticus* WH2 genes account for >99% of community-normalized RNA-Seq counts assignable to EC 3.2.1.8. Thus, *B. cellulosilyticus* WH2 virtually dictates the degree to which expressed endo-1,4-β-xylanase genes are represented within the community metatranscriptome. (C) *B. cellulosilyticus* WH2 contributes a greater number of normalized RNA-Seq counts to the community metatranscriptome in HF/HS-fed mice, despite being relatively more abundant in LF/HPP-fed animals. (D) When *B. cellulosilyticus* WH2 gene expression data are normalized independently of data from other community members (i.e., when data are ‘species-normalized’), statistically significant increases in the representation of EC 3.2.1.8 within the *B. cellulosilyticus* WH2 transcriptome become apparent in HF/HS-fed mice. (E) Breakdown of the total species-normalized counts in (D) by the *B. cellulosilyticus* WH2 gene from which they were derived. Mean values ± SEM are shown. Means for all panels were calculated from data from four animals at each time-point, except day 26 (n=2).

Figure S8. Shotgun metaproteomic analysis of cecal samples from gnotobiotic mice colonized with the model 12-member human gut microbiota. (A) Each species’ theoretical proteome was subjected to *in silico* trypsinization (see Methods). Of the resulting peptides, those specific to a single protein within our database of all predicted proteins encoded by the genomes of the 12 bacterial community members, the mouse, and 3 bacterial ‘distractors’ (*Eubacterium rectale*, *Faecalibacterium prausnitzii*, and *R. torques*) were classified as ‘unique’, while all others were considered ‘non-unique’. The ‘unique’ fraction of a species’ predicted peptides indicates how many can be unambiguously traced back to a single protein of origin if detected by LC-MS/MS. (B) Comparison of the average relative cecal abundance of each bacterial community member (dark gray bars) with the percentage of proteins within its theoretical proteome that were detected by LC-MS/MS...
(red bars), and the percentage of all genes within its genome whose expression was detected using our custom GeneChip (light gray bars). Data shown are mean values ± SEM. (C) Scatter plots illustrating Pearson’s correlation between log-transformed averages of diet-specific expression fold-differences as determined by GeneChip assay (RNA, x-axis) and LC-MS/MS (protein, y-axis) in E1. Data points within the black scatter plot represent the 448 B. cellulosilyticus WH2 genes for which reliable quantitative data could be obtained for animals in both diet treatment groups for both the GeneChip and LC-MS/MS assays (i.e., any gene for which a signal could not be detected on at least one diet treatment in at least one assay were excluded). Scatter plots in color represent the results of correlation analyses performed on subsets of genes within the black plot whose KEGG annotations fall within particular functional categories, including “Translation” (r=0.03, 59 genes), “Energy metabolism” (r=0.36, 58 genes), “Amino acid metabolism” (r=0.48, 67 genes) and “Carbohydrate metabolism” (r=0.69, 110 genes). For both (B) and (C), n=2 mice per treatment group (4 mice total).
Supplementary figures

Figure S1.

A

Bacteroides xylanisolvens JCM 15633 (AB510713)
Bacteroides xylanisolvens SD CC 1b (ADKP01000087)
Bacteroides intestinalis JCM 13266 (EU136691)
Bacteroides intestinalis JCM 13266 (AB214329)
Bacteroides intestinalis DSM 17393 (ABJL02000006)
Bacteroides intestinalis DSM 17393 (ABJL02000008)
Bacteroides intestinalis AM-1 (AB437413)
Bacteroides intestinalis DSM 17393 (ABJL02000007)
Bacteroides cellulosilyticus WH2 (BWH2_r07)
Bacteroides cellulosilyticus WH2 (BWH2_r10)
Bacteroides cellulosilyticus WH2 (BWH2_r03)
Bacteroides cellulosilyticus DSM 14838 (ACCH01000108)
Bacteroides cellulosilyticus WH2 (BWH2_r07)
Bacteroides cellulosilyticus WH2 (BWH2_r06)
Parabacteroides merdae JCM 9497 (AB238928)

B

B. cellulosilyticus WH2 (BWH2_r07)
B. cellulosilyticus WH2 (BWH2_r10)
B. cellulosilyticus WH2 (BWH2_r06)
B. cellulosilyticus WH2 (BWH2_r03)
B. cellosolysis DSM 14838 (ACCH01000108)
B. intestinalis DSM 17393 (ABJL02000006)
B. xylanisolvens JCM 15633 (AB510713)

Similarity

100% (identical)
99% (approx. strain-level cut-off)
97% (approx. species-level cut-off)
95% (approx. genus-level cut-off)
Figure S2.

A

Bacteroides distasonis (ATCC 8503)
Bacteroides caccae (ATCC 43185)
Bacteroides vulgatus (ATCC 8482)
Bacteroides uniformis (ATCC 8492)
Bacteroides thetaiotaomicron (VPI-5482)
Bacteroides ovatus (ATCC 8483)
Bacteroides cellulosilyticus (WH2)

Bacteroides cellulosilyticus WH2
Other Bacteroidetes
Figure S3.

- Fecal pellet (DNA, COPRO-Seq)
- Fecal pellet (RNA, RNA-Seq)
- Cecal contents (DNA, COPRO-Seq; RNA, GeneChip; Protein, LC-MS/MS)
- Cecal contents (DNA, COPRO-Seq)

- LF/HPP diet
- HF/HS diet

Gavage with 12-member community
Figure S4.

A) Fecal DNA yield (E1)

B) Fecal DNA yield (E2)

C) Abundance (%)

Diet regimen

HF/HS => LF/HPP => HF/HS

LF/HPP => HF/HS => LF/HPP

Experiment

E1

E2
Figure S5.
Figure S6.

A

LF/HPP-specific (3,556 genes)

HF/HS-specific (3,221 genes)

Shared (11,373 genes)

B

CAZyme diet specificity
- LF/HPP only
- Both LF/HPP + HF/HS
- HF/HS only
- Neither diet

B. vulgatus
- SC12
- WH2

B. cellulosilyticus

B. thetaiotaomicron

B. caccae
Figure S8.

A

B

C

All KEGG categories

Translation

Energy metabolism

Amino acid metabolism

Carbohydrate metabolism

r = 0.53

r = 0.03

r = 0.36

r = 0.48

r = 0.69
References


Chapter 4

Future directions
**Introduction**

The experimental system and molecular tools developed over the course of this dissertation provide a means of addressing any number of questions about how the human gut microbiota assembles, operates, and responds to environmental changes over time. Our studies of the effects of introducing probiotic strains into an established microbiota (discussed in Chapter 2) provide a framework for identifying and characterizing future probiotic strains, probiotic consortia, prebiotics, and synbiotics (combinations of pre- and probiotics). We expect that data from studies using model communities to dissect the mechanisms by which probiotic strains influence host physiology will significantly impact the ways we treat individuals suffering from malnutrition, metabolic syndrome, inflammatory disorders, and other pathophysiological states that have been linked to the gut microbiota/microbiome [1-3]. Our studies of the influence of diet change over time on the microbiota (discussed in Chapter 3) provided many valuable insights about how communities are reconfigured in terms of both their structure and function in response to an altered nutrient landscape, as well as how prominent taxa (e.g., *B. cellulosilyticus* WH2) adjust their substrate utilization strategies. Follow-up future studies can adopt a similar strategy to identify the specific components of diets that drive desirable or undesirable reconfigurations in microbiota of different human populations. At least three additional areas of research and development building on the findings reported in Chapters 2 and 3 could and should be pursued, as described below.

**Creating more representative models of the human microbiota**

**Culturing the ‘unculturable’**

Although research performed in our laboratory has recently demonstrated that a significant number of the microbes residing in a human gut microbiota represent lineages that can be cultured using traditional methods [4], the fact remains that the human gut also harbors many species that are recalcitrant to growth in the laboratory [5]. One example is the bacterial phylum TM7. Representatives from this as-yet ‘unculturable’ group were first identified by sequencing the 16S rRNA
gene in environmental samples from a German peat bog [6]. Since then, they have been found in many other habitats, including the human intestine [7]. Unculturable gut bacteria also come from phyla (divisions) with many cultured representatives. For example, some families from two of the best-studied intestinal phyla, the Firmicutes (e.g., Lachnospiraceae) and the Bacteroidetes (e.g., Porphyromonadaceae, Prevotellaceae, Rikenellaceae) are comprised of many species we cannot yet study in isolation. The defined communities of human gut bacteria characterized in this dissertation have been comprised exclusively of sequenced, culturable organisms. If monocultures of a microbe cannot be prepared or if a microbe’s genome cannot be sequenced, it is not eligible for inclusion in model community studies of the type described here. Because we know or infer from culture-independent metagenomic studies that currently unculturable organisms can play important roles in the human gut [8], finding ways to expand our inventory of culturable gut microbes will be critical as we aim to develop even more functionally and structurally representative models of human gut communities.

