Characterizing the fitness landscapes of gut symbionts in defined community and diet contexts

Meng Wu
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

Part of the Biology Commons

Recommended Citation
Wu, Meng, "Characterizing the fitness landscapes of gut symbionts in defined community and diet contexts" (2014). Arts & Sciences Electronic Theses and Dissertations. 342.
https://openscholarship.wustl.edu/art_sci_etds/342

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Human and Statistical Genetics

Dissertation Examination Committee:
Jeffrey I. Gordon, Chair
Justin C. Fay
Daniel E. Goldberg
Robi D. Mitra
Michael A. Province
Clay F. Semenkovich

Characterizing the Fitness Landscapes of Gut Symbionts in Defined Community and Diet Contexts

by

Meng Wu

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

December 2014

St. Louis, Missouri
Table of Contents

Acknowledgements ............................................................................................................................................ v
Abstract of the Dissertation .......................................................................................................................... vii

Chapter 1

Introduction

The human gut microbiota .......................................................................................................................... 2
*Bacteroides* as model human gut symbionts .......................................................................................... 4
Factors that contribute to selection of bacteria for colonization of a host (gut) habitat ................... 7
  Realized versus fundamental niches ....................................................................................................... 7
  The contributions of host genetic factors .............................................................................................. 8
Combining computational and experimental tools to study interactions between members of the gut microbiota and host .................................................................................................................. 9
  Identifying the conserved core- and species pan-genome, inter- and intra-species genomic variations by comparative genomics ................................................................................................. 9
  Identification of genes important for fitness using expression profiling and signature-tagged mutagenesis (STM) .......................................................................................................................... 11
Using gnotobiotic mouse models to study fitness determinants ............................................................ 15
References .................................................................................................................................................... 17

Chapter 2

Genomic diversity and fitness of *E. coli* strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection

Abstract ....................................................................................................................................................... 27
Introduction ................................................................................................................................................ 28
Results ....................................................................................................................................................... 30
  *E. coli* strains cluster by host instead of habitat .................................................................................. 30
  Unsupervised hierarchical clustering of isolates based on their gene and SNP content .................. 33
  Tests of relative fitness in representative urinary isolates from patient 72 ......................................... 34
  Mechanisms that could underlie the increased fitness of episode 3 strain from Patient 72 in both the bladder and gut environment ........................................................................................................ 37
Discussion .................................................................................................................................................. 40
Chapter 3

Fitness and diet responsiveness simultaneously defined in multiple Bacteroides strains in vivo

Abstract ........................................................................................................................................85
Introduction ....................................................................................................................................86
Results ...........................................................................................................................................87
  Characterizing multiple transposon mutant libraries simultaneously in vivo (multi-taxon INSeq) ..........................................................87
  Community assembly, response to diet, and resilience ...............................................................89
  Identifying core in vivo fitness determinants in four Bacteroides strains .............................91
  Differences in fitness determinants in the two B. thetaiotaomicron strains .............................93
  Identifying a diet supplement that can specifically manipulate B. cellulosilyticus abundance .................................................................................................................................95
Prospectus ....................................................................................................................................98
Acknowledgements ......................................................................................................................100
References ...................................................................................................................................101
Figure Legends .............................................................................................................................105
Figures .........................................................................................................................................108
Supplemental Materials ...............................................................................................................112
  Supplemental Methods .............................................................................................................112
  Supplemental Results .............................................................................................................118
    Reproducibility, sensitivity and precision of multi-taxon INSeq .........................................118
Evidence that the behavior of each mutant library in the artificial community was equivalent to that of the corresponding wild-type strain ........................................119

Supplemental Figure Legends .........................................................................................................................120
Supplemental Figures ..........................................................................................................................................126
Supplemental Tables ........................................................................................................................................140

Chapter 4

Future directions

Creation of clonally arrayed multi-taxon INSeq transposon mutant libraries ................................. 143
The effects of the arabinoxylan on interactions between the artificial community and host in the context of the HF/HS diet ................................................................................................................. 145
Application of multi-taxon INSeq into other model human gut symbionts ........................................ 147
References .......................................................................................................................................................... 149

Appendices

Appendix A ......................................................................................................................................................... 152
Appendix B ......................................................................................................................................................... 165
Appendix C ......................................................................................................................................................... 186
Acknowledgements

This “chapter” of my thesis is the part I have been thinking about since the first day I entered Graduate School. Everyday, I am very thankful to my colleagues, friends and family for their support and love.

I am extremely grateful and indebted to my thesis advisor, Jeffrey Gordon, for giving me the opportunity to be part of the Gordon lab and helping me grow as a scientist and as a person. His vision, perseverance, and passion for science always amazed and inspired me over the years. His generosity, encouragement, respect and care toward the students make the Gordon lab truly a family to me.

I would like to thank my thesis committee, for their time, scientific input and support: Justin Fay, for guidance in evolutionary genetics; Michael Province for advice in statistical analysis; Clay Semenkovich, for suggestions regarding metabolism; Rob Mitra, for inspiring discussions about technique challenges associated with sequencing; last but not least, my thesis committee chair, Dan Goldberg, who always reminded me to think about problems from the bacterial side.

To the past and present Gordon Lab members, thank you for being the best lab-mates I could ever imagine. Sabrina Wagoner, as the “Mom” of the lab, is always there for any questions whether it was about lab or life; David O’Donnell and Maria Karlsson, who made all the mouse experiments easy, and the trips to the mouse house so much fun; Janaki Lelwala-Guruge, who sat next to me when I joined the lab, introduced me to lab and taught me the magic of bacterial culture; Marty Meier, who made the extraction of hundreds of thousands DNA samples possible for my dissertation; Andrew Goodman, from whom I learned about the insertion sequencing (INSeq) method that served as the base on which I built my thesis work. Mark Gonzalez, Federico Rey, Ansel Hsiao and Henning Seedorf were an incredible knowledge base which I had access to whenever I have any bacterial related questions. Jeremiah Faith, Alejandro Reyes are the people I could turn in to when I faced bioinformatics challenges. Nathan McNulty, who I got direct influence from on all issues related to lab, notably labnote record keeping and critical thinking! Vanessa Ridaura,
who is super encouraging and supportive. Mark Charbonneau is the person I could always count on for help. Philip Ahern, who truly inspired me and shaped my views on scientific questions and beyond from daily conversations and discussions. Many other wonderful people in the lab made my day-to-day lab life such an inspiring and pleasant experience. There is not enough space to individually thank everyone; but I have learned invaluable things from every one of this amazing group of people, who I had the honor to share my time with.

I would also thank every staff member in the Gordon lab, CGS and DBBS whose hard work allows us to focus on graduate studies and research. Stephanie Amen, Laura Kyro, Eric Martin, Brian Koebbe, Jessica Hoisington-Lopez, Kristina Hedtkamp and Jeanne Silvestrini.

I would also thank my friends outside of lab, who made my life in graduate school colorful.

I owe my achievements, in greatest part, to my parents for their lifetime of love, unwavering support, and the sacrifice of supporting everything I do even though I was so far away from them. I also thank my grandfather, who always believes in me and encourages me to pursue my dreams.

Above all, I am so grateful for my husband, Mingchao, for all the sacrifices he has made for me, for all the love he has given me. Thank you for always being there for me.
ABSTRACT OF THE DISSERTATION

Characterizing the Fitness Landscapes of Gut Symbionts
in Defined Community and Diet Contexts

by

Meng Wu

Doctor of Philosophy in Biology and Biomedical Sciences
Human and Statistical Genetics

Washington University in St. Louis, 2014

Professor Jeffrey I. Gordon, Chair

A species’ niche is the description of all the environmental conditions required to permit a population of that species to persist, including the effects of the population on those conditions. This definition includes the species’ resource requirements, as well as stress tolerances and interactions with other species acting as competitors, predators, parasites, and mutualists. The human gut microbiota serves as a microbial ‘metabolic organ’ tasked in part with the biotransformation of many components of our diet. Relatively little is known about the factors that allow members of the human gut microbiota to persist in a habitat that experiences marked changes in its nutrient environment. Identification of these factors is important for understanding the mechanisms that determine community assembly, community responses to and recovery after various perturbations, and the food webs that link microbes to one another and to their host. Therefore, in my thesis, I developed an experimental and computational pipeline for multi-taxon insertion sequencing (multi-taxon INSeq) to identify fitness determinants in multiple species and strains of human gut Bacteroides. Libraries of tens of thousands of transposon (Tn) mutants for each of four human gut Bacteroides strains, two of which represented the same species, were generated. These libraries were then introduced into adult germ-free mice as part of a 15-member artificial defined human
gut microbiota containing 11 other wild-type bacterial species. Mice were fed diets either low in fat and high in plant polysaccharides (LF/HPP), or high in fat and simple sugars (HF/HS). Fecal samples, collected over time, were subjected to multi-taxon INSeq and my analysis pipeline, which was based on maximum likelihood estimation. A total of 86 core fitness determinants were identified across all four strains; a large fraction of these determinants were involved in various aspects of amino acid biosynthesis. Significant intra-species differences were detected in response to diet between two strains of \textit{Bacteroides thetaiotaomicron}, highlighting differential strategies facilitating their co-existence within a complex community. By combining information gleaned from \textit{in vivo} and \textit{in vitro} INSeq experiments, as well as from \textit{in vivo} and \textit{in vitro} microbial RNA-Seq, I determined that arabinoxylan, the most common hemicellulose in cereals, was able to drive expression of a polysaccharide utilization locus that represented a key fitness factor in \textit{Bacteroides cellulosilyticus} WH2 when mice consumed HF/HS diet. Supplementation of the drinking water with this glycan in turn significantly increased the representation of \textit{B. cellulosilyticus} WH2 within the defined community in the context of the high fat diet. Multi-taxon INSeq defined the changes in fitness determinants of this and the other \textit{Bacteroides} in response to arabinoxylan supplementation. Collectively, these studies revealed multiple mechanisms by which our microbial symbionts establish themselves in the gut, including species-specific, strain-specific, as well as core responses, which mapped to a variety of metabolic/nutrient processing pathways. The approach described for mapping fitness landscapes in a community context should facilitate discovery efforts aimed at identifying the niches of microbiota members, as well as ways to deliberately reshape community structure and function through dietary interventions.
Chapter 1

Introduction
The human gut microbiota

The human gut is colonized by tens of trillions of microbes, collectively known as the microbiota. The majority of organisms that form this community reside in the distal gut where microbial density reaches $10^{11}-10^{12}$ cells per milliliter contents [1]. The microbiota contains members of all three domains of life, in addition to viruses [2-4]. The gut microbiota begins to assemble at birth and achieves an adult-like configuration by 24-36 months of age [5, 6]. In healthy adults, the microbiota is dominated by members of two bacterial phyla, the Bacteroidetes and Firmicutes [7-9]. The collective genomes and constituent genes represented in this community are called the gut microbiome. The microbiota plays important roles in human biology, including defense against invasion by enteropathogens, immune homeostasis, and a variety of metabolic activities, including those related to the biotransformation of components of our varied diets.

The microbiota protects the host from enteropathogens, a process referred to as “colonization resistance”. Studies using germ-free animals showed that they are more susceptible to infection by certain bacteria (i.e., *Shigella flexneri; Salmonella spp., Streptococcus mutans*) compared to conventionally raised animals, colonized with a microbiota from birth [10-13], or gnotobiotic animals colonized with specific commensal bacteria [12, 14, 15]. Observations in human patients and animal studies showed that disruption of the structure of the microbiota by antibiotics also dramatically increases the susceptibility towards enteric infections (i.e. *Clostridium difficile, Salmonella spp.*) [16-18].

Colonization by the microbiota begins at birth and shapes both the phenotype of the mucosal and systemic immune systems. Work from germ-free animals has shown that the mucosal immune system is undeveloped in the absence of gut microbes [19, 20], with numerous facets of the immune system affected. Defects have been observed in the cellularity and organization of the spleen and secondary lymphoid organs such as the mesenteric lymph nodes and Peyer’s Patches [21], the abundance of intestinal regulatory T cells and Th17 cells [22, 23], intestinal IgA titers, cytotoxicity of intestinal epithelial lymphocytes, anti-microbial production etc [24].
Transcriptional profiling revealed myriad differences in gene expression in intestinal epithelial cells between germ-free mice and mice colonized with just a single prominent member of the gut microbiota, *Bacteroides thetaiotaomicron*. *Bacteroides fragilis* directs maturation of the immune system, including the expansion and differentiation of splenic CD4+ T cells [25] and promoting the generation of tolerogenic peripheral Treg responses in the colon [26]. Intestinal CD4+ helper T cell differentiation can also be shaped by specific members of the microbiota, with segmented filamentous bacteria promoting Th17 cell responses in the small intestine [22, 27, 28], and members of the Clostridia and Bacteroides driving Treg induction in the small intestine [23, 26, 29–32].

Accumulating evidence indicates that the gut microbiota is instrumental in the control of host energy metabolism. In this role, the gut microbiota could be viewed as a microbial “metabolic organ”. Gut microbes possess a broad array of carbohydrate active enzymes (CAZymes) that are not represented in our human genomes, allowing them to ferment a broad array of dietary polysaccharides that could otherwise not be digested. The products of microbial fermentation include short-chain fatty acids (e.g., acetate, propionate and butyrate) that serve as energy sources for host cells (e.g., colonocytes), as well as other products, including lactate, succinate and formate.

The energy provided to the host by gut microbes is quite substantial in herbivores, ranging from 30-85% of metabolizable energy in rabbits [33], horses [34], and sheep [35]. In omnivores, such as the pig, microbe-derived nutrients are responsible for 10% of the animals’ maintenance energy requirement [36]. It is estimated that up to 10% of daily energy needs in humans is supplied from the fermentation activities carried out by the gut microbiota; the contribution is greater in individuals living in non-Westernized countries where there is less consumption of processed food and more complex dietary fibers [37, 38].

Studies by our lab and others have shown that the composition of gut microbiota could affect host energy homeostasis. Backhed et al found that the germ-free (GF) mice had about 40% less total body fat than age-matched conventional-raised (CONV-R) mice, even though GF mice consumed 29% more food. Additionally, colonization was correlated with increased hepatic lipo-
genesis through the stimulation of ChREBP (Carbohydrate Responsive Element Binding Protein) and SREBP-1 (Sterol Responsive element Binding Protein) hepatic expression and enhanced glucose uptake in the intestine [39, 40]. Ley et al demonstrated that obesity can be associated with an altered gut microbiota in a genetic model of obesity (ob/ob mice had a reduction in the abundance of Bacteroidetes and a proportional increase in the abundance of Firmicutes) in a rodent model [41], and a similar phenotype has also been observed in humans [42]. To determine if the relative differences of bacterial proportions leads to different body phenotypes, Turnbaugh et al transplanted cecal microbiota from lean and ob/ob mice to GF wild-type recipients, and they found that mice harbouring the microbiota from obese mice had a fat gain, and extracted more calories from their food compared to the mice that received the gut microbiota from lean mouse donors [43]. More recent metagenome-wide association studies from different populations (Chinese, European) also showed changes in the composition of microbiota associated with other metabolic dysfunctions (e.g., type 2 diabetes) [44, 45].

These findings underscore the enormous utility in better understanding how the structure and functions of the microbiota are shaped and further, how they in turn modulate host physiology.

**Bacteroides as model human gut symbionts**

*Bacteroides* spp. are Gram-negative, obligate anaerobic bacteria that have been used as model organisms for investigating the mechanisms governing acquisition and processing of dietary polysaccharides. There are several reasons why members of the genus *Bacteroides* have become favored models: e.g., they are the primary fermenters in the intestine; they are prominently represented in the gut; they are aerotolerant and readily cultured outside their native habitat, and owing to the work of Abigail Salyers and her colleagues, tools have been available for their genetic manipulation for quite some time facilitating molecular dissection of their impacts on host biology.

A series of studies from members of the laboratory of Abigail Salyers, and other groups, demonstrated the impressive capacity of *Bacteroides* spp. to degrade a variety of complex polysac-
charides. Glycan breakdown is mediated by clusters of co-regulated genes known as polysaccharide utilization loci (PULs). *Bacteroides thetaiotaomicron* VPI-5482 was the first prominent human gut *Bacteroides* to have its genome sequenced: it contains 88 PULs comprising 866 genes that together comprised 18% of its genome [46, 47]. Subsequent work disclosed that a large investment of genome content in PULs is a common feature of gut *Bacteroides*.

The first PUL described and characterized was the starch utilization system (Sus) in the prominent human gut symbiont, *Bacteroides thetaiotaomicron* [48-50]. The Sus locus contains eight adjacent genes (*SusRABCDEFG*). The signature feature of a PUL is a pair of outer membrane proteins related to SusC and SusD, both of which are required for utilization of starch by *Bacteroides thetaiotaomicron* VPI-5482. SusD paralogs are predicted to be secreted lipoproteins peripherally associated with the outer membrane, where they play a role in polysaccharide binding. SusC paralogs are predicted to be TonB-dependent, β-barrel-type outer membrane proteins, and are likely needed for energy-dependent transport of glycans into the periplasmic space. Other features contained within PULs include ORFs encoding outer membrane and periplasmic polysaccharide lyases (PLs), glycoside hydrolases (GHs) and carbohydrate esterases (CEs). PULs also contain an upstream ORF specifying a sensor/regulator, typically hybrid two component systems or extra-cytoplasmic function sigma (ECF-σ) and genomically linked anti-σ factors.

Experiments testing the ability of 188 strains of 10 different human colon-derived *Bacteroides* species to ferment different plant polysaccharide and mucin sources discovered both species and strain-specific differences in carbohydrate utilization [51]. A recent study examined two closely related species, *B. thetaiotaomicron* and *B. ovatus*, represented in this original study. Whole genome transcriptional profiling of these isolates, in conjunction with the characterization of the growth phenotypes of derived mutants lacking PULs in minimal medium supplemented with different glycans as the sole carbon source, revealed the identify and specificity of the PULs that target plant polysaccharides represented in our diets, including hemicelluloses, pectins, and fructans [52]. Each species possesses several unique PULs: *B. ovatus* harbors PULs that enable it to use all of the common hemicelluloses, including xylan, arabinoxylan, and β-glucan, while *B.
\textit{thetaiotaomicron} does not have any. In contrast, \textit{B. thetaiotaomicron} has several PULs that allow utilization of host glycans, while \textit{B. ovatus} does not.

Another direct demonstration of how PULs help \textit{Bacteroides} to compete efficiently for the complex glycans in their habitat and occupy distinct niches (professions) involved an analysis of \textit{B. thetaiotaomicron} VPI-5482 [53]. The authors found that a single extracellular $\beta$2-6 endo-fructanase that was encoded in a single PUL was required for \textit{B. thetaiotaomicron} to utilize levan (a $\beta$2-6-linked fructan) but not inulin (a $\beta$2-1 fructan). Homology searches using the fructose-binding domain in the hybrid two component system associated with this \textit{B. thetaiotaomicron} PUL combined with a comparative genomic analysis of synteny revealed other fructan utilization PULs in \textit{B. caccae}, \textit{B. ovatus}, \textit{B. fragilis}, and \textit{B. uniformis}. However, only the \textit{B. thetaiotaomicron} genome contained a $\beta$2-6 endo-fructanase, while \textit{B. caccae} and \textit{B. ovatus} contained genes encoding enzymes that target $\beta$2-1 fructans. In followup experiments involving gnotobiotic mice co-colonized with \textit{B. thetaiotaomicron} and \textit{B. caccae} and fed a diet with inulin as the sole polysaccharide, \textit{B. caccae} displayed a competitive advantage over \textit{B. thetaiotaomicron}.

Other studies in gnotobiotic mice colonized with wild type \textit{B. thetaiotaomicron} and a mutant with deletion of five ECF- $\sigma$ transcriptional factors that activate mucin glycan utilization genes have shown that loss of the latter capacity impairs persistence in the gut and imperils mother-to-offspring transmission [47].

Unlike carbon acquisition, which has been intensively studied, the preferred nitrogen sources of \textit{Bacteroides} species in the gut remain unclear. \textit{Bacteroides} are capable of fixing NH$_3$, primarily via glutamate dehydrogenase [54]. The inability of many \textit{Bacteroides} species to substitute free amino acids, peptides, nitrate, or urea for ammonia as a nitrogen source [55] suggests that NH$_3$ may be the most common source of nitrogen for \textit{Bacteroides} in the intestine, though this remains speculative and may not apply for all species in this genus.
Factors that contribute to selection of bacteria for colonization of a host (gut) habitat

Realized versus fundamental niches

There have been a number of studies of both invertebrate and vertebrates that have focused on how microbial communities are assembled in their host habitats. For example, the light organ of the Hawaiian squid *Euprymna scoplopes* is mono-colonized with *Vibrio fischeri*. Work in the laboratory of Edward Ruby showed that a strain of *V. fischeri* recovered from Japanese pinecone fish (*V. fischeri* MJ11), could not colonize squid as efficiently as a squid-derived strain (*V. fischeri* ES114). Genome-wide comparisons revealed that *rscS*, which encoded a membrane-bound two-component sensor kinase that upregulates production of the Syp exopolysaccharide important for biofilm formation during initial colonization, is absent from *V. fischeri* MJ11. Provision of this gene *in trans* allowed the fish symbiont (*V. fischeri* MJ11) to colonize the light organ of *E. scoplopes* at levels that were comparable to that achieved by the squid’s native strain [56].

Macroecologists differentiate the conditions under which an organism actually does live (its realized niche) from the conditions under which an organism can live (its fundamental niche) [57]. Earlier work from our lab described the results of reciprocal gut microbiota transplants involving communities harvested from conventionally-raised zebrafish and conventionally-raised mouse donors and germ-free zebrafish and mouse recipients. The results of these experiments revealed that bacterial taxa from zebrafish that had not been described in the normal mouse gut microbiota could persist in the mouse gut [58]. More recent work from our lab provided evidence that the mouse intestinal tract, while highly selective, is within the fundamental niches of bacterial phylotypes that normally reside in a wide variety of foreign (non-mouse gut) environments. In this study adult germ-free C57Bl/J mice, fed a standard mouse chow high in plant polysaccharides and low in fats, were colonized with xenomicrobiota harvested from human, zebrafish, and termite guts, human skin and tongue, soil, and estuarine microbial mats. Bacteria taxa from these ‘alien’ environments colonized and persisted in the mouse gut and their capacity to metabolize dietary and host carbohydrates and bile acids correlated with colonization success [59]. CAZymes shared
across the gut-selected xenomicrobiomes included eight glycoside hydrolase families, the most prominent being GH13 (includes α-amylases and pullulanases that breakdown starch), and three carbohydrate esterase families. The prominence of GH13 is consistent with the large amount of starch in the mouse chow. The relative abundances of genes encoding glycoside hydrolases (GH) and polysaccharide lyases (PL) were also significantly correlated with microbial biomass (families GH97, GH28, and GH106 being most highly correlated). Together, these results indicate that the mouse intestinal tract, while highly selective, is within the fundamental niches of bacterial phylogenotypes derived from a wide variety of environments.

**The contributions of host genetic factors**

Using 16S rRNA amplicon sequencing to quantify variation in the bacterial composition of fecal microbiota sampled from a number of recombinant inbred (RI) mouse lines, Benson and coworkers identified 13 significant quantitative trait loci (QTL) (LOD≥3.9; P<0.05)) and five suggestive QTL (LOD≥3.5; P<0.1) that control variability in the relative abundance of specific taxa [60]. Some of these loci control individual bacterial species, some control groups of related taxa, and some have pleiotropic effects on groups of distantly related organisms. In humans, studies of monozygotic and dizygotic twins have provided evidence for host genetic effects on the representation of the gut microbiota’s dominant archaeon, the methanogen *Methanobrevibacter smithii*, other bacterial taxa associated with the hydrogen economy of the gut, as well as other functionally distinct bacterial taxa [61, 62].

Recent work using genome-wide association studies of a number of inbred strains of mice has examined responses to a diet high in fats and simple sugars but devoid of complex polysaccharides (HF/HS chow) versus one low in fat and high in plant polysaccharides [63]. They observed robust diet-induced shifts in microbiota structure: HF/HS chow fed mice had greater relative abundances of several genera belonging to the Firmicutes and lower abundances of genera in the Bacteroidetes compared to animals fed a diet rich in plant polysaccharides and low in fat, when all other environmental factors were controlled. Based on an analysis of over 100 inbred strains of mice,
they showed these shifts were robust to genetic background. This study demonstrated how diet represents a foundational shaper of the configuration of gut microbiota across a range of genetic backgrounds.

**Combining computational and experimental tools to study interactions between members of the gut microbiota and host**

Our increasing awareness of the role played by gut symbionts in health has highlighted the importance of delineating the molecular mechanisms governing how microbial communities are first established, how they respond to perturbations and subsequently restore themselves to their pre-perturbed state (resiliency), and how members of these communities influence host biology. Knowing the genetic determinants important for colonization and persistence in the face of specific perturbations, could lead to development of strategies for targeted manipulation of the microbiota that promote wellness or treat diseases where a microbiota contribution is known. This thesis has used human gut *Bacteroides* as model organisms to address this issue.

Studies of gnotobiotic mice colonized with defined consortia of sequenced members of the microbiota and fed defined diets with systematically manipulated variations in ingredient concentrations have shown how the combination of simple linear models and feature selection can be used to identify dietary factors that specifically alter the relative abundance of individual members in the community [64]. However, at the time of this study, the appropriate tools facilitating genome wide scans of the genes that are responsible for defining the fitness (as defined by abundance) of these organisms in different diets (or microbial community contexts) were not available. It was the absence of these tools that inspired and directed the studies that I describe in this thesis.

**Identifying the conserved core- and species pan-genome, inter- and intra-species genomic variations by comparative genomics**

High throughput sequencing is now a standard component of the microbiological toolbox, with thousands of publicly available bacterial genome sequences for comparison. Bacterial genome
sequencing and comparative genomic analysis have become standard methods for identifying orthologs and paralogs. Identifying genes unique to a strain/species provides the foundation for testable hypotheses about how the organism has evolved features that increase its fitness in certain environmental contexts.

In Chapter 3, the genomes of three different human gut-derived *Bacteroides* spp. and two strains representing one of these species are compared to search for orthologs shared among all four isolates as well as genes that are unique to a given isolate. The similarity of related genomes can be determined through whole genome alignment, or alignment of their predicted proteomes. Strain-level variation can be generated by several mechanisms: point mutations, gene rearrangements (e.g., inversions or deletions), and/or acquisition of DNA from other organisms via horizontal gene transfer (HGT). HGT is a common mechanism, occurring even between distantly species, through exchange of plasmids, phage, transposons, and insertion elements. Horizontally transferred DNA sequences can be identified as discrete segments (genomic islands or GEIs) [65]. Comparing the genomes of the two strains from *B. thetaiotaomicron* in Chapter 3 led to the identification of several such GEIs, including GEIs carrying PULs, suggesting a mechanism by which these two strains may have adapted to different glycan niches.

One illustration of the power of using comparative genomics is provided in Chapter 2, where we characterize the genomes of urine and fecal *E. coli* isolates, obtained from time series studies of four women with recurrent urinary tract infections, using novel SNP (single nucleotide polymorphism)-based and “OGU” (Operational Gene Unit)-based analyses. We found that *E. coli* strains from women with recurrent urinary tract infection can move between the gut and urinary tract without a fitness trade-off. *In vivo* competition studies in mouse models, using isolates taken from one of the patients with a wholesale population shift, revealed that strain that dominated her last UTI episode had increased fitness in both the gut and the bladder relative to the strain that dominated in preceding episodes. Increased fitness correlated with genes affecting nutrient utilization and virulence identified by comparative genomics.
Despite their utility, comparative genomic approaches have several limitations. While next generation sequencing technologies have provided large datasets of short reads, assembly of these short reads into long contigs that can be used to generate a finished genome sequence remains quite challenging, especially for de novo assemblies where there are no reference genomes. This challenge can be overcome by hybrid assembly methods that combine short reads with long reads generated using the latest in next generation sequencing platforms (PacBio, Oxford Nanopore). Another limitation is that current automatic annotation pipelines based on sequence homology with reference databases (e.g., UniProt, KEGG etc.) lead to numerous mis-annotations. Manual curation, coupled with experimental confirmation of function, is required for accurate annotation. Finally, the most fundamental challenge is that many genes remain unannotated; even the intensively studied model organisms such as *Escherichia coli* K12 where over 50% of its ORFs remain uncharacterized (http://www.jcvi.org). In addition, comparative genomics only yields features potentially important for characteristics (phenotypes) of interests. These limitations highlight the need for functional studies of a large number of genes across a range of species. Experimental techniques that harness the power of selective pressure to identify key loci are critical for answering the question of whether observed variations in gene structure/content revealed through comparative genomic analyses are causally related to an organisms biological properties.

**Identification of genes important for fitness using expression profiling and signature-tagged mutagenesis (STM)**

Two approaches have been used to identify factors required for growth *in vivo* and/or persistence in the intestinal tract after various perturbations (e.g., antibiotic treatment, diet change); (1) define gene expression (e.g., with microbial RNA-Seq) and/or (2) employ signature-tagged mutagenesis (STM) to identify loci genes critical for fitness.

Using model communities composed of defined collections of bacterial species/strains with sequenced genomes, microbial RNA-Seq can be used to quantify changes in community gene expression at high resolution (e.g., [66]; see Appendix B). mRNA expression data can be mapped
to catabolic and anabolic pathways to predict the metabolic response of the community (metatranscriptomics; ‘top-down’ approach) and/or the responses of individual members of the model community (‘bottom-up’ approach). The *in vivo* data can be further interpreted and expanded by RNA-Seq analysis of individual members grown *in vitro* under defined conditions (e.g., minimal medium supplemented with systematically varied carbon sources) to aid in the identification of function of genes regulated under conditions of interest (see Chapter 3 and Appendix B for an example of this approach)

However, this approach has limitations. First, transcriptional induction of a given gene does not guarantee that it is critical to the success of its host bacterium. For example, *Bacteroides spp.* express multiple PULs when gnotobiotic mice are fed a diet enriched in plant polysaccharides [52, 67] but not all of these loci may be critical for survival [68]. Secondly, genes that are not differentially expressed under different conditions may still be critical for fitness (e.g., constitutively expressed genes encoding proteins with critical functions).

A more direct/informative approach for identifying genes important for fitness under a given biological condition or conditions is to construct mutants, and compare their fitness against the parental wild-type strain. However, the process of generating targeted mutants is laborious, and single mutants must be tested individually. Signature-tagged mutagenesis (STM) was designed to enable high-throughput parallel analysis of the fitness of mutant strains. In STM, each mutant is tagged with a different DNA sequence in such a way that all tags can be co-amplified from the DNA of mixed populations [69]. This simultaneous labeling allows for identification of specific mutants before and after the entire mutant pool is subjected to a selection. Until recently, the detection of probes has typically been performed by DNA hybridization, which requires a species-specific microarray [70], or by sequencing of individual mutants [71].

Our lab recently developed a new method, named “insertion sequencing (INSeq)”, that combines existing STM methods with high-throughput parallel DNA sequencing [72]. The approach is based on a modified *mariner* transposon containing MmeI sites at each end of the trans-
poson. MmeI is a restriction enzyme that cuts DNA at a site 20–21 bp from its recognition site in the inserted transposon, capturing 16-17 bp flanking genomic DNA. In this method, the genomic regions adjacent to the transposons can be amplified by linear extension using a biotinylated primer specific for the transposon and reading outward from the insertion site. The resulting products are bound to magnetic streptavidin coated beads, digested with MmeI and barcoded with sample-specific linkers appended to each restriction fragment. After limited PCR amplification, fragment pools are analyzed using next-generation sequencing. The sequence of each read can be used to map the location of a transposon in the genome, and the resulting read count measures the relative abundance of that mutant in the overall mutant population before and after application of a selection ([68]; see Appendix A).

This approach provides a way to define the relative abundance of tens of thousands of mutants in parallel, before and after exposure to a selective environment of interest. Mutants that decrease in relative abundance in the selective condition are likely important for fitness under that condition. As INSeq library preparation can be completed in a multiplex format, experiments can involve material obtained from high-resolution time-series studies of the effects of various in vitro or in vivo selections, and from multiple biological and technical replicates. Any microbe whose genome has been sequenced and is amenable to mariner transposon mutagenesis can potentially be studied using INSeq.

Other techniques for mapping and quantifying transposon insertions by high-throughput sequencing have also been developed. They include two protocols named Tn-Seq: one version [73, 74] also uses an MmeI-adapted mariner transposon but does not include a transposon-specific, linear PCR step during library preparation. Compared with INSeq, this simpler approach to library preparation reduces the effort required for sequencing. The other Tn-Seq protocol [75] does not use an MmeI-adapted mariner transposon; instead, it uses a circularization step to capture the targeted DNA fragments. This could allow use of transposons that cannot be modified to include MmeI sites in their inverted repeats. Additional approaches, termed “high-throughput insertion tracking by deep sequencing (HITS)” [76] or “transposon-directed insertion-site sequencing (Tra-
DIS)[77] also do not require Mmel sites in the transposon. For these protocols, mutagenized DNA is randomly sheared, adapters are ligated to all fragments, and transposon-chromosomal junctions are enriched by PCR either with (HITS) or without (TraDIS) an affinity-purification step. In principle, these Mmel-independent methods permit capture of longer fragments of genomic sequence adjacent to each transposon, a feature that may be necessary for mapping insertions in more complex genomes or mixtures of genomes. Compared to these other methods, the INSeq protocol has several advantages; (i) introducing an Mmel site in the transposon allows the excised transposon/chromosomal junctions to be of uniform size, thus avoiding size preferences (bias) during PCR; (ii) the fact that INSeq is a solid-phase–based technique allows for easy separation of target sequences from ‘background’ chromosomal DNA; (iii) robotic automation of library generation makes the protocol amenable to monitoring changes in the representation of mutants during studies that require repeated, high-resolution sampling.

As with any negative selection technique, stochastic bottlenecks in the selective process need to be considered when determining the optimal number of mutants to be pooled for inclusion in an experiment [69, 71]. Because each insertion is mapped with nucleotide-level resolution, comparing the behavior of multiple mutants in the same gene highlights these selective bottlenecks. Stochastic bottlenecks can also be identified by comparing results from multiple independent biological replicates.

In my dissertation research, I modified the INSeq delivery vector to incorporate taxon-specific barcodes into the mariner transposon. Differentially tagged transposons can then be used to generate mutant libraries from different bacterial species, allowing for multiplex STM identification of fitness factors across many bacterial taxa in a defined model (human gut) community. The use of multiple taxon-specific tags allows for differentiation of mutant populations even between different strains of the same species, which allows for identification of fitness factor differences at much finer levels of taxonomy. Together, the combination of taxon-specific transposon tags, modifications to the original INSeq transposon insertion site amplification and sequencing protocol, and
associated bioinformatics analysis suite formed a new analysis pipeline for fitness determinants that I dubbed “multi-taxon INSeq”.

One major modification to the original INSeq sequencing library preparation in the multi-taxon INSeq protocol is elimination of the linear PCR step. With the taxon-specific barcode incorporated into the transposon, the length available for design of the linear PCR primer was reduced, leading to introduction of non-specific background amplicons, especially when the desired target only represented a small portion of the starting material. Instead of employing linear PCR, the multi-taxon INSeq pipeline directly digests genomic DNA adjacent to the transposon insertion site with MmeI and then uses Ampure-based size selection to separate the Tn-genomic DNA junction digestion fragment from the genomic DNA background.

I validated this new protocol using mock communities composed of mutant libraries with systematically varied abundances plus wild-type strains of other species. The results revealed that multi-taxon INSeq offers high specificity and precision while simultaneously amplifying multiple mutant libraries of different species with high reproducibility.

The development of this new pipeline led to the challenge of how to determine the significance of changes in the output population relative to the input population under multiple selective pressures (i.e., host immune system, diet). To address this, I applied a maximum-likelihood estimation on the log output to input ratio for each mutant. This analysis defined the population behavior that resembled wild-type as the neutral population, and calculated fitness indices for each mutant relative to the distribution of the reference neutral population after selection. Further details regarding the multi-taxon INSeq methodology are provided in Chapter 3.

**Using gnotobiotic mouse models to study fitness determinants**

Owing to the complexity of the microbiota, microbial-microbial and microbial-host interactions are difficult to study. Gnotobiotic mice representing a given inbred strain offers an opportunity to
do so under highly controlled environmental conditions where microbial community composition and diets are defined/constrained.

In this dissertation, two gnotobiotic mouse models have been used to examine fitness determinants in vivo: (1) mice monoassociated with *Escherichia coli* strains isolated from urine and fecal samples collected from women with recurrent UTI and chosen based on fitness predictions from comparative genomics analysis, and (2) gnotobiotic mice colonized with an artificial 15-member microbial community consisting of 14 bacterial species, with one species represented by two separate strains, all isolated from the human gut. Members of this artificial community were selected because they were phylogenetically diverse, and their sequenced genomes encoded major metabolic functions identified in the human gut microbiota. In total, the community contained 8 *Bacteroidetes* species that represent primary fermenters of dietary and host glycans, six *Firmicutes* spp., including *E. rectale* which is an efficient producer of butyrate and one species known to ferment amino acids (*Clostridium symbiosum*).

In addition to the tools we mentioned above, other methods were used to profile community structure and function: they, included a new tool to quantify the competitive indexes of two members (FitSeq, see Chapter 2), a method to quantify relative abundance of community members based on short read shotgun sequencing of community DNA (COPRO-Seq, Community PROfiling by Sequencing, see Chapter 3), and measurement of community metabolism by gas chromatography-mass spectrometry (GC-MS) and ultra high-performance liquid chromatography-mass spectrometry (UPLC-MS).
References


Chapter 2

Genomic diversity and fitness of *E. coli* strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection
Chapter 2

Genomic diversity and fitness of *E. coli* strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection

Swaine L. Chen*1,2,3+, Meng Wu*1, Jeffrey P. Henderson2,4, Thomas M. Hooton5, Michael E. Hibbing3, Scott J. Hultgren2,3, and Jeffrey I. Gordon1

1Center for Genome Sciences and Systems Biology, 2Center for Women’s Infectious Disease Research, and 3Departments of Molecular Microbiology and 4Medicine, Washington University, St. Louis, MO 63110, 5Department of Medicine, University of Miami Miller School of Medicine, Miami, FL 33136

*Contributed equally

1Current address: Dept. Medicine, Yong Loo Lin School of Medicine, National University of Singapore and Genome Institute of Singapore, Singapore 138672

This chapter corresponds to the complete and accepted version of a manuscript published as

Abstract

Urinary tract infections (UTIs) are common in women and recurrence is a major clinical problem. Most UTIs are caused by uropathogenic *Escherichia coli* (UPEC). UPEC are generally thought to migrate from the gut to the bladder to cause UTI. UPEC strains form specialized intracellular bacterial communities (IBCs) in the bladder urothelium as part of a pathogenic mechanism to establish a foothold during acute stages of infection. Evolutionarily, such a specific adaptation to the bladder environment would be predicted to result in decreased fitness in other habitats, such as the gut. To examine this concept, we characterized 45 *E. coli* strains isolated from the feces and urine of four otherwise healthy women with recurrent UTIs. Multi-locus sequence typing revealed that two of the patients maintained a clonal population in both of these body habitats throughout their recurrent UTIs, while the other two manifested a wholesale shift in the dominant UPEC strain colonizing their urinary tract and gut between UTIs. These results were confirmed when we subjected 26 isolates from two patients, one representing the persistent clonal pattern and the other representing the dynamic population shift, to whole genome sequencing. *In vivo* competition studies conducted in mouse models of bladder and gut colonization, using isolates taken from one of the patients with a wholesale population shift, and a newly developed SNP-based method for quantifying strains, revealed that the strain that dominated in her last UTI episode had increased fitness in both body habitats relative to the one that dominated in the preceding episodes. Furthermore, increased fitness was correlated with differences in the strains’ gene repertoires and their *in vitro* carbohydrate and amino acid utilization profiles. Thus, UPEC appear capable of persisting in both the gut and urinary tract without a fitness tradeoff. Determination of all of the potential reservoirs for UPEC strains that cause recurrent UTI will require additional longitudinal studies of the type described in this report, with sampling of multiple body habitats during and between episodes.
**Introduction**

More than half of all women develop at least one episode of urinary tract infection (UTI) during their lifetimes. Up to 25% of women have recurrent UTI, which is defined as two or more episodes within a 6-month period (1). The majority of community-acquired UTIs are caused by uropathogenic *Escherichia coli* (UPEC) (2). A generally accepted model for infection is that UPEC migrate from the gastrointestinal tract to the periurethral area, and eventually up the urethra into the bladder (3).

The gut and urinary tract are very distinct habitats from the perspective of their metabolic, immunologic, and microbial features. The gut is home to our largest population of microbes (4-6), while the bladder is considered a normally sterile environment, guarded by physical and biological barriers to microbial invasion (7-9). Studies of the molecular pathogenesis of UTI in a mouse model (10-12) have identified numerous virulence factors, including adhesins, toxins, iron acquisition systems, capsular structures, flagellae, pathogenicity islands, and factors important for biofilm formation (13). Among adhesins, UPEC strains typically encode a multitude of chaperone/usher pathway (CUP) pilus gene clusters. CUP pili contain adhesins at their tips that play critical roles in host-pathogen interactions, recognizing specific receptors with stereochemical specificity (14). For example, FimH, the type 1 pilus tip adhesin, binds mannosylated glycoproteins, as well as N-linked oligosaccharides of b1- and a3- integrins that are expressed on the luminal surface of the bladder epithelium (urothelium) in humans and mice (15, 16). Type 1 pilus-mediated binding can lead to invasion of UPEC into mouse and human bladder epithelial cells (17-19). Invading UPEC can be expelled from the host cell (20) or they can ‘escape’ into the cell’s cytoplasm where they replicate rapidly and form a biofilm-like structure, composed of 10⁴-10⁵ organisms, known as an intracellular bacterial community (IBC) (21, 22). Bacteria in the IBC are protected from antibiotics (23, 24) and from immune responses (11, 25). IBCs are transient; after maturation, UPEC can disperse from the IBC, exit their host cells, enter the lumen of the bladder, and subsequently invade other urothelial cells (21). One primary host defense that eliminates IBCs is exfoliation, where urothelial cells undergo an apoptotic-like cell death, detach from the underlying transitional
epithelium, and are eliminated in the urine (25, 26). Exfoliated bladder epithelial cells containing IBCs have been observed in urine collected from women with recurrent UTI but not in healthy controls or in cases of UTI caused by Gram-positive pathogens (26). However, exfoliation exposes underlying cell layers of the urothelium. Subsequent UPEC invasion of these underlying cells in mice results in formation of additional intracellular structures termed quiescent intracellular reservoirs (QIRs) (27, 28). Bacteria in the QIR are dormant, are resistant to antibiotic treatment, and elude recognition by host immune defenses. Mouse models have been used to demonstrate that bacteria in QIRs can contribute to recurrent infection after antibiotic treatment has rendered the urine sterile (23, 28).

Consistent with these adaptations for colonizing the bladder habitat, UPEC have been classified as a subset of extra-intestinal pathogenic \textit{E. coli} strains (ExPEC). ExPEC are distinguished from other gut-associated mutualistic and pathogenic \textit{E. coli} strains based on the spectrum of diseases they cause and their genomic features. ExPEC are typically members of the B2 and D subtypes of \textit{E. coli} and often carry pathogenicity islands (PAIs) encoding virulence-associated genes (29). Mutations in \textit{fimH} are postulated to be an important component in the evolution of UPEC strains, with secondary contributions from mutations affecting other loci in gut-associated \textit{E. coli} populations (30-32). In this conceptualization, mutations conferring a fitness advantage within the urinary tract are selected in this body habitat (33). The ability of multiple UPEC strains to form specialized intracellular structures such as the IBC (34) and QIR suggests a very specific adaptation to the bladder environment. Increased fitness in the urinary tract has been hypothesized to confer decreased fitness in the gut habitat of origin, so that strains successfully colonizing the urinary tract encounter a “dead-end” evolutionary path; this has been cited as an example of “source-sink” evolutionary dynamics (31, 35).

In the present study, we have used comparative genomics, \textit{in vitro} assays of growth in the presence of a broad range of potential nutrients, and \textit{in vivo} fitness tests of representative \textit{E. coli} strains obtained from both the urine and feces of women with recurrent UTIs to address several questions. How dynamic are the \textit{E. coli} populations in the gut and urinary tract of a given individu-
al sampled over a period when recurrent UTIs are experienced? Can genome-wide comparisons of gut and urinary tract isolates provide insights into whether recurrences arise from re-inoculation of gut-derived bacterial strains into the urinary tract or from intracellular reservoirs within the bladder? Are there fitness tradeoffs due to adaptation to either the gut or urinary tract environment? We observed two very different patterns: recurrent UTI caused repeatedly by the same strain, or rapid and apparently complete replacement of one strain with another in both body habitats between UTI episodes (arguing against a fitness tradeoff model). We were able to correlate replacement of one strain by another with their genomic and metabolic features. The significance of these results is discussed in the context of understanding of disease pathogenesis and designing clinical translational studies focused on new approaches for pathogen surveillance and treatment.

Results

*E. coli* strains cluster by host instead of habitat

One hundred fourteen women were enrolled in a now completed study of recurrent UTI (36); each presented with symptoms of acute cystitis. Eight of the 114 individuals had a negative enrollment urine culture, while two were lost to follow-up after the initial visit. Of the 104 remaining participants, nine had three episodes of recurrent UTI (i.e., the greatest number among enrollees). However, fecal and urine samples were collected at each episode in only four of these nine individuals. All four patients received a similar cycle of antimicrobial therapy: trimethoprim-sulfamethoxazole (TMP-SMZ) for their enrollment UTI, nitrofurantoin for their second UTI, and ciprofloxacin for their third UTI (see ref. 36 and Table S1 for further details about clinical characteristics and treatment; note that patient 56 was initially treated with TMP-SMZ during episode 2 but then switched to nitrofurantoin when her urine isolates were found to be TMP-SMZ resistant). Sequencing near full-length amplicons generated from the 16S rRNA genes in the 45 strains recovered from fecal and urine samples collected during the three UTI episodes experienced by each of these four individuals confirmed that all were *E. coli*. These 45 *E.coli* strains (Table S2), which included
one urine isolate and on average 3 fecal isolates from each of the four patients at the time of each episode of UTI, were selected for the current study to address our questions about source-sink dynamics (where do strains arise and how do they distribute themselves among different body habitats) and about the relative fitness of E. coli strains from a patient or patients that experienced a wholesale population shift between episodes.

To investigate the relatedness of the 45 E. coli strains and the relationships between strain characteristics and the body habitat from which they had been recovered, we first conducted multi-locus sequence typing (MLST) using seven well-conserved housekeeping genes: adk (adenylate kinase), fumC (fumerase isozyme C), gyrB (DNA gyrase subunit B), icd (isocitrate hydrogenase), mdh (malate dehydrogenase), purA (adenylosuccinate synthetase), and recA (recombinase A) (see Table S3 for primers used for MSLT). Two patterns emerged from the MLST analysis: (i) a “stable clonal” pattern where isolates from the same patient were nearly indistinguishable at all time points surveyed (Patients 12 and 13); and (ii) a “dynamic” pattern where isolates from the same patient included several different MLST groups during the study (Patients 56 and 72). Even among the two patients with the dynamic pattern, strains isolated at a given clinic visit tended to have identical or very similar MLST profiles, regardless of the body site from which they had been recovered (Fig. 1). In the case of patient 72, sequence typing revealed two MLST groups among her fecal and urine isolates during episode 1. During episode 2, all her fecal and urine isolates had a single MLST group assignment that was the same as one of the groups in episode 1. In episode 3, fecal and urine isolates had a single MLST group but it differed from all the MLST groups of strains recovered during episodes 1 and 2, leading us to conclude that she harbored nearly clonal E. coli populations in both body habitats, but that the population had changed between episodes 2 and 3. In contrast, only one fecal isolate, from the second UTI episode from patient 13, differed in its MLST group from the group assigned to all her other fecal and urine isolates recovered at the time of UTI episodes 1, 2 and 3, leading us to conclude from the MLST analysis that she possessed a largely clonal population across episodes.
Our initial assessment of clonality was based on MLST sequencing of a limited number of strains. While no additional strains were available to extend sampling depth to search for minor *E. coli* populations, we proceeded to collect more data on the available strains to overcome potential limitations of MLST in assessing strain relatedness (37). We chose to examine the three urine and 11 fecal strains isolated from Patient 13 (stable ‘clonal pattern’) and the three urine and nine fecal strains recovered from Patient 72 (‘dynamic pattern’) as two contrasting individual examples of colonization patterns. These isolates represent each of the three episodes of UTI experienced by each of these two women (Table S4). We sequenced PCR amplicons generated from the *fimH* gene. A maximum likelihood *fimH* gene tree is presented in Fig. S1, with the position of *fimH* alleles in the various urine and fecal isolates recovered from Patients 13 and 72 shown. All three of the urine isolates from Patient 13, representing each of her three UTI episodes, had *fimH* sequences that were identical to the *fimH* sequences in all 11 of her fecal isolates. For Patient 72, this sequence identity was true for the majority of isolates from the first two UTI episodes but not for the third UTI where the urine strain had a *fimH* sequence that was different from the strains representing the two previous UTIs. However, the *fimH* sequence of this UTI episode 3-associated urine isolate was identical to the *fimH* sequences in the contemporaneously collected four fecal isolates (Fig. S1). Compared to Patient 72’s first and second episode isolates, there were seven amino acid substitutions in the FimH allele of fecal and urine isolates from her third UTI episode: A10V, N70S, S78N, V119A, T234A, and A273G and N6T. The latter residue is in the signal sequence that is not part of the mature FimH protein assembled on the tips of pili. Positions 10, 70, 78, and 119 are in the lectin domain of the protein, while residues occupying positions 234 and 273 are in the pilin domain; all are solvent exposed with the exception of position 119 in a crystal structure of the FimC-FimH complex (15). None are in the mannose-binding pocket. However, residue 273 is positioned near the hydrophobic groove where donor strand complementation and donor strand exchange occur - processes essential for pilus biogenesis (14).
Unsupervised hierarchical clustering of isolates based on their gene and SNP content

To obtain more definitive data on strain relatedness and clonality, we performed whole genome shotgun sequencing of all six urine and all 20 fecal strains recovered from Patients 13 and 72 (60-140X coverage; see Methods). Using the Velvet (38) and AMOScmp (39) assemblers, 25/26 of the isolates yielded genome assemblies that averaged 4.98 Mbp with an average N50 contig length of 72,747 bp (Table S4).

To identify and compare the gene content of the isolates, we first compiled a database of all annotated genes from the genomes of 54 E. coli strains deposited in the NCBI RefSeq database that were classified as ‘complete’, as well as 324 other draft E. coli genomes present in RefSeq and the PATRIC database (Table S5). These genes, together with the genes predicted by Glimmer3 (41) in the assembled genomes of the newly sequenced strains from this study, were clustered using the program CD-HIT with default parameters (95% similarity) (42) to generate “OGUs” (Operational Gene Units). All raw reads for each of the 26 newly sequenced E. coli genomes were mapped to this E. coli pan-genome using BLAT with default parameters (43). The total number of raw reads mapping to a given OGU was then used as the score for that OGU. A null cutoff score was calculated by dividing the total number of reads by the total length of OGU representatives (as determined by CD-HIT); this cutoff represents the expected number of reads per OGU normalized by length if reads were randomly selected from all OGU representatives. OGUs with scores less than this cutoff were called ‘absent’; those above were called ‘present’ (44). By mapping the raw reads from each isolate and the in silico fragmented sequences from the finished UTI89 genome onto this E. coli pan-genome dataset, we identified a total of 11,151 OGUs that were present in at least one of the 26 clinical isolates from the two patients or in the finished genome of UTI89; 3,488 of the 11,151 OGUs (31.3%) were conserved in all 26 strains.

We next identified a total 295,099 SNPs among the 25 isolates at positions present in the finished UTI89 genome (see Methods, one fecal sample Ec72_E2F2 was excluded from this analysis given the problems encountered assembling its genome; see Table S4). The SNP and OGU data
were then used to more rigorously examine clonality. First, based on the SNP rate (SNPs/aligned bp), we computed a matrix of pairwise distance measurements between isolates from our clinical study of recurrent UTI and the 378 strains we used in the OGU analysis. Unsupervised hierarchical clustering based on these distance measurements (Fig. S2A) showed that the clinical isolates that we deemed clonally related to one another by MLST also clustered together in the SNP-based tree of 403 strains. With the exception of one fecal strain, the SNP rate between any pair of isolates from Patient 13 was in the noise range (< 100 total, SNP rates <0.005/bp) regardless of their date of isolation or whether they were recovered from urine or feces. In the case of Patient 72, we identified distinct groups of strains among the urine and fecal isolates, representing distinct branches on the tree. Fecal and urine strains isolated from her first and second UTI episodes clustered separately from those of the third UTI episode. Second, the OGU-based, unsupervised hierarchical clustering of the 404 strains produced the same patterns for the 26 isolates as those obtained from SNP rates (Fig. S2B). Thus, based on their SNP and OGU content, we considered UTI episodes 1-3 in Patient 13 to have all been caused by the same UPEC strain (which in turn was very similar to the strain recovered from feces and urine during UTI episode 3 in Patient 72).

Finally, we quantified the growth of each isolate from Patient 72 under 190 different culture conditions using Biolog phenotype microarrays (see Methods). Phenotype-based hierarchical clustering of her urine and fecal isolates yielded results that were virtually the same as those obtained from the MLST-, OGU-, and SNP-based comparisons (see Fig. 4A and below for additional information about the relationship between growth properties and gene content).

**Tests of relative fitness in representative urinary isolates from patient 72**

Our SNP- and OGU-based hierarchical clustering segregated the 26 isolates from current study and other isolates, including UPEC and ExPEC strains, into two major clades at the top level of the tree. This first tree division is consistent with previous phylogenetic characterizations of various other *E. coli* strains (e.g., groups A, B1, and D in clade 1 and B2 in clade 2) and we refer to this top tree division using the terms clade 1 and clade 2 for convenience (Figs. 2, 3). Comparison
of the content of known virulence factors in the 26 sequenced isolates from our study and the 54 complete *E. coli* genomes from RefSeq revealed that strains located in clade 1 had significantly fewer virulence genes compared to strains in clade 2 (see Table S6 for p-values, Chi-square tests).

We selected urine isolate Ec72_E1U1 obtained from Patient 72 during her first UTI episode and located in clade 1 as a proxy for fecal and urine samples from her first two UTI episodes, and urine isolate Ec72_E3U1 recovered from her last UTI episode and located in clade 2 as a proxy for all urine and fecal isolates from this episode and a very close strain to all urinary and fecal isolates from all three UTI episodes in Patient 13. (Figs. 2 and 3). Since her urine strain from episode 1 was replaced by the urine strain in episode 3, we asked whether the episode 3 strain had higher relative fitness compared to the episode 1 strain in both the bladder and gut. If this were true, it would provide a counter-example to the notion that there is a fitness tradeoff between the urinary tract and gut, and would suggest that the urinary tract is not necessarily a “sink” or evolutionary “dead-end” habitat for UPEC strains.

To address this question, we first turned to a well-established mouse model of UTI (45). The UTI episode 1 strain was marked with low copy plasmids containing genes conveying resistance to kanamycin (pACYC177; Kan<sup>R</sup>) or chloramphenicol (pACYC184; Chlor<sup>R</sup>). The UTI episode 3 strain was marked with pACYC184 (Chlor<sup>R</sup>). The p15A origin driving replication and segregation of these plasmids confers stable inheritance in *E. coli* and many other *Enterobacteriaceae* (46). *In vitro* competition experiments revealed that these plasmids were indeed stable in both strains, and that there was no growth (fitness) defect between the marked strains and their unmarked counterpart, as judged by quantifying colony forming units (CFUs) over the course of 8 h of growth under shaking conditions or during 48 h of growth under static conditions in LB medium. Furthermore, there was no significant difference in growth of the episode 1 strain (Ec72_E1U1) when it contained pACYC177 (Kan<sup>R</sup>) versus pACYC184 (Chlor<sup>R</sup>).

We subsequently compared the fitness of Ec72_E1U1 with Ec72_E3U1 by co-inoculating a 1:1 mixture of the Ec72_E1U1/pACYC177 (Kan<sup>R</sup>) and Ec72_E3U1/pACYC184 (Chlor<sup>R</sup>)
strains into the bladders of female C3H/HeN mice (1-3×10^7 CFU of each strain; n=5 mice). Mice were sacrificed 24h after inoculation, and the number of bladder CFU of each strain was defined by plating bladder homogenates on selective media. The marked episode 3 strain (Ec72_E3U1/pACYC184) was the only strain detectable in bladders 24h post inoculation, with the exception of one mouse where the episode 1 strain was present at a low level (10^3 CFU; Fig. 5A). In follow-up single strain infections using unmarked strains without any antibiotic resistance plasmids, the episode 1 strain was undetectable in the bladder tissue of mice 24h post-inoculation while the episode 3 strain achieved a median colonization density of 5.2 X10^4 CFU/bladder (range 4.4x10^2 - 1.12x10^8) (Fig. 5A). Confocal microscopy of bladder whole mounts, prepared 6, 12, 16, and 24 h after mice were mono-infected with the same strains containing pANT4 (encoding GFP; 47) revealed that the episode 3 but not the episode 1 strain was able to form small IBCs (Fig. S3), consistent with other reports that intracellular infection contributes to fitness during UTI. Additionally, we analyzed urine samples as well as bladder and kidney homogenates prepared from mice 24h after transurethral inoculation of ten times more CFUs (10^8) of one or the other of these strains (or the reference control UPEC strain UTI89). The results revealed barely detectable levels of the episode 1 strain, Ec72_E1U1, in bladder homogenates, although it was present in kidney and urine. In contrast, the episode 3 strain (and the control UTI89 isolate), was present in all three sample types and at significantly higher levels than the episode 1 strain (p<0.05; 2-way ANOVA and Mann Whitney U test; Fig. S4).

To test the stability of these plasmids during a much longer period of colonization in the gut, we introduced a 1:1 mixture of the episode 1 strain marked with pACYC177 (Kan^R) or pACYC184 (Chlor^R) into adult germ-free male C57Bl/6J mice using a single oral gavage (n=5 mice). Fecal samples were collected daily during the first four days after gavage, and then every 2 days for 2 weeks. Samples were plated on LB agar without and with antibiotics. Total fecal levels of E. coli ranged from 0.9-1.3 x 10^7 CFU/mg (wet weight) throughout the experiment (non-selective medium). However, levels of the Chlor^R marked strain fell throughout the experiment, while levels of the Kan^R marked strain remained constant (Fig. 5B), indicating that in contrast to short term
(24h) colonization of the bladder, pACYC184 (Chlor<sup>R</sup>) conferred a fitness disadvantage, or that this plasmid was being lost from the episode 1 strain during the 2 week period of monitoring gut colonization.

To circumvent the problem of having to mark strains with plasmids, we developed a method we named FitSeq that differentiates sequenced strains based on their SNP content and provides a digital output of their abundance (see *Methods*). To validate FitSeq, we began with an *in silico* simulation using reads from the whole genome sequencing datasets obtained from strains Ec72_E1U1 and Ec72_E3U1. Mixtures of reads were created with the fractional representation of Ec72_E1U1 set at 0.4, 0.5 and 0.6, and the observed ratios of the two strains calculated based on SNP content over a 5 order of magnitude range of input reads (10-1,000,000/strain). Fig. S5A demonstrates that 100,000 reads are more than sufficient to determine the ratio of the two strains. Next, the two strains were each grown in monoculture, genomic DNA was extracted, and the two purified DNAs mixed in a manner such that the fractional representation of strain Ec72_E1U1 was systemically varied from 0 to 1 in 0.05 increments. An Illumina sequencer was used to generate 36 nt reads from these defined mixtures. Employing 500,000 reads per sample, *in silico* simulations of the type described above and direct analysis of the defined mixtures showed excellent correlation between expected and detected representation (R<sup>2</sup>=0.999, Fig. S5B).

With these results in hand, we gavaged germ-free, adult male and female C57BL/6J mice (n=5/group) with a 1:1 mixture of the two urine strains recovered from Patient 72 during her UTI episodes 1 and 3. Fecal samples were collected as described above. FitSeq disclosed that the strain from episode 1 was rapidly outcompeted by the episode 3 strain in the guts of both male and female animals (Fig. 5C), similar to the temporal profile seen in the original human patient.

**Mechanisms that could underlie the increased fitness of episode 3 strain from Patient 72 in both the bladder and gut environment**

Our assembly of Ec72_E1U1 and Ec72_E3U1 indicated that they share 4,714 OGUs and have 1,432 and 1,969 unique OGUs, respectively. To better understand the differences in fitness of
the episode 1 and 3 strains, we undertook a more in depth genomic and phenotypic analysis. We generated a more complete assembly of their genomes after re-sequencing [150 nt x 2 (paired end) Illumina MiSeq reads; 39-42 fold coverage of each genome; N50 contig length, 108,524 bp (Ec72_E1U1) and 126,534 bp (Ec72_E3U1); Table S4]. There was a high correlation between gene coverage with the initial short read assembly and gene coverage with the longer MiSeq reads ($r^2=0.99$). In addition, BLAST searches of the new genome assemblies confirmed the presence or absence of OGUs (as defined from the earlier analysis), at both the nucleotide and predicted protein levels.

The episode 3 strain, Ec72_E3U1, contains the complete fim operon encoding type 1 pili. While the episode 1 strain, Ec72_E1U1, has a full fimH gene, it is missing the other structural genes required for assembling a functional type 1 pilus. Indeed, under laboratory growth conditions, we were unable to induce expression of functional type 1 pili in Ec72_E1U1 as measured by hemagglutination of guinea pig red blood cells. Consequently, this strain was unable to form type 1 pilus-dependent biofilms after growth in LB broth in polyvinylchloride wells. The episode 1 strain was also deficient in forming pellicle biofilms during growth in YESCA broth (note that pellicle biofilm formation is not dependent on type 1 pili; 48). In contrast, the episode 3 strain was similar to the prototypic human cystitis isolate UTI89 in assays for type 1 pilus expression and function (Fig. S6A,B), but unlike UTI89, it was not capable of pellicle biofilm formation (Table S7).

The episode 1 strain (Ec72_E1U1) was significantly depleted in genes involved in flagellar assembly function ($p<0.05$, $\chi^2$ test). Ten core flagellar assembly genes are present in the episode 3 strain and absent in the episode 1 strain: they include genes essential for formation of the MS ring ($fliF$), the C ring ($fliG$, $fliM$, and $fliN$), and the export apparatus ($fliH$, $fliI$, $fliO$, $fliQ$, $fliP$ and $fliR$) (Fig. S7). The lack of these essential components of the basal body would severely impact the ability of the flagellum to rotate, thus affecting motility. Indeed, the episode 1 strain was non-motile, while the episode 3 strain and UTI89 were motile as measured in swimming and swarming assays (Fig. S6C, Table S7). The other four strains recovered from the urine and feces of patient 72 that clustered together with Ec72_E1U1 in the MLST-, OGU- and SNP-based trees (Ec72_E1F1,
Ec72_E2U1, Ec72_E2F1, Ec72_E2F2 in Figs. 1,2,3, and S2) also lacked these ten core flagellar assembly genes (Table S8A).

The episode 3 strain possessed most of the canonical UPEC virulence-associated pathogenicity island (PAI) elements (PAI-II, PAI-III, PAI-IV), 8 chaperone-usher pilus systems, and several additional toxins and iron acquisition systems (α-hemolysin and four major siderophore systems). In contrast, the episode 1 strain was missing most of these PAIs, had only one intact chaperone-usher pilus system, and lacked all of the siderophore systems and toxins associated with UPEC (Table S6 and Fig. S8). In vitro assays confirmed the absence of α-hemolysin activity in this strain (see Ec72_E1U1 in Table S6 and Fig. S8).

The episode 3 strain was also enriched in phosphotransferase systems (PTS) relative to the episode 1 strain (p<0.05, χ² test). The genes encoding PTS-Sor-EIIA, PTS-Sor-EIIB, PTS-Sor-EIIC, PTS-Sor-EIID, which comprise L-sorbose-specific enzymes II in the phosphoenolpyruvate (PEP)-dependent PTS system (Fig. S9), were absent in the less fit episode 1 strain and present in the episode 3 strain. Three components are required in the PTS system: the two common PTS proteins, enzyme I (EI) and HPr, which transfer a phosphoryl group from PEP to the substrate-specific Enzymes II (EII) complex. Enzymes II are involved in the first step of sorbose utilization, transport of L-sorbose into the cell and phosphorylation to L-sorbose-1-phosphate. The absence of this particular sorbose PTS EII suggested that the episode 1 strain cannot use L-sorbose, a fact confirmed in the phenotypic microarray assay (Fig. 4B). L-Sorbose derived from dietary vegetables exists in both the human gut and urinary tract (49). L-Sorbose utilization is a distinctive feature of virulent E. coli, including ETEC, EIEC, STEC, EPEC, and other UPEC strains (50). The lack of the L-sorbose PTS-system was also observed in (i) the genomes of the four other strains judged to be clonal with Ec72_E1U1 based on MLST-, OGU-, SNP analyses (Ec72_E1F1, Ec72_E2U1, Ec72_E2F1, Ec72_E2F2) and (ii) the 13 reference strains in phylogenetic group A that clustered with this clonal population in the tree shown in Figs. 2 and 3 (Table S8B).
The differential representation of other genes in the genomes of these two strains suggest a genetic basis for the observed differences in their *in vitro* growth phenotypes (Fig. 4B) and fitness in the gut and bladder (Fig. 5A,C). For example, genes involved in galactose utilization were more prominent in the episode 3 strain and correlated with its higher growth rates on substrates requiring galactose metabolism (Fig. 4B). Defects in galactose utilization are known to affect colonization of the intestine. In fact, *E. coli* uses multiple sugars for growth in the intestine and multiple mutations affecting different sugar utilization pathways have an additive effect on the colonization levels of the enterohemorrhagic *E. coli* strain EDL933 in CD-1 mice (51). Peptides or amino acids are the primary carbon source for *E. coli* during UTI (52), and peptide transport, gluconeogenesis and the TCA cycle are required for UTI caused by the UPEC strain CFT073 (49,52). The phenotype microarray analysis disclosed that the more fit episode 3 strain had higher growth rates on all four dipeptides and 10 of 22 amino acids tested (Fig. 4B). Nine of these ten amino acids are glucogenic in the TCA cycle. Genes involved in gluconeogenesis and TCA cycle were enriched in the variable component of the episode 3 isolate’s genome, compared to the episode 1 isolate’s genome (Table S9).

**Discussion**

We have conducted a study analyzing the genomic features of *E. coli* strains isolated from the feces and urine of four women during recurrent bouts of UTI and assessed the relative fitness of representative strains in mouse models. We found two very different colonization patterns with respect to the dominant *E. coli* population in the gut and bladder (two patients/pattern). One pattern, exemplified by Patient 13, can have stable and seemingly clonal *E. coli* populations in the gut and urinary tract for several months over the course of multiple recurrent urinary tract infections. In contrast, other patients, illustrated by Patient 72, have a very dynamic pattern with a wholesale shift in the major population colonizing both the intestinal and urinary tract occurring over the 1 month period between her second and third episodes of UTI. Unsupervised hierarchical clustering of both genetic and phenotypic data (MLST, whole genome gene and SNP content, and *in vitro*
growth on 190 substrates) supported the clonal relationship of strains representing these dominant
\textit{E. coli} populations.

We tested the \textit{in vivo} fitness of two isolates: (1) Ec72\_E1U1, a representative of the strains present in Patient 72’s gut and bladder during her first two episodes, and (2) Ec72\_E3U1, a representative of the gut and urine isolates from the last UTI episode of Patient 72. The results revealed that the latter strain had higher fitness in both the mouse bladder and gut, consistent with the population shift documented in both the gut and bladder of patient 72. These fitness differences correlate with a number of genomic and metabolic features that provide insights about the requirements for survival in these body habitats. Nonpathogenic \textit{E. coli} strains generally contain fewer chaperone-usher pilus systems than pathogenic strains (53). We found that the less competitive episode 1 strain carries fewer pilus systems (only one of the 13 known CUP systems in \textit{E. coli}) than the episode 3 strain. Moreover, there were seven predicted amino acid differences between the FimH proteins encoded by the two isolates. FimH residues 70 and 78 define two major groups of FimH sequences (54, 55); the first UTI episode isolate (Ec72\_E1U1) has 70N and 78S, which are associated with fecal and non-UPEC strains, while the third UTI isolate (Ec72\_E3U1) has 70S and 78N, which are associated with UPEC strains.

Flagella are thought to be important for UTI pathogenesis. The flagellum consists of a basal body, hook and filament. Flagellar synthesis is a highly ordered and regulated process involving three classes of genes. Class I genes include \textit{flhDC}, which encode the FlhD/FlhC complex that functions as a transcriptional activator of flagellar class II operons. Class II genes encode the basal body and hook, as well as FliA and FlgM, which are the sigma factor and anti-sigma factor that regulate transcription of class III genes. Class III genes encode the hook-associated proteins and the filament of the flagellum (FliC), as well as proteins necessary for motility (e.g., MotA, MotB) (56). Studies of isogenic wild-type and \textit{DfliC} strains have shown that loss of the flagellar protein FliC results in reduced persistence in the urinary tract of mice, while IBC formation and dispersal are not affected (57). \textit{E. coli} strains are classically grouped by serotyping based on their LPS O-antigens and flagellar H-antigens. Nearly all \textit{E. coli} have an H-typeable flagellar antigen, includ-
ing nonmotile strains. H-typeable but nonmotile strains include sorbitol-fermenting O157 strains isolated from hemolytic uremic syndrome in which there is a 12bp deletion in \textit{flhC} (58). The urine isolate and one fecal isolate from the first two UTI episodes in patient 72, which would be typed as an H30 strain based on sequence identity to \textit{fliC} from strain HW32 (59), have deletions that eliminate a subset of flagellar class II genes. All these strains, which include the episode 1 strain, Ec72_E1U1, have an intact \textit{flhDC} operon. Thus, deletion of genes encoding flagellar structural protein represents another (alternative) route to disruption of \textit{E. coli} motility.

No common set of virulence determinants has been identified that is specific to UPEC strains and absent from \textit{E. coli} strains that have a mutualistic relationship with their host (60, 61). The general lack of known virulence determinants in the episode 1 strain from Patient 72 raises the question of how this strain was able to cause a symptomatic UTI. The episode 1 strain, Ec72_E1U1, and the closely related episode 2 strain, Ec72_E2U1, were isolated from the urine of patient 72. They were cultured from midstream urine samples using previously well-defined and validated protocols (36) that make fecal contamination highly unlikely. In a recent study, 80\% and 40\% of urine isolates collected from women with symptomatic UTI were capable of expressing functional type 1 and P pili, respectively, after \textit{in vitro} growth (60). In addition, studies of isolates from women with asymptomatic bacteriuria have found that they are enriched for UPEC strains that have lost the ability to make functional pili (62). Thus, alternate mechanisms for \textit{E. coli} colonization of the urinary tract exist. The association of UTI symptoms with the episode 1 strain is even more puzzling given its limited number of chaperone-usher pili, its \textit{fimH} sequence, its lack of PAIs, and its clustering with other non-pathogenic strains based on comparison of their sequenced genomes. Sexual intercourse has been shown to introduce bacteria into the female bladder (3). Patient 72 engaged in sexual intercourse just prior to bacteriuria and the development of symptoms (36). Therefore, it seems reasonable to propose that Ec72_E1U1, despite lacking functional P and type 1 pili, was inoculated into her bladder in sufficient quantities to maintain itself for a period of time to cause symptoms. Indeed, we found that an inoculation of $10^8$ CFU of Ec72_E1U1 into the bladders of mice was sufficient to maintain bacteriuria and kidney colonization even in the absence
of significant invasion of the bladder urothelium (*Fig. S4*). Furthermore, studies in mouse models have shown that certain UPEC strains that are defective in IBC formation can be complemented to form IBCs when there is co-infection with other UPEC (34); thus, in the context of a mixed infection, this strain may be able to persist within the bladder. In contrast, Ec72_E3U1 colonized and invaded the bladder. Our finding that the episode 3 strain Ec72_E3U1 has intact coding sequences for functional type 1 pili and flagella and a greater flexibility in utilizing carbon sources available in the gut and urinary tract emphasizes the multigenic underpinnings of virulence, provides mechanistic understanding for its observed displacement of the less fit Ec72_E1U1 between Patient 72’s UTI episodes 1 and 3, and underscores the need to examine the role of metabolic capabilities in determining the fitness of UPEC in two body habitats involved in disease pathogenesis. For example, we have shown that mutants with deletions in *sdhB* (required for conversion of succinate to fumarate in the TCA cycle) or *mdh* (catalyzes metabolism of malate to oxaloacetate; loss of *mdh* blocks the TCA cycle and glyoxylate shunt) are both attenuated in a mouse model of UTI, correlating with a decreased ability to form IBCs (63).

The observation of the same strain of *E. coli* in both urine and fecal isolates during a given UTI episode and across successive UTI episodes is somewhat surprising given the hypothesis that the urinary tract is an evolutionary dead end from which *E. coli* do not emerge to seed other habitats (31). Two possibilities that are not mutually exclusive could explain our results: (i) UPEC are very fit in both the gut and urinary tract, and UTI results from gut to bladder inoculation but the reverse never happens; (ii) UPEC transiently occupy, and in fact dominate, the gut *E. coli* population en route to the urinary tract, and because we have sampled strains only during UTI episodes, we observe the same strain in both the feces and urine. These two possibilities differ in that (i) assumes high UPEC fitness in the gut, while (ii) does not. Given our gut colonization data in gnotobiotic mice, it seems that (i) is the more likely of these two possibilities. If the urinary tract is not an evolutionary dead end, then a third possibility needs to be considered: the urinary tract may be a source for gut colonization with UPEC. The third possibility is consistent with a scenario where there is dynamic fluxing between both body habitats so that a strain originating from the gut causes
a UTI episode and UPEC from the bladder/urine subsequently re-seed the gut; this ‘cycle’ could lead to the ‘homogeneity’ that we see in our study where the one strain dominates in both habitats during a given UTI.

Overall, this is a complex issue, as gut colonization by UPEC can be consistent with the urinary tract as a dead-end; i.e., if UTI always arises from infection with gut bacteria, then human-human transmission and epidemics of UTI could be caused merely by fecal-oral transmission of bacteria from gut to gut without requiring intervening occupancy of the urinary tract. Two observations argue against this notion. First, in the case of Patient 72, we find that the relative fitness of gut and urinary isolates in mice follows their dynamics in humans, and that isolates with higher relative fitness co-exist simultaneously in both host habitats. Second, in an analysis of heterosexual couples, colonization of the gut of both partners by the same strain of *E. coli* was associated with cunnilingus (64). One explanation for this latter result involves transmission from the urinary tract of the female to the gut of the male. Foodborne transmission of extra-intestinal *E. coli* may also represent a possible route of dissemination (65, 66).

The concept in evolutionary theory that specialization to one environment is generally detrimental to fitness in another has been extensively explored in the context of microbial survival in the presence and absence of antibiotics. Fitness costs due to adaptation to antibiotic pressure are seen. However, compensatory mutations that restore fitness and no-cost adaptive mutations have been identified in numerous systems (67). Interestingly, in *C. jejuni*, fitness tradeoffs are very clear in the development of resistance to macrolide antibiotics, yet evolution of resistance to fluoroquinolones confers equal or higher overall fitness in the absence of antibiotic pressure in animal models of infection, demonstrating that fitness landscapes can be dynamic and complex (68, 69). Therefore, while some UPEC strains may suffer in their gut fitness, our data from Patient 72 indicate that a pathway to high fitness in both urinary tract and gut exists. This is interesting in light of the highly specialized intracellular infection pathway used by many UPEC in the bladder, the efficiency of which varies from strain to strain (21, 22, 34). IBC formation is multifactorial, and there is evidence that this is a selected process among UPEC strains (33). Type 1 pili mediate
binding to uroplakins (16, 25) but play an additional intracellular role in IBC formation (33, 70) and thus are key features of both UTI and IBC formation. Selection for this specialization towards uroepithelial cells is expected to decrease fitness in other habitats such as the gut, especially in a source-sink model where the urinary tract is an evolutionary dead end. We have now demonstrated that fitness in the urinary tract and gut are not necessarily inversely related (e.g., the episode 3 but not the episode 1 strain from patient 72 forms IBCs). It may be useful to consider how selection for the biofilm-like IBC in uroepithelial cells could enhance survival in the gastrointestinal tract, which is lined with a mucus layer containing polysaccharides that can serve as a nutrient repository as well as a place for attachment and establishment of syntrophic relationships with other members of the microbiota. In other words, the IBC may be a focal point for patho-adaptive changes that also increase fitness of strains in the gut.

Our findings provide a rationale for additional studies of populations of women representing different ages, genotypes, and lifestyles (including different diets and nutritional status) in order to address the question of the origins of recurrent UTI and the relative importance of dynamic and stable patterns of colonization. If UPEC are able to move freely between the gut and urinary tract, this complicates our inference of the ultimate reservoir for recurrent UTI. Therefore, further studies of the type described in this report are needed to determine the migration directions of these bacteria between different sites within an individual and between individuals. We envision these studies as a part of a translational medicine pipeline directed at developing more informed concepts about the pathogenesis of recurrent UTI, as well as more effective therapies. For example, whole genome sequencing of isolates obtained from time series studies of patients with recurrent UTI should help determine whether the majority of *E. coli* isolates obtained from the gut and urinary tract of an individual at a given time point are clonally related, and whether or not there are barriers to homogenization across different body habitats (oral, fecal/perianal, vaginal, periurethral and bladder). FitSeq and the type of animal models employed here can then be used to compare the fitness of UPEC strains that sweep to dominance in the gut and urinary tract of a human host in the setting of UTI. The results may influence standard of care: e.g., whether long-term surveil-
lance of the fecal microbiota coupled with sampling microbial communities from other body habitats can identify population shifts in patients with histories of UTI prior to the onset of UTI. This surveillance may also help define new approaches to chemoprophylaxis, either involving existing antibiotics or next generation compounds that target UPEC through novel mechanisms, such as mannosides that impede or block FimH-mediated binding of UPEC strains to mannosylated epithelial surface receptors (71).

**Materials and Methods**

**Strains.** The TOP trial was conducted using protocols approved by the Human Studies Committee of the University of Washington. Exclusion criteria included known anatomic or functional abnormalities of the urinary tract, symptoms or signs of acute pyelonephritis, chronic illness requiring medical supervision, pregnancy or planned pregnancy in the three month period following enrollment. Clinical information and strains were collected during the trial as described in Czaja et al (36).

**Amplicon sequencing.** Genomic DNA was prepared from each *E. coli* isolate using the Promega Wizard Genomic DNA kit (Promega, Madison, WI) according to the manufacturer’s instructions. Near full length amplicons from the 16S rRNA gene were generated by PCR and sequenced using the dideoxy chain termination method as described (72). Gene targets for MLST, including the *fimH* allele, were amplified with the primers listed in Table S3 in the following reaction mixture: 1x PCR buffer (Invitrogen, Grand Island, NY) supplemented with 2.5mM MgCl₂, 1.4M betaine, 1.3% DMSO, and 200μM dNTPs, 5ng of template, 12.5pmol of each primer, and 1 unit of Taq polymerase (Invitrogen, Grand Island, NY) in 50 μL total volume. Reactions were heated to 95°C in a thermocycler for 5 min, then cycled 35 times using the following conditions: 95°C for 1 min, 55°C for 1 min, and 72°C for 1-3 min (depending on the expected product size). Reactions were finished with a 10 min incubation at 72°C, cooled to room temperature, and subsequently purified using the Qiagen QiaQuick PCR purification kit according to manufacturer’s instructions. Amplicons were sequenced using standard dye-terminator capillary sequencing. Base
calling and assembly of multiple reads were performed using the programs Phred, Phrap, and Consed with default parameters.

**Biolog phenotype microarrays.** Phenotype microarrays were run according to the manufacturer’s standard protocols. Briefly, a strain was grown on solid LB agar with no antibiotic selection at 37°C overnight. Cells were scraped from this plate, resuspended in 10 mL of buffer IF 0a GN/GP Base IF (Biolog), normalized to a transmittance of 85% using the Biolog Turbidimeter, diluted 100-fold in buffer IF 0a GN/GP Base IF with Biolog Dye mix, and then pipetted into Biolog PM1 or PM2 plates. Plates were subsequently incubated at 37°C in the Biolog machine for 48 h with colorimetric measurements made every 30 min. Data were exported from the Biolog software as total AUC (area under the curve) for the 48 h assay, giving 96 AUC values for each PM plate. These values were subjected to unsupervised hierarchical clustering to determine the relatedness of the phenotype microarray profiles for each strain.

**Whole genome sequencing and assembly.** Genomic DNA libraries for the Illumina GA-II sequencer were prepared according to the manufacturer’s protocol. Each strain was sequenced (36 nt reads) using 2 lanes of the 8-lane flow cell. The output files were converted to FASTA format, ignoring quality scores, and assembled with the Velvet short read assembler (version 0.7) using custom Perl scripts to optimize k-mer length and minimum coverage parameters for both N50 length and total assembly length. AMOScmp (version 2.0.5) was used to further improve the assembly with default parameters.

Paired-end libraries with 500 bp inserts were prepared for strain Ec72_E1U1 and Ec72_E3U1 as described by the manufacturer of the Illumina MiSeq instrument. Sample-specific, 8 nt Hamming barcodes [73] were incorporated into the sequencing adapter for multiplex sequencing [8 barcodes per strain; two strains sequenced in a single MiSeq flowcell; 16 barcodes per MiSeq flow cell to get a balanced base composition during the first four cycles of the sequencing run (critical for accurate cluster calling as well as good phase and pre-phasing values)]. The output FASTQ files were assigned to each strain using the barcodes (0.68 million reads for Ec72_E1U1 and 0.78
million reads for Ec72_E3U1), and then assembled with MIRA (version 3.4.0) using default parameters (74).

**SNP analysis.** Analysis of assembled genomes for SNPs was done as described (44). UTI89 was used as the reference finished UPEC genome. All genomes were aligned against UTI89 using BLASTN with default parameters. Only the alignment with highest p-value reported by BLASTN was used for each assembled contig. The position of each SNP, based on UTI89 genome sequence, was recorded. SNP rates were calculated for each pair of assembled genomes by counting the number of sequence differences only at positions in the UTI89 reference genome where both genomes had contigs that aligned. The total number of sequence differences in overlapping regions of the genomes was divided by the total length of overlapping regions, yielding SNP rate per aligned base pair.

**Competitive tests of fitness.** All experiments involving mice were performed using protocols approved by the Washington University Animal Studies Committee. *In vivo* fitness tests in the urinary tract were performed as previously described (45). Briefly, bacteria were grown under type 1 pili-inducing conditions (two passages at 37°C in LB broth without shaking for 16-20 h, with a 1:1000 dilution between passages). These static cultures were briefly agitated to resuspend settled bacteria. Cells were collected by centrifugation (~3,000 x g for 10 min at 4°C) and resuspended in PBS to an OD_{600} of 1.0. An equal mixture of two strains was inoculated transurethrally into the bladders of 7-8 week old female C3H/HeN mice. Twenty four hours later, mice were sacrificed and their bladders were removed aseptically, placed in 1 mL of PBS, mechanically homogenized with a stainless steel electric tissue homogenizer (PRO Scientific, Oxford, CT) for 15-20 sec, and plated on LB/agar plates with and without antibiotics [kanamycin (50 mg/mL), chloramphenicol (100 mg/mL), or kanamycin plus chloramphenicol]. CFU counts were defined after a 12-18 h incubation at 37°C under aerobic conditions. Single infections were performed similarly, except that bacterial suspensions were mixed 1:1 with PBS, to achieve a final OD_{600} of 0.5 before inoculation.
**FitSeq.** To assess their relative fitness in the gut, *E. coli* strains Ec72_E1U1 and Ec72_E3U1 were grown under type 1 pili-inducing conditions. The two strains were mixed in equal proportions (2-3×10⁸ CFU/200μL/strain) and inoculated by oral gavage into germ-free 8–10-week-old C57BL/6J mice. Mice were maintained in plastic flexible film gnotobiotic isolators and fed a standard autoclaved chow diet (B&K Universal, East Yorkshire, U.K; diet 7378000) *ad libitum.* Fecal samples were collected from each mouse 1, 2, 3, 4, 6, 8, 10, 12, and 14 days after gavage. Each fecal pellet was placed in 1 mL of PBS and homogenized by vortexing. A 10 μL aliquot of the homogenate was plated directly on LB agar (n=4 plates/sample). Twenty four hours later, colonies were collected by scraping (1 mL PBS). Genomic DNA was isolated using the phenol:chloroform method (72), and then fragmented by sonication to 300~500 bp (Bioruptor sonicator for ultrasonic liquid processing; 20 cycles of 30 sec ON at high power/30 sec OFF).

Illumina sequencing libraries were prepared as described by the manufacturer. Sample-specific, 8 nt Hamming barcodes (73) were incorporated into the sequencing adapter for multiplex sequencing (n=96 samples/lane of an Illumina HiSeq 2000 flowcell; 80-100 million reads/lane; >500,000 42 nt reads/sample).

Raw-reads were assigned to each sample using the barcodes, and then mapped to the database containing the genomes of both strains using Eland (75). Only reads that could be uniquely assigned to one of the two genomes were used to score the relative representation of that strain in a given fecal sample. Counts for each strain were normalized by the informative genome size (IGS). The IGS of each strain was calculated by generating, *in silico,* a mock sample containing a 1:1 ratio of both strains, and then mapping the sample reads back to each genome: the IGS for each strain was calculated from the number of reads that mapped uniquely to that strain’s genome. The ratio of the two strains was calculated as ρ(FitSeq) after normalization by IGS.

Since bacterial cells had been harvested from LB agar plates after a 24 h incubation, to calculate the ratio (ρ) of strains in fecal samples, we needed to transform ρ(FitSeq) using the relative growth rate. To measure the relative growth rate, a 1:1 mixture of strain Ec72_E1U1 and
Ec72_E3U1 were grown on LB agar for 24 h at room temperature as above. Bacterial cells were collected and the detected ratio ($p(1:1)$) calculated by FitSeq. Knowing that the original ratio of the two strains was 1, the relative growth rate ratio ($u_1/u_2$) of the two strains could be calculated as $p$. Five independent mixtures were prepared and measured, and the average results used as the ratio for growth rate of the two strains on LB agar. Thus, we were able to calculate the ratios of two strains in fecal samples as $p$.

**Characterization of virulence related phenotypes in vitro.** Hemagglutination assays were performed on cells normalized to an OD600 of 1.0 as described previously (n=3 biological replicates performed on different days; 2 technical replicates/biological replicate) (76). For biofilm assays, shaken cultures were grown overnight at 37°C in LB were subcultured at a 1:1000 dilution into LB and grown statically in untreated polyvinylchloride 96 well plates at room temperature for 48 h. Adherent biomass was stained with 0.05% crystal violet, rinsed, solubilized with 35% acetic acid and quantified by measuring absorbance at 595 nm of the solubilized crystal violet (assays performed in duplicate; 5 technical replicates/biological replicate). Pellicle biofilms were performed by subculturing a 1:1000 dilution of an overnight LB shaken culture into Yeast Extract/Casamino Acid (YESCA) medium and incubating statically at 30°C for 72 h (pellicle assays performed twice). For motility assays, cultures were incubated statically at 37°C for 24 h in 10 mL of LB, subcultured at a 1:1000 dilution into 10 mL of fresh LB, and incubated again statically at 37°C for 24 h. Swimming motility was measured in 0.25% LB agar and swarming motility was measured in 0.6% LB agar supplemented with 0.5% glucose (swimming and swarming data shown are representative of results obtained from two separate experiments) (76). Hemolysin production was examined by plating on blood agar (assays performed in duplicate).
References


60. B. S. Norinder, B. Koves, M. Yadav, A. Brauner, C. Svanborg, Do Escherichia coli strains causing acute cystitis have a distinct virulence repertoire?, *Microb Pathogenesis* 52, 10 (2012).


Acknowledgements

We thank Jessica Hoisington-Lopez for assistance with DNA sequencing, David O’Donnell, Maria Karlsson for their assistance with mouse husbandry, plus Andrew Goodman, Andrew Kau, Mark Gonzalez, and Jeremiah Faith for their invaluable comments and help during the course of this work.

Funding: Supported by a grant from the NIH (DK064540). Data generated for Fig. S2 used a computer cluster supported by the National Research Foundation Singapore under its NRF Fellowship (NRF-RF2010-10) and the Genome Institute of Singapore (GIS)/Agency for Science, Technology and Research (A*STAR) (to SLC).

Author contributions: S.L.C., M.W. and J.I.G designed experiments; S.L.C, M.W. M.E.H. performed all experiments, whereas J.P.H. and T.M.H. conducted collection of human biospecimens; S.L.C and M.W. analyzed the data; S.L.C, M.W. S.J.H. and J.I.G wrote the paper.

Data and materials: Genome sequences from UPEC strains have been deposited in GenBank with BioProject Accession number PRJNA187034; new MLST sequences have been deposited in http://mlst.ucc.ie/mlst/dbs/Ecoli as ST2838-ST2846.
Figure Legends

**Fig. 1. MLST analysis of patient isolates.** An eBURST diagram constructed based on publicly available *E. coli* MLST sequence types. MLST types from patient isolates from the present study are represented by colored circles; the color denotes the patient while the size (diameter) of the circle indicates the proportion of strains that share that same MLST type or an MLST type that differs by only one allele.

**Fig. 2. Heatmap of SNP differences between *E. coli* strains.** SNP rates (SNPs/aligned bp) between different sequenced strains were interpreted as a distance matrix. Hierarchical clustering was done on this symmetric matrix of SNP rates. Color in the central heatmap represents SNP rate as shown in the legend at the top left. A tree based on Euclidean distance is shown to left of the central heatmap. The colors of the strains indicate their patient of origin. UTI89_finished refers to the finished UTI89 genome (40), while UTI89_illumina is the re-sequenced and re-assembled genome from a 36 nt read dataset generated for this study. The comparison between the UTI89_finished and UTI89_illumina identified the noise range of the sequencing and assembly pipeline we used.

**Fig. 3. OGU-based clustering of sequenced *E. coli* genomes.** The presence or absence of OGUs in the newly sequenced UPEC genomes characterized in the present study and 54 other publicly available *E. coli* strains were used to construct an OGU matrix. Hierarchical clustering was performed on the matrix based on Euclidean distance. There are two clades suggested by the clustering analysis. Phylogenetic group membership of the strains is indicated with bars at the right of the figure. All isolates from Patient 13 are colored red while all isolates from Patient 72 are colored green. The two strains used for *in vivo* competition experiments and phenotype microarray analyses are highlighted.

**Fig. 4. Growth phenotypes of isolates from Patient 72.** (A) Unsupervised hierarchical clustering of growth phenotypes as defined by Biolog phenotype microarrays. Squares represent fecal samples, circles denote urine samples, and the numbers inside squares and circles indicate the episode number from which the strain was derived. (B) Comparison of the growth phenotypes
of strains Ec72_E1U1 and Ec72_E3U1 from UTI episodes 1 and 3, on carbohydrate, nucleoside and amino acid substrates. Carbohydrate substrates that result in greater than 10-fold growth of the episode 3 strain are shown. See **Fig. S10** for other carbohydrates examined. Utilization of the bracketed carbohydrates requires genes involved in galactose metabolism. The color key denotes growth relative to the reference UPEC strain, UTI89.