Some ‘unculturable’ organisms may not be amenable to being grown in isolation because their metabolism requires inputs (e.g., nutrients, signaling cues) from other microbial partners, and/or researchers may not be sufficiently replicating essential aspects of their natural gut environment [9]. Model communities such as those employed here offer one avenue for characterizing the nature of such mutualistic relationships through in vivo coculture experiments. In theory, a researcher could use gnotobiotic mice colonized with defined and representative model communities as a type of culturing vessel or in vivo ‘test tube’ into which difficult or impossible-to-culture organisms could be introduced. Such organisms could come from samples where species diversity has been artificially depressed to enrich the target species population, for example through the administration of antibiotics. These low-complexity consortia could then be introduced into gnotobiotic mice harboring model communities with the aim of establishing the unculturable organism in a more defined setting. Studies of environments with low levels of species complexity, such as acid mine drainage sites, have shown that the genomes of the organisms comprising simple communities can be reconstructed from metagenomic data [10]. We anticipate that if an unculturable organism (and
a small number of other accompanying species) could be established in a gnotobiotic mouse colonized beforehand with species whose genomes are sequenced, it should be possible to reconstruct the genomes of the newly introduced organisms. Further, establishment of the target species in a model ecosystem would allow us to study its metabolic lifestyle with the same metatranscriptomic (RNA-Seq) and high-resolution metaproteomic methods used in this dissertation to characterize microbe-microbe and diet-microbe interactions, offering important insights not provided by the inspection of its genome sequence alone [11].

**Personal culture collections**

The studies carried out over the course of this dissertation have made use of carefully constructed models meant to be representative of a healthy human gut microbiota in terms of its phylogenetic structure, gene content, and functional capacity. As we strive to assemble even more representative models of the microbiota, we must account for that the fact that the way in which assemblages of organisms operate may be dictated by their composition, their co-evolution within an individual, family or population representing a given lifestyle (cultural tradition), as well as the physiological status of an individual host. Thus, a ‘one size fits all’ model community may not always be appropriate or most effective for addressing particular biological questions. For this reason, we anticipate a move in the future towards greater study of ‘personal culture collections’ [4] comprised of clonally arrayed, culturable microbes isolated from single individuals. These personal culture collections would each be constructed using a sample (e.g., feces) from an individual possessing particular age, physiological state, cultural tradition, or disease state. Armed with our current and future (see above) understanding of the culturing requirements of gut microbes, we would isolate the majority of taxa from a human subject’s microbiota and archive them as a subject-specific model microbiota. Transplantation of these culture collections into germ-free mice would then be performed to determine if consortia faithfully recapitulate (transmit) microbiota-mediated human donor phenotypes in gnotobiotic animals. Should our attempts at transferring human phenotypes to germ-free hosts using culture collections prove effective, further characterization of the microbes driving particular phenotypes could be conducted. Such follow-up studies could include the se-
sequencing of isolate genomes to gain insight about their metabolic potential and nutrient requirements, introduction of subsets of the community, mutagenesis of individual isolates prior to their inclusion in a model community, and/or replacement of a given member with by closely-related species to determine how general or specific important traits might be within a specific lineage.

**Developing new and improved tools and techniques for characterizing model communities**

Construction of more representative models of the human gut microbiota should also be accompanied by development of new and improved methods for characterizing these communities; this would allow us to interrogate their structure and operations in more detailed (comprehensive) as well as more holistic ways. Some of the methods that were developed or improved over the course of this dissertation can be scaled to characterize highly diverse model communities fairly easily. For example, COPRO-Seq can accommodate communities of nearly any complexity, provided they are comprised exclusively of sequenced species whose genomes can be differentiated from one another. Microbial RNA-Seq is also scalable and can be used to characterize complex communities if financial resources are available to sequence the metatranscriptome deeply. Although the structure and function of model communities can be probed readily with current techniques, there are also many biological parameters of the gut ecosystem that we are not yet capable of measuring with sufficient accuracy or throughput. This includes, for example, the spatial relationship between microbes and the gut’s cellular and luminal contents (the latter includes its mucus layer as well as partially digested food particles). Future efforts to develop methods for measuring some of these parameters are described below.

**Absolute quantitation of microbial biomass and taxon abundance**

In earlier chapters, we demonstrated that COPRO-Seq provides precise and accurate measurements of the proportional representation of each species within a model community. When interpreting changes in a species’ proportional abundance over time, it is important to consider that such changes can be driven by at least two factors: (i) an increase/decrease in the absolute abun-
dance of that species within the community, and (ii) an increase/decrease in the absolute abundance of other species within the community. Currently, COPRO-Seq is not capable of providing the researcher with precise measurements of absolute abundance unless additional data (e.g., measurements of total microbial cell concentrations, total community biomass) are provided. In cases where the proportional abundance of an organism is impacted to a significant degree by a perturbation, the direction of the perturbation’s effect on a particular species can be inferred with strong confidence. However, subtle changes can be much more difficult to interpret. We have therefore begun investigating approaches that can complement COPRO-Seq by providing measures of total cell concentrations or total community biomass over time, so that the absolute abundance of each species in a community can be calculated. In the simplest method, total community DNA is used as a proxy for community biomass. Experimental data suggest, however, that measurements of fecal DNA concentrations are noisy from sample-to-sample, and we cannot rule out the possibility that the composition of the diet fed to a gnotobiotic host might affect DNA extraction yields. A second method involves using fluorescence activated cell sorting (FACS) to ‘count’ the number of bacterial cells in a gut sample (e.g. feces). This approach uses DNA-specific dyes and the light scattering properties of bacterial cells to differentiate microbes in a sample from all other particles. By counting labeled cells and comparing their concentration to that of a spiked-in bead standard of known size, a reliable estimate of the total bacterial cell count in a sample can be obtained. It is also possible that the addition of other reagents might allow for the discrimination of live and dead cells. Although the latter approach for quantifying the absolute abundance of the species within a model community is still under investigation, we anticipate that it will ultimately serve as an important element of the molecular tool kit used by those studying model gut communities.

**An expanded role for INSeq in model community studies**

In Chapter 3, INsertion Sequencing (INSeq; [12,13]) was used to confirm that certain highly-expressed loci within the genome of *Bacteroides cellulosilyticus* WH2 act as important diet-specific fitness determinants *in vivo*. This approach, which was developed as part of the study described
in Appendix A, provides the researcher with a powerful tool for identifying fitness determinants in any microbial species that can be transformed with a sequencing-adapted mariner transposon delivery vector. Recently, members of our laboratory developed a version of this delivery vector that incorporates specific barcodes that, when sequenced along with captured chromosomal DNA located adjacent to an inserted transposon, can be used to distinguish different mutagenized populations of cells. This improvement on the original pSAM vector allows us to mutagenize and pool multiple species of gut bacteria when constructing model communities so that rather than identifying the genetic determinants of fitness in one species at a time, many different species can be evaluated simultaneously. Future studies leveraging this technical innovation should be performed to identify genes that are broadly important for microbial success/persistence within the gut. Such studies, identifying ‘general’ fitness determinants shared among members of a particular microbial lineage, could serve as a foundation for experiments that seek to identify other shared suites of genes needed for fitness in particular contexts (e.g., specific host physiological/pathological states or diets). Given their dominance in the human gut, role in human health [14] and the breakdown of macronutrients [15], and their genetic tractability, we see members of the Bacteroidetes as prime candidates for such studies going forward.

**New imaging techniques for evaluating gut biogeography**

Although metagenomic surveys have taught us much about the overall composition of the distal human gut microbiota, we still know surprisingly little about the microhabitats occupied by individual lineages and how the spatial distribution of a particular species is impacted by variables such as diet and invasion of the ecosystem by new taxa. Fluorescence in situ hybridization (FISH) has been used for years as a method for quantifying [16] and visualizing [17,18] microbes in a variety of ecosystems; the method is based on short, fluorescently-labeled oligonucleotide probes that hybridize to small subunit ribosomal RNAs (e.g.,16S rRNA). One of the drawbacks of traditional FISH-based methods is the inability of fluorescence microscopes to differentiate more than a handful of fluorophores simultaneously. This limitation prevents those studying microbial communities
from using more than two or three labeled oligonucleotide probes at once. Therefore, these studies typically involve low-resolution visualization of bacteria using ‘universal’ probes which label most or all bacteria, or probes designed to label taxonomically-related species (e.g., those of the same genus or class). In collaboration with researchers at the Marine Biological Laboratory, we have begun to apply a new variation of FISH to the study of defined model human gut communities in gnotobiotic mice. This method, known as combinatorial labeling and spectral imaging fluorescence in situ hybridization (CLASI-FISH; [19,20]) allows the researcher to visualize at least 15 and as many as hundreds of different species simultaneously. Our initial experiments have shown that the method is tractable in samples collected from mice, and that spatial relationships between all taxa in a model community can be quantified simultaneously. As these methods are developed further and applied to defined communities of increasing complexity, we expect to delineate new, previously unappreciated syntropic relationships between taxonomic groups and develop a greater understanding of how microbes compete and cooperate to achieve stable community configurations within the gut.

**Further study of specific taxa and ecological guilds**

Focused studies characterizing widely distributed gut taxa are important for understanding how particular lineages influence the operations of larger gut communities, and help us understand the role of keystone species within these systems. For instance, decades of research on *Bacteroides thetaiotaomicron*, whose genome was first sequenced in 2003 [21], have helped us better understand how gut microbes utilize polysaccharides [22,23], interact with the mucosal immune system [24], and adapt metabolically during postnatal host development [25]. In another example, extensive microbiological studies of *Faecalibacterium prausnitzii* and the *Roseburia* have revealed the important role these species play in the production of butyrate, an important source of energy for colonocytes [26]. Studies focused on different ecological guilds of intestinal microbes (i.e., those that exploit the same resources, often in related ways) have also taught us about their roles in gut food webs. Examples of such guilds are the methanogenic archaea [27] and acetogenic bacteria...
[28]. As our appreciation grows for the diversity of functional niches occupied by intestinal bacteria, so too will our need to better characterize the roles and capabilities of individual members of the gut microbiota.

Based on the experimental results described in Chapter 3, we would advocate strongly for further characterization of *B. cellulosilyticus* WH2, both *in vitro* and as a component of defined model human gut communities. Our draft sequence of this microbe’s genome reveals it to be the largest of any Bacteroides species to date. More importantly, it is also the most replete in carbohydrate active enzymes (CAZymes) and polysaccharide utilization loci (PULs) of any member of the Bacteroidetes sequenced so far. Though the link between its substantial carbohydrate utilization armamentarium and its strong fitness on disparate diets is still somewhat speculative, our data do suggest that this species utilizes multiple carbohydrates *in vivo*, perhaps indicative of a diversified nutrient utilization strategy. Such a strategy is likely to provide a fitness advantage to a bacterium colonizing the gut of a host whose diet is far from monotonous.

While the studies outlined in this dissertation have greatly expanded our understanding of this bacterium, further efforts to ‘crack the code’ of its carbohydrate utilization strategy are needed. Our efforts thus far have successfully identified one or more substrates that trigger the expression of at least 30 *susC/D* gene pairs (the core elements of PULs) in its genome and have revealed that its metabolic strategy focuses largely on the catabolism of diet-derived xylans. However, the substrates capable of triggering the expression of over two-thirds of this organism’s 113 PULs remain unidentified. Furthermore, we know little about the regulatory mechanisms and exact molecular moieties that control the expression of individual PULs. Delineating these regulatory networks should help us understand how gut microbes ‘decide’ what carbon sources to compete for and utilize, and which receive low priority from a metabolic perspective. INSeq-based analyses have shown that the ability to utilize xylan is critically important to *B. cellulosilyticus* WH2 *in vivo*. However, the degree to which disrupting key components of PULs required for the utilization of other substrates and the impact of disrupting the normal regulation of these loci are areas deserving of further attention.
Finally, though further characterization of carbohydrate utilization mechanisms in this prominent *Bacteroides* will be instructive, even greater insight might be gained by characterizing multiple other *Bacteroides* spp. using the same strategies employed here. Identifying the genes required for the utilization of particular glycans in a number of lineages should allow us to begin describing different PUL ‘families’ (i.e., clusters of genes with non-identical configurations/compositions that confer similar substrate utilization capabilities). Theoretically, the presence of genetic elements known to be associated with a particular PUL family could even be used as a metagenomic biomarker to indicate whether or not a given individual’s microbiota harbors the capacity to utilize particular dietary substrates. Such information could be incredibly useful as researchers begin to develop microbiome-directed therapeutics tailored to ‘supplement’ deficient activities using probiotic strategies, or ‘boost’ the representation of important lineages using prebiotic compounds. Finally, as more knowledge is acquired about the carbohydrate utilization capabilities of different species, we anticipate the development of quantitative models that allow us to predict which bacteria will prosper (or perish) under particular dietary regimens and to prescribe diets that improve the nutritional status of humans representing different cultural traditions, at various stages of their lifecycle.

**Concluding statements**

As our appreciation grows for the many ways in which our gut microbes impact human health, we must be judicious and thoughtful in the ways we address the many outstanding questions that remain. Certainly, gut microbiota studies in human subjects will continue to yield new insights about the nature of our relationship with our intestinal symbionts. However, a greater focus on the careful characterization of representative models of this complex system under controlled conditions is also warranted, and necessary, if we are to develop generalizable principles and ‘rules’ for how this important ‘microbial organ’ operates.
References


Appendices

Appendix A

Andrew L. Goodman, Nathan P. McNulty, Yue Zhao, Douglas Leip, Robi D. Mitra, Catherine A. Lozupone, Rob Knight, Jeffrey I. Gordon.

Identifying genetic determinants needed to establish a human gut symbiont in its habitat.

*Cell Host Microbe.* 2009 Sep 17; 6(3):279-89.


Appendix B


Predicting a human gut microbiota’s response to diet in gnotobiotic mice.


http://www.sciencemag.org/content/333/6038/101.long
Identifying Genetic Determinants Needed to Establish a Human Gut Symbiont in Its Habitat

Andrew L. Goodman,1 Nathan P. McNulty,1 Yue Zhao,1 Douglas Leip,1 Robi D. Mitra,1 Catherine A. Lozupone,1,2 Rob Knight,2 and Jeffrey I. Gordon1,*

1Center for Genome Sciences, Washington University School of Medicine, St. Louis, MO 63108, USA
2Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309, USA
*Correspondence: jgordon@wustl.edu
DOI 10.1016/j.chom.2009.08.003

SUMMARY

The human gut microbiota is a metabolic organ whose cellular composition is determined by a dynamic process of selection and competition. To identify microbial genes required for establishment of human symbionts in the gut, we developed an approach (insertion sequencing, or INSeq) based on a mutagenic transposon that allows capture of adjacent chromosomal DNA to define its genomic location. We used massively parallel sequencing to monitor the relative abundance of tens of thousands of transposon mutants of a saccharolytic human gut bacterium, Bacteroides thetaiotaomicron, as they established themselves in wild-type and immunodeficient gnotobiotic mice, in the presence or absence of other human gut commensals. In vivo selection transforms this population, revealing functions necessary for survival in the gut: we show how this selection is influenced by community composition and competition for nutrients (vitamin B12). INSeq provides a broadly applicable platform to explore microbial adaptation to the gut and other ecosystems.

INTRODUCTION

Our indigenous microbial communities play critical roles in shaping myriad features of our biology. The distal gut hosts the majority of our microbes; these include representatives of all three domains of life, plus their viruses. The density of organisms occupying this habitat is astonishing, exceeding $10^{12}$ cells/mL. Most phylogenetic types (phytolotypes) observed in the guts of humans and other mammals belong to just two bacterial divisions (phyla)—the Firmicutes and the Bacteroidetes (Ley et al., 2008). Microbial community (microbiota) exchange experiments indicate that gut community members are dynamically selected: for example, transplantation of a Proteobacteria-dominated zebrafish gut microbiota into germ-free mice transforms this community so that it comes to resemble a mouse gut microbiota, while transplantation of a mouse microbiota into germ-free zebrafish has the opposite effect, yielding a community that has the phylum-level characteristics of the native zebrafish microbiota (Rawls et al., 2006).

Within the Firmicutes and Bacteroidetes, hundreds to thousands of phylotypes partition available niches (professions) to create a community able to maintain itself in a continuously perfused ecosystem despite shifts in host diet, regular ingestion of foreign bacteria, intense resource competition, high bacteriophage levels, and immune surveillance. Comparisons of the sequenced genomes of cultured representatives of major gut phylogenetic lineages provide a means for identifying genomic features potentially important for colonization and competition in the gut. However, the recent surge in microbial genome sequencing projects has far outpaced development of broadly applicable tools for directly testing the role of genes in determining fitness in this habitat. The paucity of tools is unfortunate, as fundamental questions connecting genome content to function remain unexplored. For example, how are the determinants of fitness related to nutrient availability, and how closely do the genes required for fitness in vivo mirror those required for maximizing growth rate in vitro? Does community structure influence this map of genetic requirements, or is competition largely “within species”? Do the major recognized components of the host immune system play a dominant role in determining which genes are critical for symbiont fitness in vivo? To address these questions, we integrated a simple and broadly applicable genetic tool with second-generation DNA sequencers and gnotobiotic mouse models to identify fitness determinants in the genome of a human gut mutualist.