**Fig. 5. In vivo fitness of urine isolates from Patient 72’s UTI episodes 1 and 3.** (A) Urine isolates obtained from UTI episodes 1 (isolate Ec72_E1U1) and 3 (Ec72_E3U1) were introduced separately or together into the bladders of female C3H/HeN mice. Data points represent CFUs/bladder in individual mice, and horizontal bars represent the median of CFUs/bladder of that strain in the experiments. Data on the y-axis are presented in log scale: therefore all 0s were plotted as 1 for visualization. (B) The plasmid pACYC184 (ChlorR) was lost during a two-week colonization of the intestines of gnotobiotic mice or conferred a fitness disadvantage. CFU on LB agar plates without antibiotics represent all Ec72_E1U1 in the fecal samples. CFU on LB agar plates with kanamycin or chloramphenicol represent the Ec72_E1U1 carrying the corresponding antibiotic-resistance plasmid from the same fecal sample (data shown as mean +/- SEM). (C) FitSeq determination of the relative fitness of strains from Patient 72’s UTI episodes 1 and 3 in the gut of gnotobiotic mice (data shown as mean +/- SEM).
Figures

Fig. 1
Fig. 2
Fig. 3

Clade 1

Clade 2

Euclidean distance

A

B1

D

B1

B2
Fig. 4

A

B

nucleosides
intermediate amino acids
dipeptides
amino acids

Thymidine
Uridine
Adenosine
Inosine
2′-Deoxyadenosine
Tyramine
β-Phenylethylamine
N-Acetyl-L-Glutamic Acid
L-Homoserine
Hydroxy-L-Proline
L-Ornithine
Putrescine
L-Pyroglutamic Acid
Gly-Asp
Gly-Glu
Ala-Gly
Gly-Pro
D-Serine
L-Threonine
L-Glutamine
L-Aspartic Acid
L-Glutamic Acid
L-Asparagine
L-Proline
L-Leucine
D-Aspartic Acid
α-HG-γ-Lactone
D-Threonine
L-Arginine
Glycine
L-Histidine
L-Isoleucine
L-Lysine
L-Methionine
L-Phenylalanine
L-Valine
D-Alanine
L-Alanine

Control

Background
2′-Deoxy-D-Ribose
D-Saccharic Acid
D-Gluconic Acid
D-Glucose-6-Phosphate
D-Tartrate
β-Methyl-D-Galactoside
Mucic Acid
D-Galacturonic Acid
D-Galactose
D-Galactonic Acid-γ-Lactone
Dulcitol
α-D-Lactose
D-Sorbitol
D-Lactic acid
D,L-α-Glycerol Phosphate
D-Glucuronic Acid
Pyrvic Acid
D-Glucose-1-Phosphate
D-Xylose
Succinic Acid
Methylpyruvate
β-Keto-D-Gluconic Acid
Glucylic Acid
D,L-Malic Acid
Maltooltriose
L-Arabinose
Bromosuccinic Acid
β-Methyl-D-Glucuronic Acid
β-D-Allose
Maltose
Citric Acid
D-Mannose
N-Acetyl-D-Glucosamine
Citricronic Acid
Malonic Acid
Fig. 5

A

CFUs/organ

Ec72_E1U1 Ec72_E3U1

* Co-inoculation

Ec72_E1U1 Ec72_E3U1

B

CFUs/mg feces

Total bacteria  pACYC177 (Kan')  pACYC184 (Chlor')

Days after inoculation

C

%representation

Ec72_E1U1 (Female Mice)  Ec72_E1U1 (Male Mice)  Ec72_E3U1 (Female Mice)  Ec72_E3U1 (Male Mice)

Days after inoculation
Supplemental Materials

Supplemental Figure Legends

Fig. S1. Maximum likelihood fimH gene tree incorporating the UPEC strains characterized in the present study. fimH sequences analyzed in (33) are included to provide background fimH variability within E. coli. Squares indicate fecal isolates while circles are urine isolates. The number inside each symbol indicates the UTI episode from which the strain was isolated.

Fig. S2. Hierarchical clustering of 404 E. coli strains including the 26 characterized in the present study. (A) SNP-based unsupervised hierarchical clustering of the clinical isolates from the present study with 378 strains with ‘finished’, ‘complete’, or draft genomes available in the NCBI RefSeq and PATRIC databases. (B) OGU-based unsupervised clustering of 26 clinical isolates with 378 strains with finished or draft genomes available in public database. All isolates from Patient 13 are colored red while all isolates from Patient 72 are colored green.

Fig. S3. The episode 3 UTI strain Ec72_E3U1 forms IBCs. Confocal images of IBCs found in infections of the mouse bladder using the reference UPEC strain, UTI89 (A,B) and the episode 3 isolate from patient 72, Ec72_E3U1 (C,D). Bacterial strains express GFP and appear green. Epithelial cells are stained with wheat germ agglutinin (red) to show uroplakin expression. Images are taken 6 h post-infection (A,C) or 12 h post-infection (B,D). Scale bar in (A), 10mm (same in all panels).

Fig. S4. Assays of urinary tract colonization with episode 1 and 3 strains from patient 72. C3H/HeN mice were inoculated transurethrally with 1-5x10^8 CFUs of (A) UTI89, (B) Ec72_E1U1, or (C) Ec72_E3U1. 24 h after infection, urine samples were collected and mice were sacrificed. Bladders and kidney pairs were homogenized in PBS to a final volume of 1 ml and bacterial titers defined in urine and tissue homogenates (n=10 mice for each strain from patient 72, n=5 animals for the reference UTI89 strain; lines connect data points generated from a given animal).
Fig. S5. Parameter testing and validation of FitSeq. (A) *In silico* simulation of the accuracy of FitSeq at different levels of sequencing coverage. Three different defined fractional representations of Ec72_E1U1 in a mixture of Ec72_E1U1 and Ec72_E3U1 were used to test the coverage requirement. The x-axis is the total number of reads while the y-axis plots the detected representation as calculated by FitSeq. (B) Validation of FitSeq by *in silico* simulation sequencing of defined mixtures. The expected representation and the detected representation are highly correlated in both the *in silico* simulation and the defined mixtures at a coverage of 500,000 reads/sample ($R^2=0.999$).

Fig. S6. *In vitro* characterization of Ec72_U1E1 and Ec72_U3E1. (A) Biofilm formation assayed after a 48 h static incubation, at room temperature in LB media, in 96-well polyvinyl chloride assay plates. Adherent biomass was defined by staining the material in each well with crystal violet, solubilizing the crystal violet with acetic acid, and recording $A_{595}$ values for each well (data normalized to UTI89). (B) Guinea pig red blood cell hemagglutination titers, with and without mannose (determined with cells normalized to OD$_{600}$ = 1.0, as described previously (76)); (C) Swimming motility. Cultures of each strain were incubated statically at 37°C for 24 h, subcultured and incubated again under the same conditions and inoculated into 0.25% LB agar. Swim diameter was assessed 6.5 h after inoculation. Mean values ± SD are plotted.

Fig. S7. Genes encoding flagellar proteins that are present/absent in the genomes of UTI episode 1 (Ec72_E1U1) and episode 3 (Ec72_E3U1) strains. Figure adapted from the KEGG database (version 58). A red ‘X’ indicates genes that are absent in the Ec72_E1U1 but present in Ec72_E3U1.

Fig. S8. UPEC virulence-associated elements present/absent in the genomes of UTI episode 1 (Ec72_E1U1) and episode 3 (Ec72_E3U1) strains recovered from patient 72 compared to UPEC strain UTI89. Each gene is represented by a square and is colored red if present, and black if absent, as determined by the OGU-based pan-genome analysis.
Fig. S9. PTS pathway components involved in L-sorbose utilization. The components of L-sorbose-specific enzymes II (EII-A, EII-B, EII-C and EII-D) are absent in the Ec72_E1U1 strain but present in the Ec72_E3U1 strain.

Fig. S10. Comparison of growth phenotypes of Ec72_E1U1 and Ec72_E3U1 strains from UTI episodes 1 and 3. The color key indicates the level of growth during a 48 h inc
Supplemental Figures

Fig. S1.
Fig. S2A.
Fig. S2B.
Fig. S4.
Fig. S5.

A

B

Detected representation of Ec72_E1U1 (%)

Number of Reads

Detected representation of Ec72_E1U1 (%)

Expected representation of Ec72_E1U1 (%)

R² = 0.999

in silico simulation

defined mixture
Fig. S6.

A

A595 Relative to UT89

UT89   Ec72E1U1   Ec72E3U1

B

HA Titer

UT89   Ec72E1U1   Ec72E3U1

C

Swim Diameter (cm)

UT89   Ec72E1U1   Ec72E3U1
Fig. S7.
Fig. S9.

- Inner membrane
- Periplasmic space
- Cytosol

Sorbose → EII-C → EII-A → EII-B → Sorbose 1-phosphate → Phosphoenolpyruvate (PEP) → Pyruvate → Metabolism
Fig. S10.
Supplemental Table Legends

Table S1. Clinical characteristics and treatment of the four patients with three episodes of recurrent UTI.

Table S2. Summary of 45 isolates analyzed for this study, including those subjected to whole genome sequencing.

Table S3. Genes targeted for MLST and primers employed for generating PCR amplicons for DNA sequencing.

Table S4. Genome sequencing and assembly metrics for urine and fecal isolates obtained from Patients 13 and 73 during their three episodes of UTI.

Table S5. Reference genomes used for OGU- and SNP-based analyses.

Table S6. Representation of known virulence factors in the genomes of the 26 isolates from the present study and in 54 other E. coli genomes classified as ‘complete’ in the NCBI RefSeq database. (A) Distribution of virulence factors in each strain. “Reference genome” refers to the sequenced genome whose annotation was used to assign “Gene name”. (B) The number of virulence factors per virulence system (e.g., adherence, iron uptake) represented in each of the 80 genomes.

Table S7. Comparison of YESCA pellicle, Congo Red staining, and swarming phenotypes of strains EC72_E1U1, EC72_E3U1 and UTI89.

Table S8. Representation of genes involved in flagellar assembly and PTS-sorbose systems in the 26 clinical isolates from the present study and in the 54 reference complete E. coli genomes. (A) Flagellar assembly system. (B) Phosphotransferase system (PTS) for L-sorbose.

Table S9. Representation genes (OGUs) assigned to KEGG pathways in the shared and variable components of the EC72_E1U1 and EC72_E3U1 genomes.
Chapter 3

Fitness and diet responsiveness simultaneously defined in multiple *Bacteroides* strains *in vivo*
Chapter 3

Fitness and diet responsiveness simultaneously defined in multiple Bacteroides strains in vivo

Meng Wu, Nathan P. McNulty, Nicholas W. Griffin, Jiye Cheng, and Jeffrey I. Gordon

Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108
Abstract

Defining the mechanisms critical for maintaining the highly adaptive and durable relationship between the human gut microbiota and its hosts is necessary if we are to understand how this community contributes to our health status, and how it can be manipulated to promote wellness. To help address this challenge, we generated libraries of tens of thousands of transposon (Tn) mutants for each of four human gut Bacteroides strains, two of which represented the same species. These libraries were introduced into singly-caged adult germ-free mice as part of a 15-member artificial human gut microbiota containing 11 other wild-type bacterial species. Mice received one of two distinct diets, or an ordered sequence of both. Multi-taxon INsertion Sequencing (INSeq) provided a digital readout of the remarkably consistent pattern of community assembly, identified shared as well as species-, strain-, and diet-specific fitness determinants associated with a variety of metabolic/nutrient processing pathways, enabled quantitative gene-level measurement of the resiliency of the responses to diet perturbations, revealed that arabinoxylan, the most common hemicellulose in cereals, could be used to deliberately manipulate the representation of a prominent community member, Bacteroides cellulosilyticus WH2, and defined the niche adjustments of this and the other Bacteroides to arabinoxylan supplementation of a high fat diet. The approach described for mapping fitness landscapes in a community context should facilitate discovery efforts aimed at identifying the niches of microbiota members, as well as ways to deliberately reshape community structure and function through dietary interventions.
**Introduction**

A species’ niche is the description of all of the environmental conditions required to permit a population of that species to persist, including the effects of the population on those conditions (1). This definition includes the species’ resource requirements, stress tolerances, and interactions with other species that act as competitors, predators, parasites, and mutualists. Relatively little is known about the genetic factors that allow members of the human gut microbiota to occupy their niches. Identification of these factors is important for understanding the mechanisms that determine community assembly, community responses to and recovery after various perturbations, and the food webs that link microbes to one another and to their host. Discovery of these factors should spawn new approaches for intentional manipulation of the functional properties of the microbiota.

The human gut microbiota is notable for its high level of strain diversity (2-4). A current view is that strains acquired by an individual early in life persist for decades and that strains are shared among family members (4). Gut bacterial species in an individual are typically represented by multiple strains, each with somewhat divergent gene content; this feature likely contributes to the ability of the microbiota to adapt to changing conditions, but prompts the question of the degree to which given sets of strains share or compete for niche space in the gut ecosystem. Given the large effort being mobilized to identify naturally occurring human gut bacteria that can be used as next generation probiotics, answers to this question have therapeutic implications.

Deleterious mutations in genes that affect or effect interactions between an organism and its environment provide insights into the key factors that define that organism’s niches. In this report we describe an approach for simultaneously identifying genetic determinants of fitness for multiple members of a defined artificial human gut microbiota installed in gnotobiotic mice fed distinct diets monotonously or in an ordered sequence. We focus on four sequenced human gut *Bacteroides* strains: *B. cellulosilyticus* WH2, *B. ovatus* ATCC 8483, *B. thetaiotaomicron* VPI-5482 and *B. thetaiotaomicron* 7330. Culture-independent surveys of a number of human populations indicate that all three of these *Bacteroides* species are prominently represented in the guts of healthy
individuals (3, 5). All four strains contain numerous genes involved in the recognition and processing of otherwise indigestible dietary polysaccharides (6-9; Table S1). We apply genome-wide transposon mutagenesis to these three prominent human gut-derived Bacteroides species, and the two strain representatives of one of them, and colonize singly-housed, germ-free mice with all four mutant libraries together with a defined consortium of 11 other wild-type human gut bacterial species representing major phylogenetic lineages present in the microbiota. This approach allowed us to examine assembly of the 15-member artificial community, and to characterize responses to dietary perturbation, and recovery after these perturbations (resilience) at the community level, the individual species/strain level, as well as the gene level. We identify shared as well as species- and strain-specific genetic and metabolic features that impact fitness of Bacteroides in the gut environment.

**Results**

**Characterizing multiple transposon mutant libraries simultaneously in vivo (multi-taxon INSeq)**

Each library we generated from the four bacterial strains was composed of 87,000-170,000 isogenic transposon (Tn) mutants. Each mutant contained a single site of insertion of the Tn element. 75.5-91.4% of the open reading frames (ORFs) in these four Bacteroides contained a Tn insertion, allowing us to conclude that disruption of these genes did not preclude growth in the rich medium used to construct the libraries. Each ORF was covered by an average of 20 to 39 Tn insertions (Table S2; Figure S1A-D). These Tn mutant libraries were introduced into germ-free mice together with 11 wild-type species that are common constituents of the adult gut microbiota (Figure 1A). To permit simultaneous analysis of multiple mutant libraries in the same recipient gnotobiotic mouse, a mariner transposon delivery vector with Mmel sites positioned at each end of the Tn (10, 11) was modified so that it contained two taxon-specific barcodes (Figure 1B). Mmel digestion of microbial DNA prepared from the gut contents or feces of recipient gnotobiotic mice cleaves genomic DNA at a site 20-21 bp distal to the restriction enzyme’s recognition site. The site of Tn
insertion and the relative abundance of each Tn mutant in the input libraries introduced into mice and in the ‘output’ communities recovered from these animals were defined using a multi-taxon insertion sequencing (INSeq) protocol. This protocol enables sequencing of the 16-17 bp of flanking genomic sequence in addition to the taxon-specific barcode positioned within the transposon (Figure 1B). A software package (SI Methods) was used to assign sequencing reads based on the sample-specific barcode, and the taxon-specific barcode. Control experiments, using ‘mock communities’ composed of the 15-member collection of cultured human gut bacteria with different proportions of the mutant libraries (ranging from 0.5-40% of the community) revealed that this protocol had high specificity (91.6±0.84% of reads mapped to sites of Tn insertion), high sensitivity (a mutant library representing as little as 0.5% of the entire community could be characterized), and good reproducibility (R²>0.9; n=5 technical replicates/mock community) (Table S2, see SI Methods and SI Results for details).

To identify mutations that were deleterious in the gut, we calculated a “fitness index” for each gene based on the relative abundance of its INSeq reads in the output community (e.g., fecal microbiota) compared to input library. Using an approach based on the Expectation Maximization (EM) algorithm, we observed that the output to input log-ratios for the entire mutant population were bimodally distributed, with the vast majority of mutants (genes) falling in a normal distribution to the right of a much smaller distribution with much lower fitness. To identify deleterious mutations, we calculated a z-score (fitness index) for the difference between each gene’s log-ratio and the mean of the upper normal distribution (Figure 1C-E). A p-value was determined based on a normal distribution using a Z-test, and a q-value assigned after applying a multiple hypothesis testing correction (FDR). To avoid misidentifying genes as significant fitness determinants, we only considered those that were significantly depleted across biological repeats (significant q-values in 4 out of 5 samples). In silico simulations revealed that this EM algorithm-based procedure yielded a false positive rate of less than 5% for mutants in genes whose relative abundance was >0.002% (Figure S2). These simulations were verified by applying the analysis pipeline to our mock communities (Table S2). ORFs satisfying the criteria of having a significant q-value in one diet condi-
tion, and no fitness defects in any biological replicates under another diet condition, were defined as ‘diet-specific fitness factors’.

**Community assembly, response to diet, and resilience**

Adult (10-12-week-old) male germ-free C57BL/6J mice were placed on either a low-fat/high-plant polysaccharide (LF/HPP) chow or a diet where calories were largely derived from hydrogenated vegetable shortening and beef tallow, and where carbohydrates were limited to sucrose, corn starch, and maltodextrin (plus an indigestible cellulose binder) [‘high saturated fat’/’high simple sugar’ (HF/HS) diet]. Diets were started 7 d before a single oral gavage of the artificial community containing the four mutagenized strain libraries and 11 wild-type strains (Figure 2A). Control experiments used a 15-member community composed of all wild-type strains, or 15-member communities that only contained a single mutant library (Figure S3A). Groups of singly-caged mice were maintained on either the LF/HPP or HF/HS diet for the duration of the experiment, while other groups were subjected to a reciprocal set of diet oscillations (LF/HPP®HF/HS®LF/HPP or HF/HS®LF/HPP®HF/HS) (n=5 mice/diet treatment/community type; Figure 2A). Short-read shotgun sequencing (COPRO-Seq) of DNA prepared from fecal samples collected over the course of each experiment was used to define the relative abundance of each community member.

When considered as a whole, each mutant library acted similarly to the parental wild-type strain in terms of its proportional abundance in the community (Figure S3B, Table S3). In both monotonously fed treatment groups, the relative abundance of *B. cellulosilyticus* WH2 was high, although significantly greater in animals fed LF/HPP chow (36±1.6% versus 16±3.1% after 16 days on the diets; p<0.01, Mann-Whitney test). *B. ovatus* had a striking preference for conditions associated with LF/HPP chow feeding (11±0.7% versus 0.8±0.3%; p<0.01, Mann-Whitney test) (Figure S4A-B). Strain-specific differences were also evident; *B. thetaiotaomicron* VPI-5482 was more successful in the LF/HPP diet context than *B. thetaiotaomicron* 7330 (9±0.5% versus 0.02±0.04 %; p<0.05, Mann-Whitney test) while the 7330 strain had significantly higher abun-
dance than the VPI-5482 strain when animals were consuming the HF/HS diet (6±1.4% versus 3±0.5%; p<0.05, Mann-Whitney test) (Figure S4C-D).

Multi-taxon INSeq allowed us to obtain a genome-wide, gene-level view of the effects of dietary selection pressures on fitness. Analysis of the relative abundances of all *B. cellulosilyticus* mutants in fecal samples collected at 4, 10, and 16 days post-gavage in the two monotonous diet treatment groups disclosed that by day 4 the mutant population had already manifested diet-specific configurations (PERMANOVA, *p*=0.001). The relative abundances of *B. cellulosilyticus* mutants were remarkably consistent between individually-caged mice monotonously fed a given diet, including groups of mice housed in different gnotobiotic isolators (Student’s unpaired t-test; Figure S5A) This within treatment group consistency (defined by Hellinger distances) was sustained as the configuration of the mutant population evolved during the ensuing 11 days of monotonous diet consumption (Figure 2D, Figure S5B-C).

Multi-taxon INSeq of fecal samples collected at the end of 16 days of monotonous diet consumption yielded 517 HF/HS diet-specific *B. cellulosilyticus* fitness determinants and 33 LF/HPP diet-specific determinants: 270 of the 517 HF/HS diet-specific genes had KEGG Orthology (KO) assignments; among this group, there was a significant enrichment of genes belonging to the KEGG categories ‘Membrane Transport’ [e.g., an operon involved in iron transport, another involved in zinc transport, plus several Major Facilitator Superfamily transporters (glucose, mannose, xylose and phosphate)], ‘Metabolism of Cofactors and Vitamins’ (e.g., five genes in the cobalamin biosynthesis pathway, consistent with the view that the capacity to synthesize and utilize cobalamin and other substituted corrins is an important determinant of survival in the gut (10, 12), and ‘Protein Folding and Associated Processing’ (see Table S4A for the functional annotations and fitness indices of these genes).

The two groups of mice subjected to a reciprocal set of diet oscillations [LF/HPP (4d) ➔ HF/HS (6d) ➔ LF/HPP (6d) or HF/HS (4d) ➔ LF/HPP (6d) ➔ HF/HS (6d)] exhibited consistent and marked diet-specific changes in overall community structure as defined by COPRO-Seq (Figure S5B-C).
ure 2B,C; also see the patterns of change in relative abundances of *B. cellulosilyticus* and *B. ovatus* in Figure S4A,B). Followup INSeq analysis of the same fecal DNA samples used for COPRO-Seq measurements afforded an opportunity to simultaneously characterize the degree to which the aggregate collections of *B. cellulosilyticus* and *B. ovatus* mutants were able recover from a diet disturbance (an INSeq-based definition of resiliency) and the variation in such recovery between animals (an INSeq-based measure of stochasticity). The configuration of the *B. cellulosilyticus* mutant population exhibited diet-specific changes. However, during the ‘recovery phase’, when these two groups of mice were returned to the starting diet of the sequence, the *B. cellulosilyticus* mutant library shifted towards a state similar but not identical to that observed in monotonously fed hosts (Figure 2D). Of the 517 genes identified as HF/HS specific fitness determinants in monotonously fed animals, 146 had significantly higher fitness scores in animals that had been switched temporarily to LF/HPP chow and then returned to the HF/HS diet compared to mice that had only consumed the HF/HS diet [includes the 6 genes comprising the only one of its 113 PULs identified as a HF/HS diet-specific fitness determinant (see Table S4A and below)]. Similarly, 21 of the 33 genes designated as LF/HPP-specific fitness in monotonously fed mice had significantly higher z-scores in mice that had been temporarily switched to the HF/HS diet and then returned to LF/HPP chow (Table S4A). As with *B. cellulosilyticus* WH2, diet oscillations of the type A to B to A (or B to A to B) led to increased similarity in the overall patterns of fitness in the *B. ovatus* mutant library compared to those that were monotonously fed (Figure S6; Table S4B). This response to perturbation may represent an important mechanism behind the maintenance of diet-specific traits in a member or members of a microbiota harbored by hosts experiencing more varied diets (here we are using mutants as proxies for functional ‘traits’).

**Identifying core in vivo fitness determinants in four Bacteroides strains**

Genes that are conserved among the four *Bacteroides* strains and that show a significant effect on fitness in all strains in both diets contexts can be defined as a core set of *in vivo* fitness determinants for these members of this genus; as such they inform us about the resource requirements and
selective pressures these taxa experience in the gut in this community context and under these two dietary conditions.

In total, 2,204 genes are conserved among all four *Bacteroides* strains. Multi-taxon INSeq of fecal samples, collected from mice colonized for 16d and monotonously fed one or the other diet revealed 86 conserved genes with significant diet-independent effects on fitness in all four strains (*Table S5*). The fitness indices (z-scores) for these genes varied as a function of strain but less so across the different diets ($p<0.001$; two-way ANOVA). Twenty of the 86 core fitness determinants were components of KEGG pathways involved in various aspects of amino acid biosynthesis and metabolism (*Figure S7*). In follow-up studies, the *B. cellulosilyticus* mutant library was grown to stationary phase in minimal medium lacking amino acids and containing one of several carbon sources [glucose, xylose, arabinose, or wheat arabinoxylan (the most common hemicellulose in cereals)], or in control rich medium [tryptone-yeast extract-glucose (TYG) (n=6 replicates/growth condition)]. INSeq analysis of the input library and stationary phase cultures (*Table S6*) confirmed that Tn mutagenesis of the 20 core fitness determinants annotated as being involved in the biosynthesis of 11 amino acids precluded growth in any of four types of amino acid-depleted minimal media but did not affect growth in the rich medium. Together, these results emphasize the important contribution of amino acid biosynthesis to the survival of these *Bacteroides* strains in the niches they occupy in the 15-member community, at least in the two diet contexts considered in this study.

Regulators of carbohydrate utilization were also represented among the 86 shared *in vivo* fitness factors. They include homologs of *BT4338*, a Crp-like transcription factor that controls a proposed global sugar catabolic regulon (*J3*) encompassing up to 30 genes/genome that participate in utilization of arabinose, xylose, fucose, galacturonate, pectin and β-hexosamines. [Three core fitness genes are required for growth in arabinose: *araA* (L-arabinose isomerase), *araP* (arabinose transporter) and *araR* (NrtR family transcriptional regulator of arabinose utilization genes)]. Homologs of *BT3992*, an anti-sigma factor that in *B. thetaiotaomicron* functions as part of a trans-envelope signaling system that coordinately regulates expression of three polysaccharide utiliza-
tion loci (PULs) and a capsular polysaccharide synthesis (CPS) locus \(^{(14)}\) were also represented among the group of shared diet-independent \textit{Bacteroides} fitness factors.

We tested the hypothesis that genes that played a key role in determining fitness of the four \textit{Bacteroides} strains in the gut ecosystem would be expressed at higher levels in the gut than in rich medium or in minimal medium containing single carbon sources including arabinose, arabinoxylan, xylose, or glucose. To do so, we performed microbial RNA-Seq on fecal samples collected at the same time as samples for INSeq. The results revealed that only 14 of the 86 core fitness determinants had significantly higher levels of expression \textit{in vivo} under both diet conditions than under any of the \textit{in vitro} conditions examined (\(n=4\) mice/diet treatment; \(n=3\) biological replicates/\textit{in vitro} growth condition) (\textbf{Figure S8A}). The fact that expression of the remaining 72 core fitness determinants, including the regulator of the proposed sugar catabolic regulon (\textbf{Figure S8B}), was not higher \textit{in vivo} than \textit{in vitro} emphasizes the limitations of using mRNA levels as an exclusive parameter for nominating genes as critical for survival in a given (gut) environmental context.

\textbf{Differences in fitness determinants in the two \textit{B. thetaiotaomicron} strains}

\textit{B. thetaiotaomicron} VPI-5482 is one of the most-studied human gut \textit{Bacteroides} \((6, 7, 13, 15-17)\). \textit{B. thetaiotaomicron} 7330 is a strain recovered from a healthy adult [see \textbf{Figure 3A} for alignments of the two genomes, \textbf{Table S7A} for the representation of genes in KEGG categories, KEGG pathways for the two strains, and \textbf{Table S7B} for a list of the 94 PULs encompassing 777 predicted ORFs in the 7330 strain, 18 of which (130 ORFs) are unique (their protein products had \(<90\%\) similarity to any protein in the predicted VPI-5482 proteome)]. The 7330 strain is much more successful in the 15-member community context when mice are consuming the HF/HS diet. While the relative representation of each \textit{B. thetaiotaomicron} strain in the 15-member community varied by 340-fold depending upon the diet, their total relative abundance in the community was the same across the two diets (\textbf{Figure 3B}), raising the possibility that the strains might compete for a niche in this community and that occupancy of this niche by one or the other strains is highly sensitive to diet. Therefore, we used INSeq to identify factors that define their distinct dietary
niches (Table S8; includes the 236 HF/HS-diet specific fitness determinants in *B. thetaiotaomicron* 7330, listed in decreasing rank order of their FDR *q*-values. No KEGG category or pathway was significantly enriched among these 236 ORFs (Table S8B). In the absence of dietary polysaccharides, *B. thetaiotaomicron* VPI-5482 adaptively forages on host mucosal glycans suggesting that the capacity to consume both dietary and host-derived glycans is one of the mechanisms by which it is able to survive in the gut (7, 18, 19). The HF/HS diet-specific and 7330 strain-specific fitness factors include genes in two operons specifying components of cation efflux systems plus a number of glycoside hydrolases predicted to be involved in the catabolism of mucosal glycans (Figure 3C,D; Table S8B). The latter include genes (*Btheta7330_1384, Btheta7330_3308* and *Btheta7330_3311*) that encode three hexosaminidases (EC 3.2.1.52 cleaves GalNAc residues) that are also represented in the VPI-5482 genome but where they do not convey a significant fitness effect. Thus, one functional distinction between the two strains revealed by multi-taxon INSeq is that they likely use different strategies for adaptive foraging of mucosal glycans in the absence of dietary polysaccharides.

Individual *Bacteroides* species contain multiple capsular polysaccharide biosynthesis (*CPS*) loci, allowing for a large number of combinations of locus expression in different environmental contexts. We used multi-taxon INSeq to identify *CPS* loci and their component genes that were critical for fitness *in vivo* within and across species and strains as a function of diet. The two *B. thetaiotaomicron* strains vividly illustrate the unique contributions of individual loci to survival. *CPS4* was the only one of the VPI 5482 strain’s eight *CPS* loci that functioned as a significant fitness determinant in the context of the 15-member community and either of the two diets (as shown in Figure S9B, 20 of its 21 genes, covered by >1200 Tn mutants were defined as essential). In contrast, *CPS6*, a structurally distinct locus in the 7330 strain composed of 18 genes, was the only one among its six *CPS* loci critical for its fitness in either diet context (all 18 genes, covered by >750 mutants, had significant fitness indices; Figure S9C). INSeq established that this strain-specificity for *CPS* fitness effects was evident along the length of the gut (data not shown).
Identifying a diet supplement that can specifically manipulate *B. cellulosilyticus* abundance

*B. cellulosilyticus* WH2 is equipped with the largest number of carbohydrate active enzymes (CAZymes, comprising glycoside hydrolases, polysaccharide lyases, carbohydrate esterases, glycosyltransferases, carbohydrate binding modules) in any sequenced Bacteroidetes genome reported to date (9). Its 503 CAZymes are distributed among 113 PULs. Remarkably, only one of the 113 PULs in *B. cellulosilyticus* WH2 (*BACWH2_4070-76*) functioned as a significant fitness determinant in the HF/HS diet context (Figure 4A). With the exception of this PUL’s hybrid two-component system transcriptional regulator, all of its other six genes, which include a predicted xylanase (*BACWH2_4072*), a predicted β-galactosidase (*BACWH2-4071*), and a sialate O-acetylesterase (EC3.1.1.53; *BACWH2-4070*), satisfied our criteria for designation as HF/HS diet-specific fitness determinants. Foraging on mucus glycans that contain N-acetyl-O-acetylneuraminic residues would represent an adaptive response of this organism to a high fat diet that does not contain complex polysaccharides (including xylans). The fitness indices of the 117 Tn mutants that mapped to the six genes in this PUL were remarkably consistent between individually-caged animals as a function of their diet. Moreover, as noted above, the relative abundances of these mutants was higher in mice that had undergone a HF/HS-LF/HPP-HF/HS diet oscillation than in mice monotonously fed the HF/HS diet (Figure 4A).

Our previous in vitro studies of *B. cellulosilyticus* WH2 cultured in minimal medium supplemented with one of 31 distinct carbohydrate substrates, plus RNA-Seq analysis of the bacterium recovered at mid-log phase from culture in those minimal media that supported its growth, revealed that this PUL was induced by xylan and arabinoxylan (9; Figure S10A). *B. ovatus*, the only other *Bacteroides* strain in the artificial community that could grow in minimal medium containing purified arabinoxylan as the sole carbon source (Figure S10B) contains a PUL (*BACOVA_4389-94*) that is induced when this medium was supplemented with arabinoxylan (or xylan). Multi-taxon INSeq of fecal samples collected from mice harboring the 15-member community showed that unlike the xylan and arabinoxylan-inducible *B. cellulosilyticus* PUL, this PUL is not required for survival of *B. ovatus* in the HF/HS diet context (Figure S10C).
Given these results, we examined the effects of consumption of arabinoxylan purified from wheat on the relative abundances of *B. cellulosilyticus* WH2 and *B. ovatus* and their fitness determinants. Our rationale was that we could use arabinoxylan to induce expression of the one *B. cellulosilyticus* WH2 PUL that was a key fitness factor in the HF/HS diet context (this diet does not contain xylans), and that PUL induction would be accompanied by improved fitness. Prior to colonization with the 15-member community containing the INSeq libraries, one group of germ-free mice received the HF/HS diet plus drinking water supplemented with wheat-derived arabinoxylan (7.5% w/v) for seven days. Another group received the same diet but without arabinoxylan in their drinking water. All groups were then gavaged with the artificial community: the group pre-treated with arabinoxylan continued to receive the HF/HS diet and supplemented water *ad libitum* for the next 14 days and then was switched to un-supplemented water. A reciprocal treatment group received un-supplemented water plus the HF/HS diet for 14 days followed by a switch to arabinoxylan-supplemented water. A third group received un-supplemented water plus the low fat, high plant polysaccharide (LF/HPP) diet for 14 days followed by a switch to supplemented water (while being maintained on the LF/HPP chow); these animals served as a control since the targeted PUL only manifests itself as a fitness determinant in the context of the high fat, polysaccharide-deficient diet (n=5 individually caged mice per treatment group) (Figure 4B).

COPRO-Seq analysis of fecal samples collected at the end of each 14 d treatment period demonstrated that arabinoxylan produced a significant increase in *B. cellulosilyticus* WH2 abundance in mice fed the HF/HS diet (adjusted *p*-value <0.01; Student’s *t*-test) but no effect on the LF/HPP diet context (Figure 4C; n=5 singly housed mice/treatment arm). Consistent with the INSeq results showing that the arabinoxylan utilization PUL in *B. ovatus* is not a fitness determinant in the HF/HS diet context, we observed no significant effects of arabinoxylan treatment on the relative abundance of this community member (Figure 4C). We confirmed these findings in a separate experiment where two groups of gnotobiotic mice harboring the 15-member community were treated for 56 days with a HF/HS diet with or without supplementation of their drinking water with 15% arabinoxylan. In this higher dose, longer duration, monotonous diet experiment,
the abundance of *B. cellulosilyticus* WH2 increased significantly (as did levels of cecal short chain fatty acids and deconjugated bile acids), while *B. ovatus* showed no response (Figure S11A,B). Arabinoxylan treatment did not produce a statistically significant difference in total body weight (Student’s *t*-test).

We identified 294 *B. cellulosilyticus* WH2 genes that functioned as fitness factors when mice consumed the HF/HS diet without arabinoxylan supplementation that were no longer significant fitness determinants when water was supplemented with arabinoxylan (listed in Table S9A by KEGG category and ranked according to their fitness index in the unsupplemented HF/HS diet). These genes included the arabinoxylan- and xylan-inducible PUL, *BACWH2_4070-4075* (Figure S12A). (Note *B. cellulosyliticus* WH2 contains another PUL encoding a predicted xylanase; this PUL (*BACWH2_4044-4055*) is induced by xylan and arabinoxylan in vitro but is not a fitness determinant when mice are consuming the HF/HS diet ± arabinoxylan, suggesting that together these two PULs provide redundant yet distinguishable functions expressed by *B. cellulosilyticus* WH2 in the 15-member community context).

The only KEGG category significantly enriched for genes that no longer function as fitness determinants when HF/HS treated mice receive arabinoxylan is ‘Membrane transport’ (see Figure S12B for a ranking based on their *q*-scores). They encode one nucleoside H+ symporter, one ABC-type sugar transporter, two MFS-type fucose transporters, proteins involved in metal ion transport (iron, zinc), a MFS- and an ABC-type drug resistant transporter, and three ABC-type antimicrobial peptide transporters (Figure S12B). The only *B. ovatus* genes in the Membrane transport’ category that lose their significant fitness index scores with arabinoxylan treatment are two MFS-type multidrug transporters and one ABC-type antimicrobial peptide transporter (these transporters are not homologous to those identified in *B. cellulosilyticus*). Mucosa-associated fucosylated glycans are adaptively foraged by human gut *Bacteroides* spp. (7, 19, 20); these *Bacteroides* also regulate production of antimicrobial peptides by the epithelium (21, 22). Recent studies have shown that fucose derived from gut mucosal glycans can be used as a signal by pathogens to change their virulence gene expression in ways that aid invasion/colonization of the gut (23, 24). Integrating
these findings suggests that arabinoxylan supplementation reduces the need of *B. cellulosilyticus* (but not *B. ovatus*) to adaptively forage on host-derived mucosal (e.g., fucosylated) glycans, and that the diminished requirements for anti-microbial peptide transporters reflect alterations in the organisms’ interactions with the gut mucus barrier.

A total of 24 genes were identified in the *B. cellulosilyticus* WH2 genome that were significant fitness determinants when mice received the HF/HS diet plus arabinoxylan but not when animals received the HF/HS diet alone (*Table S9A*): they included (i) all genes in an operon (*BACWH2_0804-0806*) that includes a sulfotransferase and sulfate transferase (EC 3.7.74), consistent with the observed increase in deconjugated bile acids documented by UPLC-MS of cecal contents (*Figure S11B*), and (ii) several genes that highlight how increases in arabinoxylan availability enhance the importance of ammonium utilization for synthesis of amino acids and proteins in this community and diet context [an ammonium transporter (*BACWH2_4959*), plus glutamine, glutamate, and asparagine synthetases (*BACWH2_0058*, *BACWH2_4974*, and *BACWH2_4972*, respectively)].

Even though *B. ovatus* did not manifest a significant change in its relative abundance when mice received arabinoxylan supplemented water, multi-taxon INSeq revealed 31 genes that were not fitness determinants when mice consumed HF/HS chow alone but ‘acquired’ significant z-scores when arabinoxylan was introduced (for a list, see *Table S10B*). They include seven closely linked glycosyltransferase genes, including two raffinose-raffinose alpha-galactotransferase homologs (EC2.4.1.166) and a galactoside 2-L-fucosyltransferase (EC2.4.1.69)], suggesting that this organism changes its glycan utilization strategies when it encounters arabinoxylan in the gut environment.

**Prospectus**

There are several important definitions of the niche in the field of macroecology and evolutionary biology. Elton defined the niche as an animal’s “place in the biotic environment, its relations to food and enemies”(25); i.e., as the role an organism was playing in the food web. Hutchinson
(1957) described the niche as an n-dimensional hypervolume, where the dimensions are values of the environmental conditions and resources that contain the requirements of an individual or species to persist (as a population) (26); these requirements become axes in the coordinate system that describes the niche. One of the dangers of focusing on resources alone as the description of a microbe’s niche is that it places an immediate precedence on resource competition as a major mechanism for community structure, and less (or zero) focus on other interactions. For instance, the presence of predators (phage) or mutualists (co-feeding) can mediate coexistence between competing species in some situations. In our study, multi-taxon INSeq is used to provide an ‘operational’ description of an organism’s niche by determining which sets of genes allow a bacterial strain to co-exist with other strains/species under a defined set of habitat conditions (gnotobiotic mice representing a given genetic background, colonized with a given set of sequenced organisms and fed a given set of diets). In principle, this approach can be applied to gut and non-gut habitats in gnotobiotic mice representing different genetic backgrounds, harboring different microbial consortia and manipulated in various ways (diet, immune system, etc) to obtain a more comprehensive picture of how host genotype impacts niches. It also offers a way to address fundamental as well as applied questions: e.g., functionally discriminating strains when exploring issues related to the genetic foundations of opportunism, defining stochasticity during community assembly, resiliency after a perturbation, selecting lead probiotic candidates for bioremediation of perturbed/dysbiotic microbial communities, identifying compounds that affect the functional configurations of a targeted microbiota, and for characterizing the nutritional effects/value of diets and their ingredients.
**Acknowledgements**

We thank David O’Donnell and Maria Karlsson for their assistance with gnotobiotic mouse husbandry, plus Alejandro Reyes for many helpful suggestions. This work was supported in part by grants from the NIH (DK70977 and DK30292) and the Crohn’s and Colitis Foundation of America.

**Data deposition**

INSeq, COPROSeq, and microbial RNASEq datasets have been deposited in the European Nucleotide Archive (ENA).
References


102


32. F. E. Rey et al., Metabolic niche of a prominent sulfate-reducing human gut bacterium. 

Figure Legends

Figure 1. Experimental and computational pipeline for multi-taxon INSeq. (A) Mutant libraries (input populations) are pooled together with other wild-type strains to form an artificial community of sequenced members of the microbiota; this consortium is then introduced into germ-free mice. Community assembly, the effects of diet perturbation, and recovery from diet oscillations are characterized at a community-, strain- and gene-level. (B) Preparation of a sequencing library for multi-taxon INSeq. Genomic DNA from a sample is digested with MmeI and transposon-sized fragments (~1.5 kb) are isolated, ligated to adapters with a sample-specific barcode and amplified by PCR to create the final sequencing library. The sample barcode and chromosomal DNA flanking the site of transposon insertion are sequenced using read 1 sequencing primer (SP). The strain-specific barcodes are sequenced with an INSeq index SP. Note that each Tn carries two sequences, colored green and aqua that together comprise its taxon-specific barcode. (C-E) Data analysis pipeline. Sequencing reads are binned first by their sample-specific barcodes, and then by their strain-specific barcodes. Each read is then mapped to the strain’s genome to identify the insertion site from which it originated. The read count for each insertion site is tallied, normalized by sequencing depth, and the normalized sums for all insertion sites within a gene are combined. For each strain, the read count for a given gene represented in the output community is compared to the read count in the input population (panel C). The log ratio of output to input counts for each mutagenized gene in the strain is calculated and the expectation-maximization (EM) method used to define the bimodal distribution (panel D): the population on the right, delineated by the green line, is composed of mutants with normally-distributed log-ratios; the population delineated by the red line described mutants with lower log ratios that fall into a distinctly different distribution. A z-score is calculated for each gene by comparing the differences between its log-ratio and the mean of the upper normal population (panel E).

Figure 2. The effect of diet and diet oscillations on the structure of the 15 member artificial community containing 11 wild-type strains and the four mutant libraries, and the structure of the mutant population. (A) Experimental design. 10-12-week-old germ-free male C57BL/6J
mice were gavaged with the indicated consortium of 15 strains. Animals were given the LF/HPP and/or the HF/HS diet in the order shown. Fecal samples were collected at the indicated time points for INSeq and microbial RNA-Seq analyses. (B,C) Principal coordinates analysis (PCoA) of Hellinger distance measurements based on COPRO-Seq data. Each colored circle represents a fecal community sampled from an individually-caged mouse belonging to the indicated diet treatment group. (D) PCoA of Hellinger distance measurements based on the relative abundance of *B. cellulosilyticus* WH2 Tn mutants in fecal samples, as defined by multi-taxon INSeq analysis. Each colored circle represents the population of all mutants in *B. cellulosilyticus* WH2 sampled from an individual mouse in the indicated diet treatment group at a given time point. Each empty circle represents the average location of the mutant population for a given group at a given time point. The lines connect time points in a given group. The color of the lines indicates the diet. Dashed lines indicate the diet oscillation groups.

**Figure 3. HF/HS diet-specific fitness determinants in *B. thetaiotaomicron* 7330.** (A) Circos (27)-generated alignment of the *B. thetaiotaomicron* VPI-5482 and *B. thetaiotaomicron* 7330 genomes. Grey lines connect genes encoding homologous proteins present in the two genomes. The color-coded outer circle denotes the similarity between these proteins: green, >90% identity (based on Blastp alignment); blue, 70%-90% identity; red, <70% identity; black, intergenic regions. (B) COPRO-Seq analysis of the relative abundance of the two *B. thetaiotaomicron* strains in the fecal microbiota of mice sampled two weeks after gavage while consuming the HF/HS or LF/HPP diets. Mean values ± SEM are shown (n=5 individually caged mice harboring a community consisting of 11 wild-type and the four mutant libraries/treatment group). Note that the summed relative abundance of the two strains remains the same even though the relative representation of the individual strains is significantly different in the two diet contexts (p<0.001, 2-way ANOVA). (C) HF/HS diet-specific fitness determinants in *B. thetaiotaomicron* 7330 involved in degradation of glycosaminoglycans associated with the intestinal mucosa (genes highlighted in red with their EC annotations). (D) HF/HS diet-associated changes in the z-scores of 7330-strain-specific fitness determinants that are involved in transcription regulation (*Btheta7330_2549*) or are components of
operons encoding transport systems (vertical lines denote individual operons).

**Figure 4. Arabinoxylan increases the relative abundance of *B. cellulosilyticus* WH2 in vivo.**

(A) INSeq analysis reveals that all genes in PUL *BACWH2_4070-6* have significant fitness indices (z-values) in the HF/HS diet. *BACWH2_4076*, encoding a hybrid two-component system regulator (HTCS), is the only gene in this PUL that has a significant fitness effect on the LF/HPP diet. Functional annotations for genes in the PUL are shown together with direction of their transcription. Fitness indices for each gene in the different diet contexts (orange, HF/HS, green LF/HPP) are plotted as mean values ± SEM. The horizontal dashed line indicates the cutoff for significance (*p*<0.05; z-test with FDR correction). (B) Experimental design. Adult C57BL/6J germ-free mice were gavaged with a consortium containing 11 wild type strains plus the four *Bacteroides* INSeq libraries. Animals were fed the HF/HS or LF/HPP diets with or without supplementation of their drinking water with 7.5% arabinoxylan (n=5 individually-caged mice/group). (C) The relative abundance of *B. cellulosilyticus* WH2 and *B. ovatus* was defined by COPRO-Seq analysis of fecal samples collected at the indicated time points. Mean values ± SEM are plotted. ***, *p* <0.001 compared to the reference group A at 14 days post-gavage (dpg); ***, *p* <0.001 for within group comparisons of the 30 dpg versus 14 dpg fecal sample (Student’s t-test after FDR correction). *B. ovatus*, the only other *Bacteroides* strain in the community that exhibited significantly increased growth in minimal medium supplemented with arabinoxylan (Figure S10B) did not manifest a significant change in relative abundance *in vivo* when arabinoxylan was added to the drinking water (ns, not significant).
Figures

Figure 1.