Mariner transposon mutagenesis is an attractive forward genetic strategy for connecting phenotype to gene because stable random insertions can be generated in a recipient genome without specific host factors: the ability of these transposons to serve as mutagenic agents is well established in members of all three domains of life (Lampe et al., 1996; Mazurkiewicz et al., 2006). After alignment of the inverted repeat (IR) sequences that delimit mariner family transposons, we noted that a single G-T transversion at a nonconserved position would create a recognition sequence for the type IIs restriction enzyme Mmel. When directed to this location, Mmel would cleave 16 bp outside of the transposon, capturing a genomic fragment that identifies the insertion site. Moreover, if the genome sequence of the recipient organism were known, the short genomic DNA sequences captured by this Mmel digestion would be sufficient to uniquely map transposon location. We reasoned that in a mixed population of transposon mutants produced in a given recipient bacterial species, the relative abundance of each Mmel-liberated genomic fragment, identified after limited PCR amplification and massively parallel sequencing, would in
We constructed pSAM, a sequencing-adapted mariner transposon delivery vector with three major features: an antibiotic resistance cassette flanked by Mmel-modified mariner IRs, a multiple cloning site immediately upstream of the himar1C9 mariner transposase (Lampe et al., 1999), and machinery for replication in the donor strain and transfer by conjugation (see Figures S1A–S1C available online). B. thetaiotaomicron was chosen as the recipient species to test this approach for several reasons (Figure S2A). First, it is highly adapted to life in the distal human gut (Zocco et al., 2007). A prominent member of the gut microbiota, this mutualist is richly endowed with a broad arsenal of genes encoding glycoside hydrolases and polysaccharide lyases not represented in the human genome (Xu et al., 2003). These genes are incorporated together with genes encoding nutrient sensors and carbohydrate transporters into 88 polysaccharide utilization loci (PULs) representing 18% of the organism’s genome (Martens et al., 2008). Thus equipped, B. thetaiotaomicron functions as a flexible forager of otherwise indigestible dietary glycans, as well as host glycans when dietary polysaccharides are not available (Martens et al., 2008).

Using this insertion-sequencing (INSeq) approach, we subjected a mutagenized population of the prominent human gut symbiont, Bacteroides thetaiotaomicron, to varying selective pressures: in vitro growth in continuous flow chemostats, monoassociation of wild-type and knockout germ-free mice lacking major branches of their innate or acquired immune systems, and as one component of several defined in vivo communities of human gut-derived microbes. The results provide evidence that human gut symbionts, like their pathogenic counterparts, possess dedicated mechanisms critical for interaction with their host and each other. The relative importance of these mechanisms is not static but instead is shaped by other members of the microbiota. As a strategy for functional characterization of newly sequenced genomes in general, and of the human gut microbiome in particular, INSeq extends existing techniques in several important aspects.

RESULTS

We found that transfer of the Mmel-modified mariner transposon into the genome of B. thetaiotaomicron occurs with high efficiency (Figure S2). To identify the site of transposon insertion and the relative abundance of each mutant in an otherwise isogenic population, we developed a straightforward procedure to extract the two 16 bp genomic sequences adjacent to each

figure caption: Figure 1. Mapping and Quantifying Tens of Thousands of Transposon Insertion Strains by High-Throughput INSeq

(A) A negative selection scheme for identification of genes required for colonization in vivo. Mutants in genes important for competitive growth (red) are expected to decrease in relative abundance in the output population.

(B) Preparation of an INSeq library. Genomic DNA is extracted from the mutagenized bacterial population, digested with Mmel, and separated by polyacrylamide gel electrophoresis (PAGE). Transposon-sized fragments are appended with double-stranded oligonucleotide adapters by ligation. Limited cycles of PCR create the final library molecules for sequencing.

(C) Map of insertion sites in the B. thetaiotaomicron genome. An arrow marks the origin of replication.

(D) Reproducibility of library preparation and sequencing protocols. Technical replicates were prepared and sequenced from a single transposon mutant population. Each point represents the abundance of insertions in a single gene; the coefficient of determination, $R^2$, on log-transformed abundance values is 0.92.
transposon, append sequencing adapters to these fragments, and separate the desired molecules from genomic background (Figure 1B and Supplemental Experimental Procedures). Sequencing these tags using an Illumina Genome Analyzer II produced ~8 million raw reads from a single flow cell lane: ~90% of these reads contained the transposon (Figure S2D). We designed a software package, MapSAM, to filter out low-quality sequences, quantify and pair reads generated from either side of an insertion, and assign these paired reads to a specific location in the target genome (Figure 1C). Examination of technical replicates indicated that the library preparation, sequencing, and mapping strategies were highly reproducible (Figure 1D and Figure S3). The proportion of reads that were unambiguously mapped (98%) matched predictions from an in silico model of random transposon insertion (Figure S4A); no insertion sequence bias beyond the known “TA” dinucleotide requirement (Bryan et al., 1990) was apparent (Figures S4B and S4C).

We first characterized a mutant population containing ~35,000 B. thetaiotaomicron transposon insertion strains. Insertions were well distributed across the genome at an average density of 5.5 insertions/kb (Figure 1C). After filtering out insertions in the distal (3’) 10% of any coding region (because such insertions could possibly still permit gene function), we found that 3435 of the 4779 predicted open reading frames in the genome (72%) had been directly disrupted in the mutant population. Inclusion of genes disrupted by upstream (polar) mutations in a predicted operon increased this number to 78%; rarefaction analysis suggested that this population is approaching saturation (Figure S4D).

To identify B. thetaiotaomicron genes unable to tolerate transposon insertion, we generated and mapped a second, independent mutant population. We combined both data sets and applied a Bayesian model to account for the number of informative insertion sites in each gene (Lamichhane et al., 2003). The results yielded a conservative list of 325 candidate essential genes for growth under anaerobic conditions when plated onto rich (tryptone-yeast extract-glucose; TYG) medium (Table S1). These genes were significantly enriched (Benjamini-Hochberg corrected p < 0.05) for Clusters of Orthologous Groups (COG) categories representing cell division (category D), lipid transport/metabolism (l), translation/ribosomal structure/biogenesis (J), and cell-wall/membrane biogenesis (M). This is consistent with genome-wide mutagenesis studies of Escherichia coli and Pseudomonas aeruginosa (Baba et al., 2006; Jacobs et al., 2003). For nonessential genes, insertion frequency showed some correlation ($R^2$ of log-transformed values = 0.33) with GeneChip-defined expression levels during mid-log phase growth of the parental wild-type strain in batch fermentors containing TYG medium (Figure S4E). The reason for this relationship is not known, although studies of mariner transposition in vitro suggest that the enzyme has a preference for bent or bendable DNA (Lampe et al., 1998).

Combinatorial Mapping of Individual Insertion Strains from an Archived Mutant Collection

A mutant population of this complexity contains insertions in most of the coding potential of the genome and can facilitate forward genetic approaches for identifying genotypes connected with a phenotype of interest. These mixed populations, however, are less amenable to reverse genetics: specific genotypes are not individually retrievable. Arrayed transposon mutant collections, in which strains of known genotype are archived individually, provide an important avenue for retrieval and further study of specific strains of interest. To date, such collections have been created by using a strain-by-strain procedure that typically consists of cell lysis, removal of cellular debris, multiple rounds of semirandom or single-primer PCR, DNA cleanup, and individual Sanger sequencing of each amplicon. As an alternative, we developed a combinatorial technique for simultaneously mapping thousands of individually archived transposon mutant strains in parallel (Figure 2A, Figure S5, and Supplemental Experimental Procedures).

This approach consists of three basic steps: (1) culturing and archived storage of randomly picked mutant colonies in individual wells of 96-well (or higher density) plates, (2) placement of each of these strains into pools in unique patterns, and (3) sequencing of these pools by INSeq in order to associate each transposon insertion location with a strain in the original arrayed multiwell plates. Because $n$ pools can contain $2^n$ unique presence/absence patterns, a small number of pools can uniquely identify a large number of strains. To this end, a benchtop liquid-handling robot was used to distribute archived transposon mutant strains across a subset of pools in a pattern selected to minimize the likelihood of mistaking one strain for another or incorrectly mapping clonal strains (Figure 2B and Supplemental Experimental Procedures). Libraries were then prepared from each pool using the same method described in Figure 1B, except that a pool-specific, barcoded adaptor (Table S2) was used in the ligation step. These libraries were combined into a single sample that can be sequenced with an Illumina Genome Analyzer II using just one lane of the instrument’s eight-lane flow cell. Reads were first mapped to the reference B. thetaiotaomicron genome to determine insertion sites; to assign an insertion site to a specific archived strain, the pool-specific barcodes associated with a given insertion location on the chromosome were then matched with the patterns assigned to the strains in the original archived set of plates.

Using this strategy, we were able to identify the insertion coordinates for over 7000 individually archived B. thetaiotaomicron transposon mutant strains in parallel (Table S3). To verify the accuracy of these assignments, we first used ELISA to test strains predicted to have lost reactivity to two monoclonal antibodies of known specificity (Peterson et al., 2007). We also amplified transposon-genome junctions of test strains by semirandom PCR and sequenced the amplicons. In total, 179 of 183 strains tested (98%) produced the anticipated results (Figure 2C), confirming that combinatorial barcoding and INSeq can be used to efficiently and economically generate archived, sequence-defined, mutant collections.