A 15-member defined community with four members represented by tens of thousands of mutants

Diet 1  
Diet 2  
Diet 1  
Diet 2  
Diet 1  
Diet 2

Community assembly, diet perturbation/recovery at community-, species/strain-, and gene- level

B

Digestion

Size selection

Ligation

Amplification

Sequencing

Gnotobiotic mouse

C

Log transformation and EM estimation

D

E

Fitness index (z-score)

Log ratio (Output/Input)

Frequency

Input Counts

Output Counts

100 101 102 103 104

0 1 2 3 4

−3 −2 −1 0 1 2 3

−7 −6 −5 −4 −3 −2 −1 0 1 2 3

Log transformation and EM estimation

Z-test

5 4 3 2 1 0 −1 −2 −3 −4 −5

Fitness index (z-score)
Figure 2.

A

B

C

D

Fecal samples collected for INSeq
Fecal samples collected for Microbial RNASeq

Days post gavage

Days post gavage

LF/HPP->HF/HS->LF/HPP

LF/HPP

HF/HS

HF/HS

LF/HPP

LF/HPP->HF/HS->LF/HPP

B. ovatus ATCC 8483T::INSeq
B. cellulosilyticus WH2::INSeq
B. thetaiotaomicron VPI-5482::INSeq
B. thetaiotaomicron 7330::INSeq
Bacteroides caccae ATCC43185T
Bacteroides uniformis ATCC 8492
Bacteroides vulgatus ATCC 8482
Parabacteroides distasonis ATCC 8503
Eubacterium rectale ATCC 33656
Ruminococcus obeum ATCC 29174
Collinsella aerofaciens ATCC 25986
Clostridium scindens ATCC 35704
Clostridium symbiosum ATCC 14940
Clostridium spiroforme DSM 1552
Dorea longicatena DSM 13814

PC1 (44%)

PC2 (23%)

PC1 (41%)

Days post gavage
Figure 3.

A. *B. thetaiotaomicron* VPI-5482

B. *B. thetaiotaomicron* 7330

C. Chondroitin sulfate

Heparan sulfate

Keratan sulfate

D. HF/HPP LF/HPP gene annotation

<table>
<thead>
<tr>
<th>Gene</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Btheta730_2672</td>
<td>contains a lipocalin_6 domain</td>
</tr>
<tr>
<td>Btheta730_2549</td>
<td>contains a HTH_XRE domain</td>
</tr>
<tr>
<td>Btheta730_0384</td>
<td>putative TeiR transcriptional regulator</td>
</tr>
<tr>
<td>Btheta730_0385</td>
<td>outer membrane efflux protein</td>
</tr>
<tr>
<td>Btheta730_0386</td>
<td>cation efflux system protein</td>
</tr>
<tr>
<td>Btheta730_0387</td>
<td>putative cation efflux transporter</td>
</tr>
<tr>
<td>Btheta730_0979</td>
<td>cation efflux system protein, AcrB/AcrD/AcrF family protein</td>
</tr>
<tr>
<td>Btheta730_0980</td>
<td>cation efflux system protein</td>
</tr>
<tr>
<td>Btheta730_0981</td>
<td>conserved hypothetical protein</td>
</tr>
<tr>
<td>Btheta730_4874</td>
<td>V-type ATP synthase subunit E</td>
</tr>
<tr>
<td>Btheta730_4875</td>
<td>conserved hypothetical protein</td>
</tr>
<tr>
<td>Btheta730_4876</td>
<td>V-type ATP synthase subunit A</td>
</tr>
<tr>
<td>Btheta730_4877</td>
<td>V-type ATP synthase subunit B</td>
</tr>
<tr>
<td>Btheta730_4878</td>
<td>V-type ATP synthase subunit D</td>
</tr>
<tr>
<td>Btheta730_4879</td>
<td>V-type ATP synthase subunit I</td>
</tr>
<tr>
<td>Btheta730_4880</td>
<td>V-type sodium ATP synthase subunit K</td>
</tr>
</tbody>
</table>
Figure 4.

A

B

C

111
Supplemental Materials

Supplemental Methods

Whole genome transposon mutagenesis of *Bacteroides* spp.

We introduced taxon-specific barcodes into the mutagenesis vector (pSAM_Bt or pSAM_Bcell-WH2) ([Figure 1B](#)) by PCR amplification using vector DNA as template and the various primer pairs described in Table S10A. Amplification conditions were as follows: initial denaturation at 94 °C for 2 min, followed by 25 cycles of denaturation (94 °C for 15 sec), annealing (58 °C for 30 sec), and amplification (58 °C for 90 sec). The vector was then digested with KpnI and BamHI, and the linear product ligated to the amplicon, yielding the barcoded Tn mutagenesis vector.

Whole genome transposon mutagenesis of *B. ovatus* (ATCC 8483T), *B. thetaiotaomicron* 7330 and *B. cellulosilyticus* WH2 was performed using the respective vectors and a modified version of a previously published protocol (10). To improve conjugation efficiency, we combined 100 mL of a mid-log phase culture of one of the *Bacteroides* strains that had been grown in TYG medium under anaerobic conditions (75% N₂, 20%CO₂, 5% H₂), with 25 mL of a mid-log phase culture of *E. coli* λ17, harboring the respective Tn mutagenesis vector, grown under aerobic conditions in LB medium supplemented with 100 µg/mL of ampicillin. Immediately after the cultures were combined, they were passed through a 0.2µm diameter cellulose nitrate filter membrane (Thermo Scientific). The membrane, containing retained bacteria, was then placed on top of brain-heart infusion (BHI) agar supplemented with 10% horse serum (Colorado Serum Co.) and the cells allowed to mate for 6 h at 37°C under aerobic conditions. The conjugates were harvested from the filter by scraping, resuspended in TYG medium and plated on BHI agar supplemented with erythromycin (50 µg/mL) and gentamicin (200 µg/mL), after which the plates were incubated for 2 d at 37 °C under anaerobic conditions. 87,000-170,000 isogenic transposon mutants/strain were pooled in TYG medium containing 30% glycerol and aliquots were stored at -80 °C.
Gnotobiotic mice

All experiments involving mice were performed using protocols approved by the Animal Studies Committee of Washington University School of Medicine. Germ-free 10-12-week-old male C57BL/6J mice were individually caged in flexible plastic gnotobiotic isolators. Animals were given autoclaved LF/HPP diet (B&K chow #73780000, Zeigler Bros, Inc.) or HF/HS diet (Harlan Teklad TD96132) sterilized by gamma irradiation (20-50 kGy). Mice were fed ad libitum and maintained under a strict 12 h light cycle (lights on at 0600h, off at 1800h).

Mice were placed on the LF/HPP or HF/HS diet for one week and then received a single gavage of 300 µL of MegaMedium 1.0 (Table S11) containing an equivalent mixture of *B. cellulosilyticus* WH2::INSeq, *B. ovatus*::INSeq, *B. thetaiotaomicron* VPI-5482::INSeq and *B. thetaiotaomicron* 7330::INSeq transposon libraries, plus 11 other wild-type species (*Bacteroides caccae* ATCC43185T, *Bacteroides uniformis* ATCC 8492, *Bacteroides vulgatus* ATCC 8482, *Parabacteroides distasonis* ATCC 8503, *Eubacterium rectale* ATCC 33656, *Ruminococcus obeum* ATCC 29174, *Clostridium scindens* ATCC 35704, *Clostridium symbiosum* ATCC 14940, *Clostridium spiroforme* DSM 1552, *Dorea longicatena* DSM 13814, and *Collinsella aerofaciens* ATCC 25986). [The 11 wild-type strains had been grown separately in MegaMedium under anaerobic conditions to an OD\textsubscript{600} of 1.0 prior to pooling of 4 mL aliquots of each monoculture. Two mL of each INSeq library (OD\textsubscript{600} = 10) was then combined with the 44 mL pool of the 11 wild-type strains. The 300 µL aliquot of this mixture used for the gavage contained ~10\textsuperscript{8} colony forming units. In separate control experiments, mice received a gavage containing 14 wild-type strains and only one INSeq mutant library. The design of single diet and two diet oscillation experiments, including the time points when fecal samples were collected, are described in Figures 2, 4, S3 and S11.]

**COMMunity PROfiling by Sequencing (COPRO-Seq)**

Total DNA was isolated from mouse fecal samples, or samples obtained from the cecum and mid-colon at the time of sacrifice, using phenol:chloroform:isoamyl alcohol and then purified further with the QIAquick PCR Purification kit (Qiagen). The concentration of DNA was measured (Qubit
Quant-IT dsDNA BR protocol;Life Technologies) and adjusted to 5 ng/µL in 100 µL EB buffer. DNA was subsequently fragmented by sonication to 150-350 bp (median size ~250 bp) and the products were purified using the MinElute PCR purification kit (Qiagen). DNA libraries for sequencing on Illumina HiSeq 2000 or MiSeq instruments were prepared using a multistep protocol that involved end-repair, A-tailing, and ligation to customized sample-specific barcoded Illumina adapters. The resulting adapter-ligated DNA was size-selected (250-350 bp) using agarose gel electrophoresis and then used as a template for enrichment PCR (cycling conditions: 98 °C for 30 sec, followed by 18 cycles of 98 °C for 10 sec, 65 °C for 30 sec, 72 °C for 45 sec, and then 72 °C for 5 min). The resulting amplicons were pooled and subjected to multiplex sequencing (50 nt unidirectional reads; 96 samples/pool; 150-200 million reads/pool). Datasets were analyzed using custom Perl scripts (9).

**Multi-taxon insertion sequencing**

DNA purified from a sample containing the transposon mutant library or libraries (either the input library, or feces collected from gnotobiotic mice at various times after their colonization with the artificial community containing one or four of the libraries, or cecal or mid-colonic contents obtained at the time of sacrifice) was digested with MmeI. The resulting digestion products were purified using a ratio of AMPure beads designed to recover ~1.5 kb fragments (the size fraction that would include transposons and chromosomal DNA sequences flanking their sites of insertion in the genomes of host strains). The purified fragments were ligated to double-stranded adapters containing a sample-specific barcode, and the ligation products were amplified by limited cycles of PCR using the primers shown in Table S10 and the following conditions: 94 °C for 2 min, followed by 22 cycles of 94 °C for 15 sec, 65 °C for 1 min, 68 °C for 30 sec, and then 68 °C for 5 min. PCR products containing sample-specific and taxon (strain)-specific barcodes were purified from each reaction, quantified (Qubit), combined in equimolar amounts into a pool (n=24-48 samples; final concentration, 10 nM DNA in 50 µL of EB buffer), and subjected to multiplex sequencing with an Illumina HiSeq 2000 instrument (150-200 million unidirectional 50 nt reads/pool; note that the sample-specific barcodes and chromosomal DNA adjacent to the site of transposon in-
Insertion were sequenced using the Illumina read 1 primer while the strain-specific barcodes were sequenced using a customized INSeq index primer shown in Table S10). For further details of the library preparation and sequencing protocol, see https://gordonlab.wustl.edu/SuppData/Multi_taxon_libraryprepprotocol.

A software package previously used for single taxon INSeq was modified to process and analyze data generated from multi-taxon INSeq (available at https://gordonlab.wustl.edu/SuppData/Multi-taxon_analysis_pipeline). Briefly, Illumina 50 nt reads and the 7 nt INSeq index reads (see Figure 1B) in FASTQ format were first split by sample using the 4- or 7-nt sample-specific barcodes (Figure 1B), and then further binned by strain using the 4-nt strain-specific barcodes. The remaining portion of the read contains 43-46 nt from the Tn and 16-17 nt from chromosomal DNA adjacent to the site of Tn insertion. After trimming the Tn sequences, reads were mapped to the genome of its assigned strain to identify the insertion site from which it originated (no mismatches were allowed). Read counts from each insertion site (i.e., reads that mapped to the 5’ or 3’ termini of the Tn) were tallied, normalized by sequencing depth, and the normalized sums for all insertion sites within a gene were combined. Methods used for calculating a fitness index for genes targeted by a Tn are described in the main text.

To validate that the number of sequencing reads reflect the relative abundances of the mutants when its library represented a small portion of the artificial community, or when the community contained multiple mutant libraries whose abundances varied over a wide range, we created a series of “spike-in” communities consisting of a fixed and consistent amount of DNA isolated from a pool of fecal samples collected from mice colonized for 16 d with the 15 wild-type strains and fed the HF/HS or LF/HPP diet plus varying concentrations of one or more of the INSeq libraries. In total, we created four one species spike-in communities (one each where the abundance of B. thetaiotaomicron VPI-5482::INSeq, B. thetaiotaomicron 7330::INSeq, and B. ovatus::INSeq was 1%, and one community where the abundance of B. cellulosilyticus::INSeq was 10%), plus three multi-taxon INSeq communities (Mock_Even, where each of the four INSeq libraries each represented 10% of the total DNA, and Mock_LF/HPP and Mock_HF/HS where each of the four
mutant libraries represented the observed abundance of the respective wild-type strain on these diets) (Table S2). Once the mock community DNA was prepared, it was subjected to the multi-taxon INSeq library preparation, sequencing, and analysis procedures described above.

**In vitro analysis of the B. cellulosilyticus WH2 INSeq library**

Four milliliters of the *B. cellulosilyticus::INSeq* library were mixed with 4 mL of 2X minimal medium (Table S11). Aliquots (0.5 mL) of this mixture (OD$_{600}$~5) were immediately introduced into 250 mL of minimal medium supplemented with a single carbon source [5 mg/mL of glucose, xylose, arabinose, or wheat arabinoxylan (Megazyme, P-WAXYL, Low Viscosity; ~8cSt)], or into TYG medium. The resulting cultures (n=2 for each type of media) were incubated at 37 °C in an anaerobic Coy chamber until they reached stationary phase. Three aliquots (technical replicates) were withdrawn from each culture, bacterial cells were pelleted by centrifugation at 3,100 x g for 10 min at 4 °C, and the cell pellets were stored at -80 °C until they were subjected to INSeq analysis.

**Microbial RNA-Seq**

Fecal samples collected and immediately frozen at -80 °C. Total fecal RNA was extracted from each sample using phenol-chloroform-isoamyl alcohol (25:24:1, pH 4.5; Ambion). Protocols for (i) removing residual DNA, (ii) depleting small RNAs (e.g., tRNA) plus ribosomal RNA (5S, 16S, and 23S rRNA) from the sample, (iii) synthesizing double-strand cDNA, and (iv) multiplex sequencing with an Illumina HiSeq 2000 instrument (8-10 samples pooled; 200 million unidirectional 50 nt reads/pool) have been described previously (28, 29). Reads were mapped to the genomes of community members using the SSAHA2 aligner with default parameters (30). The number of mapped raw read counts was normalized across samples using the DESeq package in R (31).
Gas chromatography-mass spectrometry (GC-MS) and ultra high performance liquid chromatography-mass spectrometry (UPLC-MS) of cecal metabolites.

Details of the methods used for targeted GC-MS of short chain fatty acids have been described elsewhere (32). Briefly, 10 μL of a mixture of internal standards (20 mM acetic acid-13C2-D4, propionic acid-D6, butyric acid-13C4, lactic acid-3,3,3-D3 and succinic acid-13C4), 20 μL of 33% HCl (v/v) and 1 mL of diethyl ether was added to the cecal contents of a given animal. The mixture was homogenized, centrifuged and the organic fraction was removed. The diethyl ether extraction was repeated. After combining the two ether extracts, 60 μL of the extract was derivatized with 20 μL N-tert-butyldimethylsilyl-N-methyltrifluoroacetamide (MTBSTFA) for 2 h at room temperature. Derivatized samples (1 μL) were injected with 15:1 split into an Agilent 7890A gas chromatography system, coupled with a 5975C mass spectrometer detector (Agilent, CA). The quantity of each metabolite was determined from the spike-in internal standards and a calibration curve. Quantities were normalized to the initial weight of the sample.

To perform targeted UPLC-MS of bile acids, frozen cecal samples were spiked with 1 vol/wt of internal standard (200 mg/ml cholic acid 13C1 in water) and then homogenized with 20 vol/wt of cold methanol. Samples were subsequently placed at -20 °C for 1 h and then centrifuged (20,800 x g for 10 min at room temperature). The resulting supernatant (200 μL) was collected and dried in a SpeedVac at room temperature. Dried samples were resuspended in 100 μL of 95:5 water:ethanol, allowed to clarify (5 min at room temperature), and were then subjected by centrifugation (20,800 x g for 10 min at 4 °C). The resulting supernatant was subjected to UPLC-MS using a Waters Acquity I Class UPLC system (Waters Corp., Milford, MA) coupled to an LTQ-Orbitrap Discovery (Thermo Fisher Corporation). Mobile phases used for negative ionization were (A) consisted of 5 mM ammonium bicarbonate in water and (B) 5 mM ammonium bicarbonate in 95/5 acetonitrile/water. Bile acids standards were subjected to the same protocol and used to build standard calibration curves for sample measurements.
Supplemental Results

Reproducibility, sensitivity and precision of multi-taxon INSeq

For each of the four mutant libraries, technical replicates of the input mutant population were prepared, sequenced using multi-taxon INSeq, and the resulting datasets analyzed with the multi-taxon INSeq pipeline. The coefficient of determination ($R^2$) of the replicates was 0.995, 0.996, 0.972 and 0.997 for *B. thetaiotaomicron* VPI-5482::INSeq, *B. thetaiotaomicron* 7330::INSeq, *B. ovatus*::INSeq and *B. cellulosilyticus* WH2::INSeq, respectively (Figure S13), thereby establishing the highly reproducible nature of the multi-taxon INSeq protocol and mapping strategies.

As noted in *Supplemental Methods*, we created a series of mock communities where the representation of each INSeq mutant library was varied from 1-10%. Figure S14 plots read counts obtained from each library and from the corresponding mock communities. The Spearman correlation coefficient between the abundances of mutants in the ‘starting’ input library and when it was portrayed as a community member at a level of 1% was 0.996 (*B. thetaiotaomicron* VPI-5482), 0.990 (*B. thetaiotaomicron* 7330 strain), 0.992 (*B. cellulosilyticus* WH2), and 0.976 (*B. ovatus*).

For mock communities containing all four mutant libraries at equivalent levels of 10% (Mock_Even), or at the abundances documented for the corresponding wild-type strains in mice containing a 15 member all wild-type community characterized 16 d after colonization and consumption of the LF/HPP or HF/HS diets (Mock_LF/HPP, and Mock_HF/HS), the correlation coefficients were all >0.95 (Table S2).

We also performed simulations where the read depth of each input library was systematically varied from 10,000 -1,000,000, and the false positive rate for a gene having a significant fitness index (z-score) was calculated. The results revealed a false positive rate of <5% for genes whose relative abundance in a library is >0.002% (20 reads/million) at a read depth of ≥30,000 for the library (Figure S2). The results of these simulations were validated using experimental data generated from the mock communities (Table S2).
Evidence that the behavior of each mutant library in the artificial community was equivalent to that of the corresponding wild-type strain

COPRO-Seq analysis was used to compare the relative abundances of the four \textit{Bacteroides} strains at the end of each diet phase in five different community contexts: one containing 15 wild-type strains; three containing 14 wild-type strains plus one of the mutant INSeq libraries (\textit{B. thetaiotaomicron} VPI-5482, \textit{B. thetaiotaomicron} 7330, and \textit{B. ovatus}), plus a community containing 11 wild-type strains and the four mutant libraries.

The results revealed that in each diet context, there were no statistically significant differences in the abundances of a given \textit{Bacteroides} strain in any of the five community contexts (ANOVA; \textbf{Figure S3B}, Table S3A,B).

Microbial RNASeq was performed on fecal samples collected at the time of sacrifice from mice harboring artificial communities composed of 15 wild-type strains or 14 wild-type strains plus the \textit{B. ovatus}::INSeq library (see \textbf{Figure S3A}; average 20 million reads/sample). The presence or absence of the \textit{B. ovatus} mutant library had no significant effect on the community metatranscriptome as judged by principal coordinates analysis based on pairwise comparisons of transcript abundances using the Hellinger distance metric (\textbf{Figure S3C}). In addition, the Spearman correlation coefficient between the level of expression of each \textit{B. ovatus} gene in the all wild-type community and in the community with the mutant library was 0.939 (\textbf{Figure S3D}).

Fecal samples collected from mice containing each of the three types of communities with 14 wild-type strains and one mutant library at the end of the first diet phase of the experiment shown in \textbf{Figure S3A} were subjected to INSeq analysis and the results compared to mice consuming the same diet but colonized with the community containing 11 wild-type strains plus the four mutant libraries. The correlation coefficient of the relative abundances of mutants was 0.977, 0.973 and 0.913 for \textit{B. ovatus}::INSeq, \textit{B. thetaiotaomicron} VPI-5482::INSeq, and \textit{B. thetaiotaomicron} 7330::INSeq, respectively, indicating the individual mutants behave similarly when the community harbors one or four mutant libraries.
Supplemental Figure Legends

**Figure S1.** Distribution of Tn mutants in the genomes of each of the four *Bacteroides* strains included in the artificial human gut community. Key; Track 1 (innermost circle), GC skew with yellow denoting values greater than 0 and purple values less than 0; Track 2, histogram of the number of Tn insertions at each site in the targeted genome (normalized to one million INSeq reads from the library); Track 3, another depiction of the distribution of Tn insertions, showing genes with relative abundances >0.002% in the entire community; Track 4, all genes in the genome are plotted with those represented in polysaccharide utilization locus (PULs) colored green and those in capsular polysaccharide synthesis (CPS) loci colored red.

**Figure S2.** Estimating the precision of multi-taxon INSeq in identifying significant fitness determinants in the four *Bacteroides* strains. See Supplemental Methods and Supplemental Results for details of the *in silico* simulations.

**Figure S3.** Comparison of data obtained from gnotobiotic mice harboring different artificial communities and subjected to diet oscillations. (A) Experimental design showing community types in each treatment group and diet oscillation protocols. (B) PCoA analysis of Hellinger distances based on COPRO-Seq datasets generated from fecal samples collected from mice in this series of experiments and in the experiments described in Figure 2. Mean values ± SEM of the positions of all communities from all mice sampled at the end of the indicated diet period, on PC1 of the ordination plot are shown. PC1 explains 44% of the variance in the entire dataset. n.s., not significant based on 2-way ANOVA. (C) Mean values ± SEM for the positions of communities on PC1 based on Hellinger distance measurements of fecal meta-transcriptomes, defined by RNA-Seq assays of biospecimens collected at end of each diet treatment period. n.s., no significant difference (2-way ANOVA) in the meta-transcriptomes of fecal communities composed of all wild-type strains compared to sampled communities containing 14 wild-type strains and the *B. ovatus* INSeq library (n=5 mice harboring each community type; all mice in both groups fed the following sequence of diets; LF/HPP→HF/HS→LF/HPP; fecal samples were collected at the end
of the last diet period). (D) Correlation of gene expression levels for *B. ovatus* as a member of the 15 wild-type strain community or as a member of the 14-wild-type strain community containing the *B. ovatus* INSeq library. The same fecal samples used for the PCoA analysis shown in panel C were used for the Spearman correlation analysis plotted in this panel. Each circle represents a transcript from a gene. The x- and y-axes plot the number of reads (expressed as reads per kb per million reads, RPKM) for each gene. (E-G) Spearman correlation values of the abundances of Tn mutants for each targeted gene in the indicated strains in mutant libraries represented in the different community contexts indicated. Data are based on multi-taxon INSeq of fecal samples collected from animals at the end of the first diet phase shown in panel A of this figure and in Figure 2 (n=5 mice/treatment group). Each dot represents a targeted gene. The number of INSeq reads assigned to each gene with Tn insertions has been normalized per million reads/INSeq library.

**Figure S4. Changes in the relative abundances of Bacteroides strains in the artificial human gut community as a function of diet.** Fecal samples were collected from groups of mice harboring the 11 wild type strains and four Bacteroides INSeq libraries (n=5 individually housed mice/treatment group; mean values ± SEM, based on COPRO-Seq analysis, are plotted)

**Figure S5. Hellinger distances of the relative abundance of B. cellulosilyticus WH2 Tn mutants in fecal samples within and across the treatment groups.** Sampling occurred at 4, 10 and 16 days post gavage (dpg). For each panel, the first four columns represent the Hellinger distance (variation in the configuration of mutant population as judged from the relative abundances of its members) present in fecal samples collected from members of the indicated treatment group at the indicated time point. The next two columns represent the Hellinger distance between fecal samples in mice belonging to one of the monotonous diet treatment groups compared to respective diet oscillation group. The last column represents the Hellinger distance between fecal samples in mice belonging to the two monotonous diet treatment arms of the experiment. Mean values ± SEM are plotted. *** p<0.001, Student’s t-test with FDR correction; ns, not significant.
Figure S6. Consistent diet-specific configurations observed in the population of *B. ovatus* mutants. PCoA of Hellinger distance measurements based on the relative abundances of *B. ovatus* Tn mutants in fecal samples. Each colored circle represents the population of all mutants in *B. ovatus* sampled from an individual mouse in the indicated diet treatment group at a given time point. Each empty circle represents the average location of the mutant population along PC1/PC2 in a given treatment group at a given time point. Lines connect time points within a given group. The color of the lines indicates the diet. Dashed lines are used to delineate the diet oscillation groups.

Figure S7. Mapping core *Bacteroides* fitness determinants to amino acid biosynthetic pathways. Red arrows define steps in the indicated KEGG amino acid biosynthetic pathways. Enzymes catalyzing these steps are identified by their enzyme commission (EC) numbers. Genes encoding ECs identified as diet-independent *in vivo* fitness factors shared among the four *Bacteroides* strains are highlighted by grey boxes. Black arrows outline steps in the KEGG ‘Central Carbohydrate Metabolism Pathways’ that are linked to the indicated amino acid biosynthetic pathways.

Figure S8. Comparison between INSeq measurements of fitness indices and microbial RNA-Seq measurements of gene expression levels for the 86 identified core fitness determinants. Data are from *B. cellulosilyticus* WH2. Fecal samples were collected from mice that harbored the 11 wild-type strains and the four mutant libraries and were members of two monotonous diet treatment groups shown in Figure 2. INSeq or microbial RNA-Seq data from all mice in both groups (n=8 animals) were used for the analysis. In addition, microbial RNA-Seq data were generated from monocultures of *B. cellulosilyticus* WH2 at mid-log phase of growth in minimal medium containing either glucose, xylose, arabinose, or wheat arabinoxylan as the sole carbon source, or in TYG medium (n=3 biological replicates/growth condition). (A) Plot showing fold-difference in expression of the 86 core fitness determinants *in vivo* versus *in vitro* and their fitness indices *in vivo*. The locus designation for the *B. cellulosilyticus* homolog of each of identified 86 core fitness determinants is shown. *, Expression of the gene is significantly greater *in vivo* versus *in vitro* (p<0.05; Student’s t-test with FDR correction). (B) Microbial RNA-Seq showing expression
of three core fitness determinants encoding three transcription factors in vivo versus in vitro. n.s., not significant; ****, \( p < 0.0001 \) (Student’s t-test, FDR correction).

Figure S9. Strain-specific fitness determinants in B. thetaiotaomicron. (A) Annotation of genes in CPS4 and CPS6 in the two B. thetaiotaomicron strains. Homologs represented in the two CPS loci are connected using grey bars. (B) All but one of the 21 genes that comprise the CPS4 locus in B. thetaiotaomicron VPI-5482 are significant fitness determinants in both diet contexts. (C) All 18 genes in CPS6 of B. thetaiotaomicron 7330 are significant fitness factors in both diet contexts. Genes highlighted in red are unique to the respective strain. Mean values ± SEM for fitness indices are shown (n=5 individually caged mice/monotonous diet treatment group; fecal samples obtained at the end of the experiment outlined in Figure 2.

Figure S10. Effects of arabinoxylan on the Bacteroidetes strains present in the 15-member artificial community. (A) Microbial RNA-Seq documenting that expression of all genes in PUL BACWH2_4070-6 (except its linked HTCS regulator) is significantly greater when B. cellulosilyticus is cultured in a minimum medium containing xylan or arabinoxylan as the sole carbon source compared to minimal medium containing glucose, arabinose, or xylose. Read counts (expressed on log scale) were normalized using DESeq. ***, \( p < 0.001 \) [ANOVA with multiple test correction (Tukey-Kramer)]. (B) Comparison of growth characteristics of the indicated community members cultured under anaerobic conditions in minimal medium supplemented with 0.5% (wt/vol) of wheat arabinoxylan. Mean values ± SEM are shown for triplicate monocultures of each strain. Only B. cellulosilyticus and B. ovatus show significant growth under these conditions. (C) Fitness indices of genes in the B. ovatus arabinoxylan-induced PUL BACOVA_4389-94 as defined by multi-taxon INSeq analysis of fecal samples obtained from mice harboring the artificial community containing 11 wild-type strains and the four mutant libraries and monotonously fed either the LF/HPP or HF/HS diet (n=5 mice/treatment group; all animals sampled on the last day of the experiment shown in Figure 2). Mean values ± SEM are shown. Dashed line indicates the threshold cutoff level used to define a significant \( z \)-score.
Figure S11. The effects of arabinoxylan supplementation of the drinking water for 2 months in mice colonized with a consortium of 15 wild-type strains and fed the HF/HS diet. (A) Experimental design. (B) The relative abundances of *B. cellulosilyticus* WH2 and *B. ovatus* in fecal samples collected at 14 dpg were defined by COPRO-Seq. Mean values ± SEM are plotted. ***, p<0.001; n.s., not significant (Student’s t-test). (C) GC-MS analysis of short chain fatty acids (SCFA) and UPLC-MS analysis of bile acids in cecal contents harvested at the end of the experiment from mice in both treatment arms (n=5/group). *, p<0.05; **, p<0.01; ***, p<0.001 (Student’s t-test). Arabinoxylan supplementation produces an increase in the abundance of *B. cellulosilyticus* but not *B. ovatus* in the fecal microbiota, significant increases in cecal levels of all three SCFA assayed, significant decreases in conjugated bile acids (tauro-β-muricholic acid and taurocholic acid), plus significant increases in the deconjugated bile acids (β-muricholic acid, α-muricholic acid and cholic acid).

Figure S12. Arabinoxylan supplementation greatly impacts the fitness indices of genes in one but not both arabinoxylan-inducible *B. cellulosilyticus* PULs, *BACWH2_4070-4076* and *BACWH2_4044-4055*. Multi-taxon INSeq was performed on fecal samples obtained 14 dpg from mice belonging to groups A and B in Figure 4B (n=5/treatment arm). Mean values ± SEM are plotted. (A) For most genes in the *BACWH2_4070-4076* PUL, z-scores are significant on the HF/HS diet in the absence of arabinoxylan supplementation but not when mice receive drinking water supplemented with arabinoxylan. The exceptions are *BACWH2_4076* encoding the linked HTCS, and *BACWH2_4073* which specifies a hypothetical protein; these two genes retain their fitness effects under both diet conditions. In contrast, none of genes in the other arabinoxylan-inducible PUL, *BACWH2_4044-4055*, have a fitness effect whether or not mice receive arabinoxylan. (B) The effect of arabinoxylan on the fitness indices of genes that encode various membrane transporters.

Figure S13. Reproducibility of multi-taxon INSeq Illumina library preparation and sequencing protocols. Technical replicates were prepared and sequenced from each mutant library. Each
Figure S14. Sensitivity and reliability of INSeq library preparation and sequencing protocols. Technical replicates were prepared from each mutant library and subjected to the multi-taxon INSeq sequencing and analysis pipeline. Each point represents the abundance of insertions in a single gene. The coefficient of determination (R²) on log-transformed abundance values is indicated.
Supplemental Figures

Figure S1.
Figure S2.
Figure S3.

A. Feces (INSeq) and Feces (RNA-Seq) from different dietary groups:
- 15 WT strains
- 14 WT strains + B. ovatus::INSeq
- 14 WT strains + B. thetaiotaomicron VPI-5482::INSeq
- 14 WT strains + B. thetaiotaomicron 7330::INSeq

Legend:
- Orange circles: Feces (INSeq)
- Green line: LF/HPP diet
- Cyan circles: Feces (RNA-Seq)
- Red line: HF/HS diet

B. Principal Component (PC1) analysis:
- Diet 1
- Diet 2
- Diet 3

C. Comparison of principal components for 15 WT strains and 14 WT strains + B. ovatus::INSeq:
- ns

D. Scatter plot of B. ovatus::INSeq:
- Wild type strain correlation: R = 0.939

E. Scatter plot of 14 WT strains + 1 mutant library:
- R = 0.977

F. Scatter plot of 14 WT strains + B. thetaiotaomicron VPI-5482::INSeq:
- R = 0.973

G. Scatter plot of 14 WT strains + B. thetaiotaomicron 7330::INSeq:
- R = 0.913
Figure S4.

A. *B. cellulosilyticus WH2*

B. *B. ovatus*

C. *B. theta VPI-5482*

D. *B. theta 7330*
Figure S5.

A 4dpG

B 10dpG

C 16dpG

HF/HS vs (HF/HS → LF/HPP → HF/HS)

LF/HPP vs (LF/HPP → HF/HS → LF/HPP)

HF/HS vs LF/HPP

Legend:
- Orange: HF/HS
- Green: LF/HPP
- Pink: HF/HS → LF/HPP → HF/HS
- Blue: LF/HPP → HF/HS → LF/HPP
- Gray: HF/HS vs LF/HPP

Statistical significances:
- ns
- ***
Figure S6.

- PC1-Percent variation explained 40%
- PC2-Percent variation explained 19%

Legend:
- Green: LF/HPP
- Orange: HF/HS
- Blue: LF/HPP->HF/HS->LF/HPP
- Pink: HF/HS->LF/HPP->HF/HS
Biosynthesis Of Amino Acids
Figure S8.

A

Fitness index (z-score)  
Fold-Differences in expression  
(in-vivo / in-vitro)

-10 0 10 20 30 40 50 60

B

-10 0 10 20 30 40 50 60

in-vitro  
in-vivo

Normalized reads

An anti-sigma factor  
(BACWH2_4009)

A predicted global transcriptional regulator  
(Crp-like)

A RNA polymerase sigma factor  
(ECF-type)

****  
****

****  
****
Figure S9.

A

B. theta VPI-5482 CPS4

B. theta 7330 CPS6

upxY upxZ

Export protein KpsD Flippase Glycosyl transferase (GT) Dehydratease/Dehydrogenase
Epimerase Capsule Biosynthesis protein CDP-abequose synthase Reductase
Colanic acid biosynthesis glycosyl transferase WcaI Hypothetical protein

B

B. theta VPI-5482 CPS4

Fitness index (Z-score)

LT/HPP

HF/HS

Export protein

Transcription regulator

Polysaccharide biosynthesis

C

B. theta 7330 CPS6

Fitness index (Z-score)

LT/HPP

HF/HS

Export protein

Transcription regulator

Polysaccharide biosynthesis
Figure S10.

A. Fitness index (z-score) for B. ovatus against different categories of sugar molecules.

B. Growth curve of B. caccae ATCC43185T, B. thetaiotaomicron VPI-5482, B. thetaiotaomicron 7330, B. cellulosilyticus WH2, B. uniformis ATCC 8492, and Parabacteroides distasonis ATCC 8503.

C. Fitness index for B. ovatus ATCC 8483T against hybrid two-component system (C), carbohydrate esterase (CE), glycoside hydrolase (GH), SusD-like, SusC-like, and hypothetical genes.
Figure S11.

A

- B. cellulosilyticus WH2
- B. ovatus

Arabinoxylan

Percent relative abundance

0 10 20 30 40 50

Arabinoxylan - +

B

SCFA

Cecal bile acids

umol/g wet cecal content

ng/g wet cecal content

Acetate

Butyrate

Propionate

Tauro-muricholic acid sulfate

Tauro-β-muricholic acid

α-Muricholic acid

β-Muricholic acid

Cholic acid

Chenodeoxycholic acid

Arabinoxylan

AX

AX
Figure S12.

A. Bar chart showing the activity of different enzymes under various conditions. The x-axis represents the different treatments, while the y-axis shows the fitness index (z-score).

B. Table listing genes and their functions under different conditions.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>BACWH2_4083</td>
<td>MFS, Nucleoside H+ symporter</td>
</tr>
<tr>
<td>BACWH2_4968</td>
<td>ABC-type sugar transporter</td>
</tr>
<tr>
<td>BACWH2_0829</td>
<td>MFS, L-fucose transporter</td>
</tr>
<tr>
<td>BACWH2_1233</td>
<td>MFS, L-fucose transporter</td>
</tr>
<tr>
<td>BACWH2_0056</td>
<td>Fatty Acyl CoA Transporter</td>
</tr>
<tr>
<td>BACWH2_0334</td>
<td>iron complex transport system ATP-binding protein</td>
</tr>
<tr>
<td>BACWH2_0214</td>
<td>DUF3875 superfamily</td>
</tr>
<tr>
<td>BACWH2_2745</td>
<td>zinc transport system ATP-binding protein</td>
</tr>
<tr>
<td>BACWH2_4802</td>
<td>phosphate transport system ATP-binding protein</td>
</tr>
<tr>
<td>BACWH2_2688</td>
<td>FtsX-like permease family</td>
</tr>
<tr>
<td>BACWH2_1403</td>
<td>iron complex transport system ATP-binding protein</td>
</tr>
<tr>
<td>BACWH2_3636</td>
<td>MFS, drug resistance transporter</td>
</tr>
<tr>
<td>BACWH2_3328</td>
<td>ABC-type multidrug transport system ATP-binding protein</td>
</tr>
<tr>
<td>BACWH2_1163</td>
<td>ABC-type antimicrobial peptide transport system, ATP-binding protein</td>
</tr>
<tr>
<td>BACWH2_0605</td>
<td>ABC-type antimicrobial peptide transport system, ATP-binding protein</td>
</tr>
<tr>
<td>BACWH2_2086</td>
<td>ABC-type antimicrobial peptide transport system, ATP-binding protein</td>
</tr>
<tr>
<td>BACWH2_2746</td>
<td>ABC-type metal ion transport system, periplasmic component/surface adhesin</td>
</tr>
<tr>
<td>BACWH2_2088</td>
<td>ABC transporter system for protein secretion</td>
</tr>
<tr>
<td>BACOVA_1376</td>
<td>MFS, multidrug resistance protein B</td>
</tr>
<tr>
<td>BACOVA_2585</td>
<td>ABC-type antimicrobial peptide transport system, permease component</td>
</tr>
<tr>
<td>BACOVA_3833</td>
<td>MFS transporter, multidrug resistance protein B</td>
</tr>
</tbody>
</table>
Figure S13.

A. B. thetaiotaomicron VPI-5482::INSeq

B. thetaiotaomicron 7330::INSeq

C. B. cellulosilyticus WH2::INSeq

D. B. ovatus::INSeq

Correlation coefficients:

Replicate 1 Replicate 2
R²=0.972 R²=0.995

Replicate 1 Replicate 2
R²=0.996 R²=0.997
Figure S14.

A. B. thetaiotaomicron VPI-5482::INSeq

B. B. thetaiotaomicron 7330::INSeq

C. B. cellulosilyticus WH2::INSeq

D. B. ovatus::INSeq

Mutant library represented 1%

Mutant library represented 100%

Read counts/million

Mutant library represented 1%

Mutant library represented 100%

Read counts/million

R² = 0.996

R² = 0.990

R² = 0.976

R² = 0.992
Supplemental Tables

Table S1. CAZymes encoded in the genomes of the four Bacteroides strains subjected to whole genome transposon mutagenesis.

Table S2. INSeq library characteristics and validation of the multi-taxon INSeq pipeline. (A) Characteristics of the four Bacteroides INSeq libraries. (B) Application of multi-taxon INSeq to mock communities.

Table S3. COPRO-Seq analysis of fecal microbiota sampled over time from mice belonging each treatment group. (A) Mice described in Figure 2 harboring the artificial community containing 11 wild-type strains plus four mutant libraries. (B) Mice described in Figure S3 harboring either a 15-member community composed of all wild-type strains or one of three mutant libraries containing 14 wild-type strains and one mutant library.

Table S4. Diet-independent and diet-specific fitness determinants identified in B. cellulosilyticus and B. ovatus. (A) B. cellulosilyticus WH2. (B) B. ovatus.

Table S5. A core group of fitness determinants identified by multi-taxon INSeq as being shared among the four Bacteroides strains.

Table S6. Fitness determinants identified in B. cellulosilyticus WH2 at stationary phase after growth in TYG medium or in minimal medium containing single carbon sources.

Table S7. Comparison of B. thetaiotaomicron VPI-5482 and 7330 genomes. (A) Summary of the total number of genes assigned to KEGG categories, KEGG pathways. (B) PULs identified in the 7330 strain.

Table S8. Diet-independent and diet-specific fitness determinants identified in the two Bacteroides thetaiotaomicron strains. (A) B. thetaiotaomicron VPI-5482. (B) B. thetaiotaomicron 7330.
Table S9. Fitness determinants in *B. cellulosilyticus* WH2 and *B. ovatus* when mice are consuming a HF/HS diet with or without arabinoxylan supplementation. (A) *B. cellulosilyticus* WH2. (B) *B. ovatus*

Table S10. Primers and adapters used for the indicated steps of the multi-taxon INSeq procedure. (A) Vector construction. (B) Preparation of Illumina sequencing library. (C) Illumina index sequencing.

Table S11. Media used for bacterial culture in this study. (A) MEGAMedium 1.0. (B) Minimal Medium.
Chapter 4

Future directions
Creation of clonally arrayed multi-taxon INSeq transposon mutant libraries

After identifying “core” fitness genes, species/strain-specific and diet-specific fitness determinants from the studies described in Chapter 3, a natural move forward is to further study specific mutant strains of interest in isolation, after their recovery from the mixed population of mutants used for multi-taxon INSeq. However, the creation of individually targeted mutants by molecular cloning would be laborious and time-consuming. INSeq allows the array of the transposon mutant library into a collection of isolated individual mutants in multi-well plates, then simultaneously map thousands of individually archived transposon mutant strains in parallel to define their genetic background [1, 2]. Applying this method to create clonally arrayed transposon mutant libraries for the multi-taxon INSeq libraries generated in the study described in Chapter 3 would dramatically expand our ability to perform mechanistic follow-up studies, as well as confirm the effects of specific mutations on the host; for example, we could perform direct competition of a collection of mutants against the wild-type parental strain in gnotobiotic mice under defined diet conditions to validate our INSeq results. Moreover, the study of mutants in isolation would facilitate further dissection of the precise mechanisms by which a particular gene mediated the fitness of a strain of interest.

There are several fitness determinants identified from the multi-taxon INSeq that are of particular interest. One group includes two transcriptional regulators identified as core fitness determinants in *Bacteroides*. One of these factors is a homolog of BT4338, a putative Crp-like transcription factor that controls a proposed global sugar catabolic regulon [3]. By retrieving the individual mutant(s) from an arrayed INSeq library, we can compare their growth phenotypes to wild-type *in vitro* under different nutrient conditions: e.g., minimal medium with arabinose, xylose, fucose and other carbon sources selected based on metabolic predictions made from the *in vivo* INSeq data. This, combined with transcriptional analysis to define the regulon affected by this transcription factor, could provide a better map of the metabolic pathways affected by the regulator, and offer insights into their importance for fitness in the gut environment. Homologs of BT4720, an extracytoplasmic function (ECF)-type sigma factor is another core *Bacteroides* fitness determinant of interest since expression of this factor is dramatically upregulated (~55 fold
increase) \textit{in vivo} compared to \textit{in vitro} growth. ECF-type sigma factors regulate PULs involved in glycan degradation and capsular polysaccharides biosynthetic loci \cite{4}. By retrieving strains containing mutations in this gene, we could identify the genes/pathways regulated by this core fitness determinant in order to gain new insights about critical genes/pathways that allow \textit{Bacteroides} spp to survive in the gut.

Another set of genes that would be interesting to further characterize are represented in two related CPS loci identified as strain-specific fitness determinants in \textit{Bacteroides thetaiotaomicron}. The ability to synthesize multiple capsular polysaccharides is likely an important biological property for long-term survival of \textit{Bacteroides} in a complex competitive intestinal ecosystem that is under selective pressure from the immune system \cite{5-7}. We have shown that for the two strains of \textit{Bacteroides thetaiotaomicron} examined using multi-taxon INSeq, only one CPS locus per strain was important for fitness regardless of which of the two diets were administered. Comparative genomic analysis revealed these two loci are in synteny in terms of their genomic locations in the two \textit{Bacteroides thetaiotaomicron} strains, and that they share homologous transcriptional regulators. However, the genes involved in the assembly of monosaccharides into polymers are different between the two loci, suggesting that the capsules they synthesize have different structures. To study this, capsular polysaccharides could be extracted from the wild-types and their CPS mutants and the monosaccharide composition of their extracellular polysaccharides analyzed by high pH anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD) after acid hydrolysis \cite{8}. Retrieving and characterizing Tn mutants in the transcriptional regulators contained within these loci could identify how they (and perhaps other loci) are regulated by these factors, whether these patterns of regulation are shared between the two \textit{B. thetaiotaomicron} strains, and shed light on why these genes are such important fitness determinants in the gut environment.
The effects of the arabinoxylan on interactions between the artificial community and host in the context of the HF/HS diet

The ‘spread’ of diets high in fat and protein but low in fruits and vegetables, may be associated with the increase in the incidence of diseases such as metabolic syndrome, inflammatory bowel disease (IBD), and cancer in countries undergoing ‘Westernization’. Epidemiologic studies of IBD in China, Korea and Puerto Rico have documented that the incidence of IBD is rising in these areas [9-11]: a systematic review of nineteen studies, encompassing 2,609 IBD patients and over 4,000 controls revealed a positive association between high intake of total fat, mono- and disaccharides, and meat and increased risk of IBD, while high fiber and fruit/vegetable intake were associated with decreased IBD risk [12]. However, much more work is needed to define the mechanisms that underlie the deleterious effects of Western diets.

In the studies described in Chapter 3, when arabinoxylan was added to the drinking water of mice consuming a HF/HS diet, we observed a diminished requirement for fucose transporters in *B. cellulosilyticus* WH2, which remained at a relatively high relative abundance, but not in *B. ovatus*, which became almost undetectable in the community on HF/HS diet. Fucose cleavage from mucosal fucosylated glycans is an important nutrient source for *Bacteroides* spp. when dietary fibers are deficient, such as in the case of the HF/HS diet [13-16]. The differential requirement for fucose transporter genes between these two species may be related to the capacities of these *Bacteroides* species to adaptively forage host glycans, and their fitness on the HF/HS diet. There was also a diminished requirement for anti-microbial peptide and other multidrug transporters in both *B. cellulosilyticus* WH2 and *B. ovatus*. The shared requirement for multidrug and anti-microbial peptide transporters raises the question of whether consumption of HF/HS diet changes interactions between the host (notably components of its gut mucosal barrier) and these members of the microbiota. These interactions include the physical juxtaposition of the organisms with the mucus that overlies the gut epithelium.
Although we identified arabinoxylan-associated changes in the overall abundance of *B. cellulosilyticus* WH2, we did so using fecal samples; we did not interrogate abundance along the length of the gut nor did we examine the physical location of the organism across the width of the intestine. Techniques providing information regarding the spatial organization of microbial cells in the gut thus need to be applied to decipher whether arabinoxylan shapes localization as well as abundance of members of the gut microbiota. Fluorescence *in situ* hybridization (FISH), a commonly used technique to address such issues, is limited in its ability to characterize the locations of members of complex microbial communities by the number of fluorophores that can be differentiated simultaneously. As a result, these studies typically employ “universal” probes that target most or all bacteria, or probes designed to label taxonomically-related species; In a collaboration between the Gordon lab and Gary Borisy lab, a new high resolution imaging technique, ‘combinational labeling and spectral imaging fluorescence *in situ* hybridization’ (CLASI-FISH; [17, 18]), was applied to define the spatial structure of defined model human gut communities in gnotobiotic mice. This method allows us to visualize at least 15 different species simultaneously, with the potential to possibly visualize as many as hundreds of species. By applying CLASI-FISH to gut tissue sections prepared from mice colonized with the 15-member community and fed the HF/HS diet with/without arabinoxylan supplementation, we expect to be able to observe how arabinoxylan changes the biogeographical features of community members. A testable hypothesis is *B. cellulosilyticus* WH2 is more dominant in the mucus layer when mice are consuming HF/HS diet without arabinoxylan due to their ability and need to consume host glycans.

If we observe greater representation of *B. cellulosilyticus* in the inner mucus layer when arabinoxylan is not included in the drinking water (or diminution in the thickness of the mucus layer as a result of adaptive foraging of mucus glycans), we could further characterize barrier function by using FACS to define the microbes that are targeted by the IgA response, and as well as expression of epithelial cell-derived antibacterial markers (i.e., α-defensins, *Reg3g* (RegIIIg), *Pla2g2a* (phospholipase A2) using laser capture microdissection of epithelial cells and host RNA-Seq.
Application of multi-taxon INSeq into other model human gut symbionts

The results we obtained from studies with gnotobiotic mice show how multi-taxon INSeq provides a simultaneous readout of the genetic underpinnings of the niches of multiple members represented in a microbial community inhabiting the gut ecosystem as a function of dietary perturbations. This approach could be applied into other models where we are interested in identifying potential mechanisms to allow for targeted changes in the structure and function of the microbiota.

A series of studies in the Gordon lab has been focused on examining interrelationships between obesity and gut microbiota. The microbiota of obese humans have reduced bacterial diversity compared to lean individuals [19] as well as reduced gene diversity [20, 21]. Transplanting the intact uncultured or the culturable component of the fecal microbiota obtained from twin pairs stably discordant for obesity into groups of germ-free mice can recapitulate the composition of the donor’s communities and differences in body composition [22]. Co-housing of mice harboring the transplanted “lean” microbiota prevented development of an increased adiposity and body mass phenotype in the mice transplanted with the microbiota of an obese donor, and transformed the metabolic profile of the resulting microbiota into a lean-like state. This process was associated with invasion of members of Bacteroidales (including B. cellulosilyticus) from the lean community into the guts of cagemates harboring the obese donor microbiota. Invasion and phenotypic rescue were diet-dependent, occurring with the diet representing low in fat and high in fruits and vegetables, but not with the diet high in fat and low in fruits and vegetables. By applying multi-taxon INSeq to the invading “lean” Bacteroidales spp. and identifying the key fitness determinants that allows them to establish within “obese” communities, we could identify key nutrients that support this process. This application of multi-taxon INSeq follows the paradigm established in Chapter 3, in which we identified one PUL important for B. cellulosilyticus WH2 on the HF/HS diet, and then used the predicted carbohydrate preference of this PUL in order to design a dietary supplement that specifically boosted the levels of B. cellulosilyticus WH2 on HF/HS by 3-fold. Multi-taxon INSeq is a particularly attractive technique for studying the rescue of the obesity phenotype by Bacteroides observed by Ridaura, et al, since the majority of the invasion species are Bacteroides and
B. cellulosilyticus is one of the leading invasive taxa. By assessing the genes which are essential to allow members of the Bacteroides to invade ‘obese’ communities, several strategies could be designed to promote growth of bacteria that help to confer a lean state.

Finally, many additional taxa within the gut microbiota have been implicated as effectors of host phenotypes. For example, Faecalibacterium prausnitzii and other Clostridia strains have anti-inflammatory effects [23, 24]; Ruminococcus obeum restricts the virulence of the diarrhea-inducing pathogen Vibrio cholerae [25]; Bifidobacterium longum, an efficient utilizer of milk oligosaccharides commonly incorporated into probiotic formulations, improves lactose tolerance and prevents food allergies [26, 27]. The ability of INSeq to be adapted to any bacterium where we are able to express a functional mariner transposase and a selectable marker means that our multi-taxon INSeq pipeline have broad utility in identifying bacterial genes that mediate many processes, opening up new avenues for the exploration of the role of the microbiota in host health and disease.
References


Appendices

Appendix A

Andrew L. Goodman, Meng Wu, and Jeffrey I. Gordon,

Identifying microbial fitness determinants by insertion sequencing using genome-wide transposon mutant libraries.

Identifying microbial fitness determinants by insertion sequencing using genome-wide transposon mutant libraries

Andrew L. Goodman¹,², Meng Wu¹ & Jeffrey I Gordon¹

¹Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri, USA. ²Present address: Section of Microbial Pathogenesis and Microbial Diversity Institute, Yale School of Medicine, New Haven, Connecticut, USA (A.L.G.). Correspondence should be addressed to A.L.G. (andrew.goodman@yale.edu) or J.I.G. (jgordon@wustl.edu).

Published online 17 November 2011; doi:10.1038/nprot.2011.417

Insertion sequencing (INSeq) is a method for determining the insertion site and relative abundance of large numbers of transposon mutants in a mixed population of isogenic mutants of a sequenced microbial species. INSeq is based on a modified mariner transposon containing MmeI sites at its ends, allowing cleavage at chromosomal sites 16–17 bp from the inserted transposon. Genomic regions adjacent to the transposons are amplified by linear PCR with a biotinylated primer. Products are bound to magnetic beads, digested with MmeI and barcoded with sample-specific linkers appended to each restriction fragment. After limited PCR amplification, fragments are sequenced using a high-throughput instrument. The sequence of each read can be used to map the location of a transposon in the genome. Read count measures the relative abundance of that mutant in the population. Solid-phase library preparation makes this protocol rapid (18 h), easy to scale up, amenable to automation and useful for a variety of samples. A protocol for characterizing libraries of transposon mutant strains clonally arrayed in a multiwell format is provided.

INTRODUCTION

Development of the protocol

By exposing populations of transposon mutants to selective conditions in vitro or in vivo, it is possible to identify changes in the representation of mutants, thereby allowing identification of genes and pathways that are key to fitness under the conditions being examined. In principle, this approach could be applied to any microbe whose genome has been sequenced and that is amenable to basic genetic manipulation (insertion of foreign DNA, antibiotic selection). INSeq combines existing methods for signature-tagged mutagenesis¹ with new techniques for high-throughput DNA sequencing such that the precise location and relative abundance of tens of thousands of mutants can be determined in parallel. This approach is based on the modification of mariner, the broad host-range, randomly integrating, transposable element, to allow short fragments of genomic DNA adjacent to the site of transposon insertion to be captured, sequenced and quantified². Mutants that decrease in relative abundance in a selective condition are likely to be in genes that are important for fitness under that condition; mutants that increase in abundance under the selection high-light genes that could be deleterious. The protocol described here includes substantial improvements from our earlier publication³; because library preparation can now be performed in 96-well, bead-based format, reagent costs and sample losses are minimized, and large numbers of samples can be barcoded and sequenced in a single sequencing run.

Applications of the method

As INSeq library preparation can be completed in multiplex format, experiments can involve material obtained from high-resolution time-series studies of the effect of in vitro or in vivo selections, and from multiple biological and technical replicates. Any microbe whose genome has been sequenced and that is amenable to mariner transposon mutagenesis could potentially be investigated with INSeq. This method can also be used to efficiently determine the insertion site of transposon mutant strains that have been clonally arrayed in a multiwell format (Box 1). This latter technique relies on a combinatorial pooling approach: a small number of pools are defined, and a liquid-handling robot is used to place each mutant strain into a subset of these pools in a unique pattern. The pools are then characterized by INSeq, and the presence/absence pattern of a given insertion site across the pools reflects the unique pattern associated with the corresponding bacterial strain in the multiwell collection. Alternative pooling patterns that use larger numbers of pools to further reduce the likelihood of mistaking one strain for another can also be considered⁴–⁵. These alternatives would be appropriate if many mutants in the population were likely to carry insertions at the same location in the genome, but they require greater effort in preparing libraries because of the increased number of pools. Furthermore, these alternate techniques were not developed specifically for mapping transposon insertions and may require modification in order to accomplish this goal.

Limitations and alternative techniques

In bacterial genomes, a high percentage (>90%) of insertions can be precisely localized using INSeq. With more complex organisms, the short (16–17 bp) fragments of adjacent chromosomal sequence captured and quantified by MmeI digestion may be insufficient to uniquely identify transposon locations. As with any negative selection technique, stochastic bottlenecks in the selective process need to be considered when determining the optimal number of mutants to be included in an experiment. Because each insertion is mapped with nucleotide-level resolution, comparing the behavior of multiple mutants in the same gene highlights these selective bottlenecks.

Other techniques for mapping and quantifying transposon insertions by high-throughput sequencing are also available⁶–¹⁰. They include two protocols named Tn-Seq: one version⁶–¹⁰ also uses an
Box 1 | Combinatorial pooling and mapping of clonally arrayed transposon mutant libraries by INSeq ● TIMING ~1.6 h per 96-well tray

Timing estimates assume the use of an EpMotion liquid-handling robot. Additional time is required for preparing the arrayed mutant collection by manually or robotically picking transposon-mutagenized colonies into 96-well trays and conducting INSeq library preparation on the pooled strains.

1. Prepare EpMotion programs. Two sets of pooling patterns are provided in the analysis package INSeq_analysis/Arrayed_library directory (Supplementary Data 1): ‘16384_strings.txt’ is the list of 24-bit strings with a minimum Hamming distance of 6; and ‘13000_strings.txt’ is the subset of these that are less likely to be mistaken for each other if multiple archived strains happen to transposate at the identical genome coordinate. The larger set can be used to map up to 170 96-well trays, whereas the smaller set is preferable for arrayed libraries of under 135 trays. A script called write_dws.pl translates these pooling patterns into programs for the EpMotion robot. Each program will pool 5 trays (20 programs are needed for a 100-tray mutant library, for example). To create EpMotion .dws files, go to the data analysis package INSeq_analysis/Arrayed_library directory (Supplementary Data 1) and type:

```
perl write_dws.pl <pooling pattern> <number of programs>
```

This will create an output directory called ‘Write_dws_output_X_programs’. Create a new application on the EpMotion, and then import the .dws files in this directory into the new application and confirm that they are functioning properly. (Note that the provided script is designed for Midwest Scientific 96-well culture trays (cat. no. TP92696) and the EpMotion Thermorack for 24 cryotubes (cat. no. 960002491). Alternatives may require changes to the program). The write_dws.pl script will also create a map file that will be used in the data analysis step. Although the software provided is designed for the EpMotion instrument, analogous programs could be written for other liquid-handling robots if they can be run from a text-format file.

2. Prepare the mutant collection by manually or robotically picking transposon-mutagenized colonies into 96-well trays containing 250 µl of culture medium and appropriate antibiotics. Incubate under standard conditions to allow cultures to grow to turbidity, and note any wells with poor or no growth.

3. For each 96-well culture tray, prepare two 96-well archive trays by adding 30 µl of sterile culture medium plus 40% (vol/vol) glycerol to each well. Transfer 30 µl of the turbid culture to the corresponding well in each of these archive trays, mix well and seal with foil lid before storing archive trays at ~80 °C. Turbid cultures may need to be mixed before dispensing into archive trays. To recover strains from the archive trays, wipe down the foil lid with 70% (vol/vol) ethanol and puncture the well of interest with a sterile pipette tip. Draw a small amount of frozen material onto the tip of the pipette without allowing the entire plate to thaw. Restreak this material onto agar plates, and place a small foil patch over the targeted well in the archive tray before returning it to the −80 °C freezer.

4. Seal culture trays in groups of five with Parafilm and store at 4 °C.

5. On the EpMotion, initiate the Pool_group_1.dws program. Set up the robot deck as shown in the graphical interface with tips, input culture trays nos. 1–5 and a cryovial rack containing 24 2-ml screw-cap cryovials (lids removed and vials labeled with program and pool number). In the cryovial rack, tubes in the top row should be labeled 1–6, the tubes in the second row 7–12, etc. While the program is running, monitor the waste tip bucket to ensure that it does not overflow. When the program is complete, cap the 2-ml output cryovials and store them at −20 °C.

▲ CRITICAL STEP Each program takes ~8 h to complete. Consider adding a growth inhibitor if your mutants are likely to grow during this process. If the robot is performing without interventions required, programs can be run overnight. Also, if input trays have been stored at 4 °C, cultures may need to be resuspended manually before being placed in the robot. See Supplementary Movie 2 for a demonstration of the pooling procedure.

6. Run pooling programs on each subsequent set of five culture trays as described in Step 5.

7. When pooling is complete, thaw all cryovials, vortex well and combine half of the volume of each cryovial with same number from each pooling program to create 24 final pools. Regardless of the number of programs run, there should be 24 pools at this point. Save the remainder of each cryovial at −20 °C.

▲ CRITICAL STEP Be sure to combine like pools and not to combine different pools from a single program.

8. Prepare INSeq libraries from each of the 24 pools (Steps 1–78), associating a different barcode with each pool.

9. Run the data analysis pipeline as described in the arrayed library section of the README.txt included in the analysis package.

▲ CRITICAL STEP The cutoffs for presence or absence of a strain in a pool may need to be determined empirically by examining the distribution of reads across pools for a given data set. As a starting point, normalize counts per pool for each strain by the overall number of reads in the pool; for each strain, set the presence or absence of a pool to 0 if there are less than three reads associated with that pool and to 1 if there are three or more reads associated with the pool.

Mmel-adapted mariner transposon but does not include a transposon-specific, linear PCR step during library preparation. Compared with INSeq, this simpler approach to library preparation could reduce the effort required for sequencing. The other Tn-Seq protocol does not use an Mmel-adapted mariner transposon; instead, it uses a circularization step to capture the targeted DNA fragments. This could allow the use of transposons that cannot be modified to include Mmel sites in their inverted repeats. Additional approaches, termed ‘high-throughput insertion tracking by deep sequencing (HITS)” or ‘transposon-directed insertion-site sequencing (TraDIS)” also do not require Mmel sites in the transposon. For these protocols, mutagenized DNA is randomly sheared, adapters are ligated to all fragments, and transposon-chromosomal junctions are enriched by PCR either with (HITS) or without (TraDIS)
an affinity-purification step. In principle, these Mmel-independent methods could allow the capture of longer fragments of genomic sequence adjacent to each transposon, a feature which may be necessary for mapping insertions in more complex genomes.

The INSeq protocol we describe below has several advantages compared with these other methods: (i) introducing an Mmel site in the transposon allows the excised transposon/chromosomal junctions to be of uniform size, thus avoiding size preferences during PCR; (ii) the linear PCR step at the beginning of the protocol enriches the target and increases the efficiency of library preparation; (iii) the solid-phase-based technique separates target sequences from chromosomal DNA background, making it easy to scale up to a 96-well plate format; (iv) robotic automation makes the protocol amenable to monitoring changes in the representation of mutants during studies that require repeated, high-resolution sampling; and (v) descriptions for modifying the INSeq protocol to map arrayed mutant libraries are provided in Box 1. If other transposons that cannot be modified with Mmel sites are used, alternative protocols such as the Tn-seq method described by Gallagher et al., HITS or TraDIS are preferable.

Experimental design

Overview. An overview of the INSeq protocol is provided in Figure 1. Steps 1–16 describe the procedure used for extraction of DNA. In Steps 17–20, DNA is further purified using columns. Steps 21–23 describe a linear PCR procedure that is designed to enrich the transposon/chromosomal junction regions by using a primer complementary to the transposon-specific sequence. Steps 21–23 are necessary for two reasons. First, they add a biotin tag to each target molecule, which facilitates high-throughput, low-volume solid-phase library preparation. Second, the linear PCR enriches for the desired target molecules early in the procedure, minimizing the amount and cost of enzymes needed for subsequent reaction steps. In Steps 24–36, the linear PCR product is bound to magnetic beads, allowing the PCR product to be separated from background chromosomal DNA by affinity capture. Second strand synthesis (and creation of the double-stranded Mmel site) takes place in Steps 37–43; because Mmel requires two sites for efficient DNA cleavage, a second Mmel recognition site is provided in trans via the double-stranded DNA (dsDNA) fragment M12. Steps 44–52 use Mmel to cut the dsDNA fragment to a uniform length (64–65 bp). The product of this reaction consists of the amplified transposon region generated by the linear PCR, plus 16–17 bp of chromosomal DNA. In Steps 53–63, sequencing adapters are ligated to the transposon/chromosomal junction to introduce the complementary sequence for the sequencing primer used by the Illumina Genome Analyzer/HiSeq instrument. By varying the four-nucleotide barcode sequences in the adapter, multiple samples can be sequenced in a single lane of an Illumina flow cell. With limited cycles of PCR (Steps 64–78), targets are enriched in the linear range. PCR products are then purified by gel extraction to remove extra primers and adapters, and they are then sequenced. Data analysis is described in Steps 79–86; detailed documentation of data analysis tools is available in the README.txt in the INSeq_analysis package (Supplementary Data 1). Box 1 describes an additional protocol for combinatorial pooling and mapping of clonally arrayed transposon mutant libraries by INSeq.

Starting material. Steps 1–16 describe the procedure used for extraction of DNA from the transposon library (the ‘input mutant population’) and from biological specimens containing mutants that survive a given selection (the ‘output mutant population’). These steps for DNA isolation were originally developed for mouse feces and for samples obtained from the more-proximal regions of the mouse gastrointestinal tract such as the cecum, and can also be used for in vitro mutant populations. Transposon mutant libraries can be introduced into a wide variety of biological systems (e.g., various animal or environmental habitats) or characterized in many different in vitro formats (e.g., chemostats, tissue culture). In these cases, users should isolate crude DNA using protocols designed for the specific system chosen for study. Whatever the system, experiments need to be designed so that the amount of input and output material is sufficient for library preparation. For example, if the mutagenized species constitutes 100% of a microbial community, then the initial purification steps as described in the PROCEDURE ideally should yield >500 ng of DNA. If the mutagenized species represents only 50% of the population, then the amount of starting material should be doubled.

Oligonucleotide sequence design. The linear PCR primer BioSamA is 5′-biotinylated and encodes a TEG spacer to reduce steric hindrance during the solid-phase enzymatic steps. The primer also encodes a transposon-specific region with a 26-bp sequence complementary to sequences found on both sides of the transposon (underlined in Supplementary Table 1). Adapters encode 4-bp barcodes (marked in bold in Supplementary Table 1) that allow
sample multiplexing and produce a 2-bp NN overhang (complementary to the overhang produced by Mmel digestion) for ligation. For each adapter, two complementary oligonucleotide sequences are provided in Supplementary Table 1: LIB_AdaptT_x and LIB_AdaptB_x, where x is the unique barcode specified by the adapter). These pairs of oligonucleotide sequences are annealed to each other (AdaptB_x, where x is the unique barcode specified by the adapter). Replicates and controls. Biological and technical replicates should be performed. The number of biological replicates required depends on the experimental system under study and should be determined empirically. Technical replicates can include identical samples associated with different sample-specific barcodes. The number of technical replicates should be determined by the user; two per biological sample is likely to be sufficient. No-DNA negative controls will help identify cross-contamination between samples; these controls should be run in parallel with DNA-containing samples. Cross-contamination can occur if beads transfer from tube lids to the users’ gloves to other samples: this can be avoided with careful technique.

MATERIALS

REAGENTS

- Phenol:chloroform:isoamyl alcohol (25:24:1), pH 7.9 (Applied Biosystems, cat. no. AM9712) • CAUTION This mixture is hazardous; handle properly during experiments.
- Sodium acetate (3 M), pH 5.5 (Applied Biosystems, cat. no. AM9740)
- Isopropanol (100%, Fisher Scientific, cat. no. A416-4) • CAUTION Isopropanol is flammable; handle properly during experiments.
- Ethanol (100%, Pharmco-AAPER, cat. no. 111AC5200) • CAUTION Ethanol is flammable; handle properly during experiments.
- TE buffer (10 mM Tris–HCl, 1 mM EDTA, pH 7; Ambion, cat. no. AM9861)
- NaCl (NaCl powder, Fisher Scientific, cat. no. S271-10)
- Tris base (Trizma base, Sigma-Aldrich, cat. no. T1503-1KG)
- EDTA (EDTA powder, Sigma-Aldrich, cat. no. S657-500)
- RNase A (Qiangen, cat. no. 19101)
- QIAquick PCR purification kit (Qiagen, cat. no. 28106)
- Platinum PtDNA polymerase (Invitrogen, cat. no. 11708-021)
- Deoxyribonucleotide triphosphates (dNTPs; 10 mM, Invitrogen, cat. no. 18427-013)
- Dynabeads M-280 streptavidin (Invitrogen, cat. no. 112-05D)
- Klenow fragment (exo-; 5 U) (New England Biolabs, cat. no. M0202T)
- T4 DNA ligase (2,000,000 U ml⁻¹; New England Biolabs, cat. no. M0212T)
- Agarose (Roche, cat. no. 11388911-001)
- B+W buffer (2×)
- Glycerol (J.T. Baker, cat. no. 2136-01)
- Magnesium chloride (MgCl₂)

For the arrayed library mapping protocol given in Box 1 only
- Glyceraldehyde (G.T. Baker, cat. no. A0186-01)

EQUIPMENT

- Electronic 200-µl multichannel pipettor with variable speed control
- Heat block or water bath set to 50 °C
- Refrigerated microcentrifuge
- Magnetic particle concentrator (MPC) for 1.5-ml tubes (Invitrogen, cat. no. 123-21D)
- MPC for PCR tubes
- Electronic 200-µl eight-channel pipettor with variable speed control
- Heated oven or water bath set to 50 °C
- Vacuum evaporator
- Thermocycler capable of a ramp rate of 0.1 °C s⁻¹
- Electrophoresis equipment for agarose gels
- Non-UV gel illuminator
- Invitrogen QuBit DNA spectrophotometer (or similar, for DNA concentrations ≤ 1 ng µl⁻¹)
- DNA sequencing facility with Illumina Genome Analyzer/HSeq instrument

For the arrayed library mapping protocol given in Box 1 only
- EpMotion 5075PC liquid-handling robot with TS50 single-channel dispensing tool and rack for 24 cryotubes • CRITICAL Eppendorf, cat. no. 960002491 was used, but temperature control is not necessary. Note that with the EpMotion, each set of five plates in the arrayed library takes ~8 h to pool. Higher-throughput liquid-handling robots are likely to be faster.
- EpMotion tips (50-µl tips, nonfiltered, 1 × 96 tips per culture tray (Eppendorf, cat. no. 96005200)
- Culture trays, 96 well (BD Falcon, cat. no. 354550)
- 1× LoTE buffer (2×) Combine 3 mM Tris and 0.2 mM EDTA; adjust pH to 7.5 with concentrated HCl.
- 2× Buffer A (2×) Buffer A is prepared using 200 mM NaCl, 200 mM Tris, and 20 mM EDTA; pH is adjusted to 8 with concentrated HCl • CRITICAL This and all other buffers can be stored for months at room temperature (20–25 °C).
- 4× Buffer B (2×) Buffer B is prepared using 2 M NaCl, 10 mM Tris, and 1 mM EDTA; pH is adjusted to 7.5 with concentrated HCl. Buffer B can be run on a computational cluster or a desktop/laptop computer under a UNIX system with 4 GB memory, perl and Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) installed. Basic understanding of Unix and perl are required for data analysis.
- 5× Buffer C (2×) Buffer C is prepared using 250 mM NaOH in dH₂O; pH is adjusted to 8–8.5 with boric acid.
- 10× Buffer D (2×) Xylene cyanol DNA dye (10×) Dye is prepared using 20% (wt/vol) Ficoll 400, 0.1 M disodium EDTA and 0.25% (wt/vol) xylene cyanol.
- 10× Buffer E (2×) Bromophenol blue DNA dye (10×) Dye is prepared using 20% (wt/vol) Ficoll 400, 0.1 M disodium EDTA and 0.25% (wt/vol) bromophenol blue.
**PROCEDURE**

Isolation of crude DNA ● **TIMING 4 h**

1. Add 250 µl of zirconium beads to the sample (e.g., ~10⁹ colony-forming units of an *in vitro* mutant population or <600 mg of cecal/fecal contents from mice colonized with a mutant population) in a 2-ml screw-top vial (cryovial). These beads help with cell disruption. Note that this procedure describes the processing of a single sample, but multiple samples can readily be processed in parallel.


3. Use the BeadBeater on the ‘homogenize’ setting for 2 min. Rest the sample on ice for 2 min and repeat bead-beating for 2 min.

4. Centrifuge the tubes in a refrigerated microcentrifuge (4 °C, 6,800g, 3 min).

5. Transfer the aqueous phase (~600 µl) to a Phase Lock gel tube (prespun as described in the manufacturer’s instructions).

6. Add an equal amount (600 µl) of phenol:chloroform:isoamyl alcohol to the Phase Lock gel tube and mix by inversion.

7. Centrifuge the tube in a microcentrifuge (room temperature, 18,000g, 5 min).

8. Transfer the aqueous phase to a microcentrifuge tube and discard the organic phase.

9. Add 600 µl of cold (−20 °C) 100% isopropanol to the microcentrifuge tube.

10. Add 60 µl (one-tenth volume) of 3 M sodium acetate (pH 5.5), and mix thoroughly by vortexing.

11. Incubate at −20 °C for at least 1 h.

   **PAUSE POINT** The sample can be stored overnight or longer at −20 °C at this point.

12. Centrifuge the tube in a refrigerated microcentrifuge (4 °C, 18,000g, 20 min). Carefully decant and discard the supernatant.

13. Wash the pellet with 500 µl of 100% ethanol and centrifuge it in a refrigerated microcentrifuge (4 °C, 18,000g, 3 min). Carefully decant and discard the supernatant.

14. Remove any excess ethanol by gently tapping the tube upside-down on a laboratory tissue.

15. Evaporate any remaining supernatant in a vacuum evaporator (no heat, check at 5-min intervals until dry).

16. Resuspend the pellet in 200 µl of TE buffer (pH 7). Incubate in an oven or water bath at 50 °C for 30 min, vortexing every ~10 min. The pellet will dissolve much faster once it has been dislodged from the wall of the microcentrifuge tube.

**RNase treatment and cleanup of crude DNA using QIAquick columns ● **TIMING 1 h**

17. Transfer 100 µl of the crude DNA sample to a new tube. The remainder should be stored at −20 °C or −80 °C.

18. Add 0.5 µl of RNase A and incubate the tube at room temperature for 2 min. RNase treatment ensures accurate DNA quantitation in subsequent steps.

19. Clean up the sample using QIAquick PCR purification columns according to the manufacturer’s instructions. Make sure that all the ethanol-containing wash buffer is removed from each column before elution by pipetting around the inner rim with a 10-µl tip, if necessary. Elute each sample in 52 µl of buffer EB (part of QIAquick kit).

20. Measure the DNA concentration of each sample using a spectrophotometer (either a UV-based or dye-based system is acceptable).

   **PAUSE POINT** The sample can be stored for months at −20 °C or −80 °C at this point.
**PROTOCOL**

**Linear PCR ● TIMING 2 h**

21| Assemble the linear PCR reactions on ice.

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume per sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>dH₂O</td>
<td>to 100 µl</td>
</tr>
<tr>
<td>Pfx buffer</td>
<td>10 µl</td>
</tr>
<tr>
<td>10 mM dNTPs</td>
<td>2 µl</td>
</tr>
<tr>
<td>50 mM MgCl₂</td>
<td>2 µl</td>
</tr>
<tr>
<td>BioSamA (1 pmol µl⁻¹)</td>
<td>5 µl</td>
</tr>
<tr>
<td>Clean DNA (from Step 20)</td>
<td>0.5–2 µg</td>
</tr>
<tr>
<td>Pfx polymerase</td>
<td>1 µl</td>
</tr>
</tbody>
</table>

22| Split the reaction into 2 × 50 µl in PCR tubes and run them on a thermocycler as follows: 94 °C for 2 min, followed by 50 cycles of 94 °C for 15 s and 68 °C for 1 min.

23| Pool the tubes containing the same DNA sample, run them over a QIAquick PCR cleanup column according to the manufacturer’s instructions, and elute them in 50 µl of buffer EB.

**Bind linear PCR products to beads ● TIMING 1 h**

24| Resuspend streptavidin-coated beads by shaking.

25| Add beads (32 µl per sample) to a new microcentrifuge tube (1 ml maximum; use multiple tubes if a larger volume is required).

26| Place the tube on the MPC for 1–2 min.

27| Carefully remove the supernatant with a pipette.

28| Remove the tube from MPC and add 1,000 µl of 1× B&W buffer; gently resuspend by pipetting.

29| Repeat Steps 26–28 twice for a total of three washes.

30| Remove the final wash and add 2× B&W buffer (52 µl per sample). Aliquot into PCR strip tubes (one tube per sample, 50 µl per tube).

31| Add the entire volume of one sample from Step 23 to the tube.

32| Incubate at room temperature with gentle mixing for 30 min.

33| Place the tube on the MPC for 2 min.

34| Carefully remove the supernatant with a pipette. **▲ CRITICAL STEP** To avoid disturbing the beads, set the electronic multichannel pipettor to its slowest setting, place the end of the tip against the opposite side of the tube from the beads, and slowly move the tip downward as the supernatant is removed. See Supplementary Movie 1 for a demonstration.

**TROUBLESHOOTING**

35| Remove the tube from the MPC and add 100 µl of 1× B&W buffer; gently resuspend by pipetting.

36| Repeat Steps 33–35 twice, but resuspend beads in 100 µl of LoTE buffer each time. **■ PAUSE POINT** The sample can be stored at 4 °C overnight at this point.

**Second strand synthesis ● TIMING 1 h**

37| Denature the sample by heating in a thermocycler: 95 °C for 2 min, then chill quickly to 4 °C.
38| Prepare second strand mix on ice.

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume per sample (µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dH_2O</td>
<td>16</td>
</tr>
<tr>
<td>10× hexanucleotide mix</td>
<td>2</td>
</tr>
<tr>
<td>10 mM dNTPs</td>
<td>1</td>
</tr>
<tr>
<td>Klenow (exo-)</td>
<td>1</td>
</tr>
</tbody>
</table>

39| Collect the beads with the MPC, carefully discard the supernatant, remove the tube from the MPC, and then gently resuspend the sample in 20 µl of second strand mix.

40| Incubate the samples in a thermocycler at 37 °C for 30 min. Mix by gently tapping the tube every 10–15 min.

41| Add 100 µl of LoTE buffer to the sample, collect the beads in the MPC, and then carefully discard the supernatant.

42| Repeat Step 41.

43| Resuspend the beads in 100 µl of LoTE buffer.

■ PAUSE POINT The sample can be stored at 4 °C overnight at this point.

**MmeI digestion ● TIMING 2.5 h**

44| Prepare 50 µM double-stranded M12 oligonucleotide by combining the following in a new PCR tube.

| M12_top (100 µM in EB)          | 15 µl      |
| M12_bot (100 µM in EB)          | 15 µl      |
| 1 M NaCl                        | 1.5 µl     |

45| Anneal oligonucleotides in a thermocycler: 95 °C for 5 min; cool to 4 °C at a rate of 0.1 °C s⁻¹; store in 5-µl aliquots at −20 °C for future use.

46| Prepare MmeI buffer mix on ice.

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume per sample (µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dH_2O</td>
<td>16.8</td>
</tr>
<tr>
<td>10× NEBuffer 4⁺</td>
<td>2</td>
</tr>
<tr>
<td>32 mM SAM⁺</td>
<td>0.08</td>
</tr>
<tr>
<td>M12 dsDNA</td>
<td>0.2</td>
</tr>
</tbody>
</table>

*Part of the MmeI restriction enzyme package from NEB.

47| Collect the beads from Step 43 with the MPC, carefully discard the supernatant, remove the tube from the MPC, and then gently resuspend each sample in 19 µl of MmeI buffer mix.

48| Add 1 µl of MmeI to the sample.

49| Incubate in the thermocycler at 37 °C for 1 h. Gently mix the sample every 10–15 min.

50| Add 100 µl of LoTE buffer to the sample, collect the beads in the MPC, and then carefully discard the supernatant.

51| Repeat Step 50.

52| Resuspend the beads in 100 µl of LoTE buffer.

■ PAUSE POINT The sample can be stored at 4 °C overnight at this point.
**Protocol**

**Linker ligation** ● **TIMING 2.5 h**

53| Prepare a 50 µM stock of barcoded, double-stranded sequencing adapters (one barcode sequence per sample) by combining the following in a new PCR tube.

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume per sample (µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIB_AdaptT_(barcode) (100 µM in EB)</td>
<td>15 µl</td>
</tr>
<tr>
<td>LIB_AdaptB_(barcode) (100 µM in EB)</td>
<td>15 µl</td>
</tr>
<tr>
<td>1 M NaCl</td>
<td>1.5 µl</td>
</tr>
</tbody>
</table>

54| Anneal oligonucleotides in a thermocycler: 95 °C for 5 min; cool to 4 °C at a rate of 0.1 °C s⁻¹.

**PAUSE POINT** Adapters can be stored for months in 5-µl single-use aliquots at -20 °C for future use.

55| If they are frozen, thaw dsDNA sequencing adapters on ice (one barcode per sample). Dilute each 50 µM stock of dsDNA sequencing adapter to 5 µM in ice-cold 1× T4 DNA ligase buffer.

**CRITICAL STEP** If the adapters are not thawed on ice, they can dissociate into single-stranded DNAs.

56| Prepare ligation mix on ice.

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume per sample (µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dH₂O</td>
<td>16.4</td>
</tr>
<tr>
<td>10× T4 DNA ligase buffer</td>
<td>2</td>
</tr>
</tbody>
</table>

57| Collect the beads from Step 52 with the MPC, carefully discard the supernatant, remove the tube from the MPC and gently resuspend the sample in 18.4 µl of ligation mix.

58| Add 0.6 µl of 5 µM dsDNA sequencing adapter (from Step 55) containing a unique barcode to the sample. Record which barcode is associated with the sample.

59| Add 1 µl of T4 DNA ligase to the sample.

60| Incubate in a thermocycler at 16 °C for 1 h. Gently mix every 10–15 min.

61| Add 100 µl of LoTE buffer to the sample, collect the beads in the MPC and carefully discard the supernatant.

62| Repeat Step 61.

63| Resuspend the beads in 100 µl of LoTE buffer.

**PAUSE POINT** The sample can be stored at 4 °C overnight at this point.

**PCR and final purification** ● **TIMING 4 h**

64| Assemble the PCR mix on ice.

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume per sample (µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dH₂O</td>
<td>31.5</td>
</tr>
<tr>
<td>10× Pfx buffer</td>
<td>10 (used at 2× standard concentration)</td>
</tr>
<tr>
<td>10 mM dNTPs</td>
<td>2</td>
</tr>
<tr>
<td>50 mM MgCl₂</td>
<td>2</td>
</tr>
<tr>
<td>5 µM LIB-PCR5</td>
<td>2</td>
</tr>
<tr>
<td>5 µM LIB-PCR3</td>
<td>2</td>
</tr>
<tr>
<td>Pfx polymerase</td>
<td>0.5</td>
</tr>
</tbody>
</table>

65| Collect the beads from Step 63 with the MPC, carefully discard the supernatant, remove the tube from the MPC and gently resuspend the sample in 50 µl of PCR mix on ice.
Run on a thermocycler, as follows, and prepare a 2% (wt/vol) agarose gel (1–2 lanes per sample plus 2 ladder lanes per gel; GelGreen dye at 1:10,000 dilution; wide-tooth comb) while PCR is running. Run at 94 °C for 2 min followed by 18 cycles of: 94 °C for 15 s, 60 °C for 1 min, 68 °C for 2 min and then 68 °C for 4 min.

**PAUSE POINT** The sample can be stored at 4 °C overnight at this point.

Collect the beads on the MPC and transfer the supernatant to a new PCR tube.

Prepare the DNA ladder.

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-bp DNA ladder</td>
<td>4 µl</td>
</tr>
<tr>
<td>Bromophenol blue loading dye</td>
<td>1× final concentration</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Add xylene cyanol loading dye to the sample (supernatant from Step 67) to 1× final concentration.

**CRITICAL STEP** Xylene cyanol is used because bromophenol blue can migrate near the same position as the desired 125-bp PCR product, thereby obscuring the band in the gel. The use of bromophenol blue in the ladder lanes will provide an approximate position of the samples in the gel during electrophoresis.

Load the DNA ladder onto the first and last lanes of the gel.

Load the sample into its own lane of the gel. If the wells are too small for the full volume, split the sample over two lanes.

Run the gel for 30 min at 200 V.

**CRITICAL STEP** Time and voltage may need to be adjusted on the basis of gel size and buffer composition; these parameters work well for 7-cm gels run in borate buffer.

Use a non-UV gel illuminator to excise the band at ∼125 bp, minimizing surplus agarose in the gel fragment. Place the gel fragment in a microcentrifuge tube (if two lanes were used per sample, the similar fragments can be placed in the same tube as long as the total weight remains under 300 mg).

**TROUBLESHOOTING**

Clean up the sample using QIAquick gel purification columns according to the manufacturer’s instructions (including the isopropanol step for small fragments). Be certain that all of the ethanol-containing wash buffer is removed from the column before elution by pipetting around the inner rim with a 10-µl tip, if necessary. Elute in 32 µl of buffer EB. Gel purification ensures that primer-dimers or other erroneous products do not contribute to the DNA quantification or sequencing in subsequent steps.

Quantify the sample on a QuBit or similar spectrophotometer.

**CRITICAL STEP** DNA concentration is typically 0.1–5 ng µl⁻¹, which is below the reliable detection limit of many UV absorbance-based spectrophotometers.

All samples should subsequently be normalized to the same concentration. If all samples are ≥0.83 ng µl⁻¹, adjust an aliquot of each sample to 10 nM in EB buffer. On the basis of an expected fragment length of 125 bp, the formula for a 10 nM normalization is volume of DNA (µl) = 10 × final volume (µl) / (DNA concentration (ng µl⁻¹) × (10⁶) × (1/649) × (1/125)). If sample concentrations are <0.83 ng µl⁻¹, adjust an aliquot of each sample to 1 nM in EB buffer. Tip: the formula for a 1 nM normalization is the volume of DNA (µl) = 1 × final volume (µl) / (DNA concentration (ng µl⁻¹) × (10⁶) × (1/649) × (1/125)).

Combine an equal volume of each sample into a single tube.

**CRITICAL STEP** Required volumes may vary by sequencing facility. Be sure to inform the sequencing facility of whether the sample is at 10 or 1 nM.
Submit for Illumina sequencing. The optimum loading concentration may vary by sequencing center or by Genome Analyzer model; we currently use 6 pM for GA-IIx and HiSeq machines.

Data analysis ● TIMING ~3 min per million raw reads

Download the data analysis package (INSeq_analysis.zip; Supplementary Data 1) and sample data package (INSeq_demo.zip; downloadable from http://gordonlab.wustl.edu/SuppData.html) and save each to your own directory.

Bowtie is used to map sequences to the reference genome. Download this mapping software from http://bowtie-bio.sourceforge.net/index.shtml and install in your system. Edit the ‘config.txt’ file in the data analysis package (INSeq_analysis) by changing the bowtie_dir variable to the path where bowtie was installed. Detailed instructions are provided in the README file in the INSeq_analysis folder (see Supplementary Data 1).

Go to the uncompressed data analysis package (INSeq_analysis; see Supplementary Data 1) and then to the ‘indexes’ directory. From this location, create a new directory with the organism’s name (e.g., BthetaVPI_5482). Put the genome file in .fasta format and the annotation files (in .ptt format) in the new directory. Go into the new directory and construct a bowtie index by typing the command:

```
“path to bowtie directory”/bowtie-build <fasta file> <name of index, same as the new directory>
```

Create an experiment directory where you would like to store your analysis results. For example:

```
mkdir /Users/mwu/Documents/Experiment_1
```

Create a mapping file that contains the barcode for each sample in tab-delimited format in the experiment directory:

```
<barcode> <Sample Name>
```

From the experiment directory, run the analysis pipeline by running the wrapper script INSeq_pipeline.pl with the usage:

```
perl “path for analysis package”/INSeq_pipeline.pl -i <the raw reads file> -m <Barcodes mapping file> -s <indexed genome name> -d <length_disrupt_percent (max=1)>
```

Required arguments: -i gives the input raw reads file, -m gives the mapping file and -s gives the name of the indexed genome to which the reads should be mapped.

Optional arguments: -d gives the region of the gene in which insertions are expected to disrupt gene function. The default is 1, which means that when the insertion falls anywhere within the gene (100%), the gene’s function will be considered to be disrupted. Setting the -d argument to 0.9, for example, would exclude insertions in the distal 10% of the gene when calculating the total number of reads/insertions for that gene.

The INSeq_pipeline.pl will generate several mappingjobsXXX.job under the experiment directory. On a desktop/laptop, run them in series by typing sh mappingjobsXXX.job for each job, or (on a cluster) run in parallel. Several output files will be generated:

- `INSEQ_experiment.scarf_assigned.txt`: reads are assigned to different samples on the basis of the sample-specific barcode.
- `INSEQ_experiment.scarf.log`: contains some statistics for the analysis process, including the total number of reads, the percentage of reads being mapped and trimmed (that have the transposon), how many insertions in the sample (coverage) with the number of raw reads (reads after filtering in the normalization step) and the scale factor being used for normalization.

Several output files will be placed in the ‘results’ folder:

- `INSEQ_experiment.scarf_Samplename.bowtiemap`: a raw mapping output file from bowtie.
- `INSEQ_experiment.scarf_Samplename.bowtiemap_processed.txt_chromosomenname`: a text file containing the processed mapping output, with the format:
<chromosome name> <insertion position> <reads mapped to the left side of the insertion> <reads mapped to the right side of the insertion> <the total number of reads mapped to that position>

**INSEQ_experiment.scarf_Samplename.bowtiemap_processed.txt_chromosomename_filter_cpm.txt**: a text file with positions having more than three total reads, and normalized to counts per million reads.

**INSEQ_experiment.scarf_Samplename.bowtiemap_processed.txt_chromosomename_filter_cpm.txt_mapped**: a text file that lists genes mapped by insertions, only when the insertions located at the proximate XX percentage (specify by the -d option) are considered to interrupt the function of genes. The format is:

<gene name> <the total number of unique insertions in the gene> <the sum of normalized read counts in that gene> <gene annotation from the ptt file>

## Troubleshooting

86 | *Cleanup*: as there are many intermediate files generated along the pipeline, if storage space is a big concern, deleting them is optional. You can run `clean_up.sh` from the experiment folder by using command `sh clean_up.sh`, which will remove the sorted barcode assigned reads file, the folder 'bcsortedseqs' and the mappingjob files.

**Troubleshooting**

Troubleshooting advice can be found in **Table 1**.