Identification of Genes Required for Fitness In Vitro

To identify genes that contribute to exponential growth in nutrient-rich conditions in vitro, we maintained a 35,000-strain mutant population in this growth phase (OD$_{600}$ 0.1–0.4), under anaerobic conditions, in chemostats that were continuously supplied with fresh TYG medium. Output populations were
sampled from four independent chemostats; two after ∼15 hr of continuous exponential growth and two after ∼45 hr of growth. DNA was prepared from each of these populations, and transposon-adjacent genomic fragments were identified by INSeq. After removing insertions in the 3’ 10% of each gene, a z test was applied to identify genes that show a significantly altered representation from the overall distribution of output/input abundance ratios after q value correction for multiple hypothesis testing (q < 0.05; see the Supplemental Experimental Procedures). To test this approach for identifying fitness determinants, we also examined insertions in intergenic “neutral loci” (Figure S6A). None of these 80 control regions passed the statistical cutoff for underrepresentation (three increased in abundance). In contrast, 477 genes (~14% of the genes represented in the input population) showed a statistically significant change in abundance after in vitro selection (265 underrepresented/212 overrepresented; Figure S6B and Table S4). Consistent with selection for maximal growth rate in rich medium, the list of factors important for fitness under these conditions was significantly enriched in genes annotated as being in COG category C (energy production and conversion).

As a proof of principle, we asked whether an observed enrichment in a broad functional group (COG category), represented among genes required for fitness, could be altered by manipulating environmental conditions. To do so, we harvested exponentially growing cells from mutant populations grown in minimal defined medium in the presence or absence of exogenous amino acids and quantified the abundance of transposon mutants by INSeq (n = 4 replicate populations, each assayed independently). Gratifyingly, the set of genes required for fitness specifically in the amino acid-depleted condition was most highly enriched (p < 0.0005) in COG category E (amino acid transport/metabolism) (Figure S7 and Table S5).

Genes Required for Establishment of B. thetaiotaomicron within the Distal Gut of Monoassociated Gnotobiotic Mice

To survey the B. thetaiotaomicron genome for genes critical for fitness in a mammalian gut ecosystem, we colonized germ-free mice with a single gavage of approximately 10^8 colony-forming units (CFUs) of the 35,000-strain mutant population (n = 15 animals representing three independent experiments, each involving a cohort of five 8- to 12-week old C57BL/6J males; experiments were performed ~3 months apart). The five animals in each cohort were caged individually in a shared gnotobiotic isolator and fed a standard, autoclaved, polysaccharide-rich, low-fat chow diet ad libitum. The relative abundance of each mutant strain in the cecal bacterial population was defined at the time of sacrifice 14 days after gavage; this interval between gavage and sacrifice encompassed several cycles of turnover of the mucus layer and the underlying gut epithelium, and is sufficient to allow mobilization of innate and adaptive immune responses (Peterson et al., 2007).

All recipients of the gavage harbored equivalent levels of B. thetaiotaomicron at the time of sacrifice (10^{11}–10^{12} CFU/mL cecal contents as quantified by plating and by qPCR). Moreover, the relative representation of mutants was consistent...
between the ceca of individual mice (Tables S9–S11) and for most genes reflected their abundance in the input population (Figure 3A). However, compared to the input population, mutants in 370 genes showed significantly (q < 0.05) altered representation (90 overrepresented/280 underrepresented) in all three cohorts of mice (Table S6; note that the largest category of genes identified in this screen encode hypothetical or conserved hypothetical proteins). Only one of the “neutral intergenic” controls described above was significantly underrepresented in these populations (two were overrepresented).

While the smaller group of 90 genes that produce a competitive advantage in the cecum when mutated (highlighted in Table S6 in blue and green) are not significantly enriched in any broad predicted functional (COG) categories, only half can be explained by corresponding behavior in chemostats containing rich medium. The underrepresented genes show a similar trend: half (146) of the 280 genes critical for in vivo fitness can be predicted from growth defects in rich medium; the remainder (134/280), which are highlighted in Table S6 in yellow, do not show a defect after prolonged exponential growth in vitro. These include loci with diverse predicted functions, including assembly of polysaccharide- and protein-based surface structures (BT1339-55, BT1953-7), synthesis and utilization of vitamin B12-dependent cofactors (BT2090-1, BT2760), and an nrf-like oxidoreductase complex (BT0616-22) (e.g., Figure S8).

To confirm that the requirement for these genes in vivo could not be explained by general growth defects, we analyzed individual mutant strains retrieved from the archived strain collection. This collection contained sequence-defined, single-insertion transposon mutants in ~70% of the genes designated as critical for survival in the distal gut in vivo (~80% if predicted polar effects are included) (Table S3). After validating the site of selected transposon insertions by semirandom PCR and Sanger sequencing, we determined the exponential doubling time of representative strains individually (Figure 3B). Strains carrying transposon insertions in genes uniquely required in vivo had a slower doubling time when cultured in vitro growth rate similar to wild-type B. thetaiotaomicron, further suggesting that the critical function of these genes in vivo cannot be simply explained by a necessity for sustaining exponential growth in rich medium. In contrast, mutants that exhibited a competitive defect in the 35,000-strain population both in vitro and in vivo had a slower doubling time when cultured individually, suggesting that these genes play a basic role in bacterial cell physiology.

An earlier report from our group used a targeted mutagenesis strategy to disable expression of genes encoding capsular polysaccharide (CPS) 4 in this organism; this strain was rapidly displaced by wild-type B. thetaiotaomicron after initial inoculation as a 1:1 mixture into germ-free mice (Peterson et al., 2007). The 35,000-strain transposon mutant population recapitulated this observation (Figure 3C and Figure S9). The transposon mutant population included over 1100 independent insertions across 143 genes that span all eight CPS loci encoded in the B. thetaiotaomicron genome. None of the other CPS loci
were required for fitness in vivo, indicating that CPS4 plays a unique role for *B. thetaiotaomicron* in the gut environment of monoassociated gnotobiotic mice fed a standard plant polysaccharide-rich chow diet. Moreover, our observation that transposon inactivation of genes in any single PUL did not confer a competitive disadvantage in vivo is consistent with *B. thetaiotaomicron*’s capacity for adaptive foraging of a broad range of glycans present in this diet (a total of 5137 distinct transposon inactivation of genes in any single PUL did not confer a competitive disadvantage in vivo is consistent with *B. thetaiotaomicron*’s capacity for adaptive foraging of a broad range of glycans present in this diet (a total of 5137 distinct transposon

**The Impact of Host Genotype and Community Structure on Selection In Vivo**

We next asked whether a broad view of the mutant population as a whole could help address some basic questions in mammalian gut microbial ecology described in the Introduction: are the genetic determinants of fitness influenced by the bacterial community (microbiota) context; do inter-specific competition and intra-specific competition play distinct roles in shaping the selective pressures on a genome; is this selection primarily maintained by elements of the host immune system?

To explore these questions, we manipulated two features of the host habitat: (1) the immune system, by introducing the *B. thetaiotaomicron* mutant population into germ-free mice with genetically engineered defects in innate or adaptive immunity (*Myd88*+/− and *Rag1*+/−, respectively); or (2) microbial composition, by including this mutant population as one component of three different types of defined communities, one consisting of six other sequenced human gut-associated Bacteroidetes, another composed of eight sequenced human gut-associated Firmicutes and Actinobacteria, and a third consisting of all 14

**Figure 4. Identification of Environmental**

**and Microbial Factors that Determine the Fitness Landscape for**

**B. thetaiotaomicron**

(A) The relative abundance of each mutant in the *B. thetaiotaomicron* population was evaluated in multiple host genotypes (wild-type C57Bl/6; *Rag1*−/−; *Myd88*−/−) and microbial contexts.

(B) qPCR assays of cecal microbial community composition in gnotobiotic mice at the time of sacrifice. The ~36,000-strain *B. thetaiotaomicron* population is indicated with arrows.

of these representatives of the human distal gut microbiota (n = 4–5 animals per treatment group per experiment; **Figure 4A** and **Table S7**). Animals were sacrificed 14 days after gavage and their cecal contents harvested. qPCR assays of cecal DNA, using species-specific primers, revealed that gavage with a given multispecies input community yielded consistent cecal “output” community compositional profiles (**Figure 4B** and **Table S8**.

Each of these treatments shifted the *B. thetaiotaomicron* population from its input distribution (**Tables S9**–**S12**). To search for functional trends in these shifts, we assigned the genes underrepresented in output populations to functional (COG) categories. The in vivo fitness determinants were significantly enriched in different predicted functions compared to essential genes, or to the genes required for maximal exponential growth in rich medium in vitro (**Figure 5**). For example, the predicted essential genes are most prominently enriched in COG categories J (translation, ribosome structure/biogenesis) and D (cell-cycle control and cell division), neither of which is enriched among the in vivo fitness determinants. Instead, the genes required in vivo are biased toward energy production/conversion (category C) and amino acid and nucleotide transport/metabolism (COG categories E and F, respectively). This represents an expansion beyond the single category (C) enriched after selection for maximal growth rate in vitro and is consistent across all in vivo treatment groups.

Closer examination of enriched COGs and carbohydrate-active enzyme (CAZy) families highlights the role of polysaccharide synthase for competitive fitness in this environment; overrepresented functions include UDP-glucose-4-epimerases and glycosyltransferases (specifically, GT2 and GT4 families [Cantarel et al., 2009] (**Table S13**). Together, these observations suggest that at the level of statistical enrichment of broad functional groups, the in vivo fitness requirements were distinct from those derived in vitro but that these enrichments were consistent across in vivo treatment conditions. Many of the annotated fitness determinants (e.g., CPS4 and the mf-like oxidoreductase) followed this pattern: dispensable in vitro but critical across all in vivo conditions tested.