<table>
<thead>
<tr>
<th>Step</th>
<th>Problem</th>
<th>Possible reason</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>Beads are lost during wash steps</td>
<td>Electronic multichannel pipettor is too fast</td>
<td>If alternate pipettors are not available, perform this step manually</td>
</tr>
<tr>
<td>73</td>
<td>Bands are faint or inconsistent between samples</td>
<td>Ligation adapters not properly annealed, or have barcodes that prevent proper annealing</td>
<td>Use technical replicates (same input DNA with different sample-specific barcodes) to determine whether the barcode or the input DNA is the problem</td>
</tr>
<tr>
<td>78</td>
<td>Illumina sequencing produces fewer reads than expected</td>
<td>Sequence characteristics prevent proper function of Illumina base-calling software</td>
<td>Provide the sequencing facility with information about common and variable bases in your expected sequences on the basis of the barcodes used</td>
</tr>
<tr>
<td>81</td>
<td>Error message `bowtie-build: No such file or directory'</td>
<td>Command did not specify the location of bowtie</td>
<td>Make sure to include the entire path for bowtie, e.g., `/Users/XXX/bowtie-0.12.7/bowtie-build Bt.fna BtVPI 5482'</td>
</tr>
<tr>
<td>84</td>
<td>Analysis pipeline cannot recognize the raw reads file</td>
<td>The raw reads file is not in the scarf format</td>
<td>Convert your raw reads file to scarf format</td>
</tr>
<tr>
<td>85</td>
<td>Error message `bowtie: No such file or directory'</td>
<td>The analysis package could not locate bowtie</td>
<td>Check the ‘config.txt’ in Step 80; make sure the bowtie_dir variable is the path where your bowtie installation is located. For example, bowtie_dir=&quot;/Users/XXX/bowtie-0.12.7/&quot;</td>
</tr>
<tr>
<td></td>
<td>Error message 'Could not locate a bowtie index corresponding to base name 'XXX&quot;'</td>
<td>The analysis package could not find the bowtie index. The analysis package is looking for the index file located in 'INSEQ_analysis/indexes/XXX/XXX' (XXX is the 'indexed genome name' specified in Step 84 by argument -s)</td>
<td>Check the location and the name of your index file. Make sure it is UNDER the 'INSeq_analysis/indexes' directory (it should be in the position exactly inside of the 'indexes' folder), and check that the name of the index, which is the prefix of the .X.ebwt file, is consistent with the name of the directory. If not, remake the ebwt file using bowtie-build and repeat Step 81</td>
</tr>
</tbody>
</table>
TABLE 1  Troubleshooting table (continued).

<table>
<thead>
<tr>
<th>Step</th>
<th>Problem</th>
<th>Possible reason</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>The gene-mapped results show all genes have zero insertions, whereas the</td>
<td>The analysis package matches the gene annotation file (.ptt) to chromosome (.fna) by names. If the naming is not consistent, the pipeline can not locate the insertions to genes correctly.</td>
<td>Remove any spaces in any names, and rename the header names of the .fna files to match the .ptt files, or rename the .ptt files to match the header names of the .fna files.</td>
</tr>
</tbody>
</table>

● TIMING
Steps 1–16, Isolation of crude DNA: 4 h
Steps 17–20, Cleanup of crude DNA using QIAquick columns: 1 h
Steps 21–23, Linear PCR: 2 h
Steps 24–36, Bind linear PCR product to beads: 1 h
Steps 37–43, Second strand synthesis: 1 h
Steps 44–52, MmeI digestion: 2.5 h
Steps 53–63, Linker ligation: 2.5 h
Steps 64–78, PCR and final purification: 4 h
Steps 79–86, Data analysis: initial analysis requires ~3 min per million raw reads on a 3.06 GHz Intel Core 2 Duo processor with 4 GB memory

Box 1, arrayed library preparation: ~1.6 h per 96-well tray, plus time for picking colonies into trays and INSeq library preparation

ANTICIPATED RESULTS
The INSeq analysis pipeline produces several output files with data at the level of individual insertion locations, genes and (optionally) operons. Good libraries should have >95% of reads that contain the transposon sequence, consistent read counts generated from both sides of each insertion and consistent data from biological and technical replicates. A real data set from a successful library preparation is provided in the INSeq_demo files.

Note: Supplementary information is available via the HTML version of this article.

ACKNOWLEDGMENTS This work was supported by grants from the National Institutes of Health (DK30292, DK70977, and DK064540 to J.I.G.; F32AI078628 and K01DK089121 to A.L.G.) and by the Crohn’s and Colitis Foundation of America. We thank L. Kyro and J. Eberle for their assistance in making the movies that accompany this article.

AUTHOR CONTRIBUTIONS A.L.G., J.I.G. and M.W. designed the experiments and software. A.L.G., M.W. and J.I.G. wrote the manuscript.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Published online at http://www.natureprotocols.com/. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

Appendix B

Nathan P. McNulty, Meng Wu, Alison R. Erickson, Chongle Pan, Brian K. Erickson, Eric C. Martens, Nicholas A. Pudlo, Brian D. Muegge, Bernard Henrissat, Robert L. Hettich, Jeffrey I. Gordon

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

Effects of Diet on Resource Utilization by a Model Human Gut Microbiota Containing Bacteroides cellulosilyticus WH2, a Symbiont with an Extensive Glycobiome

Nathan P. McNulty1, Meng Wu1, Alison R. Erickson2,3, Chongle Pan3, Brian K. Erickson3, Eric C. Martens4, Nicholas A. Pudlo4, Brian D. Muegge1, Bernard Henriussat5, Robert L. Hettich2,3, Jeffrey I. Gordon1*

1 Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri, United States of America, 2 Graduate School of Genome Science and Technology, University of Tennessee–Oak Ridge National Laboratory, Knoxville, Tennessee, United States of America, 3 Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, United States of America, 4 Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, Michigan, United States of America, 5 Architecture et Fonction des Macromolécules Biologiques, CNRS and Aix-Marseille University, Marseille, France

Abstract
The human gut microbiota is an important metabolic organ, yet little is known about how its individual species interact, establish dominant positions, and respond to changes in environmental factors such as diet. In this study, gnotobiotic mice were colonized with an artificial microbiota comprising 12 sequenced human gut bacterial species and fed oscillating diets of disparate composition. Rapid, reproducible, and reversible changes in the structure of this assemblage were observed. Time-series microbial RNA-Seq analyses revealed staggered functional responses to diet shifts throughout the assemblage that were heavily focused on carbohydrate and amino acid metabolism. High-resolution shotgun metaproteomics confirmed many of these responses at a protein level. One member, Bacteroides cellulosilyticus WH2, proved exceptionally fit regardless of diet. Its genome encoded more carbohydrate active enzymes than any previously sequenced member of the Bacteroidetes. Transcriptional profiling indicated that B. cellulosilyticus WH2 is an adaptive forager that tailors its versatile carbohydrate utilization strategy to available dietary polysaccharides, with a strong emphasis on plant-derived xylans 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 of xylan utilization systems. Introduction of a B. cellulosilyticus WH2 library comprising >90,000 isogenic transposon mutants into gnotobiotic mice, along with the other artificial community members, confirmed that these loci represent critical diet-specific fitness determinants. Carbohydrates that trigger dramatic increases in expression of these two loci and many of the organism’s 111 other predicted PULs were identified by RNA-Seq during in vitro growth on 31 distinct carbohydrate substrates, allowing us to better interpret in vivo RNA-Seq and proteomics data. These results offer insight into how gut microbes adapt to dietary perturbations at both a community level and from the perspective of a well-adapted symbiont with exceptional saccharolytic capabilities, and illustrate the value of artificial communities.


Academic Editor: Jonathan A. Eisen, University of California Davis, United States of America

Received March 26, 2013; Accepted July 12, 2013; Published August 20, 2013

Copyright: © 2013 McNulty et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The majority of this work was supported by grants from the NIH (DK30292, DK70977). Other sources of support included the Crohn’s and Colitis Foundation of America, the European Research Council under the European Union’s Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. 322830, and an ONRL Laboratory Director’s Exploratory Seed Money grant (for metaproteomics). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: CE, carbohydrate esterase; COPER-Seq, community profiling by sequencing; EC, enzyme commission; ECF, extracellular function; GH, glycoside hydrolase; GT, glycosyltransferase; HF/HS, high-fat/high-sugar; HTCS, hybrid two-component system; INSeq, insertion sequencing; LF/HPP, low-fat/high-poly, minimal medium; MPC, magnetic particle collector; NBC, normalization by community; NBS, normalization by species; PCoA, principal coordinates analysis; PL, polysaccharide lyase; PoMA, percentage of maximum achieved; PSMs, peptide-spectrum matches; PTS, phosphotransferase system; PUL, polysaccharide utilization locus; RFO, raffinose family oligosaccharide.

* E-mail: jgordon@wustl.edu

Introduction
A growing body of evidence indicates that the tens of trillions of microbial cells that inhabit our gastrointestinal tracts extend our biological capabilities in important ways. Microbial enzymes process many compounds that would otherwise pass through our intestines unaltered [1], and cases of particular nutrient substrates favoring the growth of particular taxa are being reported [2–5]. Changes in diet are therefore expected to lead to changes in the composition and function of the microbiota [6–10]. 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
Recent studies have focused on better understanding the polysaccharide utilization loci (PULs) that are distributed throughout the genome [23,24]. These genes are organized into polysaccharide utilization loci (e.g., hybrid two-component systems, extracytoplasmic factors), which include processing components of our diet. The way this incredible machine assembles itself and operates remains mysterious. One approach to understanding its properties is to create artificial communities composed of a limited number of sequenced human gut bacterial species and to install them in the guts of germ-free mice that are then fed different diets. In this report, we adopt this approach. We describe the genome sequence of a new gut bacterial isolate, Bacteroides cellulosilyticus WH2, which is equipped with an unprecedented number of carbohydrate active enzymes. Deploying four different “omics” technologies, we characterize the response to diet, the relative stability, and the temporal dynamics of a 12-species artificial bacterial assemblage (including B. cellulosilyticus WH2) implanted in germ-free mouse guts. We also combine high-throughput substrate utilization screens and RNA-Seq to generate reference data analogous to a “Rosetta stone” in order to decipher what types of carbohydrates B. cellulosilyticus encounters and uses within the gut, and how it interacts with other organisms that have similar and/or distinct “professions.” This work sets the stage for future ecological and metabolic studies of more complex assemblages that more fully emulate the properties of our native gut communities.

Author Summary

Our intestines are populated by an almost unimaginably large number of microbial cells, most of which are bacteria. This species assemblage operates as a microbial metabolic organ, performing myriad tasks that contribute to our well-being, including processing components of our diet. The way this incredible machine assembles itself and operates remains mysterious. One approach to understanding its properties is to create artificial communities composed of a limited number of sequenced human gut bacterial species and to install them in the guts of germ-free mice that are then fed different diets. In this report, we adopt this approach. We describe the genome sequence of a new gut bacterial isolate, Bacteroides cellulosilyticus WH2, which is equipped with an unprecedented number of carbohydrate active enzymes. Deploying four different “omics” technologies, we characterize the response to diet, the relative stability, and the temporal dynamics of a 12-species artificial bacterial assemblage (including B. cellulosilyticus WH2) implanted in germ-free mouse guts. We also combine high-throughput substrate utilization screens and RNA-Seq to generate reference data analogous to a “Rosetta stone” in order to decipher what types of carbohydrates B. cellulosilyticus encounters and uses within the gut, and how it interacts with other organisms that have similar and/or distinct “professions.” This work sets the stage for future ecological and metabolic studies of more complex assemblages that more fully emulate the properties of our native gut communities.

identity of all taxa and genes comprising the system are known. Within these assemblages, 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 systems also afford an opportunity to tightly control experimental variables to a degree not possible in human studies and have proven useful in studying microbial invasion, microbe–microbe interactions, and the metabolic roles of key ecological guilds [11–15]. Studies aiming to better understand community-level assembly, resilience, and adaptation are therefore likely to benefit from a focus on such defined systems. However, the limited taxonomic and functional representation within artificial communities of modest complexity requires that caution be exercised when extrapolating results to more complex, naturally occurring gut communities (see Prospects).

Culture-independent surveys of the healthy adult gut microbiota consistently conclude that it is composed primarily of members of two bacterial phyla, the Bacteroidetes and Firmicutes [16–21]. 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 [22]. 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 (GHs) and polysaccharide lyases (PLs), 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 [23,24]. 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 [25,26]. Little is known, however, about the metabolic strategies adopted by multiple competing species in more complex communities, how dietary changes lead to reconfigurations in community structure through changes in individual species, or whether dietary context influences which genes dominant species rely on to remain competitive with other microbes, including those genes that are components of PULs.

Here, we adopt a multifaceted approach to study an artificial community in gnotobiotic mice fed changing diets in order to better understand (i) the process by which such a community reconfigures itself structurally in response to changes in host diet; (ii) how aggregate community function, as judged by the metatranscriptome and metaproteome, is impacted when host diet is altered; (iii) how the metabolic strategies of its individual component 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 an assemblage containing related and unrelated species.

Results and Discussion

Sequencing the Bacteroides cellulosilyticus WH2 Genome

Though at least eight complete and 68 draft genomes of Bacteroides spp. are currently available [27], 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 (Figure 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 [28] revealed an extraordinary complement of 503 CAZymes comprising 373 GHs, 23 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 [29]. The B. cellulosilyticus WH2 CAZome is enriched in a number of GH families even when compared with prominent representatives of the gut Bacteroidetes (Figure 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 (Figure 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 [30]. 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 [30–32]. 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 suc- and susD-like genes in *B. thetaiotaomicron* VPI-5492 revealed an unprecedented number of sucC/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 [25,33]. Analysis of deeply sampled adult human gut microbiota datasets indicates that *B. cellulosilyticus* strains are common, colonizing approximately 77% of 124 adult Europeans characterized in one study [18] and 62% of 139 individuals living in the United States examined in another survey [20]. 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,” Figure S3). Two groups of 10–12-wk-old male germ-free C57BL/6j animals were moved to individual cages within different diets at 2-wk intervals as described in Figure S3. COPRO-Seq data from E1 and E2 were ordinated in the same multidimensional space. For each species chosen for inclusion in this microbial assemble meeting four criteria: (i) it was a member of one of three bacterial phyla (i.e., Bacteroidetes, Firmicutes, or Actinobacteria), (ii) it was identified as a prominent member of the human gut microbiota in previous culture-independent surveys, (iii) it could be grown in the laboratory, and (iv) its genome had been sequenced to at least a high-quality draft level. Species were also selected for their functional attributes (as judged by their annotated gene content) in an effort to create an artificial community that was somewhat representative of a more complex human microbiota. For example, although more than half of the species in the assemble were Bacteroidetes predicted to excel at the breakdown of polysaccharides, several were also prominent inhabitants of the human gut that are thought to have limited carbohydrate utilization capabilities (e.g., *Firmicutes* from Clostridium cluster XIVa). Some attributes for the 12 strains included in the artificial community are provided in Table S4.

For 2 wk, 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)) [12]. Over the course of 6 wk, diets were changed twice at 2-wk intervals, such that each group began and ended on the same diet, with an intervening 2-wk period during which the other diet was administered (Figure 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; Figure 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 only exception (4% of the diet by weight versus 46.3% for the other carbohydrates 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/}

![Figure 1. COPRO-Seq analysis of the structure of a 12-member artificial human gut microbial community as a function of diet and time.](https://doi.org/10.1371/journal.pbio.1001637.g001)
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 energy for saccharolytic members of the artificial community.

To evaluate the impact of each initial diet and subsequent diet switch on the structural configuration of the artificial community, we performed shotgun sequencing (community profiling by sequencing; COPRO-Seq) [11] of DNA isolated from fecal samples collected throughout the course of the experiment, as well as cecal contents collected at sacrifice. The relative abundances of the species in each sample (defined by the number of sequencing reads that could be unambiguously assigned to each microbial genome after adjusting for genome uniqueness) were subjected to ordination by principal coordinates analysis (PCoA) (Figure 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). The overall structure of the artificial community 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. A structural reconfiguration also took place over the course of ~5 d following transition to the second diet phase. Notably, the two treatment groups underwent a near-perfect inversion in their positions along PC1 after the first diet switch; the artificial community in animals switched from a LF/HPP to HF/HS diet took on a structure like that which arose by the end of the first diet phase in animals consuming the HF/HS diet, and vice versa. The second diet switch from phase 2 to 3 resulted in a similar movement along PC1 in the opposite direction, indicating a reversion of the artificial community’s structure to its originally assembled structure in each treatment group. These results, in addition to demonstrating the significant impact of these two diets on the structure of this 12-member artificial human gut community, also suggest that an assemblage of this size is 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 E1 (experiment 2, “E2’’). In this follow-up experiment, fecal samples were collected more frequently than in E1, providing a dataset with improved temporal resolution. Ordination of E2 COPRO-Seq data by PCoA showed that (i) for each treatment group in E2, the artificial community assembles in a manner similar to its counterpart in E1; (ii) structural reconfigurations in response to diet occur with the same timing as in E1; and (iii) the quasi-equilibria achieved during each diet phase are highly similar between experiments for each treatment group (compare Figures 1B and S5A). As in E1, cecal data for each E2 treatment group overlap with their corresponding fecal samples, and DNA yields from E2 fecal samples vary substantially as a function of host diet (Figure S4B).

COPRO-Seq provides precise measurements of the proportional abundance of each member species present in the artificial community. Data collected in both E1 and E2 (Table S5) revealed significant differences between 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 (Figure 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, rules-based fashion and with minimal stochasticity in this artificial community. A species’ relative abundance immediately after colonization (i.e., 24 h 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) (E1, \( R^2 = 0.23 \); E2, \( R^2 = 0.27 \)), suggesting that early dominance of the founder population was not strongly tied to relative success in the assembly process.

In mice initially fed a HF/HS diet, four Bacteroides spp. (Bacteroides caccae, B. cellulosilyticus WH2, B. thetaiotaomicron, and Bacteroides vulgatus) each achieved a relative abundance of \( >10\% \) by the end of the first diet phase (day 13 postgavage), with B. caccae attaining the highest levels (37.1±4.9% and 34.2±5.5%; group mean \( \pm SD \) in E1 and E2, respectively). In animals fed the plant polysaccharide-rich LF/HPP Chow during the first diet phase, B. cellulosilyticus 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, Dorea 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 relative abundance to the range observed for each strain, we also normalized each species’ representation within the artificial community at each time-point to the maximum proportional abundance 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 (Figure 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 taxa, 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 (Figures 1C and S4C). Among the diet-insensitive taxa, B. thetaiotaomicron showed the most stability in its representation (Figures 1C and S4C), consistent with its reputation as a versatile forager. Paradoxically, B. cellulosilyticus WH2 was both diet-sensitive and highly fit on its less-preferred diet; although this strain clearly achieved higher levels of representation in animals fed the LF/HPP diet, it also maintained strong levels of representation in animals fed the HF/HS diet (Figures 1C and S4C).

When taking into account the abundance data for all 12 artificial 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) (E1, \( R^2 = 0.77 \); E2, \( R^2 = 0.84 \)), suggesting that the intervening dietary perturbation had little effect on the ultimate outcomes for most species within this assemblage. However, one very low-abundance strain (D. longicatena) achieved significantly different maximum percentage abundances across the two treatment groups in each experiment, suggesting that steady-state levels of this strain may have been impacted by diet history. In mice initially fed the LF/HPP diet, D. longicatena was found to persist throughout the experiment at low levels on both diet regimens. In mice initially 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 persist within a gut over long periods of time.

**The Cecal Metatranscriptome Sampled at the Time of Sacrifice**

These diet-induced reconfigurations in the structure of the artificial community led us to examine the degree to which its members 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 artificial human gut microbiome (see Materials and Methods). Total RNA was collected from the cecal contents of each animal in E1 at the time of sacrifice and hybridized to this GeneChip. The total number of genes whose expression was detectable on each diet was remarkably similar (14,929 and 14,594 detected in the LF/HPP and HF/HS treatment groups, respectively).

A total of 11,373 genes (24.3%) were expressed on both diets in members of CAZymes that are required for the breakdown of plant-derived polysaccharides such as cellulose, hemicelluloses, and pectins. These CAZymes are encoded within the genome of this organism and 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 each diet treatment group.

Among taxa demonstrating obvious diet preferences (as judged by relative abundance data), *B. caccae* and *B. cellulosilyticus* WH2 provided examples of CAZyme-level responses to diet change that were different in several respects. Our observations regarding the carbohydrate utilization capabilities and preferences of *B. caccae* are summarized in Text S1. However, our ability to evaluate shifts in *B. caccae*’s metabolic strategy in the gut was limited by its very low abundance in animals fed LF/HPP chow (i.e., our mRNA and subsequent protein assays were often not sensitive enough to detect low abundance in animals fed LF/HPP chow (i.e., our mRNA and subsequent protein assays were often not sensitive enough to exhaustively sample *B. caccae*’s metatranscriptome and metaproteome). In contrast, the abundance of *B. 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 (Figure 2A). Twenty-five of the 50 most expressed CAZymes on the LF/HPP diet were significantly up-regulated compared to the HF/HS diet; of these, seven were members of the GH43 family (Figure 2B). The GH43 family consists of enzymes with activities required for the breakdown of plant-derived polysaccharides such as cellulose, hemicelluloses, and pectins. These enzymes are encoded within the genome of this organism and allow for a comprehensive analysis of its expressed genes and proteins.

![Figure 2](image-url)
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 up-regulated genes was also exclusively up-regulated 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. doi:10.1371/journal.pbio.1001637.g002

as hemicellulose and pectin. Inspection of the enzyme commission (EC) annotations for the most up-regulated GH43 genes shows that they encode xylan 1,4-β-xylanidases (EC 3.2.1.37), arabian endo-1,3-β-L-arabinosidases (EC 3.2.1.99), and α-L-arabinofuranosidases (EC 3.2.1.53). 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 Bacillus cellulosilyticus WH2 genome making the list. Strikingly, of the 43 predicted genes with putative GH43 domains in the Bacillus cellulosilyticus WH2 genome, none were up-regulated on the “Western”-style HF/HS diet.

The most highly expressed Bacillus 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 nonreducing ends of raffinose family oligosaccharides (RFOs, including raffinose, stachyose, and verbascose), galacto(rgluco)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 [34] as well as terminal sialic acid and sulfate [35]. Hydrolysis of the α2 linkage connecting terminal fucose residues to the galactose-rich extended core is thought to be catalyzed in large part by GH95 and GH29 enzymes [36]. The Bacillus 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 mucus by bacteria is usually carried out by GH33 family enzymes. Bacillus 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 Bacillus cellulosilyticus WH2 CAZymes. Therefore, utilization of host glycans by Bacillus cellulosilyticus WH2, if it occurs, likely requires partnerships with other members of the artificial 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 Bacillus cellulosilyticus WH2 CAZymes, 12 were significantly up-regulated 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 amylase (one of the two components of starch) and maltodextrin, both prominent ingredients in the HF/HS diet.

GenChip-based profiling of the E. coli cecal communities provided a snapshot of the metatranscriptome on the final day of the final diet phase in each treatment group. The analysis of Bacillus 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 that are not metabolized by the host.

Community-Level 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 artificial 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 (Figure S3).

We began with a “top-down” analysis in which every RNA-Seq read count from every gene in the artificial 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 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 d to acclimate to each diet). The 157 significant changes were identified and subjected to hierarchical clustering by EC number to determine which functional responses occurred with similar kinetics. The results revealed that in contrast to the rapid, diet-induced structural reconfigurations observed in this artificial community, community-level changes in microbial 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 were those where the representation of EC numbers in the transcriptome increased/decreased dramatically within 1–2 d of a diet switch; “gradual” responses were those where the representation of EC numbers changed notably, but slowly, between the two diet transition points; and “delayed” responses were those where significant change did not occur until the end of a diet phase (Figure 3, Table S7). EC numbers associated with reactions important in carbohydrate metabolism and transport were distributed across all three of these response types for each of the two diets. Nearly all genes encoding proteins with EC numbers related to amino acid metabolism that were significantly up-regulated on HF/HS chow binned into the “rapid” or “gradual” groups, suggesting this diet put immediate pressure on the artificial microbial 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 up-regulated on HF/HS diet were spread more evenly across these three response types (Figure 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 at the community level, 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 assemblage of microbes where all of the genes in component members’ genomes 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
purple markers, while those with relevance to carbohydrate metabolism and/or transport are denoted by
end of a diet phase. EC numbers specifying enzymatic reactions
are those where significant expression changes did not occur until the
but slowly, between the two diet transition points. "Delayed" responses
"Gradual" responses are those where expression changed notably,
heatmap visualization. "Rapid" responses are those where expression
mean-centered, and subjected to hierarchical clustering, followed by
significant EC numbers (out of 1,021 total tested) were log-transformed,
demonstrating a statistically significant difference in their representa-
tion in the metatranscriptome between the final days of the first two
diet phases were identified using a model based on the negative
action in the metatranscriptome (Figure S7C). This dramatic reduction in the extent to
which B. cellulosilyticus WH2 contributes to the metatranscriptome in HF/HS-fed mice “masks” the significant up-regulation of EC
B. cellulosilyticus U following the first diet shift (day 13 versus day 27; Mann–Whitney
activity was reduced to a statistically significant degree as a result
first diet switch (LF/HPP
p = 0.03; Figure S7D). A further breakdown of endo-1,4-β-
xylanase up-regulation in B. cellulosilyticus WH2 when mice are
switched to the HF/HS diet reveals that most of this response is
driven by two genes, BWI2_4068 and BWI2_4072 (Figure S7E).
Our realization that we were unable to correctly infer the direction
of one of the most significant diet-induced gene expression changes
in the second most abundant strain in the artificial community
when inspecting functional responses at the community level
provides a strong argument for expanding the use of microbial
assemblages comprised exclusively of sequenced species in studies
of the gut microbiota. This should allow the contributions of
individual species to community activity to be evaluated in a
rigorous way that is not possible with microbial communities 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 [37,38]. For
these reasons and others, much work has been dedicated to
applying shotgun proteomics techniques to microbial ecosystems
in various environments [39,40]. 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 artificial
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
Materials and 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 semiquantitative peptide/protein
data that might validate and enrich our understanding of

**Figure 3. Top-down analysis of fecal microbiome RNA expres-
son 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 microbial community functions triggered by a change in diet. After aligning reads to genes in the defined artificial 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 representa-
tion in the metatranscriptome between the final days of the first two
diet phases were identified using a model based on the negative
binomial distribution [57]. 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 d of a diet switch.
“Gradual” responses are those where expression changed notably,
but slowly, between the two diet transition points. “Delayed” responses
are those where significant expression changes did not occur 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.

doi:10.1371/journal.pbio.1001637.g003
functional responses identified at the mRNA level, particularly with respect to the niche (profession) of CAZyme-rich *B. cellulosolyticus* 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 nonunique 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 *in silico* tryptic digest. This simulated digest took into account both the potential for missed tryptic cleavages and the peptide mass range that could be detected using our methods. The results (Figure S8A, Table S8) demonstrated that for an artificial community of modest complexity, the proportion of peptides within each strain's theoretical peptidome that are "unique" (i.e., assignable to a single protein within the theoretical metaproteome) varies substantially from species to species, even among those that are closely related. We found the lone representative of the *Actinobacteria* in the artificial community, *Collinsella 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 artificial community member evaluated, despite having a proteome composed of mostly unique peptides. On the other hand, *B. cellulosolyticus* WH2 (5,244 predicted protein-coding genes) contributed 241,473 (4.5%) unique peptides, the highest of any member despite a high fraction of nonunique peptides (18.4%) within its theoretical peptidome. The evolutionary relatedness of the *Bacteroides* components of the artificial community 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* 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 peptidome 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 artificial 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 (Figure 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, both this finding and the higher number of mouse proteins detected in samples from HF/HS-fed animals are consistent with the results of our fecal DNA analysis, which indicated that the HF/HS diet supports lower levels of gut microbial biomass than the LF/HPP diet (Figure S4A,B). Second, a breakdown of all detected microbial proteins by species of origin (Figure S8B) revealed that the degree to which we could inspect protein expression for a given species was dictated largely by its relative abundance and the diet to which it was exposed.

Our ability to detect many of *B. cellulosolyticus* WH2's expressed transcripts and proteins in samples from both diet treatment groups allowed us to determine how well RNA and protein data for an abundant, active member of the artificial community might correlate. These data also allowed us to evaluate whether or not the types 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. cellulosolyticus* 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 (Figure 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 (Figure S8C, red plot; \( r = 0.03 \)). Correlations for other categories of *B. cellulosolyticus* WH2 genes, such as those involved in energy metabolism (Figure S8C, green plot; \( r = 0.36 \)) and amino acid metabolism (Figure S8C, orange plot; \( r = 0.48 \)), were also poorer than the correlation for the complete set of detectable genes. In contrast, the correlation for the 110 genes with predicted involvement in carbohydrate metabolism was quite strong (Figure 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 classification 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 nonunique peptides so that deeper and more thorough surveys of the microbial protein landscape can be performed and evaluated alongside more robust transcriptional datasets.

**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. cellulosolyticus* WH2 genes in this study fell within two putative PULs. One PUL (BWH2_4044–55) contains 12 ORFs that include a dual suc/C/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 (Figure 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. Our mRNAs-
Figure 4. Two xylanase-containing \textit{B. cellulosilyticus} WH2 PULs demonstrating strong diet-specific expression patterns \textit{in vivo}. (A) The PUL spanning \textit{BWH2\_4044–55} includes a four-gene cassette comprising two consecutive \textit{susC/D} pairs, multiple genes encoding GHs and 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-\textit{b}-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 \textit{BWH2\_4072–6} contains a \textit{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) Heatmap visualization of GeneChip expression data for \textit{BWH2\_4044–55} and \textit{BWH2\_4072–6} showing marked up-regulation 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 \textit{BWH2\_4044–55} and \textit{BWH2\_4072–6} PULs that were detectable in the same material used for GeneChip quantitation in panel (B). Bars represent results (mean \pm SEM) from two technical runs per sample. For each MS run, the spectral counts for each protein were normalized against the total number of \textit{B. cellulosilyticus} WH2 spectra acquired. (D) Comparison of \textit{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 \textit{B. cellulosilyticus} WH2 on each diet, as measured by
level analysis disclosed that BWH2_4047 was the most highly expressed B. cellulosilyticus WH2 susD homolog on this diet. Likewise, BWH2_4046/4, the two susC-like genes within this PUL, were the 2nd and 4th most highly expressed B. cellulosilyticus WH2 susC-like genes in LF/HPP-fed animals, and exhibited expression level reductions 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 (Figures 2A and 4B) but not for neighboring genes. The same trends were obvious and amplified when we quantified protein expression (Figure 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) (Figure 4A) demonstrated a strong but opposite diet bias, in this case exhibiting significantly higher expression in animals consuming the HF/HS “Western”-like diet. Our mRNA-level analysis showed that this xylanase was the second most highly expressed B. cellulosilyticus WH2 CAZyme in animals consuming this diet (Figure 2A). As with the previously described PUL, shotgun metaproteomics validated the transcriptional analysis (Figure 4B,C); with the exception of the gene encoding the PUL’s presumed transcriptional regulator (BWH2_4076), diet specificity was substantial, with protein-level fold changes ranging from 10–33 across the locus (Table S10).

Recent work by Cann and co-workers 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 [41]. The “core” gene cluster of the prototypical xylan utilization system they described consists of two tandem repeats of susC/susD homologs (susABC/D), a downstream hypothetical gene (susE) and a GH10 endoxylanase (syn10c). 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 utilization capabilities of B. ovatus ATCC 8483 also identified two PULs involved in xylan utilization (BACOVA_03417–20) whose gene configurations differ from those described in Prevotella spp. [25]. Interestingly, the five genes encoded by the smaller xylanase-containing PUL described above (BWH2_4027–51) are homologous to the products of the last five genes in BACOVA_03417–20 (i.e., BACOVA_03417–20). The order of these five genes in these two loci is also identical. The similarities and differences observed when comparing the putative xylan utilization systems encoded within the genomes of different Bacteroidetes illustrate how its members may have evolved differentiated strategies for utilizing hemicelluloses like xylan.

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 an artificial community similar to the one used in E1 and E2 (see Materials and Methods). The wild-type B. cellulosilyticus WH2 strain used in our previous experiments was replaced with a transposon mutant library consisting of over 90,000 distinct transposon insertion mutants in 91.5% of all predicted ORFs (average of 13.9 distinct insertion mutants per ORF). The library was constructed using methods similar to those reported by Goodman et al. ([42]; see Materials and 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). The INSeq results revealed clear, diet-specific losses of fitness when components of these loci were disrupted (Figure 4D). Additionally, as observed in E1 and E2, expression of each PUL was strongly biased by diet, with genes BWH2_4044–55 demonstrating up-regulation 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. The four genes, three were susC or susD homologs (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_4072–6. Notably, disruption of BWH2_4076, which is predicted to encode a hybrid two-component regulatory system, had negative consequences on either diet tested, indicating that 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 that 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. cacaee

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 (MM) supplemented with one of 46 different carbohydrates [25]. 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 glyans (4 of 7) and one cellulosaccharide (cellulose). The seven substrates that did not support growth included three esoteric carbohydrates (carrageenan, porphyran, and algic acid), the simple sugar N-acetylmuramic acid, two host glyans (keratan sulfate and mucin O-glycans), and fungal cell wall-derived β-mannan. B. cellulosilyticus WH2 clearly grew more robustly on some carbohydrates than others. Excluding simple sugars, fastest growth was achieved on dextran (0.099±0.048 OD₆₀₀ units/h), laminarin (0.095±0.014), 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. cellulosilyticus degrades cellulose [43], the WH2 strain failed to demonstrate any growth on MM plus cellulose (specifically, Solka-Floc 200 FCC from International Fiber Corp.) after 5 d. Maximum cell density was achieved with amylopectin (1.17±0.02 OD₆₀₀ units), dextran (1.12±0.20), cellulose (1.09±0.08), laminarin (1.08±0.08), and xyloglucan (0.99±0.04). Total B. cellulosilyticus WH2 growth (i.e., maximum cell density achieved) on host-derived glyans was typically very poor, with only two substrates achieving total growth above 0.2 OD₆₀₀ units (chondroitin sulfate, 0.50±0.04; glycojen, 0.99±0.02). The disparity between total growth on plant polysaccharides versus host-derived glyans, including O-glycans that are prevalent in host mucin, indicates a preference for diet-derived saccharides, consistent with our in vivo mRNA and protein expression data.

We also determined how the growth rate of B. cellulosilyticus WH2 on these substrates compared to the growth rates for other prominent gut Bacteroides spp. After subjecting B. caccae to the same phenotypic characterization as B. cellulosilyticus WH2, we combined our measurements for these two strains with previously published measurements for B. thetaotaomicron and B. ovatus [25]. The results underscored the competitive growth advantage B. 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. cellulosilyticus WH2 grew fastest on six while B. ovatus grew fastest on two (Table S11). B. caccae and B. thetaotaomicron, 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. cellulosilyticus WH2 grew fastest on the greatest number of polysaccharides (11 of 26) and tied with B. caccae for the greatest number of monosaccharides (6 of 15). B. thetaotaomicron and B. caccae did, however, outperform the other two Bacteroides tested with respect to utilization of host glycanas in vivo, 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 xyloglucan, 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 MM (Table S12) plus one of 16 simple sugars or one of 15 complex sugars (Table S13) (see Materials and Methods; n=2–3 cultures/substrate; 5.2–14.0 million raw Illumina HiSeq reads generated for each of 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 (Figure 5A). The consistency of this clustering illustrates that (i) technical replicates within each condition exhibit strong correlations 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/arabinobioxyan (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 [44]).

Importantly, these findings suggested that by considering in vitro profiling data alongside in vivo expression data from the artificial community, it might be possible to identify the particular carbohydrates to which B. cellulosilyticus WH2 is exposed and responding 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, MM plus glucose (MM-Glc). The results revealed a dynamic PUL activation network in which some PULs were activated by a single substrate, some were activated by multiple substrates, and some 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 susG- and susD-like genes in the cassette were up-regulated an average of >100-fold relative to MM-Glc across all technical replicates (Figure 5B). At least one susCD 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 polysaccharide 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 up-regulation of five susC/D-like pairs (BWH2_0865/6, 0867/8, 4044/5, 4046/7, and 4074/5). 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 the heatmap in Figure 5B). In mice fed the polysaccharide-rich LF/HPP chow, B. cellulosilyticus WH2 up-regulates three susC/D pairs (BWH2_2717/8, 4044/5, 4046/7) whose expression is activated in vivo by arabinoxylan. The most significantly up-regulated susC/D pairs (BWH2_1736/7, 2514/5, 4074/5) in mice fed the HF/HS diet rich in sugar, corn starch,
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 cultures 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 heatmap 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...
and maltodextrin are activated in vitro by amylopectin, ribose, and xylan/arabinoxylan, respectively. All three PULs identified as being up-regulated at the RNA level in LF/HPP-fed mice were also found to be up-regulated at the protein level (Figure 5B). Two of the three PULs up-regulated at the mRNA level in HF/HS-fed mice were up-regulated at the protein level as well. The presence of an amylopectin-activated PUL among these two loci is noteworthy, given the significant amount of starch present in the HF/HS diet. The up-regulation of four other PULs in HF/HS-fed animals was only evident in our LC-MS/MS data, reinforcing 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 Figure 4A). Another one of the 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 Figure 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 vitro 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 an artificial 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 “leveraging” results obtained from in vitro assessments of individual species’ 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 artificial human gut microbial communities in gnotobiotic mice also allows us to evaluate the technical limitations of current molecular approaches for characterizing native communities. For example, the structure of an artificial 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. Short read cDNA sequences transcribed from total microbial community RNA can also often be assigned to the exact species and gene from which they were derived, and the same is also often true for peptides derived from particular bacterial 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 [2,45]. 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 [46]. We anticipate that the study of sequenced 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.

The recent development of a low-error 16S ribosomal RNA amplicon sequencing method [LEA-Seq] and the application of this method to the fecal microbiota of 37 healthy adults followed for up to 5 years indicated that individuals in this cohort contained 195 ± 48 bacterial strains representing 101 ± 27 species [47]. Furthermore, stability follows a power-law function, suggesting that once acquired, most gut strains in a person are present for decades. New advances in the culturing of fastidious gut microbes may one day allow us to capture most (or all) of the taxonomic and functional diversity present within an individual’s fecal microbiota as a clonally arrayed, sequenced culture collection, providing a perfectly representative and defined experimental model of their gut community. In the meantime, first-generation artificial communities of modest complexity such as the one described here offer a way of studying many questions related to the microbiota. However, the limited complexity and composition of our 12-species artificial community, and the way in which it was assembled in germ-free mice, make it an imperfect model of more
complex human microbiota. Native microbial communities, for example, are subject to the influence of variables that are notably absent from this system, such as intraspecies genetic variability and exogenous microbial inputs. There are also taxa (e.g., Proteobacteria, Bifidobacteria) and microbial guilds (e.g., butyrate producers) typical of human gut communities that are absent from our defined assembly that could be used to augment this system in order to improve our understanding of how their presence/absence influences a microbiota’s response to diet and a spectrum of other variables of interest. These future attempts to systematically increase complexity should reveal what trends, patterns, and trajectories observed in artificial assemblages such as the one reported here map or do not map onto natural communities.

Finally, one of the greatest advantages of studying defined assemblages in mice is that they afford us the ability to interrogate the biology of key bacterial species in a focused manner. The artificial community we used in our experiments included B. 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 artificial 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 symbiotics.

Materials and Methods

Ethics Statement

All experiments involving mice used protocols approved by the Washington University Animal Studies Committee in accordance with guidelines set forth by the American Veterinary Medical Association. Trained veterinarians from the Washington University Division of Comparative Medicine supervised all experiments. The laboratory animal program at Washington University is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International (AAALAC).

B. cellulosilyticus WH2 Genome Sequencing

A strain of B. cellulosilyticus designated “WH2” (see Figure S1A,B) was isolated from a human fecal sample during an iteration of the Microbial Diversity Summer Course overseen by A. Salyers (University 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 10 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 [48,49]. Gene calling was performed using the YACOP metatool [50]. 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 S4. Cells were grown in supplemented TYG (TYGm; [42]) at 37 °C under anaerobic conditions in a Coy anaerobic chamber (atmosphere: 75% N2, 20% H2, 5% CO2). After reaching stationary phase, cells were pelleted by centrifugation and resuspended in TYGm medium supplemented with 20% glycerol. Individual aliquots containing 400–500 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 strain 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 (AGAGTTTGTATCAGTGGTGCAG [51]) and 1391R (GAGGGGCTTGWGTGAC [52]), sequencing the entire amplicon with an ABI 3730 capillary sequencer (Applied Biosystems, Foster City, 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 S3. 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. To do so, an aliquot of cells from each overnight culture or set of frozen stocks was used to prepare a 10-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 artificial community in the germ-free animals. Total CFUs per gavage were estimated at 8.0 × 10⁸ and 4.2 × 10⁹ 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. For each experiment, two groups of 10–12-wk-old male germ-free C57BL/6J mice were maintained in flexible film 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 2 wk each (see Figure 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 sterile ‘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 2 wk, after which each treatment group was switched to the alternative diet. Two weeks later, the mice were switched back to their original starting diet and were retained on this diet until the time of sacrifice.

DNA and RNA Extraction
DNA and RNA were extracted from fecal pellets and cecal contents as previously described [11].

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 [11] using short-read (36 nt) data collected from an Illumina sequencer (data were generated using a combination of the Genome Analyzer I, Genome Analyzer II, and Genomes Analyzer IIx platforms). After demultiplexing each barcoded pool, reads were trimmed to 25 bp and aligned to the reference genomes. An abundance threshold cutoff of 0.003% was set for determining an artificial community members’ presence/absence, 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 intrasample percentage. In order to make changes in abundance over time more easily comparable between species with significantly different relative abundances, these percentages were also in some cases normalized by the maximum abundance (%) observed for a given species across all time-points from a given animal. This transformation resulted in a value referred to as the percentage of maximum achieved (“PoMA”) 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 [53]. Data from both E1 and E2 were combined to generate a single table-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. 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 [54].

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 purpose 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 five of seven 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 [55] to identify differentially expressed genes. Genes were generally considered significantly differentially expressed in cases meeting the following three criteria: $p < 0.01$, PPDE($<0.05$) $\geq 0.99$, and $|fold-change| > 2$.

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 [14]. Illumina libraries were prepared [11] from 26 fecal samples obtained from the second diet oscillation experiment (four 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 [56]. Normalization of the resulting raw counts was performed using the DESeq package in R [57]. Raw counts derived from the 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 [58] and GENE-E. Heatmap visualizations of expression data were prepared using JavaTreeview [59] 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, 100 mM Tris-HCl (pH 8.0), 10 mM dithiothreitol (DTT)), lysed mechanically by sonication, incubated at 95°C for 5 min, and centrifuged at 21,000 × 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 8 M urea and 100 mM Tris-HCl (pH 8.0), reduced by incubation in DTT (final concentration of 10 mM) 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 s on, 10 s 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 100 mM Tris-HCl and 10 mM CaCl2 (pH 8.0) to a final urea concentration below 4 M. Proteolytic digestions were initiated with sequencing grade trypsin.
An 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 considered 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 for the Growth of Bacteroides spp. on Various Carbohydrates

The ability of B. cellulosityticus WH2 and B. 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 [23]. Growth measurements were collected in duplicate (two wells per substrate) over the course of 3 d at 37°C under anaerobic conditions. A total of three independent experiments were performed for each species tested (n = 6 growth profiles/substrate/species). Total growth (A_{tot}) was calculated from each growth curve as the difference between the maximum and minimum optical densities (OD_{600}) observed (i.e., A_{max} - A_{min}). Growth rates were calculated as total growth divided by time (A_{tot}/(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 B. cellulosityticus WH2 Gene Expression During Growth in Defined MM Containing Various Carbohydrates

RNA-Seq. To characterize the impact of select mono- and polysaccharides on the *in vitro* gene expression of *B. cellulosityticus* WH2, cells were cultured in 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 and 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 (MM-Glc). Anaerobic conditions were generated within each individual culture tube using a previously described method [64] 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% (v/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 were 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 *B. cellulosityticus* WH2’s logarithmic phase of growth. Finally, 50 μL of each “acclimatization” culture were inoculated into triplicate 10 mL volumes of medium of the same composition, and the 90 “harvest” cultures were grown anaerobically at 37°C. At mid-log phase, 5 mL of cells were immediately preserved in Qiagen RAPNaprotect Bacteria Reagent according to the manufacturer’s instructions. Cells were then pelleted, RAPNaprotect 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 (200 mM NaCl, 20 mM EDTA), 210 μL of 20% SDS, and 500 μL of acid phenol-chloroform-isomyl 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 placed on ice, 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 RNA was extracted as described above for fecal and cecal samples.

Identification of Diet-Specific Fitness Determinants within the B. cellulosilyticus WH2 Genome Using Insertion Sequencing (INSseq)

Whole genome transposon mutagenesis of B. cellulosilyticus WH2 was performed using protocols originally developed for B. thetaitaoenicron [42,46], with some modifications. Initial attempts to transform B. cellulosilyticus WH2 with the pSAM_Bt construct reported by Goodman 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 (ermG) with the promoter for the gene encoding EF-Tu in B. cellulosilyticus WH2 (BW2_3183, dramatically improved the number of resistant clones recovered after transformation. The resulting library consisted of 93,458 distinct isogenic mutants, with each mutant strain containing a single randomly inserted modified mariner transposon. Of all distinct isogenic mutants, with each mutant strain containing a single randomly inserted modified mariner transposon. Of all predicted ORFs, 91.5% had insertions covering the first 80% of each gene (mean, 13.9 distinct insertion mutants per ORF).

At 11 wk 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, animals received a single gavage containing the B. cellulosilyticus identified in the fecal and cecal samples. B. cellulosilyticus WH2 to other Bacteroidetes species. (A) Near full-length 16S rRNA gene sequences from the B. cellulosilyticus WH2 isolate, its closest relatives (two strains of Bacteroides xylanisolvens, three strains of Bacteroides intestinalis, and the type strain of B. cellulosilyticus), and Parabacteroides distasonis (the latter was included as an outgroup) were aligned against the SILVA SEED using the SINA aligner [65]. The 3’ 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 [66]. Sequences in the trimmed alignment used to generate the tree shown correspond to bases 22–1498 of the Escherichia coli 16S rRNA gene [67]. Parenthetical identifiers indicate the locus tag for B. cellulosilyticus WH2, whose genome contains four copies of the 16S rRNA gene or GenBank 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 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 closest relatives (two strains of Bacteroides xylanisolvens, three strains of Bacteroides intestinalis, and the type strain of B. cellulosilyticus), and Parabacteroides distasonis (the latter was included as an outgroup) were aligned against the SILVA SEED using the SINA aligner [65]. The 3’ 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 [66]. Sequences in the trimmed alignment used to generate the tree shown correspond to bases 22–1498 of the Escherichia coli 16S rRNA gene [67]. Parenthetical identifiers indicate the locus tag for B. cellulosilyticus WH2, whose genome contains four copies of the 16S rRNA gene or GenBank 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 GH families identified in the B. cellulosilyticus WH2 genome compared to their representation in other sequenced Bacteroidetes species. Enumeration of the GH repertoire of B. cellulosilyticus WH2 relative to (A) the six other Bacteroidetes species included in the artificial microbial community described in Figure 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 in any of the other Bacteroidetes to which it is being compared. An asterisk following a GH family number indicates that 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.