This functional category enrichment analysis depends critically on genome annotation (~50% of genes were not assignable to COG categories), while not accounting for...
To further evaluate the robustness of the clustering algorithm, we conducted a principal coordinates analysis on these 48 transposon mutant populations (Figure 6B and Figure S10). Similar to the hierarchical clustering dendrograms, the first principal coordinate separates in vitro from in vivo mutant populations. The second coordinate separates these populations by microbial context: Bacteroidetes-containing communities shape a *B. thetaiotaomicron* mutant population that is distinct from that produced in monooassociations, or in the Firmicutes+Actinobacteria consortia, further indicating that although the same broad functional categories are enriched among *B. thetaiotaomicron* fitness determinants under a range of in vivo conditions, these mutant populations are additionally shaped by changes in microbial community composition.

We applied a random forest classifier (Breiman, 2001) to identify genes that were responsible for the observed separation of mutant populations in monooassociated mice from those mutant populations present in Bacteroidetes cocolonized mice. This machine-learning algorithm serves to estimate the importance of predictor variables (i.e., genes) for differentiating between classes (i.e., the monooassociation versus Bacteroidetes cocolonized groups, which were distinguished by unsupervised clustering and principal coordinates analysis). This approach identified a total of 220 genes as important for differentiating these groups (Table S14). Mutants in 144 of these predictor genes had lower output/input ratios in the monooassociations: in other words, these genes were more important when other Bacteroidetes were not present. Mutants in 76 genes had lower output/input ratios in mice cocolonized with other Bacteroidetes. These 76 genes, which provide a signature of functions under increased selection for *B. thetaiotaomicron* in the presence of other Bacteroidetes, are enriched for components of amino acid biosynthetic pathways, suggesting that although these functions are required in all in vivo conditions, other Bacteroidetes may outcompete *B. thetaiotaomicron* for exogenous amino acids.

**The Functional Requirement for a Vitamin B12-Regulated Locus Is Modulated by Community Composition**

We identified 165 independent transposon insertions, mapping to five adjacent genes (BT1957-53), that conferred a drastic fitness disadvantage during monooassociation of germ-free mice yet had no impact on exponential growth in vitro. Moreover, their effect on fitness was influenced by community context: the Bacteroidetes-only community exacerbated the competitive defect, while the Firmicutes+Actinobacteria consortium fully nullified the requirement for these genes. Introducing all 14 of these species resulted in an intermediate phenotype (Figures 7A and 7B).

Examination of *B. thetaiotaomicron* transcriptional profiles (NCBI GEO archive) disclosed that expression of genes in this locus (spanning BT1957–BT1949) is strongly upregulated in vivo compared to growth in vitro under a variety of conditions. Moreover, in two closely matched experiments conducted in defined minimal medium that differed in five components, expression was modulated >10-fold (Table S15). This observation was validated by qRT-PCR assays of BT1954 and BT1956 (data not shown). Systematic addition of each variable...
component revealed that locus transcription is induced in response to reduced levels of vitamin B12 (Figure 7C). Moreover, the ABC transporter encoded by BT1952-50 shares homology with the BtuFCD B12 acquisition system of S. typhimurium LT2 (23%, 35%, and 25% identity, respectively).

Vitamin B12 is critically involved in normal mammalian physiology yet is synthesized exclusively by microbes (Krautler, 2005). Because the complete genome sequence of each member of the defined microbial communities used in our experiments was known, we were able to use BLAST to identify homologs to known B12 synthesis, transport, or utilization genes in the synthetic human gut microbiomes (Tables S16 and S17). As described for the related species Porphyromonas gingivalis (Roper et al., 2000), members of the Bacteroidetes community (including B. thetaiotaomicron) were missing some or all of the genes necessary for synthesis of B12 or its direct precursors but encoded predicted transporters and likely have an obligate B12 requirement for growth. In contrast, the Firmicute/Actinobacteria group contained several members that harbored complete B12 biosynthetic pathways. To test these predictions, we attempted to culture each species on defined medium in the presence and absence of vitamin B12. While the Bacteroidetes were auxotrophic (Table S16), the B. thetaiotaomicron system did not exhibit noticeably restructured mutant populations compared to wild-type animals. It is possible that examination of microbial populations in closer contact with the host mucosa, or populations from host animals exposed to intentional immune stimulation, would aid the identification of genes differentially required for fitness in response to mutation.

This approach (INSeq) for functional genome-wide analysis of organisms for which a genome sequence (and possibly little else) is known is generally applicable and extends existing techniques in several important ways. First, a single transposon replaces the sets of individually barcoded variants needed for signature-tagged mutagenesis (Hensel et al., 1995). Second, high-throughput sequencing provides a general alternative to the species-specific DNA microarrays required for hybridization-based mutant profiling (Mazurkiewicz et al., 2006). Third, this sequencing-based strategy identifies the precise genomic locations and provides a “digital” count-based abundance readout of individual insertions in both coding and noncoding regions. In this way, independent insertions with shared behavior serve to validate gene-level fitness effects. Finally, because mariner family transposon activity has been demonstrated in Bacteria, Archaea, and Eukarya, this method is generalizable. Further, the barcoded pooling strategy used to create a sequence-defined archived strain collection allows for retrieval of individual strains of interest for follow-up studies of the impact of individual gene disruptions on various microbial functions and adaptations. In this way, a forward genetic tool (a mutagenized cell population that can be screened for phenotypes en masse) can also serve as a platform for reverse genetics (a collection of isogenic, sequence-defined mutations in most of the coding potential of the target genome).

Surprisingly, mice lacking major branches of the immune system did not exhibit noticeably restructured B. thetaiotaomicron mutant populations compared to wild-type animals. It is possible that examination of microbial populations in closer contact with the host mucosa, or populations from host animals exposed to intentional immune stimulation, would aid the identification of genes differentially required for fitness in response to mutation.
S16 and S17 for annotations) that are able to grow in defined medium lacking B12 ("demonstrated prototrophs"); light blue, organisms with a predicted complete biosynthetic pathway able to grow on rich medium but not on defined medium with or without B12 ("predicted prototrophs"); dark green, species without a complete pathway whose growth in defined medium requires B12 ("demonstrated auxotrophs"); light green, species without a complete biosynthetic pathway unable to grow on defined medium with or without B12 ("predicted auxotrophs"); black, species that do not grow on defined medium with or without B12 but that possess a B12-independent methionine synthase (the presence of such an enzyme implies the absence of a B12 requirement). The relative proportions of prototrophs and auxotrophs shown represent the average in each community in vivo as determined by species-specific qPCR of cecal contents.

(E) The B. thetaiotaomicron fitness requirement for BT1957-49 correlates with levels of the B12-prototrophic species Ruminococcus obeum in the microbial community. Each point represents an individual mouse containing a defined multispecies microbiota in addition to the B. thetaiotaomicron community. The relative abundance of R. obeum (determined by qPCR analysis of cecal contents at the time of sacrifice) is plotted against the average output/input ratio of B. thetaiotaomicron transposon mutants in genes BT1957-3 in each individual.

host immune surveillance. INSeq already can be applied to relatively small amounts of starting material: further decreases should help address the largely unexplored question of the relationship between the activities of the innate and/or adaptive immune systems and the biogeography of the microbiota.

Our study underscores how selection is shaped by microbial context. Because the INSeq strategy specifically targets transposon-adjacent chromosomal fragments, it is possible to monitor changes in the structure of a mutagenized population (in either wild-type or genetically manipulated gnotobiotic mice) even if this population constitutes a small fraction of the gut microbiota, ranging from the characteristics, determinants, and ecologic principles underlying its initial assembly, to the genomic correlates of niche partitioning among its members, to the genetic and metabolic factors that determine the persistence...
and i...e and enteropathogen species
(degree to which the functions required for persistence of gut mutualists overlap with those of pathogens [Hendrixson and DiRita, 2004; Lalouli et al., 2005; Liu et al., 2008; Shea et al., 2000]).

In summary, INSeq can be applied to a variety of phylotypes to identify factors that shape their adaptations to myriad environments, and further, to readily retrieve mutants in the genes encoding such factors. As such, INSeq, and the rapidly evolving capacity for massively parallel DNA sequencing that supports its application, should be a useful platform for microbial genetics, genomics, and ecology.

EXPERIMENTAL PROCEDURES

Bacterial Culture Conditions
Escherichia coli S-17 λpir strains (Cowles et al., 2000) were grown at 37°C in LB medium supplemented with carbenicillin 50 mg mL⁻¹ where indicated in the Supplemental Experimental Procedures. B. thetaiotaomicron VPI-5482 (ATCC 29148) was grown anaerobically at 37°C in liquid TYG medium (Hod...man et al., 1977) or on brain-heart-infusion (BHI; Becton Dickinson) agar supplemented with 10% horse blood (Colorado Serum Co.). Antibiotics (gentamycin 200 μg mL⁻¹ and/or erythromycin 25 μg mL⁻¹) were added as indicated in the Supplemental Experimental Procedures. Other human gut-derived species were cultured in supplemented TYG (TYG₆; see the Supplemental Experimental Procedures).