(TIF)

Figure S3 Design and sampling schedule for experiments E1 and E2. In each experiment, two groups of C57BL/6J germ-free mice were gavaged at 10–12 wk of age with a 12-member artificial human gut microbiota community (the day of gavage, referred to as day 0, is denoted by a large black arrow). Over time, animals were fed diets low in fat and high in plant polysaccharides (LF/HPP, bold green) or high in fat and simple sugar (HF/HS, bold orange) in alternating fashion. Fecal pellets and cecal contents were collected as indicated for profiling community membership and gene expression (sample types are denoted by a circle’s color and the methods applied to each sample are indicated in parentheses within the sample key). Values shown along the time course indicate the number of days since gavage of the artificial community into germ-free animals.

(EPS)

Figure S4 COPRO-Seq analysis of the proportional representation of component taxa in the 12-member artificial community 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 E1. (B) DNA yields from samples collected in experiment E2. (C) COPRO-Seq quantitation of the 12 bacterial species comprising the assemblage used to colonize germ-free mice in experiments E1 and E2. 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 E1 and E2, 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.

(EPS)

Figure S5 Further COPRO-Seq analysis of the relative abundance of components of the 12-member bacterial community as a function of diet and time. (A) Plot of the ordination results for experiment 1 (E1) from the PCoA described in Figure 1B. COPRO-Seq data from E1 and E2 were ordinated in the same multidimensional space. For clarity, only data from E1 are shown (for the E2 PCoA plot, see Figure 1B). Color code: red/blue, feces; pink/cyan, cecal contents. (B) Heatmap representation of the relative abundance data from E1 normalized to each species’ maximum across all time-points within a given animal (“Percentage of maximum achieved (PoMA)”). Each heatmap cell denotes the mean for one treatment group (n = 7 animals), and each treatment group is shown as its own heatmap.

(EPS)

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 ≥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 taxa that maintained a proportional representation in the cecal microbiota that was >5% on each diet.

(EPS)

Figure S7 Dissecting the 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 human gut microbiome were binned by EC number annotation and normalized (i.e., data were “community-normalized” at the level of ECs), a significant decrease in the representation of EC 3.2.1.8 in the metatranscriptome was observed when comparing the first time-point of the diet phase (day 13, LF/HPP diet) and the final time-point of the second diet phase (day 27, HF/HS diet) (*Mann–Whitney U* test, p = 0.03). (B) Transcribed *B. cellulosilyticus* WH2 genes account for >99% of community-normalized RNA-Seq counts assignable to EC 3.2.1.8 (note how counts at the community level in panel (A) compare to those attributable to *B. cellulosilyticus* WH2 in panel (B)). Thus, *B. cellulosilyticus* WH2 essentially dictates the degree to which expressed endo-1,4-β-xylanase genes are represented within the metatranscriptome. (C) *B. cellulosilyticus* WH2 contributes a greater number of community-normalized RNA-Seq counts to the metatranscriptome in LF/HPP-fed mice than in HF/HS-fed animals. (D) When *B. cellulosilyticus* WH2 gene expression data are normalized independently of data from other taxa (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 panel (D) by the *B. cellulosilyticus* WH2 gene from which they were derived. For all panels (A–E), 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). In each of the first four panels (A–D), the differences between day 13 and day 27 were deemed statistically significant by *Mann–Whitney U* test (p = 0.03 for each of the four tests performed).

(EPS)

Figure S8 Shotgun metaproteomic analysis of cecal samples from gnotobiotic mice colonized with the 12-member artificial community. (A) Each species’ theoretical proteome was subjected to in silico trypsinization (see Materials and 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 assemblage members, the mouse, and three bacterial “distractors” (*E. rectale*, *F. 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 assemblage 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 the Pearson correlation coefficient (r) between log-transformed averages of diet-specific fold-differences in expression 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 was 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 “Transla-
tion” (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).

Table S1 Sequencing statistics for B. cellulosilyticus
WH2.
(XLSB)

Table S2 B. cellulosilyticus WH2 genome features with relevance to carbohydrate metabolism.
(XLSB)

Table S3 Composition of the 12-member artificial community inoculated by oral gavage into germ-free
animals.
(XLSB)

Table S4 Bacterial strains included in this study.
(XLSB)

Table S5 COPRO-Seq quantitation of the relative abundances of artificial community members over time.
(XLSB)

Table S6 GeneChip measurements of cecal gene expression for the 12 bacterial species comprising the artificial human gut microbial community studied in experiment E1.
(XLSB)

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

Table S8 Summary of theoretical peptidome statistics.
(XLSB)

References

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

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

Table S11 Growth of B. cellulosilyticus WH2 and B. cacae on a panel of structurally diverse carbohydrates.
(XLSB)

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

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

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

Text S1 Supplementary results.

Acknowledgments
We thank Maria Karlsson, David O’Donnell, Sabrina Wagoner, and Su Deng for superb technical assistance, as well as Nathan VerBerkmoes, Alejandro Reyes, and Federico Rey for helpful suggestions throughout the course of this study.

Author Contributions
The author(s) have made the following declarations about their contributions: Conceived and designed the experiments: NPM JIG. Performed the experiments: NPM MW ARE NAP. Analyzed the data: NPM. Contributed reagents/materials/analysis tools: ECM BH RLH JIG. Wrote the paper: NPM ARE RLH JIG.
Appendix C

Alejandro Reyes, Meng Wu, Nathan P. McNulty, Forest L. Rohwer, and Jeffrey I. Gordon

Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut.

Gnotobiotic mouse model of phage–bacterial host dynamics in the human gut

Alejandro Reyes¹,², Meng Wu³, Nathan P. McNulty³, Forest L. Rohwer⁴, and Jeffrey I. Gordon⁵

¹Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108; and ²Department of Biology, San Diego State University, San Diego, CA 92186

Contributed by Jeffrey I. Gordon, October 22, 2013 (sent for review August 29, 2013)

Bacterial viruses (phages) are the most abundant biological group on Earth and are more genetically diverse than their bacterial prey-hosts. To characterize their role as agents shaping gut microbial community structure, adult germ-free mice were colonized with a consortium of 15 sequenced human bacterial symbionts, 13 of which harbored one or more predicted prophages. One member, Bacteroides cellulosilyticus WH2, was represented by a library of isogenic transposon mutants that covered 90% of its genes. Once assembled, the community was subjected to a staged phage attack with a pool of live or heat-killed virus-like particles (VLPs) purified from the fecal microbiota of five healthy humans. Shotgun sequencing of DNA from the input pooled VLP preparation plus shotgun sequencing of gut microbiota samples and purified fecal VLPs from the gnotobiotic mice revealed a reproducible nonsimultaneous pattern of attack extending over a 25-d period that involved five phages, none described previously. This system allowed us to (i) correlate increases in specific phages present in the pooled VLPs with reductions in the representation of particular bacterial taxa, (ii) provide evidence that phage resistance occurred because of ecological or epigenetic factors, (iii) track the origin of each of the five phages among the five human donors plus the extent of their genome variation between and within recipient mice, and (iv) establish the dynamic in vivo fitness advantage that a locus within a B. cellulosilyticus prophage conveys upon its host. Together, these results provide a defined community-wide view of phage–bacterial host dynamics in the gut.

The human gut is home to tens of trillions of microbial cells representing all three domains of life, although most are bacteria. These organisms collaborate and compete for functional niches and physical locations (habitats). Together, they form a continuously functioning microbial metabolic “organ.” The microbial diversity, interpersonal variation, and dynamism of the gut microbiota make the task of identifying the factors that define community configurations extremely challenging.

In some ecosystems, phages maintain high bacterial strain level diversity through lysis of their host strains (constant diversity dynamics model; refs 1, 2). The resulting emptied niche is filled with either an evolved resistant bacterial strain or a taxonomically closely related bacterial species. These dynamics have been observed in open marine environments (1). In contrast, a recent study of 37 healthy adults indicated that a person’s fecal microbiota was remarkably stable, with 60% of bacterial strains retained over the course of 5 y (3). Stability followed a power law dynamic that when extrapolated suggests that most strains in an individual’s gut community are retained for decades (3). In a metagenomic analysis of virus-like particles (VLPs) purified from the fecal microbiota of healthy adult monozygotic twins and their mothers, sampled over the course of a year, viral community structure exhibited high interpersonal variation. In contrast, the viral (phage) population within an individual was very stable over time, both at the level of sequence conservation and relative abundance (4). These observations, as well as other reports (5–7), suggest that temperate lifestyles, rather than a predator–prey relationship, dominate the phage–host bacterial cell dynamic in the distal guts of healthy humans.

To improve our understanding of viral–bacterial host dynamics, we constructed a gnotobiotic mouse model containing a simplified defined artificial community composed of 15 prominent human gut-derived bacterial taxa whose genomes had been sequenced (Dataset S1). This 15-member artificial community was used as bait for a staged attack that involved oral gavage of VLPs purified from human fecal samples. This system allowed us to (i) test whether phage populations would mount a simultaneous attack on susceptible members of the microbial community or whether such an attack would be nonsimultaneous (i.e., have an identifiable sequence), (ii) document the capture of previously unknown viruses present in the VLP preparations by members of the artificial community, and (iii) track induction of native prophages.

Results and Discussion

Attacking a 15-Member Artificial Human Gut Microbiota with VLPs Isolated from the Fecal Microbiota of Healthy Adult Humans. Our experimental design consisted of three groups of germ-free C57BL/6J mice (n = 5 per group). Each group was kept in a separate gnotobiotic isolator, where each mouse was individually caged. The first group was gavaged with the 15-member artificial community at 8 wk of age. Three weeks later, they were each gavaged with a pool of VLPs (p-VLP) isolated from fecal samples obtained from five healthy humans (“live p-VLP group”). A “heat-killed p-VLP group” was also colonized with the artificial microbe from the fecal microbiota of healthy adult monozygotic twins and their mothers, sampled over the course of a year, viral community diversity model; refs 1, 2). The resulting emptied niche is filled with either an evolved resistant bacterial strain or a taxonomically closely related bacterial species. These dynamics have been observed in open marine environments (1). In contrast, a recent study of 37 healthy adults indicated that a person’s fecal microbiota was remarkably stable, with 60% of bacterial strains retained over the course of 5 y (3). Stability followed a power law dynamic that when extrapolated suggests that most strains in an individual’s gut community are retained for decades (3). In a metagenomic analysis of virus-like particles (VLPs) purified from the fecal microbiota of healthy adult monozygotic twins and their mothers, sampled over the course of a year, viral community structure exhibited high interpersonal variation. In contrast, the viral (phage) population within an individual was very stable over time, both at the level of sequence conservation and relative abundance (4). These observations, as well as other reports (5–7), suggest that temperate lifestyles, rather than a predator–prey relationship, dominate the phage–host bacterial cell dynamic in the distal guts of healthy humans.

To improve our understanding of viral–bacterial host dynamics, we constructed a gnotobiotic mouse model containing a simplified defined artificial community composed of 15 prominent human gut-derived bacterial taxa whose genomes had been sequenced (Dataset S1). This 15-member artificial community was used as bait for a staged attack that involved oral gavage of VLPs purified from human fecal samples. This system allowed us to (i) test whether phage populations would mount a simultaneous attack on susceptible members of the microbial community or whether such an attack would be nonsimultaneous (i.e., have an identifiable sequence), (ii) document the capture of previously unknown viruses present in the VLP preparations by members of the artificial community, and (iii) track induction of native prophages.

Significance

A consortium of sequenced human gut bacteria was introduced into germ-free mice followed by a “staged” phage attack with virus-like particles purified from the fecal microbiota of five healthy adult humans. Unique phages were identified attacking microbiota members in nonsimultaneous fashion. Some host bacterial species acquired resistance to phage attack through ecological or epigenetic mechanisms. Changes in community structure observed after attack were transient. Spontaneous induction of prophages present in seven bacterial taxa was modest, occurring independently of the phage attack. Together, these results reveal a largely temperate phage–bacterial host dynamic and illustrate how gnotobiotic mouse models can help characterize ecological relationships in the gut by taking into account its most abundant but least understood component, viruses.


The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the European Nucleotide Archive (Project ID PRJEB4370).

1Present address: Departamento de Ciencias Biológicas, Universidad de los Andes, Bogotá 111711, Colombia.

2To whom correspondence should be addressed. E-mail: jgordon@wustl.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1319470110/-/DCSupplemental.
Neither the bacterial gavage nor the p-VLP inoculum contained components that appeared to compromise gut barrier/immune function or perturb overall health status. At the time of sacrifice, none of the treatment groups exhibited any significant differences in total body weight or adiposity (epididymal fat pad weight as a percentage of total body weight; $P = 0.3957$ and $P = 0.4794$, respectively; Kruskal–Wallis test; Dataset S1). FACS analysis did not reveal any differences between the groups in the CD4$^+$ and CD8$^+$ T-cell compartments of their spleens or mesenteric lymph nodes (MLNs), as judged by CD44 and CD62L

The weight of feces, increased linearly and abruptly in the 4 d following introduction of the artificial community in both the live and heat-killed p-VLP groups (Fig. S2A). Fecal DNA concentrations correlated significantly with fecal bacterial cell counts (determined by flow cytometry; Spearman correlation, 0.843; Fig. S2B).

Microbial biomass, defined as nanograms DNA/milligram wet weight of feces, increased linearly and abruptly in the 4 d following introduction of the artificial community in both the live and heat-killed p-VLP groups (Fig. S2A). Fecal DNA concentrations correlated significantly with fecal bacterial cell counts (determined by flow cytometry; Spearman correlation, 0.843; Fig. S2B).

Because the genome sequences of the 15 bacterial taxa were known, we used community profiling by sequencing (COPRO-Seq; ref. 8), a method based on short read (50 nt) shotgun sequencing of total fecal community DNA, to quantify the relative abundance of each taxon as a function of time after initial colonization and after the staged VLP attack [2,544,433 ± 96,255 (mean ± SEM) reads per sample; Dataset S1]. Principal coordinates analysis of a Hellinger distance matrix constructed from the COPRO-Seq datasets showed that most of the variation in composition over time occurred during the period of initial community assembly (Fig. S3A). Changes in the relative abundance of community members also occurred following gavage of the live but not the heat-killed p-VLP preparation (Fig. 1A–C and Fig. S3B–N).

**A Nonsimultaneous Pattern of Change in the Abundances of Five Phages.** To identify which exogenously administered VLP-associated viruses might be causing the observed structural rearrangements in community configuration, we modified our previously reported method for purifying VLPs (4) so that it could be applied to mouse fecal samples. We then sequenced DNA isolated from the purified VLP preparations ($n = 27$; two fecal pellets/VLP preparation, each amplified by multiple displacement amplification (MDA); 49,819 ± 6,983 (mean ± SEM) pyrosequences per sample; Fig. S1 and Dataset S1). To discriminate between activation of endogenous prophage in members of the artificial community versus exogenous viruses derived from the p-VLP preparation, reads generated from either the input VLPs or total mouse fecal DNA were mapped to the sequenced genomes of bacterial community members and to the mouse genome. We used reads without

Fig. 1. 1. Sequential changes in the relative abundance of two members of the 15-member artificial human gut microbiota and correlation with the appearance of two previously undescribed phages. (A) Relative abundance plot for each bacterial species as a function of time for either the live p-VLP or the heat-killed p-VLP treatment groups. Mean values ± SEM are shown ($n = 5$ mice). The color key next to the plot indicates the identity of each bacterial species. (B and C) Plots of the relative abundance (fraction of the total community; mean ± SEM; $n = 5$ animals per treatment group) of B. caccae and B. ovatus in the fecal microbiota of gnotobiotic mice as a function of time before and after gavage with live purified VLPs pooled from the fecal microbiota of five human donors or a control heat-killed version of the same p-VLP preparation (time of gavage indicated by the upward pointing arrow; $t = 0$ on the x axis refers to the time of introduction of the 15-member artificial community into germ-free animals). The change in abundance of these Bacteroides spp. occurs in a reproducible sequence among individually caged mice that received live but not heat-killed p-VLPs. (D and E) Changes in the abundance of two phages, derived from the p-VLP sample, in the fecal microbiota of recipient gnotobiotic mice. Differences in the time course of change in bacterial and viral abundances are highlighted by the gray shading (lighter for B. caccae and ϕHSC01). Insets in D and E are assembled genome sequences for ϕHSC01 and ϕHSC02. The location of genes on the positive strand (green) and negative strand (red) are shown; those that have significant sequence similarity to known viral genes are colored blue (blastp $e$-value < 10$^{-5}$; Dataset S1). The inner plot represents GC skew based on 200-bp windows (yellow, G/C ratio is greater than the average for the genome; purple, ratio is lower than the average).
significant matches to either dataset to characterize viral genomes not represented in the starting 15-member artificial community.

In total, five viral genomes, none of which have been described previously, were assembled and annotated from these analyses. These viruses were detected in the gut communities of mice that had received the live p-VLP preparation but not in the heat-killed p-VLP group (Fig. 1 D and E, Fig. S4 A–H, and Dataset S1). Rather than finding a concurrent attack on all susceptible members of the model human gut microbiota, we observed a nonsimultaneous pattern of change in the abundances of these viruses with corresponding changes in the representation of community members.

A DNA virus with a circular 37-kb genome, human synthetic community phage 01 (ϕHSC01), was the first to significantly increase in abundance. It not only encodes typical phage proteins (e.g., terminase, tail protein, DNA polymerase; Dataset S1), but also a protein containing a Bacterioidetes-associated carbohydrate-binding often N-terminal (BACON) domain (Profam ID: PF13004) postulated to target glycoproteins and possibly host mucin (9). From the time of its first detection in feces 24 h after animals were gavaged with the live p-VLP preparation, the marked increase in abundance of ϕHSC01 over the course of the next 2 d correlated with a decrease in the abundance of Bacteroides caccae (R² = –0.446; P = 3.2 × 10⁻⁵ after Bonferroni correction; Fig. 1 B and D). No other community member showed a statistically significant inverse correlation, suggesting that this bacterium is a host for ϕHSC01. The drop in the relative abundance of B. caccae was abrupt, occurring over the course of 1 d between days 2 and 3 after VLP gavage. The 74.5 ± 3.7% decrease relative to pretreatment levels was followed by a recovery to 75.8 ± 5.2% of the pre-VLP gavage values within 5–6 d (Fig. 1). This fourfold decrease was independently validated using quantitative (q)PCR (from 3.4 × 10⁵ to 7.8 × 10⁴ genome equivalents/mg of fecal pellet; see SI Methods). The spike in viral abundance, just like the coincident reduction in B. caccae abundance, was remarkably consistent in terms of its magnitude and time course among the individually caged members of this treatment group.

**Evidence That Phage Resistance Occurs Because of Ecological or Epigenetic Mechanisms.** To determine whether B. caccae’s recovery after viral attack was based on acquisition of identifiable fixed changes in its genome, we performed deep shotgun sequencing of total fecal community DNA isolated from samples obtained 9–23 d after gavage of the live p-VLP preparation. Pooling sequencing reads from these four groups of samples allowed us to assemble the B. caccae genome at an average coverage of 30-fold per treatment group. This analysis revealed that the genome of B. caccae is acquired due to phase variants or down-regulation of a phage receptor, making B. caccae resistant to viral attack without any discernible mutations in its genome (epigenetic mechanism). Another explanation is that mutations occur in one or more regions of the genome that are difficult to sequence and/or assemble.

### Nonsimultaneous Detection of Viruses and Community Rearrangements

The second virus to show an increase in abundance was ϕHSC02. All four predicted proteins encoded by this phage with a 6.2-kb genome (Fig. 1E) exhibit significant similarity to the Alpavirinae, a recently described subfamily identified from Bacteroidetes prophages, that targets the Microviridae (previously considered to be exclusively lytic) (11, 12). The changes in relative abundance of ϕHSC02 best correlated with a change (reduction) in the abundance of Bacteroides ovatus. Expansion of this virus and the attendant decrease in B. ovatus were first detected 2 d after the “crash” of B. caccae (i.e., within 5 d of gavage with the live p-VLP preparation) and coincided with the onset of recovery of the B. caccae population (Fig. 1 B–E).

The six- to eightfold reductions in relative representation of these two Bacteroides species took place during the 7-d period after gavage with live p-VLPs and were followed by a rise in abundance over a 7–8-d interval (Fig. 1 B and D). As these organisms increased their representation, we documented transient decreases in Bacteroides cellulosolvens, as well as the two previously described phages (i.e., HSC01 and HSC05) that still persisted in the community. At the same time, levels of Parabacteroides distasonis, Clostridium symbiosum, Clostridium scindens, and Ruminococcus obeum rose. These successional changes in bacterial abundance were limited to the group of mice that had received the live p-VLPs (Fig. S3).

The rise and fall of these organisms in the live p-VLP group occurred during a time period when three other previously undescribed viruses appeared: ϕHSC03 (153.4 kb), ϕHSC04 (104.2 kb), and ϕHSC05 (95.7 kb). These viruses, initially detected 7 d after gavage of the live p-VLP preparation, subsequently increased in abundance to approximately equivalent levels and persisted during the remaining 14 d of the experiment (Fig. S4 C–H). Unlike the distinctive negative correlation between ϕHSC01 and B. caccae abundance, and subsequently ϕHSC02 and B. ovatus abundance, the simultaneous appearance and rise of ϕHSC03, ϕHSC04, and ϕHSC05, their subsequent persistence, and the coincident complex patterns of change in the abundances of bacterial community members during this later period of the experiment made it difficult for us to assign candidate bacterial hosts to these three previously undescribed phages. Therefore, the same live p-VLPs used for the in vivo staged phage attack were also used for in vitro “attacks” of monocultures of each of the species present in the 15-member artificial community (SI Methods). There was no enrichment of any of the five previously undescribed viruses in any of the cultures (n = 2 independent experiments), suggesting that the bacterial host susceptibilities and requirements for infection with these phages are not recapitulated under these in vitro conditions (SI Results and Dataset S1). This observation highlights the value of gnotobiotic animal models for isolation of previously undescribed gut viruses.

### Tracking the Origin of Each of the Five Phages Among the Human Donors as Well as Their Genome Variation Between and Within Recipient Mice

To determine whether ϕHSC01–ϕHSC05 were distributed among all of the human VLP donors or whether they were unique to particular individuals, we generated a hybrid assembly using reads from the original VLP-derived viromes from each of the five donors as well as from the p-VLP preparation used for gavage (SI Methods). The hybrid assembly yielded 159 contigs greater than 2 kb. Based on read distribution and percent identity in contigs, we concluded that ϕHSC01 and ϕHSC04 originated from a twin in family 2 (F2T1.2), whereas ϕHSC02 and ϕHSC03 originated from the twin in family 4 (F4T1.2) (Fig. S5A). ϕHSC05 was observed in four of the five individuals used to construct the VLP pool. Mapping reads from each human donor fecal virome to ϕHSC05 revealed that the
virus recovered from mice likely came from individual F3T1.2 because the average percent identity of reads from this person’s virome mapping to the ϕHSC05 genome was equivalent to the percent identity obtained from VLPs isolated from mouse fecal samples (Fig. S5B). Nonetheless, analyzing VLP DNA purified from fecal samples obtained from each recipient mouse 9 d and 17 d after gavage of the live p-VLPs, and from cecal samples obtained at sacrifice, revealed animal-to-animal variations in ϕHSC05 genome structure (Fig. S5C). These genomic variations could have been present in the human donor fecal virome and distributed to different mouse recipients. However, we also observed variations in ϕHSC05 genome structure within mice over time (Fig. S5C), raising the possibility of a red queen dynamic (evolution of the viral genome in response to evolution of the bacterial host over time). It is important to note that none of the other four phages (ϕHSC01–4) showed this type of variation between animals, or within individual animals as a function of time (Fig. S5 D–G), indicating that ϕHSC05’s genome evolution is not a general feature in mice harboring this artificial community. Together, our findings not only illustrate the utility of gnotobiotic mice for identifying candidate bacterial hosts for human gut-associated phage, but also for identifying differences in properties among related virote types derived from different human gut virome.

**Distribution and Persistence of Phages Within the Gut.** Intestinal transit time in mice is on the order of several hours (13). We collected intestinal contents from the proximal and distal small intestine, cecum, and colon, as well as a fecal sample from each mouse in the live and heat-killed p-VLP treatment groups at the time of their sacrifice (25 d after the staged p-VLP attack). All gut samples were processed for COPRO-Seq analysis, and an aliquot of cecal contents was used to isolate VLPs for subsequent shotgun sequencing of viral DNA (Dataset S1). The results revealed no detectable phages in any of the gut segments from mice that received the heat-killed p-VLP preparation (Fig. S6A). In the live p-VLP treatment group, neither ϕHSC03, ϕHSC04, nor ϕHSC05 exhibited significant differences in their relative abundances between the distal small intestine and distal colon, and between luminal contents and feces (Fig. S64). Moreover, at the time of sacrifice, there were no significant biogeographical differences in the relative abundance of bacterial species within or between members of the two treatment groups (Fig. S6D).

The fact that ϕHSC03, ϕHSC04, and ϕHSC05 first appeared in VLPs from the live p-VLP treatment group 7 d after the single gavage of p-VLPs suggests that an intra- and/or extracellular compartment/reservoir exists that harbors components of the administered human fecal phage population. Pseudolysogeny, a state where phages exist in a host bacterial cell without multiplying or synchronizing their replication with the host (14) could represent one potential mechanism for persistence. Hypervariable domains, including C-type lectins, have been identified in gut-associated phages (15), suggesting that extracellular sequestration with binding to mucus or epithelial cell surface glycans could represent another potential mechanism. Barr et al. found that end-revealed of phages in mucus occurs through binding of Ig-like domains exposed on phage capsids to carbohydrate residues present in the mucin glycoprotein component of mucus, thereby creating a form of antimicrobial defense that could protect mucosal surfaces (16). COPRO-Seq analysis of cecal samples obtained from mice in the germ-free treatment group that lacked the 15-member artificial community and were gavaged with the live p-VLP preparation alone revealed no detectable phages in the cecum at the time of sacrifice (Dataset S1), supporting the notion that persistence of ϕHSC03–ϕHSC05 may be dependent upon the presence of bacteria. [Note that members of the gut microbiota, including *Bacteroides* spp. represented in the artificial community, are known to impact the mucus layer and mucosal glycans through a variety of means (17–19).] Although we cannot completely rule out the possibility that ϕHSC03 and ϕHSC04 were first detected at later time points because there had been a selection for mutants that could replicate better, this seems unlikely; their late appearance was a common feature in mice receiving the live p-VLP preparation, and, as noted above, there was no obvious variation in their viral genomes between animals and within a given mouse over time (Fig. S5 F and G).

**Prophage Activation in the 15-Member Artificial Community.** Thirteen of the 15 bacterial taxa in the artificial community had predicted prophages in their genomes (Dataset S1). To verify these predictions and to assess the capacity of these prophages to undergo induction, we used reads obtained from shotgun pyrosequencing of DNA isolated from two sources: (i) VLPs purified from fecal samples collected at weekly intervals from animals gavaged with live p-VLP preparation and (ii) VLPs purified from cecal samples obtained at sacrifice. Instead of mapping randomly throughout bacterial genomes (implying a background level of bacterial DNA contamination in the purified VLPs), VLP reads mapped to one or more of the predicted prophages. In this way, we identified 10 prophages derived from seven bacterial genomes that had the capacity to undergo induction in vivo (Fig. S7).

*B. cellulosilyticus* WH2 has two prophages, one of which (prophage 1) exhibited the greatest fold-induction among these 10 prophages. Prophage 1 has a lambdaoid genome architecture (Fig. 24) with syntenic arrangements identified in other *Bacteroides* genomes (Fig. S9). WH2 prophage 1 induction was observed 7–9 d after initial gavage of the 15-member artificial community prior to introduction of either live or heat-killed p-VLPs. Induction occurred at the end of the period of initial bacterial community assembly, right after microbial biomass reached its peak (Fig. S2A), suggesting a potential role of bacterial density in the induction process. Induction of prophage 1 correlated with a decrease in the relative abundance of its bacterial host (Fig. 2 B and C). The other *B. cellulosilyticus* prophage (prophage 2) did not exhibit significant levels of induction at any time point surveyed during the experiment (Fig. 2D and Fig. S5B; also see SI Results showing that prophage 2 was induced in a few mice).

*B. cellulosilyticus* WH2 was represented in the 15-member artificial community by a library of 93,458 isogenic mutants, with each mutant strain containing a single randomly inserted modified mariner transposon (Tn) (91.5% of predicted ORFs had insertions covering the first 80% of each gene with an average of 13.9 insertions per ORF). Because the modified Tn had engineered recognition sites for the type II restriction endonuclease MmeI at its ends, 16 bp of flanking chromosomal DNA could be excised together with the Tn to form a new genome. A community DNA was sequenced (20). This makes it possible to use high-throughput sequencing to define the precise location and abundance of each transposon mutant in the library (SI Methods).

Comparing the number of reads for each mutant in an “output” population after a given selection to the number of reads generated from an “input” population provides information about the effect each transposon insertion had on the fitness of the organism under the selection condition applied (20, 21). Applying this Tn INsertion Sequencing (INSeq) analysis to DNA prepared from fecal samples collected before, during, and after prophage 1 induction showed a dramatic enrichment for transposons located within a ~600bp intergenic region positioned between the ORFs encoding the prophage’s putative Rha protein (22) and cI repressor at the time of its induction 5–9 d after introduction of the 15-member artificial community (Fig. S9 A and B and SI Results). The intergenic region upstream of cI in phage lambda is an extremely well-studied transcriptional regulatory region; the right operator (OrR) with its three sites that competitively bind the repressor and Cro proteins, constitutes a carefully regulated switch between lysogenic and lytic cycles (23, 24). Upon RecA activation, the repressor is cleaved and the prophage is induced (25). Thus, accumulation of mutations in this region could have important consequences on the regulation and lifecycle of the lambdaoid prophage 1 and its bacterial host.

In some mice, enrichment of strains with Tn inserts in the cI-rha intergenic region was observed as early as 2 d after gavage.
of the 15-member artificial community (before prophage induction). Enrichment did not reflect clonal expansion of a single mutant strain within a given animal, but rather expansion of 1 or more of 10 independent mutants, each harboring a single transposon insertion within this intergenic region. The number and sites of these insertion mutants varied between animals (Fig. S9A). Moreover, no Tn insertions were observed within the ORF encoding the putative cI repressor or in the region 100 bp immediately upstream of the ORF in the input library, nor in any of the output fecal samples (Fig. S9A), suggesting an essential role for the repressor and the upstream OR region in bacterial host fitness (Fig. S9 C and D).

Control experiments were carried out for 20 d (the time before p-VLP gavage) using the same 15-member artificial community but where B. cellulosilyticus WH2 was represented by the wild-type strain rather than by a library of Tn mutants. Although community assembly and structure were highly similar to that observed when the artificial community contained the Tn mutant library (compare Fig. S8C with Fig. 1A), neither prophage 1 nor prophage 2 were induced, and no drop in B. cellulosilyticus WH2 abundance was observed (Fig. S8D). These findings suggest that Tn mutations in the cI-rha intergenic region facilitate prophage induction in the mouse model.

We defined the time course of clonal expansion of bacteria containing Tn inserts in this intergenic region to quantify the fitness effects of disrupting this part of the prophage genome. Reads mapping to this locus represented 77.9 ± 4.6% (mean ± SEM) of all Tn reads in fecal samples collected between 5 d and 9 d after bacterial gavage (range, 61.3–99.9%). A sliding window analysis was performed to determine if any other 600-bp region of the B. cellulosilyticus WH2 genome containing a Tn insertion went through a clonal expansion analogous to that documented for the cI-rha intergenic region during the first 31 d of the experiment in mice belonging to the live and heat-killed p-VLP treatment groups. The results revealed that on average any given 600-bp window with a Tn decreased its abundance over time, usually to less than 0.001% of the B. cellulosilyticus WH2 population. Only 5–10% of the windows other than the cI-rha intergenic region exhibited any enrichment over time, with less than 0.1% of the windows reaching levels >1% of the population (Fig. S9B). However, all of these other enrichments occurred 11 d or more following gavage when the bacterial host population was recovering from prophage induction (Fig. S9B). Importantly, strains with the Tn-containing cI-rha intergenic region selected for before and during prophage 1 induction subsequently maintained high relative abundance (~4%) in the B. cellulosilyticus population in both the live and heat-killed p-VLP treatment groups (Fig. S9B).

These results indicate that prophage 1 induction is restricted in time (i.e., nonrecurring over the course of the experiment) and insensitive to the attack of other members of the artificial community by exogenous human fecal phage. Together, the data demonstrate how disruption of an intergenic region, located just upstream of the predicted OR region that functions as a switch between lysogenic and lytic cycles in other lambda phages, and between a putative repressor and antiterminator, is capable of conferring a fitness advantage to its bacterial host strain before and independently of prophage induction.

Prospectus. Gnotobiotic mice containing defined consortia of sequenced human gut bacterial symbionts provide a tractable system that is more realistic than in vitro approaches for characterizing phage–bacterial host dynamics. These mice disclosed that: (i) a deliberately executed phage attack with a mixture of diverse human phages did not result in a simultaneous attack of all susceptible members of the artificial human gut microbiota, but rather was manifested by a succession of changes in the abundance of a subset of its bacterial taxa; (ii) phage resistance can occur through an ecological or epigenetic mechanism (i.e., without changes in bacterial CRISPR elements or bacterial genes encoding cell surface markers); (iii) one phage that was widely distributed among the five human donors but only reached low abundance in mice exhibited variations in its genome sequence in gnotobiotic animals over time, raising the possibility of red queen dynamics in this case; and (iv) phages contain important in vivo fitness determinants for their host strains; these determinants can reside in regulatory regions responsible for prophage induction (as illustrated by B. cellulosilyticus WH2).

The model gut microbial community that these animals harbored was remarkably resilient with several fold changes in the relative abundance of different taxa occurring for only brief periods of time. Nonetheless, the identified targets of the prophage attack (B. caccae and B. ovatus) did not fully return to preattack levels in the artificial community or to levels seen in mice receiving the heat-killed p-VLPs that were sampled at corresponding time points, suggesting modest long-lasting effects. The transient nature of the changes in community structure observed after p-VLP attack, and the fact that prophage induction was modest in most bacterial hosts, support the view that the phage–
bacterial host dynamic in this simplified defined gut ecosystem is predominately temperate rather than lytic (4).

Our findings extend previous work analyzing the dynamics of well-characterized T4 and T7 phages in gnotobiotic mice mono-colonized with *Escherichia coli*, where their in vitro behavior was a very limited predictor of their in vivo behavior (26). Using a more complex defined microbiota, we also observe a diverse range of viral dynamics: (i) exogenous viruses rising in abundance soon after their introduction (with a corresponding decrease in abundance of their putative bacterial hosts) followed by depletion of these viruses from the community without any obvious trace of genetic resistance or adaptation; (ii) exogenous viruses that survive at undetectable levels for almost a week before increasing their abundances, and in one case, displaying genetic variability over time; and (iii) basal levels of induction of 10 prophages with only one prophage achieving a level that produced detectable alteration in the abundance of its bacterial host. Duerkop et al. also showed that prophage induction differed in vitro and in vivo, and how under the appropriate conditions induction could provide a fitness benefit to the bacterial host (27), further highlighting the importance of a temperate lifestyle in the gut, and why prophages are so widely distributed in gut bacterial genomes without necessarily being induced at significant levels. Our findings are also consistent with the previous finding that T7 phages are capable of surviving at undetectable levels for 1 wk in germ-free animals before they rise in abundance after gavage of a bacterial host (26); this capacity to maintain infectivity has potential implications for preventive phage therapy.

We envision a future where gnotobiotic mouse models allow “personalized” assessment of phage–bacterial interactions. Complex mixtures of VLPs, isolated from previously frozen fecal samples obtained from human donors representing various ages, plexus mixtures of VLPs, isolated from previously frozen fecal samples obtained from human donors representing various ages, can be introduced into recipient mice harboring a defined collection of human gut community members. The community can be used as a “filter” to identify and assemble the genomes of phages present in the human donor viromes and link them to bacterial hosts. Miniaturization of methods for preparing VLPs from mouse fecal pellets provides a way for purifying these phages and at the same time verifying that they have lytic activity. The system is capable of distinguishing very closely related viropotypes present in multiple human gut microbiota based on their differential ability to establish themselves in recipient gnotobiotic mice. These attributes not only provide a discovery pipeline that complements metagenomic surveys of the human gut virome by identifying phages “buried” in large gut microbiome datasets, but facilitate identification of phages that can be used as experimental tools to deliberately manipulate model microbial communities or that can be considered as candidate therapeutic agents.

**Methods**

Protocols for the recruitment of human subjects and sampling of their fecal microbiota were approved by the Washington University Human Research Protection Office. All experiments involving mice were performed with protocols approved by the Washington University Animal Studies Committee. Procedures for gnotobiotic mouse husbandry, introduction of VLPs purified from human fecal samples into mice, sampling fecal microbiota from gnotobiotic mice, isolation of total DNA from mouse feces and intestinal contents, quantification of microbial cells in fecal samples by flow cytometry, preparation of DNA libraries for Illumina HiSeq or MiSeq sequencing, preparation of viral samples from fecal DNA samples, 454 pyrosequencing of VLP-derived DNA, COPRO-Seq analysis, assembly and annotation of viral genomes, cross-contig comparisons, INSeq analysis of fitness determinants present in the *B. cellulosilyticus WH2* prophage, PCR quantification of *B. caccae* abundance in the fecal microbiota of gnotobiotic mice, CRISPR analysis, PCR determination of Tn insertions in uninuced and induced *B. cellulosilyticus WH2* prophage 1, in vitro induction of *B. cellulosilyticus WH2* prophages, and in vitro assays for bacterial host tropism of phages represented in the pooled human fecal VLP preparation are described in SI Methods.

**ACKNOWLEDGMENTS.** We thank Dave O’Donnell and Maria Karlsson for their assistance with gnotobiotic mouse husbandry, Philip Ahern for help with immune characterization of tissues, Martin Meier for technical support with robotics, and members of the J.I.G. laboratory for their many helpful suggestions during the course of this study. This work was supported by grants from the National Institutes of Health (DK30292, DK078669, and DK07866951) and the Crohn’s and Colitis Foundation of America. A.R. is the recipient of an International Fullbright Science and Technology Program award.

**References**

Supporting Information

Reyes et al. 10.1073/pnas.1319470110

**SI Methods**

**Gnotobiotic Mouse Husbandry.** Germ-free mice belonging to the C57BL/6j inbred strain were maintained in plastic gnotobiotic isolators under a strict 12-h light/dark cycle and fed a standard, autoclaved, low-fat, plant polysaccharide-rich Chow diet (B&K Universal) ad libitum. Three groups of age-matched germ-free animals (n = 5 per group) were individually caged in three separate isolators. At ~8 wk of age, mice from two groups were colonized with a single gavage of 300 μL of supplemented tryptophan, yeast extract, glucose (TYG) medium (1) representing 15 sequenced human gut-derived bacterial taxa pooled from monocultures (~6 × 10^9 cfu per strain). One of the 15 strains, Bacteroides cellulosilyticus WH12, was represented as a library of 10^8 VLP/g frozen fecal sample; 10^10 p-VLPs, either live or heat-killed by incubation at 95 °C for 15 min, were pooled, human-derived VLPs. To ensure that the gastric pH would not affect the viability of the viral particles, mice were fasted for 12 h. Each mouse was then gavaged with 100 μL of a 500:210 mixture of 2x buffer A (200 mM NaCl, 200 mM Tris, 20 mM EDTA) and 20% SDS were added, followed by 563 μL of phenol:chloroform:isoamyl alcohol (25:24:1; Ambion). The tubes were capped, mixed using a Bio-Spec Minibeadbeater-96 (4 min), and then centrifuged at 3,200 × g for 4 min. A total of 480 μL of the aqueous phase from each sample was subsequently transferred to an Abgene P-DW-11-C 96-well plate with the Genesis series robot. Using the Biomek FX laboratory automation workstation (Beckman Coulter), 180 μL of the stored aqueous phase was transferred along with 720 μL of a 675:45 mixture of Qiagen buffer PM and 3 M sodium acetate (pH 5) to a QiaQuick 96 plate stacked on a Nunc 260251 plate. After mixing 10 times by pipetting, the plate was centrifuged at 3,200 × g for 4 min. The plate was washed twice with Qiagen PE buffer (900 μL per well) and centrifuged at 3,200 × g for 2 min. Following another centrifugation step to remove leftover ethanol, the multiwell plate was placed on a vacuum manifold to remove residual liquid from the membranes. To elute the DNA, 100 μL of buffer EB (Qiagen) was added to each well, and the plate was left standing for 2 min before centrifugation at 3,200 × g for 2 min. Purified DNA and leftover aqueous phase were stored at −20 °C. DNA concentration was measured using the Qubit Quant-IT dsDNA BR protocol as recommended by the manufacturer (Life Technologies).

**Quantifying Microbial Cell Counts in Fecal Samples by Flow Cytometry.** To quantify the number of microbial cells per fecal pellet and to compare the results with fecal DNA yields, we used the Bacteria Counting Kit (B7277; Life Technologies) and the manufacturer’s protocol, with some modifications. In brief, each frozen fecal pellet was weighed and transferred to a sterile 2-mL screw cap tube; 300 μL of sample buffer [10mM TrisHCl, 1mM, EDTA (1xTE), 0.9% wt/vol NaCl] was added, and the pellet was resuspended by vortexing for 5 min. The slurry was allowed to settle for 5 min at room temperature and a 10-μL aliquot of the clarified supernatant was transferred to a new tube. The remaining suspension was stored at −20 °C for extraction of total community DNA. Serial dilutions of the clarified sample were performed to a final dilution of 1:10,000 in 200 μL of sample buffer containing 1× SYTO BC dye and the kit’s standard counting control beads at a concentration of 250 beads/μL. The sample was then passed through a nylon mesh (60-μm pore diameter) to remove large particles that could potentially clog the cytometer [note that the control beads (6-μm diameter) are larger than bacterial cells and the number of beads present was not affected by passage through the nylon mesh]. The flow-thru was then split into three equal parts and each replicate was counted for 1 min in the Cytosics FC 500 MPL low cytometer (Beckman Coulter). Forward and side scatter data were collected along with fluorescence data in the 525-nm channel. Gates were drawn on the total bacterial cell population, the SYTO BC positive population, and the control...
beads. Quantifying the control beads allowed us to calculate the number of bacterial cells (positive fluorescent cells) per milligram (wet weight) of fecal material.

To extract DNA from these samples, we added zirconia and steel beads to the resuspended pellet described above, along with 300 μL of extraction buffer (118 mM NaCl, 435 mM Tris, 44 mM EDTA, 547 mM SDS), so that the final concentration of components was equivalent to 2× buffer A (see above) and 20% SDS. At this point, the protocol for extracting total community DNA was performed exactly as described above.

Preparation of DNA Libraries for Sequencing on Illumina Instruments (HiSeq 2000 or MiSeq). One hundred microliters of total community DNA in TE buffer (pH 7.0; 5 ng/μL), was fragmented by sonication in thin-walled 0.2-mL 8-strip PCR tubes using a BioruptorXL multisample sonicator (Diagenode) set on “high”; sonication occurred over the course of 20 min using successive cycles of 30 s “on” followed by 30 s “off.” Sonicated samples were subsequently cleaned up using the MinElute 96 PF 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 nucleic-free sterile water. For end repair and A-tailing, 20 μL of sonicated DNA was mixed with 5 μL of a mixture containing 2.5 μL of T4 DNA ligase buffer (NEB), 1 μL of 1 mM dNTPs (NEB), and 0.5 μL of each of the following enzymes: T4 polymerase (3 units/μL; NEB), T4 poly(nucleotide) kinase (10 units/μL; NEB), and Taq polymerase (5 units/μL; Life Technologies). The solution was mixed by vortexing and then incubated for 30 min at 25 °C followed by 20 min at 75 °C. Customized Illumina adapters containing maximally 4- to 6-bp barcodes (Dataset S1) were ligated to the A-tailed DNA in a 27-μL reaction by adding 1 μL of 25 μM adapter mix plus 1 μL of T4 DNA ligase (2,000,000 units/mL; NEB). Adapter mix was prepared by mixing 12.5 μL of a 100-μM stock of each adapter oligo (constituting the forward and reverse strands) and 25 μL of oligo buffer (1× TE, 0.1 M NaCl), incubating the mixture at 95 °C for 1 min, then slowly decreasing the temperature (0.1 °C per second) until reaching 4 °C. After 30 min incubation at 16 °C, 2.5 μL of 50 mM EDTA was added to stop the ligation reaction. Sets of 24 samples, all harboring different adapter sequences, were pooled, and the pool was purified using MinElute PCR purification columns according to the manufacturer’s instructions (final elution volume, 15 μL). A 10-μL aliquot of the purified pool was subjected to 2% agarose gel electrophoresis. DNA (~200 bp) was excised and purified using the QIAquick Gel Extraction kit and MinElute purification columns (Qiagen; final elution volume, 12 μL). Finally, 1 μL of the size-selected library was used as the template in an enrichment PCR [19 cycles of 98 °C for 10 s, 67 °C for 30 s, and 72 °C for 30 s, using Phusion HF Master mix (NEB) and Illumina’s standard amplification primers (Dataset S1) in a final volume of 25 μL]. The PCR product was purified using MinElute PCR purification columns and DNA was quantified using the Quant-it dsDNA High-Sensitivity (HS) Assay kit (Life Technologies). Library concentrations were then normalized, and an equimolar pool was subjected to multiplex sequencing with either an Illumina HiSeq 2000 or MiSeq instrument (2.5 pM/lane).