Genetic Techniques
DNA purification, PCR, and restriction cloning were performed by using standard methods. Primer sequences are provided in Table S19. pSAM DNA purification, PCR, and restriction cloning were performed by using Genetic Techniques a double-stranded DNA adaptor bearing a 3'-NN overhang. PAGE-purified transposon-sized fragments were extracted from the gel and ligated to a transposon-specific and an adaptor-specific primer. The 125 bp product adaptor-ligated library molecules were PCR amplified for 18 cycles using a transposon-specific and an adaptor-specific primer. The 125 bp product was purified by PAGE and sequenced using an Illumina Genome Analyzer as described in the user’s manual. Sequence images were converted into raw reads using Illumina software with default settings. Filtered, normalized, and mapped sequencing results from all samples are provided in Tables S9–S12 and S18.

Gnotobiotic Husbandry
All experiments using mice were performed using protocols approved by the animal studies committee of Washington University. Germ-free mice were maintained in gnotobiotic isolators and fed a standard autoclaved chow diet (B&K Universal, East Yorkshire, UK) ad libitum. Animals were sacrificed 14 days after gavage and fecal contents frozen immediately at –80°C.

ACCESSION NUMBERS
All of the sequencing results from this study are available from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) at accession number GSE17712.

SUPPLEMENTAL DATA
Supplemental Data include 10 figures, Supplemental Experimental Procedures, Supplemental Protocol, Supplemental MapSAM Software, Supplemental References, and 19 tables and spreadsheets and can be found with this article online at http://www.cell.com/cell-host-microbe/supplemental/S1931-3128(09)00281-9.

ACKNOWLEDGMENTS
We thank David O'Donnell, Maria Karlsson, Sabrina Wagoner, Nicole Koropatkin, Daniel Peterson, Pankaj Pal, Laura Langton, Jessica Hoisington-Lopez, Xuhua Chen, Laura Kyro, and James Dover for assistance, plus Gary Stormo, Jay Shendure, Ryan Kennedy, and the Gordon laboratory for helpful suggestions. This work was supported by National Institutes of Health grants DK30292 and 1F32AI07628-01 (to A.G.).

Received: March 30, 2009
Revised: June 8, 2009
Accepted: August 13, 2009
Published: September 16, 2009

REFERENCES


any XMRV isolates with the same or nearly the same sequences identified elsewhere originated from this event (23).

References and Notes


Predicting a Human Gut Microbiota from Dietary Ingredients

Jeremiah J. Faith, Nathan P. McNulty, Federico E. Rey, Jeffrey I. Gordon*

The interrelationships between our diets and the structure and operations of our gut microbial communities are poorly understood. A model community of 10 sequenced human gut bacteria was introduced into gnotobiotic mice, and changes in species abundance and microbial gene expression were measured in response to randomized perturbations of four defined ingredients in the host diet. From the responses, we developed a statistical model that predicted over 60% of the variation in species abundance evoked by diet perturbations, and we were able to identify which factors in the diet best explained changes seen for each community member. The approach is generally applicable, as shown by a follow-up study involving diets containing various mixtures of pureed human baby foods.

Owing to its many roles in human health (1–3), there is great interest in deciphering the principles that govern the operations of an individual’s gut microbiota. Current estimates indicate that each of us harbors several hundred bacterial species in our intestine (4, 5), and different diets lead to large and rapid changes in the composition of the microbiota (6, 7). Given the dynamic interrelationship between diet, the configuration of the microbiota, and the partitioning of nutrients in food to the host, inferring the rules that govern the microbiota’s responses to dietary ingredients represents a challenge (8).

Gnotobiotic mice colonized with simple, defined sequences of representative bacteria (9–12). These studies have focused on small communities exposed to a few perturbations. We used gnotobiotic mice harboring a 10-member community of sequenced human gut bacteria to model the response of a microbiota to changes in host diet. Our goal was to predict the absolute abundance of each species in this microbiota on the basis of knowledge of the composition of the host diet.

Furthermore, we wanted to gain insights into the niche preferences of members of the microbiota and to discover how much of the response of the community was a reflection of their phenotypic plasticity.

The 10 bacterial species were introduced into germ-free mice to create a model community with representatives of the four most prominent bacterial phyla in the healthy human gut microbiota (fig. S1A) (13). Their genomes encode major metabolic functions that have been identified in anaerobic food webs, including the ability to break down complex dietary polysaccharides not accessible to the host (Bacteroides thetaiotaomicron, Bacteroides ovatus, and Bacteroides caccae); consume oligosaccharides and simple sugars (Escherichia coli, Escherichia coli septic); and ferment amino acids (Clostridium symbiosum and E. coli). We also included two species capable of removing the end products of fermentation: a H2-consuming, sulfate-reducing bacterium (Desulfuromonas piger) and a H2-consuming acetogen (Blautia hydrogenotrophica).

To perturb this community, we used a series of refined diets in which each ingredient represented the sole source of a given macronutrient (casein = protein, corn oil = fat, corn starch = polysaccharide, and sucrose = simple sugar) and in which the concentrations of these four ingredients were systematically varied (fig. S1, B and C, and table S1). Each individually caged male C57Bl/6J mouse was fed a randomly selected diet, with diet switches occurring every 2 weeks (n = 13 animals; fig. S1D shows the variation of diet presentation between animals). Shuffling sequencing of total fecal DNA allowed us to determine the absolute abundance of each community member, based on assignment of reads to the various species’ genomes, in samples obtained from each mouse on days 1, 2, 4, 7, and 14 of a given diet period (13).

To measure the absolute abundance of each species in the model community, we determined the concentration of each of the four perturbed diet ingredients. We used a linear model where $y_i = \beta_0 + \beta_{casein}X_{casein} + \beta_{starch}X_{starch} + \beta_{sucrose}X_{sucrose} + \beta_{oil}X_{oil}$

where $y_i$ is the absolute abundance of species $i$; $X_{casein}$, $X_{starch}$, $X_{sucrose}$, and $X_{oil}$ are the amounts (in grams per kilogram of mouse diet) of casein,
cornstarch, sucrose, and corn oil, respectively, in a given host diet; \( \beta_0 \) is the estimated parameter for the intercept; and \( \beta_{\text{casein}}, \beta_{\text{starch}}, \beta_{\text{sucrose}}, \text{and} \beta_{\text{oil}} \) are the estimated parameters for each of the perturbed diet components. Because each mouse underwent a sequence of three diet permutations presented in different order, and each of the diet periods covered all of the 11 possible diets (fig. S1D), we were able to use two of these three diet intervals to fit the model for Eq. 1 (13 mice \( \times \) 2 diets per mouse = 26 samples per bacterial species); we then measured our ability to predict the abundance of each bacterial species for the 13 samples in the remaining (third) diet (13). Averaging this cross-validation from all three subsets, the model explained over 61\% of the variance in the abundance of the community members (abundance-weighted mean \( R^2 = 0.61 \); see table S2 for species-specific \( R^2 \)).

Although the cross-validation provided evidence that the response of this microbiota was predictable from knowledge of these diet ingredients, a more conclusive validation of the model would be its ability to make predictions for new diets. Therefore, we designed six additional diets with new combinations of the four refined ingredients. Using a design similar to the first experiment, eight different 10-week-old gnotobiotic male C57Bl/6J mice harboring the 10-member community were each given a randomized sequence of diets selected from the six new diets (shaded diets L to Q in fig. S1B) or one of the previous diets (fig. S1E). Fitting the model parameters with the data from the first experiment, we were able to explain 61\% of the variance in the abundance of the community members on the new diets, showing virtually equivalent results to the cross-validation procedure (table S2).

These results indicate that the linear model explains the majority of the variation in abundance of each community member by using only a knowledge of the species in the community and the concentrations of casein, cornstarch, sucrose, and corn oil in the diet, without having to explicitly consider the effects of microbe-microbe or microbe-host interactions or diet order. We also tested several other models, including adding interactions between the variables, quadratic terms, and interactions with quadratic terms (13). After correcting for the number of parameters in the model by using Akaike information criterion, the linear model was still the best-performing.

To further dissect the community response to these diet perturbations, we needed to infer which set of diet ingredients is associated with the abundance of each community member. Feature selection algorithms assume that the response from the first and second set of mice given the refined diets (fig. S1, D and E) are shown in purple and green, respectively, and the mean and standard error for all diets at a given concentration of casein are shown in red and tan, respectively.

**Fig. 1.** Total community abundance (biomass) and the abundance of each community member can best be explained by changes in casein. (A) The total DNA yield per fecal pellet increased as the amount of casein in the host diet increased (shown are mean \( \pm \) SEM for each tested concentration of casein). (B) Changes in species abundance as a function of changes in the concentration of casein in the host diet were also apparent for all 10 species; seven species (such as \( B. \) caccae) were positively correlated with casein concentration, whereas the remaining three species (such as \( E. \) rectale) were negatively correlated with casein concentration. Data points
variable (in this case, the abundance of each organism) is potentially affected by only a fraction of the variables in the model and use statistical methods to choose the subset of variables that most informatively predict the abundance of each species. Using stepwise regression as a feature selection procedure with the equation above, all species in our 10-member community had the diet variable $X_{\text{casein}}$ significantly associated with their abundance (table S3).

*E. coli* and *C. symbiosum* were the only bacteria with more than one variable significantly associated with their abundance (casein and sucrose for *E. coli*, and casein and starch for *C. symbiosum*). Further exploring this finding, we found casein highly correlated with the yield of total DNA per fecal pellet across all diets (Figs. 1A and 2). A component of casein, presumably amino acids and/or nitrogen, limits the biomass of the community: This resource limitation was observed even for combinations of three additional refined protein and two additional fat sources (soy, lactalbumin, egg-white solids, olive oil, and lard; $n = 9$ different diets given to another group of 9 C57Bl/6J male mice) (fig. S2 and table S4). However, the observed changes in species abundance are not a simple consequence of a constant relative abundance of each community member that is scaled upwards as casein is increased: Three community members (*E. rectale*, *D. piger*, and *M. formatexigenes*) decreased in absolute abundance by $-25$ to $50\%$ from the low- to high-casein diets, even though total community biomass tripled (Fig. 1B, fig. S3, and table S5). Similar changes in species abundance and total community DNA levels were observed when casein concentrations were altered in gnotobiotic mice harboring a nine- or eight-member subset of the original community (minus *B. hydrogenotrophica* or minus *D. piger* and *B. hydrogenotrophica*) (table S6).

Microbial RNA-seq was used on fecal RNA samples—prepared from mice on each diet (mean = $2.1 \pm 0.7$ replicates per diet) (table S7) (13)—to determine whether perturbations in diet ingredients correlated with underlying changes in mRNA expression by community members. Each of the 36 RNA-seq data sets was composed of 36-nucleotide-long reads (3.20 $\pm$ 1.35 $\times$ 10$^6$ mRNA reads per sample). Transcript abundances were normalized for each of the 10 species to reads per kilobase per million (RPKM) (14). After correcting for multiple hypotheses, we found no statistically significant changes in gene expression within a given bacterial species as a function of any of the diet perturbations (13). Although community members do not appear to significantly alter their gene expression, they do respond by increasing or decreasing their absolute abundances (Fig. 2), adjusting the total available transcript pool in the microbiota for processing dietary components. For example, as casein levels are increased across the diets *B. caccae* increases its contribution to the gene pool/community transcriptome; so, the number of transcripts per unit of casein remains roughly constant.

Because RNA-seq provides accurate estimates of absolute transcript levels (15), we used transcript abundance information as a proxy to predict the major metabolic niches occupied by each community member. For species positively correlated with casein, we found high expression of mRNAs predicted to be involved in pathways using amino acids as substrates for nitrogen, as energy and/or as carbon sources. In contrast, the three species that negatively correlated with dietary casein concentration showed no clear evidence of high levels of expression of genes involved in catabolism of amino acids (13). The changes in abundance of the negatively correlated species (such as *E. rectale*) can be explained by competition with another member of the community that increases with casein (fig. S4) (13, 16).

The power of the refined diets we used lies in the capacity to precisely control individual diet variables and to aid data interpretation from more complex diets. To test whether the modeling framework we used generalizes to diets containing food more typically consumed in human diets, we created 48 meals consisting of random combinations and concentrations of four ingredients selected from a set of eight pureed human baby foods (apples, peaches, peas, sweet potatoes, beef, chicken, oats, and rice) (table S8). The meals were administered for periods of 7 days to the same eight gnotobiotic mice that we used for the follow-up refined diet experiments described above and in fig. S1E (13). Each mouse received a sequence of six baby-food diets. The order of presentation of the baby-food diets was varied between animals (table S8) (13). We measured the absolute abundance of each bacterial community member on days 1, 5, 6, and 7 for each diet. Using the linear modeling approach described above (13), we were able to explain over half of the variation in species abundance using only knowledge of the concentrations of the pureed foods present in each meal ($R^2 = 0.62$). We used stepwise regression to identify the type of pureed food (or foods) present in a given mixed meal that was most significantly associated with changes in each bacterial species (Fig. 3 and table S9).

Defining the interrelationship between diet and the structure and operations of the human gut microbiome is key to advancing our understanding of the nutritional value of food, for creating new guidelines for feeding humans at various stages of their life span, for improving global human health, and for developing new ways to manipulate the properties of the microbiota so as to prevent or treat various diseases. The experiments and model described above highlight the extent to which host diet can explain the configuration of the microbiota, both for refined diets in which all of the perturbed diet components are digestible by the host and for human diets whose ingredients are only partially known. These models can now be tested by using larger defined gut microbial communities representing those of humans living in different cultural settings, and with more complex diets, including various combinations of food ingredients that they consume.
Oxytocin Selectively Gates Fear Responses Through Distinct Outputs from the Central Amygdala

Daniele Viviani,1 Alexandre Charlet,1 Erwin van den Burg,1* Camille Robinet,1* Nicolas Hurni,1 Marios Abatis,1 Fulvio Magara,2 Ron Stoop†,1,2†

Central amygdala (CeA) projections to hypothalamic and brain stem nuclei regulate the behavioral and physiological expression of fear, but it is unknown whether these different aspects of the fear response can be separately regulated by the CeA. We combined fluorescent retrograde tracing of CeA projections to nuclei that modulate fear-related freezing or cardiovascular responses with in vitro electrophysiological recordings and with in vivo monitoring of related behavioral and physiological parameters. CeA projections emerged from separate neuronal populations with different electrophysiological characteristics and different response properties to oxytocin. In vivo, oxytocin decreased freezing responses in fear-conditioned rats without affecting the cardiovascular response. Thus, neuropeptidergic signaling can modulate the CeA outputs through separate neuronal circuits and thereby individually steer the various aspects of the fear response.

Fear can be severely immobilizing but can also be a major driving force for some of humans’ most heroic acts. In both cases, the internal emotional experience may be similar, although it may lead to substantially different behavioral outcomes (1–3). Studies on human emotions often use autonomic nervous system parameters to assess arousal, because of the role of our internal organs in the emotional state (4, 5). Projections from the central nucleus of the amygdala (CeA) to the hypothalamus and different brain stem nuclei coordinate behavioral and physiological fear expression (6). It has been postulated that different fear responses, characterized by more active or passive behavioral coping strategies, can be triggered by a neuronal switch within the CeA (7). The question thus arises whether fear responses only vary in intensity, or whether different qualities of fear responses exist, reflected in different associations between behavioral and physiological components. We investigated whether a neurophysiological basis for such a distinct regulation could be found in the CeA.

†To whom correspondence should be addressed. E-mail: rstoop@unil.ch

*These authors contributed equally to this work.

1Centre for Psychiatric Neuroscience, Department of Psychiatry, Lausanne University Hospital Center, University of Lausanne, CH-1015 Lausanne, Switzerland. 2Department of Physiology, University of Lausanne, CH-1015 Lausanne, Switzerland.

Fig. 1. Distinct neuronal populations of the CeM project to PAG and DVC. (A and B) Coronal views of injection sites (A) in PAG of fluorescent green and (B) in DVC of red microspheres (left), corresponding CeM labeling (right). Scale bars: left, 1 mm; right, 500 μm. IPAG, lateral PAG; vIPAG, ventrolateral PAG; sp5, spinal trigeminal tract. (C) Separately labeled CeM neurons after conjections of green and red microspheres in respective PAG and DVC (left) versus colabeled neurons after injections of both microspheres in DVC (right, scale bar, 50 μm). (D) Quantification of colabeled (coloc) CeM neurons (see also table S2). (E) Electrophysiological characteristics of CeM→PAGs, “PAG” and CeM→DVCs “DVC.” (*P < 0.05; **P < 0.01, n = 22 to 43 neurons) Error bars indicate SEM.

References and Notes
12. A. L. Goodman et al., Cell Host Microbe 6, 279 (2009).
13. Materials and methods are available as supporting material on Science Online.
15. X. Fu et al., BMC Genomics 10, 161 (2009).

Acknowledgments: We are indebted to D. O’Donnell, M. Karlsson, and S. Wagner for their help with various aspects of gnotobiotic mouse husbandry and to B. Mickelson, J. Mogno, A. Goodman, N. Griffin, H. Seedorf, G. Simon, J. Chase, and B. Cohen for their many helpful suggestions during the course of this work. This work was supported by grants from NIH (DK30292 and DK70977) and the Crohn’s and Colitis Foundation of America. COPRO-seq and microbial RNA-seq data are available in the Gene Expression Omnibus (accession GSE26687). Processed data can be obtained at http://gordonlab.wustl.edu/modeling_microbiota/

Supporting Online Material
www.sciencemag.org/cgi/content/full/science.1206025/DC1
Materials and Methods
SOM Text
Figs. S1 to S5
Tables S1 to S13
References
24 March 2011; accepted 28 April 2011
Published online 19 May 2011.
10.1126/science.1206025