Preparation of VLP DNA from Mouse Fecal Samples. VLP purification and DNA extraction were performed as described previously (2), with some modifications. Because a single mouse fecal pellet had too little viral mass for efficient viral purification, pairs of fecal pellets obtained from either one or two mice in a given treatment group at the same time point (30–100 mg) were resuspended in 400 μL of SM buffer (filter-sterilized, 0.02-μm pore diameter; Whatman). After homogenization by vortexing for 5 min, samples were centrifuged twice at 2,500 × g for 10 min at 4 °C to remove large particles and bacterial cells. The resulting supernatant was filtered once through a 0.45-μm pore diameter Millipore filter (Millipore) and twice through 0.22-μm pore diameter Millex filters (Millipore). The volume of the filtrate was adjusted to 200 μL with SM buffer if needed. Each sample was treated with 20 μL of lysozyme (100 mg/mL) for 30 min at 37 °C, followed by incubation for 10 min with 0.2 volumes of chloroform.

The sample was then centrifuged at 2,500 × g for 5 min at room temperature. The aqueous phase was collected and incubated with 3 units of DNaseI (Sigma) and 20 μL of 10× DNaseI buffer (50 mM MgCl2, 10 mM CaCl2) for 1 h at 37 °C, after which time enzyme activity was inactivated by incubation at 65 °C for 15 min. To isolate DNA, VLPs were incubated with 10 μL of 10% SDS and 1 μL of proteinase K (Sigma, 20 mg/mL) for 20 min at 56 °C, after which 35 μL of 5 M NaCl and 28 μL of 10% (wt/vol) cet-3ytrimethylammonium bromide (CTAB)/0.7 M NaCl were added, followed by incubation at 65 °C for 10 min. The sample was then mixed with an equal volume of phenolchloroform/isoamyl alcohol (25:24:1), vortexed, and centrifuged at 8,000 × g for 5 min at room temperature. The resulting aqueous phase was mixed with an equal volume of chloroform and spun at 8,000 × g for 5 min at room temperature. The resulting aqueous phase was passed through a Qiagen MinElute purification column (elution volume, 30 μL).

Multiple displacement amplification (MDA) was performed with Illustra GenomiPhi V2 (GE Healthcare Life Sciences), according to the manufacturer’s instructions (n = 4 independent reactions/sample to prevent single amplification bias; reactions were subsequently pooled and the DNA product was purified using a Qiagen DNeasy purification kit (elution volume, 75 μL)).

The 454 Shotgun Pyrosequencing of VLP-Derived DNA. Libraries were generated for 454 shotgun pyrosequencing using a protocol similar to that used for the Illumina libraries. The protocol differed slightly in that sonication was performed for 8 min instead of 20 min. Each library was purified using Agencourt AMPure XP beads (Beckman Coulter) and quantified using the recommended 454 rapid library barcode-labeled adapters and a plate reader (Synergy2; Biotek). A total of 24 independent adapters were synthesized containing 24 different multiplex identifiers (MID) barcodes. After quantification, normalized pools of 24 samples were sequenced using 454 FLX Titanium chemistry. Initial quality filtering of the raw data consisted of parsing the sequenced reads by their MID followed by removal of short reads (under 100 bases). Reads shorter than 35 nt after removing the adapter sequence, or with a Hamming distance between any two barcodes of at least 2, requiring a perfect match to a barcode sequence allowed us to efficiently and accurately assign the reads to specific samples. After dividing the sequences by sample, we were able to run the downstream analysis in parallel. This downstream processing consisted of first screening for adapter sequences that indicated the presence of either short inserts or adapter dimer [screening was performed using Cross_match (version 1.090518; www.phrap.org) with slight modifications to the default parameters (--gap1_only -minmatch 6 -minscore 10 -gap_init -3 --screen)]. Next, reads were either trimmed back when an adapter sequence was seen or removed entirely in the event that (i) the sequence was shorter than 35 nt after removing the adapter sequence, or (ii) more than three N’s were found anywhere in the read. The “clean” reads were used to query the 15 sequenced bacterial community.
genomes represented in the artificial community, as well as their predicted prophages, using FR-HIT v0.7 (4) with default parameters except that the minimum percent identity was set to 90% (-c 90). The mapping file was parsed as described previously (5). In particular, reads mapping uniquely to a single genome were used to determine relative abundances. Subsequently, exact ties or reads mapping equally to more than one genome were weighted, based on the relative abundance of the genomes involved. Raw counts were then normalized to reads/ kb/million mapped reads. Reads that did not map to any bacterial genome were then mapped against other known plasmids from these bacterial species, PhiX174 (used as the internal control on the Illumina sequencer) and to the reference mouse genome. All remaining reads were used for de novo hybrid assembly to identify exogenous viruses (see below).

Normalized counts for the different bacterial species and virotypes present in each sample were used to calculate relative abundances or genome equivalents (GE; the latter by normalizing the percentage of mapped reads to the length of the genome and the DNA yield/sample). Matrices of either relative or absolute abundances were generated for principal coordinates analysis (PCoA) and generation of 3D plots using scripts in QIIME (version 1.3.0-dev; ref 6). Statistical analysis and figure generation were performed in Prism (v6).

Assembly and Annotation of Exogenous Viral Genomes. Reads generated from either VLP-derived DNA (454 FLX pyrosequencing using titanium chemistry) or from total community DNA (Illumina HiSeq. 2000 instrument) that did not map to any of the reference genomes were pooled together and submitted for a hybrid assembly using MIRA V3.4.0 (7) with minimum overlap of 70 nt (454) or 20 nt (Illumina) and a minimum relative score of 90%. Note that raw data from 454 pyrosequencing of VLPs were parsed with essentially the same pipeline used for Illumina reads except for the adapter-mapping step that is incorporated by default in 454 sfftools. Given that 454 FLX reads are longer than Illumina reads, FR-HIT parameters used were 90 for percent similarity (-c) and 40 for minimal length of the alignment (-m). A de novo assembly was performed for the first round. After assembly, contigs were analyzed individually using Tablet v1.12.09.03 (8) to identify potential chimeras in the assembly, and checked using blastn for overlapping ends either within contigs (indicating a complete, circular plasmid genome) or between contigs (indicating a potential link between two contigs). Raw reads were then mapped to the contigs using mapping-based assembly in MIRA, which allows extension of previously assembled contigs via the incorporation of new reads at the edges of these contigs. The process was repeated five times, yielding in the final iteration a total of five large circular contigs covering most of the unmapped reads from both the 454 and Illumina datasets. These exogenous viral genomes were used in conjunction with the reference bacterial genomes as an updated reference dataset for analysis of the relative abundances of the viruses and host bacteria from the VLP and total fecal community DNA shotgun sequencing datasets.

The five viral genomes were annotated first with Glimmer v3.02 (9) trained on all ORFs predicted from viral ref_seq (National Center for Biotechnology Information, NCBI). Predicted ORFs were then blasted (blastp, threshold <10−5, no low redundancy filter) against COG (STRING v9.0), KEGG (v58), ACLAME (v0.4), CDD (online Batch CD-search), NCBI nr (retrieved 13/08/2012), and Phantome (retrieved 01/09/2012).

Cross-Contig Comparison. To determine which human donors were the source of the viral genomes assembled from the mouse fecal VLP and COPRO-Seq datasets, we first retrieved FASTA files of pyrosequencing datasets that we had previously generated from each of the five purified human donor VLP DNA preparations used to create the input pool for our mouse experiments (GenBank ID SRX028824; ref 2). The sequences were pooled together with 454 FLX titanium sequences obtained from the p-VLP input (71,546 reads). The assembly pipeline consisted of the following steps: (i) CD-HIT v4.6 (10) was used to cluster reads at 90% global identity from each cluster and the top five sequences were taken as representatives; (ii) these reads were used for de novo assembly using Newbler v2.6 (454 Life Sciences) with default parameters; (iii) FR-HIT was used to map all raw reads to the assembled contigs at 90% identity; (iv) reads that did not map were pooled and reassembled using Newbler; (v) contigs >500 bp after both rounds of assembly were pooled together along with sequences from the five previously undescribed viral genomes and all raw reads were mapped using MIRA in mapping assembly mode, which allows for extension at the edges of contigs; (vi) the extended contigs were assembled using Phrap; (vii) the Phrap output files “contigs,” “singletons,” and “problems” were concatenated, renamed, and used to map the reads using FR-HIT; (viii) chimeras were identified as sudden drops in coverage given by the FR-HIT mapping (contigs were split at these chimeric junctions); (ix) contigs were sorted by size and the program Megablast was then used to compare “all-against-all” and to identify contigs that were fully contained within another contig and contigs with overlapping ends; and (x) the final set of contigs over 2 kb was used as a reference set for mapping (FR-HIT) all raw reads from each of the fecal VLP-derived viruses from each of the individual human fecal samples used to generate the input pooled (p)-VLPs for the staged viral attack, the p-VLPs used to gavage the mice, and mouse fecal VLP DNA. A matrix of reads per kilo base pair of contig sequence per million reads of sample was generated. The matrix was then log transformed, and a heat map was built using R (11).

INSeq Analysis of Fitness Determinants Present in the B. cellulosilyticus WH2 Prophage. Whole genome transposon mutagenesis of B. cellulosilyticus WH2 was performed using protocols described in earlier publications (1, 12). Total fecal community DNA was isolated from fecal pellets obtained from mice belonging to the live and heat-killed VLP treatment groups at time points between 3 and 13 d after they had received a single gavage of this library of staged viral attack, the p-VLPs used to gavage the mice, and mouse fecal VLP DNA. A matrix of reads per kilo base pair of contig sequence per million reads of sample was generated. The matrix was then log transformed, and a heat map was built using R (11).
was performed in a final volume of 50 μL using 32 μL of the cleaned up sample DNA, 10 μL 10x Pfx polymerase buffer, 2 μL 10 mM dNTPs, 0.5 μL 50 mM MgSO₄, 2 μL of 5 μM amplification primers (forward primer: 5’ CAACTGAAAGGCGCATACG 3’ and reverse primer: 5’ AATGATACGCGACTACACGAC TCTTTCCCTACTACGA 3’), and 1.5 μL Pfx polymerase (2.5 units/μL; Life Technologies) and 22 cycles of denaturation at 94 °C for 15 s, annealing at 65 °C for 1 min and extension at 68 °C for 30 s. The 134-bp PCR product from each reaction was purified ([4% (wt/vol) metaphere gel; MinElute gel extraction kit; Qiagen]) in a final volume of 20 μL, and was quantified (Qubit dsDNA HS assay kit; Life Technologies). Reaction products were then combined in equimolar amounts into a pool that was subsequently adjusted to 10 nM and sequenced (Illumina HiSeq 2000 instrument).

Illumina 50-nt short reads were separated by sample using 4- or 7-nt sample-specific barcodes. The remaining read contains either the 5′ or 3′ end of the transposon (Tn) sequence along with 16–17 nt of flanking genomic DNA. After trimming the Tn sequences, reads were mapped to the B. cellulosilyticus WH2 genome, allowing up to a 1-nt mismatch. The read counts derived from mapping the 5′ or 3′ termini of the transposon were summed, and a normalized count of reads per million for each insertion position was generated. For the sliding window analysis, normalized read counts were added up for every 600-bp window throughout the reference genome, while sliding the window in increments of 100 bp. The final distribution of counts per window for any given sample obtained at any given time point was analyzed using the quantile function on R (11) to determine the different abundance levels. Final plots of the quantiles over time were generated using Prism (v6).

PCR Quantification of Bacteroides caccae Abundance in the Fecal Microbiota of Gnotobiotic Mice. B. caccae abundance was evaluated using total fecal community DNA isolated from samples obtained 22 or 23 d after gavage of the 15-member artificial community (i.e., 2 and 3 d following gavage with p-VLPs). Purified B. caccae genomic DNA was used to generate a standard curve. The standard curve and amplification plots for the samples, generated employing the PCR primer set dpg_11 (Dataset S1), were used to calculate the number of genome copies per milligram of fecal pellet obtained at each selected time point from members of both the live and heat-killed p-VLP treatment groups.

Clustered Regularly Interspaced Short Palindromic Repeats Analysis. The genome sequences of the bacterial species introduced into gnotobiotic mice were searched using the program CRISPRFinder (Clustered Regularly Interspaced Short Palindromic Repeats Finder) (13). All Illumina HiSeq reads generated for COPROSeq analyses of fecal DNA samples collected throughout the course of the experiment from all animals that received p-VLPs were screened for CRISPR repeats using tre-agrep (v0.8) for fuzzy string matching. Any read covering the repeat with three or fewer mismatches was retrieved for each CRISPR locus we had identified (Dataset S1). Cross_match was subsequently used to screen each locus for all individual spacer sequences. The number of reads mapping per spacer was then scored. The number of reads mapping per spacer was then scored. Accumulation of new spacers should be identifiable as sequences without similarity to known spacers but present at abundances similar to known spacers. Therefore, reads mapping to the repeat but not to the spacer were separated and clustered at 90% identity using CD-Hit-est (after masking for known repeat and spacers). The size of the cluster should be indicative of the relative abundance of the novel spacer. In every case, >85% of the sequences were present as singletons at an abundance that was below the average for known spacers, and within what would be expected for sequencing artifacts (SI Results and Dataset S1).

PCR Determination of the Tn Insertions in Uninduced and Induced B. cellulosilyticus WH2 Prophages. 1. Total DNAs from mouse 1 (M1)–M4 plus M7 (negative control) isolated 2–13 d after gavage of the 15-member bacterial consortium were used to confirm the presence and enrichment of Tn insertions at one of the 10-distinct locations identified within the ~600-bp cl-rha genomic region (Fig. S9A, black arrow). Primers were designed to amplify: (i) the prophage sequence at the insertion site (primer “Intergenic F” in Dataset S1); (ii) the junction between the Tn and the prophage at the insertion site (primer “Intergenic + Tn F” in Dataset S1); (iii) the host bacterial genome 165 bp upstream of the att site, outside of the prophage (primer “Genome R” in Dataset S1); (iv) putative phage integrase, 117 bp downstream of the att site, inside the prophage region (primer “Integrate R” in Dataset S1). PCRs were performed so that amplification with any combination of primers would yield products of approximately the same size (~2,300 bp). The combination of primers Intergenic F and Genome R functioned as a positive control, amplifying uninduced prophages whether or not they contain a Tn insertion. Primers Intergenic F and Integrate R should only yield products when there was recombination at att sites. A product obtained with primers Intergenic + Tn F and Genome R would indicate that a Tn insert was present in the prophage, whereas products obtained with Intergenic + Tn F and Integrate R would indicate that the Tn was present in induced phages. Amplifications were performed in 25-μL reaction mixtures that contained 1× Platinum Taq HiFi PCR buffer (Life Technologies), 0.3 μM PCR primers, 2 mM MgSO₄, 0.2 mM dNTPs, 0.1 mg/mL BSA, 0.01% Tween 20, 0.5 M betaine (Sigma-Aldrich), 0.02 units/μL Platinum Taq (Life Technologies), and total fecal community DNA (10 ng). Amplification was performed using 30 cycles of denaturation at 94 °C for 30 s, annealing at 54 °C for 1 min, and elongation at 68 °C for 2 min.

In Vitro Induction Assay of B. cellulosilyticus WH2 Prophages. B. cellulosilyticus WH2 was inoculated in 5 mL of TYG medium. The culture was allowed to grow overnight at 37 °C under anaerobic conditions. Aliquots were taken and diluted 1:100 into seven tubes containing 5 mL of TYG medium. These cultures were grown until they reached early log phase (OD₆₀₀ = 0.1) before adding a commercial brand of soy sauce [Kikkoman; 1:10 or 1:50 dilution (vol/vol)], or mitomycin C (Sigma; final concentration 0.0025% or 0.005%). Control cultures contained no additions. All cultures were incubated at 37 °C until the no-treatment control reached late log phase (OD₆₀₀ = 1). At this point, bacterial cells were pelleted by centrifugation (3,000 × g, 10 min at 4 °C), the resulting supernatant was filtered through a 0.22-μm pore diameter filter and virus particles and nucleic acids were precipitated by adjusting the filtrate to 0.5 M NaCl, 10% PEG-8000, and incubating at 4 °C for 16 h. The solution was centrifuged at 18,000 × g for 30 min at 4 °C, and the resulting pellet was suspended in SM buffer (2). DNA was extracted from viral particles, MDA was performed, and the resulting DNA subjected to shotgun sequencing as described above (MiSeq, single-end 76-nt-long reads).

In Vitro Assays for Bacterial Host Tropism of Phages Represented in the Pooled Human Fecal VLP Preparation. The following procedure was performed for each of the 14 members of the artificial community (of the two Bacteroides thetaiotaomicron strains present, only VPI-5482 was characterized). Initial cultures were started from frozen stocks in supplemented TYG medium (TYG₄; ref 1) and incubated overnight at 37 °C in prereduced Balch tubes under anaerobic conditions (20% CO₂, 75% N₂, and 5% H₂). At midlog phase, each of the 14 monocultures was split; one-half remained as a noninfected control used to follow the growth of the other half of the culture, which was inoculated with 50 μL of the same p-VLP preparation that had been introduced.
into gnotobiotic mice. After a 5-h incubation, bacterial cells from each culture that received the human p-VLP preparation were pelleted (3,000 × g for 10 min at 4 °C). Fifty microliters of each supernatant was used to inoculate a fresh midlog phase culture of the corresponding bacterial species. These second round viral enrichment cultures were processed following the same protocol used for the first round. Precipitation of viral particles and isolation of VLP DNA were performed exactly as described above in Preparation of VLP DNA from Mouse Fecal Samples. Purified DNA was subject to PCR amplification with five different sets of primers designed to specifically anneal to each of the five previously undescribed viral particles identified in gnotobiotic mice (Dataset S1). Moreover, each primer set was designed to generate a different length of amplicon. PCR was performed using ThermoPrime 1.1 ReddyMix (Thermo Scientific; 1× final concentration) with 10 pmol of primer mix, 2 μL of purified DNA, and the following cycling conditions: initial denaturation for 5 min at 95 °C followed by 30 cycles of 95 °C for 30 s, 50 °C for 30 s, and 72 °C for 1 min. A second independent in vitro liquid enrichment experiment for viral particles was performed as described above with the exception that the VLP inoculum was reduced fivefold and the incubation was extended from 5 h to 16 h for each round of enrichment. Following MDA amplification, each of the VLP-derived DNA preparations was sequenced (Illumina MiSeq, single end, 78-nt-long reads).

SI Results

A Search for Fixed Mutations in the B. caccae Genome That Could Potentially Confer Viral Resistance After the Staged Attack with Pooled Human VLPs. Reads derived from total community shotgun sequencing that mapped to the B. caccae genome were selected from fecal samples collected 11–21 d after introduction of the 15-member artificial human gut bacterial community and 11–25 d following p-VLP gavage. Reads were initially grouped into four datasets: (i) live p-VLPs, pre-VLP gavage; (ii) live p-VLPs post-VLP gavage; (iii) heat-killed p-VLPs; pre-VLP gavage, and (iv) heat-killed p-VLPs post-VLP gavage. A mutation conferring resistance should be fixed in the population represented in dataset ii. All of the reads from i to iv were pooled and used for assembly in mapping mode (MIRA), while keeping track of the dataset from which the reads originated. Thirty-fold coverage of the genome was obtained on average across each of the four datasets. We identified no single nucleotide substitution specific to any given dataset and fixed in over 10% of the B. caccae population (as defined by the sequence coverage at a given nucleotide position).

A total of nine loci were identified where there was low or no coverage in one of the four datasets, suggesting possible insertions or deletions (InDels). These loci were selected after filtering for repetitive regions that can mask true SNPs due to variation between the repeats and for regions with an overall coverage (added for all four datasets) of less than 25-fold. We designed a pair of PCR primers flanking each of the nine loci to amplify a product between 270 and 370 bp (Dataset S1). DNA from each of the five mice in each treatment group were selected as templates for the PCR: (i) fecal DNAs from all mice in the live p-VLP treatment group that had been sampled on day 17 following gavage with the 15-member artificial community (live p-VLP, pre-VLP gavage group); (ii) fecal DNAs from all mice in this treatment group prepared from samples obtained 21 d after viral gavage (live p-VLP, post-VLP gavage group); (iii) fecal DNAs from all mice from the heat-killed p-VLP treatment group sampled 17 d postbacterial gavage (heat-killed p-VLP, pre-VLP gavage group); and (iv) fecal DNAs from all mice in this heat-killed p-VLP treatment group sampled 21 d after viral gavage (heat-killed p-VLPs, post-VLP gavage group).

Amplifications were performed using PCR primers described in Dataset S1. After amplification, a standard dissociation curve was generated to assess the specificity of the amplification, followed by a 10-min incubation at 68 °C to ensure that the amplified products had an additional adenine at the 3′ terminus (A-overhang). PCR products were subsequently purified, quantified, and pooled by treatment group and primer set (100 ng/sample). A total of 500 ng of amplified DNA for each treatment group and primer set was added to a ligation reaction mixture that contained Illumina adapters (Dataset S1). The resulting adapter-ligated DNAs were purified using AMPure beads as described in SI Methods for the INSeq protocol, quantified, and a 4-ng aliquot was used per treatment per primer set for enrichment PCR, as described above for preparation of Illumina libraries. Barcoded amplicons were subjected to 2% (wt/vol) agarose gel electrophoresis to confirm their molecular weight and purified. After quantification, the different products were pooled in equimolar amounts and sequenced using an Illumina MiSeq instrument (paired end, 250-nt reads). Sequencing reads were split by barcode; for each sample, reads were trimmed by quality and the paired-end products were assembled using FLASH (14) (note that assembly of the paired-end reads is expected under these conditions because the largest PCR product is 370 bp and reads are 242 bp after removing the 8-bp barcode at each end of the amplicon). Even after quality trimming of the reads, overlaps >30 bp were observed, allowing assembly of >80% of the data. Once the amplicons were assembled, they were pooled with the reference sequence and CD-Hit-est (10) was used to generate 100% identical clusters of sequences from per-locus per cluster was generated; the percentages of reads that were different from the reference cluster never exceeded 5% of the total reads, indicating that no single mutation was fixed in B. caccae at the time points and conditions surveyed.

CRISPR Analysis. CRISPRs play a role in conveying resistance to foreign DNA through a mechanism that involves accumulation of spacers in the bacterial genome that map to short segments in the invading DNA (15). Only six of the 15 bacterial species comprising our artificial community contained CRISPR regions; five were members of the Firmicutes; the other, Parabacteroides distasonis, is a member of the Bacteroidetes (Dataset S1). Our analysis of the sequencing datasets indicated that none of the CRISPR loci appeared to accumulate new spacer sequences (Dataset S1). It is important to note that the number of reads that contain a given CRISPR repeat will be proportional to the length of the locus (number of spacers) and to the relative abundance of the bacterial species in the artificial community. We could not make any conclusions about CRISPR loci in Dorea longicatenata due to its low abundance throughout the experiment. For all other species, except Clostridium spiroforme, the average number of reads mapping to each individual CRISPR spacer was greater than 100 (Dataset S1). A newly incorporated spacer should be covered with a similar or significant fraction of the read coverage of other spacers. However, for all of the bacterial species analyzed, reads with similarity to a CRISPR repeat but not to any given spacer, cluster as singletons (i.e., >85% reads are singletons), suggesting that they correspond to sequencing artifacts and not to accumulation of novel spacers. Nonetheless, this finding has to be interpreted cautiously because none of these bacterial species reached a relative abundance greater than 10% at any time point after p-VLP gavage, thereby limiting our coverage at particular time points.

PCR Determination of the TN insertions in Uninduced and Induced B. cellulosilyticus WH2 Prophage 1. A PCR assay was designed to discriminate an integronic Tn insertion within the uninduced prophage from one within the genome of an induced lytic phage (for details of the assay see Fig. S9/4 and SI Methods). Results from the assay indicated that whereas the lytic phage contained the Tn insertion, it was also present in the uninduced prophage, indicating that Tn insertions do not prevent prophage induction. The PCR assay did not allow us to precisely quantify the proportion
of Tn inserts in lytic viruses versus the prophage. Therefore, we cannot formally rule out the possibility that persistence of mutants with Tn insertions in this region was due to reduced spontaneous induction and reduced host bacterial cell lysis.

**In Vitro Induction of *B. cellulosilyticus* WH2 Prophages.** To further characterize the biology of the lambdoid prophage and understand its genetic regulation, we attempted to identify conditions for its induction in vitro. Three different treatments were evaluated, each at two different concentrations: mitomycin C, hydrogen peroxide (to generate an oxidative stress; ref 16), and soy sauce (has been reported to induce prophage in bacterial hosts present in the mouth; ref. 17). When VLP DNA was purified from the culture supernatants and sequenced, we found that 99.48 ± 0.28% (mean ± SD) of the reads from each sample mapped to the *B. cellulosilyticus* WH2 chromosome (includes prophage 1 and prophage 2) and its two plasmids. The ratio of the abundance of the two *B. cellulosilyticus* WH2 plasmids was comparable for all conditions tested. The fold-induction of *B. cellulosilyticus* WH2 prophage 2 (the ratio of relative abundance of the phage genome to bacterial chromosome) after treatment with 0.0025% hydrogen peroxide was 17-fold higher than observed in the no-treatment control. This was the only condition that produced a higher level of prophage induction than the control (Fig. S8B). However, prophage 2 is not induced in vivo. Prophage 1, which is induced in mice, was never present at a higher relative abundance than the corresponding bacterial chromosome. We concluded that conditions that promote induction in vivo differ significantly from the ones we tested in vitro. Previous work with other gut bacterial species has shown that in vitro and in vivo conditions for induction of their prophage can be quite distinct (2, 18).

**In Vitro Enrichment for Viral Particles.** To further establish the bacterial hosts of the phage identified during the staged in vivo attack of the 15-member artificial community, we turned to an in vitro liquid culture system. Performing a large screen based on plaque assays of multiple potential host bacterial strains in anaerobic chambers can be problematic due to space constraints and other technical challenges. Therefore, we turned to liquid monocultures. Our protocol involved exposure of midlog phase cultures to the same p-VLP preparation that we used for the staged phage attack in mice. After the first round of incubation, an aliquot of the supernatant was taken and used to infect a fresh monoculture. At the end of this second round, bacterial cells were pelleted, the supernatant was passed through 0.22-μm diameter filters, and viral particles together with DNA were precipitated from the filtrate. DNA was further purified and assayed by PCR using primers specific for each of the five previously undescribed phage genomes, and by shotgun sequencing (n = 14 monoculture experiments, five virus-specific PCR assays/culture, and two independent experiments). Enrichment was defined as an increase in the relative abundance of a virus in a given culture compared with any other culture. The results revealed no enrichment for any of the phages in any of the monocultures. In the case of *B. caccae*, *Parabacteroides distasonis*, *B. thetaiotaomicron*, and *C. spiriforme*, the yield of DNA from the monocultures was too low for sequencing (even after MDA). Therefore, we assumed that none of these organisms served as hosts for the previously undescribed phage. Sequencing the remaining samples [212,408 ± 135,200 (mean ± SD) 78-nt-long reads per sample; MiSeq] revealed that the majority of reads generated from each culture mapped to bacterial genomic DNA isolated from the cellular debris that passed through the 0.22-μm filter (Dataset S1). Importantly, in cases where the cultured bacterial species harbored an inducible prophage identified in the mouse experiments, 93.11% ± 6.8% (mean ± SD) of the reads mapped to that prophage sequence, indicating some, albeit very low, levels of basal induction occurred in vitro. We concluded that the susceptibilities and/or requirements for infection of these bacterial species with any of the five phages differ substantially within the intestine compared with our in vitro culture conditions.

**Identification of a Putative *B. cellulosilyticus* WH2 Prophage 1 Right Operator Region and Its Relationship to Enriched Tn Insertions.** As noted in the main text, the genomic features of *B. cellulosilyticus* WH2 prophage 1 and the identification of cos sites used for linearization of its genome upon induction led us to classify it as lambdoid type. Lambdoid replication machinery generates concatemers of the genome that are linearized at cos sites before being packed inside the viral capsid. The absence of reads spanning the potential cos sites (Fig. 2A) emphasizes the quality of the VLP purification. Principal control of the switch between lysogenic and lytic cycles in phage lambda is provided by the repressor, cl, which binds to the right operator region (OR3) located immediately upstream of the cl gene. The repressor binds with different affinities to three different binding sites (OR1–OR3) within the OR3; these sites finely regulate expression of early lytic genes and cl itself. In *silico* comparison of the predicted ORFs of prophage 1 and other lambdoid phages yielded a candidate repressor (BACWH2_5232) with 54% amino acid similarity to lambda cl. BACWH2_5232 contains a helix-turn-helix DNA binding domain, a peptidase S24 LexA-like protein domain, and significant similarity to a pfam01381 domain that is conserved in cl and cro. The gene immediately upstream of BACWH2_5232 (BACWH2_5233) is present in the opposite orientation: its predicted protein product has no significant similarity to cro, but rather to rha (43% identity over 50% of the protein), pfam09669 (phage regulatory protein rha) and pfam03374 (phage antirepressor protein KilAC domain).

The intergenic region between these two genes would be expected to contain the OR3 sequence. Given that there is essentially no nucleotide sequence similarity between this region and the lambda genome, we performed a secondary structure prediction of the regions comprising the intergenic region and two flanking genes (Fig. S9 C and D). The analysis showed that lambda OR3 has a 63-nt almost perfect stem loop located just 11 nt upstream from the *cl* start codon. We also identified a 66-nt stem loop starting 20 nt upstream from the putative start site of the BACWH2_5232 repressor homolog. All of the Tn insertions that were enriched were positioned upstream of the predicted OR3; no Tn inserts were observed in BACWH2_5232 or the predicted OR3 in the intergenic region, even in the input *B. cellulosilyticus* WH2 mutant library, suggesting that disruption of either the repressor or the OR3 is detrimental to the fitness of the bacterial host.

Measurements of fecal microbial biomass in mice gavaged with live or heat-killed human VLPs. (A) Abundance and genome annotation of human phages identified in the fecal microbial communities of gnotobiotic mice. (A–H) The assembled annotated phage genomes and their changes in abundance in the fecal microbiota of mice belonging to the live and heat-killed VLP treatment groups are shown (mean values ± SEM; n = 5 animals per group). The upward pointing arrow at day 20 of the experiment indicates the time when the live or heat-killed p-VLP preparation was administered. The circular representation of each phage genome illustrates the location of genes on the positive (green) and negative (red) strands. Genes significantly similar to known viral genes are noted (blue; blastp E value <10−5; Dataset S1). The inner plot represents GC skew based on 200-bp windows (yellow, GC ratio is greater than the average for the viral genome; purple, ratio is lower than the average). Note that A and B correspond to Insets in Fig. 1 D and E but here they have been enlarged to facilitate visualization of genome features.
Fig. S5. Human donor assignment and viral genome variation in phages ϕHSC01, ϕHSC02, ϕHSC03, ϕHSC04, and ϕHSC05. (A) Heat map of the cross-assembly between different input human p-VLP samples. To determine whether the viruses identified in gnotobiotic mice were distributed among all of five human VLP donors or whether they were unique to particular individuals or subset of individuals, a cross-assembly was generated that included reads from each fecal VLP-derived virome from each individual human fecal sample used to generate the input p-VLP preparation for the staged viral attack, reads from the p-VLPs introduced into mice by gavage, and reads generated from VLPs purified from mouse fecal samples. Shown are the normalized abundances of the different contigs greater than 2 kbp (n = 159), including those assembled from the human phages identified in gnotobiotic mice. Each column presents data for the five individual human donor fecal VLP-derived viromes plus the p-VLP sample introduced into gnotobiotic mice. (FX, family number; M, mother of twin pair; TX, cotwin 1 or 2 in a given monozygotic twin pair; X, one of 3 time points where feces were collected from the human donor over the course of a year). Abundance is shown as the log (10) transformation of RPKM (reads per kb of contig per million pyrosequencer reads; see SI Methods and main text for further details). Rows representing the five previously undescribed viruses isolated and their corresponding human donor are enlarged for visualization purposes. The name of each phage is positioned to indicate the human donor where its abundance and nucleotide sequence conservation is greatest. (B) Individual pyrosequencing reads obtained from each of the five human donor VLP-derived viromes used for the pooled “input” VLP preparation were mapped to the assembled viral genome for ϕHSC05. In the case of the human VLP donor F3T1.2, 87.5% of her virome had 98.45 ± 0.02% identity (mean ± SEM) to the assembled ϕHSC05 genome. In the case of the other individuals, ϕHSC05 detected in mice was present at significantly lower abundance (e.g., for F4T1.2, 0.01% of VLP-derived reads) and/or had lower percent sequence identity. The panel labeled “mice” shows the results of mapping reads from all mouse fecal samples described in C. (C) Percent identity of pyrosequencer reads, generated from VLP DNA purified from individual mouse fecal or cecal samples, to the assembled ϕHSC05 genome. Fecal samples were obtained on days 29 and 37 from mice in the live p-VLP treatment group [mouse (M)1–M4]. Cecal samples were collected at the time of sacrifice on day 45 of the experiment. The dashed boxes highlight loci in the ϕHSC05 genome that show sequence variation between different mice and/or over time within animals. For animal-to-animal variation, compare locus 1, which is 100% conserved in M1 and M4 but only 90–95% conserved in M2 and M3 (red arrows). Variation within mice over time can be observed in locus 8, which is 100% conserved in the viromes of animals M3 and M4, 9 d after viral gavage but subsequently undergoes sequence changes over the remaining 16 d of the experiment (red arrows). Each locus is numbered as follows: locus 1 encompasses genes ϕHSC05_01 to ϕHSC05_05 (includes a predicted phage secretion activator protein); locus 2, ϕHSC05_09 to ϕHSC05_11 (all hypothetical proteins); locus 3, ϕHSC05_24 and ϕHSC05_25 (best blast hits to a DNA helicase and a DNA polymerase-associated protein, respectively); locus 4, ϕHSC05_26 (putative DNA polymerase); locus 5, ϕHSC05_56 (hypothetical protein); locus 6, ϕHSC05_67 (hypothetical protein); locus 7, ϕHSC05_68 to ϕHSC05_70 (includes a putative tail tape measure protein); and locus 8, ϕHSC05_73 to ϕHSC05_75 (all hypothetical proteins). (D–G) Mapping data comparable to that shown in C for the other four phage genomes. The human donor harboring a given phage was identified based on the analysis shown in A. Note that ϕHSC02 is a single-stranded phage that is overamplified with MDA. The red arrow points to a locus that appears to show sequence variation. However, this variation occurs in a homopolymeric region and may be a pyrosequencing artifact.

Fig. S6. Relationship between the abundance of bacterial species and ϕHSC03, ϕHSC04, and ϕHSC05 along the length of the gut of gnotobiotic recipients of the live p-VLP preparation, at the time of sacrifice. (A) Viral abundance for each mouse in the indicated treatment group as defined by shotgun sequencing of DNA isolated from different gut locations. The five columns for each mouse (M1–M10) are organized from left to right as follows: proximal small intestine (PSI), distal small intestine (DSI), cecum (Ce), colon (Co), and fecal sample (Fe). Note that no virus was detected in mice that had received the heat-killed human VLP preparation, at the time of sacrifice. The data are derived from individual mouse fecal or cecal samples, to the assembled viral genome for the five phages that are induced. The number of reads mapping to bacterial genomes outside the prophage region was low in all mice. Therefore, all hits to any bacterial genome (outside of a prophage region) are depicted in black. Each bar represents data generated from VLPs prepared from two fecal pellets obtained from one or two mice within a given treatment group at a given time point (Dataset S1). Cecal samples were collected at the time of sacrifice and are numbered according to the mouse ID in each treatment group. Relative abundance is defined as the number of VLP reads mapping to any of the specific prophage genomes (or elsewhere on the bacterial genome) divided by the total number of VLP reads that did not map to any of the 5 exogenous phages.

Fig. S7. Shotgun 454 pyrosequencing data generated from VLPs isolated from mouse fecal samples validate predicted prophage regions in bacterial genomes. Mapping of VLP-derived shotgun sequencing reads to the genomes of the 15-member artificial community assembled in gnotobiotic mice allows identification of prophages that are induced. The number of reads mapping to bacterial genomes outside the prophage region was low in all mice. Therefore, all hits to any bacterial genome (outside of a prophage region) are depicted in black. Each bar represents data generated from VLPs prepared from two fecal pellets obtained from one or two mice within a given treatment group at a given time point (Dataset S1). Cecal samples were collected at the time of sacrifice and are numbered according to the mouse ID in each treatment group. Relative abundance is defined as the number of VLP reads mapping to any of the specific prophage genomes (or elsewhere on the bacterial genome) divided by the total number of VLP reads that did not map to any of the 5 exogenous phages.
Fig. 58. Genomic characteristics, conservation, and results of in vivo and in vitro induction assays for \textit{B. cellulosilyticus} WH2 prophages. (A) Subset of genes in \textit{B. cellulosilyticus} WH2 prophage 1 that are conserved in the genomes of four other human gut \textit{Bacteroides}. Only genes conserved in synteny with \textit{B. cellulosilyticus} WH2 prophage 1 are shown, along with the average similarity for their protein products. Each box represents a predicted ORF. Colored boxes represent the products of ORFs that are characteristic of this and other lambdoid phages. Starting from the cos sites, the prophage 1 genome contains a group of genes that are conserved and organized in a way common to lambdoid viruses of the Siphoviridae family (1); the cos site is followed by ORFs encoding small and large terminases, portal protein, protease, the major head protein, several small capsid structural proteins, the major tail protein, and tape measure protein, ending with an integrase next to the att site. The intergenic region positioned between the putative \textit{rha} homolog and \textit{cl} (\textit{lambda} repressor) is highlighted. Arrows drawn on top of genes indicate the relative positions and directionality of PCR primers designed to identify the presence of Tn inserts in the prophage and/or the induced lytic virus. The red arrow indicates a primer designed to anneal at the junction between the Tn and the intergenic region between \textit{cl-rha} (see SI Methods for details). Searching the genomes of 38 human gut \textit{Bacteroides} (Dataset S1) revealed four other bacterial species, including another strain of \textit{B. cellulosilyticus} (DSM 14838), that contained homologs of all or most of the viral structural proteins in a syntetic arrangement. Moreover, homologs of six genes in \textit{B. cellulosilyticus} prophage 1 are present in 35–37 of the 38 \textit{Bacteroides} genomes. Thus, our identification of prophage 1 in \textit{B. cellulosilyticus} revealed lambdoid viruses that had not been associated previously with \textit{Bacteroides} spp. (B) In vitro assays for \textit{B. cellulosilyticus} WH2 prophage induction. Three different treatments at two different concentrations of various known prophage-inducing agents were evaluated in monocultures of \textit{B. cellulosilyticus} WH2 grown anaerobically. DNA extracted from the culture supernatants was sequenced. Mapping reads to either of the two prophages or the rest of the bacterial chromosome showed no detectable or very low levels of spontaneous induction of prophage 1 under any of the in vitro conditions tested (prophage 1 is induced in vivo). Hydrogen peroxide at a concentration of 0.0025% produces a 48-fold induction of prophage 2 (17-fold greater than the no-treatment control). (C and D) Prophage induction is not observed in wild-type \textit{B. cellulosilyticus} WH2 in the context of the 15-member artificial community. (C) Relative abundance (mean ± SEM) of \textit{B. cellulosilyticus} WH2 and its two associated prophages in the fecal microbiota of mice colonized with a 15-member community that contains wild-type \textit{B. cellulosilyticus} WH2 rather than a library of Tn mutants. Prophage induction was not observed at any time during the 20-d experiment. See Fig. 2 for comparison with the community containing transposon-mutagenized \textit{B. cellulosilyticus} WH2.


Fig. 59. INSeq reveals clonal selection for Tn mutants with insertions in the region between genes encoding the Rha protein and \textit{cl} homolog in \textit{B. cellulosilyticus} WH2 prophage 1. (A) Heat map of log-transformed, normalized abundance of reads (reads per million, RPM) that mapped to a site of transposon insertion in the BACWH2_5232 gene (\textit{rha} homolog) or the intergenic region between the \textit{rha} homolog and BACWH2_5233 (\textit{cl} homolog). Each row represents a time point 3–9 d after gavage with the 15-member bacterial community (before subsequent gavage with either live or heat-killed p-VLPs). All mice were individually caged. Fecal samples collected from five mice in each treatment group were characterized (M1–M5 in the live p-VLP group and M6–M10 in the heat-killed p-VLP control group). Each column represents the total number of normalized reads, in a window of 20 bp, obtained with a given fecal DNA sample. Tick marks between nucleotides 7,078,320 and 7,080,520 of the \textit{B. cellulosilyticus} WH2 genome represent 100-bp increments. The bottom four rows labeled “input” represent the read distribution for four technical replicates of the INSeq analysis for the input \textit{B. cellulosilyticus} WH2 mutant library. Note that no Tn mutants were identified in BACWH2_5232 (\textit{cl} homolog). There is no change in the representation of Tn insertion mutations in BACWH2_5233 (\textit{rha} homolog) in the output library. In contrast, a marked increase in the representation of Tn insertions at 10 sites in the intergenic region is seen over the time period sampled. The number and abundance of Tn mutants represented in each mouse differ. The increase in representation of Tn mutants correlates with the time before and during induction of prophage 1. The 600-bp \textit{cl-rha} intergenic region is shown as a thin green line. The black arrow indicates the Tn insertion site used for PCR assay of whether the mutation was present in the prophage and/or induced virus (SI Methods and SI Results). (B) Insertions in a 600-bp intergenic region provide a fitness advantage to the bacterial host that is maintained over time. The number of Tn reads per million obtained by INSeq analysis of fecal DNA was calculated for a 600-bp intergenic region between the ORFs encoding the \textit{cl} regulator and putative \textit{Rha} protein. The relative abundance of bacterial cells harboring this Tn insertion as a fraction of the total \textit{B. cellulosilyticus} WH2 population is represented by the RPM value at any given time point. Mean values ± SEM for the mice in each treatment group are plotted for 13 different time points sampled during the first 31 d of the experiment. Live p-VLPs or heat-killed p-VLPs were introduced at day 20 (downward arrow). A sliding window of 600 bp was used, at intervals of 100 bp, to scan the whole \textit{B. cellulosilyticus} WH2 genome (except the 600-bp \textit{cl-rha} intergenic region) to quantify the abundance of mutants containing Tn insertions. The resulting distribution of reads is plotted for a given time point with lines at the 25th, 50th, 75th, 99th, and 100th quantiles (100th = maximum value of Tn reads observed within a 600-bp window). Shaded areas represent the area where 25–75% of the data (RPM/600-bp window) falls (i.e., the second and third quartiles). In general, the abundance of Tn mutations falls over time with less than 10% of the Tn-containing windows increasing their abundance. The relative abundance of bacterial strains harboring Tn-containing \textit{cl-rha} intergenic mutants was maintained at significantly higher levels than all other mutants. (C and D) Computational analysis of the intergenic region affected by Tn insertions in \textit{B. cellulosilyticus} WH2 prophage 1. (C) Secondary structure prediction using Mfold (1) for the region containing \textit{cl} (green line with arrow) and \textit{cro} (red line with arrow) plus the corresponding intergenic region in phage lambda (GenBank accession: NC_001416). Zoomed in at Right is the intergenic region containing the right operator (Ou) with its three binding sites (OR1–OR3). (D) The same secondary structure analysis was performed to identify the corresponding regulatory region upstream of the \textit{cl} homolog (green line with arrow) in \textit{B. cellulosilyticus} WH2 prophage 1. The 600-bp intergenic region significantly enriched for Tn insertions is highlighted with a thick green line (see corresponding region in A). Each Tn insertion site is shown with a red arrow. Note that only one stem loop is positioned outside of the Tn-enriched intergenic region. This stem loop (labeled predicted operator region) has a similar length and location as the known operator region in phage lambda (zoomed in at Right).


Fig. 59

Other Supporting Information Files

Dataset S1 (XLSX)
Live p-VLP group

Heat-killed p-VLP group

Germ-free group

Model community

Bacteroides cellulosilyticus WH2
Parabacteroides distasonis
Bacteroides thetaiotaomicron VPI-5482
Bacteroides thetaiotaomicron 7330
Bacteroides uniformis
Clostridium scindens
Clostridium spiroforme
Clostridium symbiosum
Ruminococcus obeum
Collinsella aerofaciens
Parabacteroides distasonis
Bacteroides thetaiotaomicron
7330
Clostridium symbiosum
Bacteroides ovatus
Bacteroides vulgatus
Dorea longicatena
Eubacterium rectale

Fecal pellet (COPRO-Seq)

Fecal pellet (viral DNA)
Panel A: Time course of DNA yield in Live VLP group (blue) and Heat-killed VLP group (red). The x-axis represents time in days (d), and the y-axis represents DNA yield in ng/mg of feces.

Panel B: Scatter plot showing the relationship between microbial cells per fecal pellet and DNA concentration (µg/fecal pellet). The x-axis represents microbial cells per fecal pellet, and the y-axis represents DNA concentration. A linear trend line is depicted.
Human donor ID

- F2M.2
- F2T1.2
- F3T1.2
- F4T1.2
- F4T2.2
- VLP input

Log_{10} (RPKM)

- HSC01
- HSC04
- HSC02
- HSC03
- HSC05

- A
Figure A shows a heatmap of differential gene expression in response to post-bacterial gavage. The vertical axis represents different samples (M1 to M10), while the horizontal axis represents genomic positions from 7,078,320 to 7,080,520. The color intensity indicates the log10 RPM (reads per million) values, with a scale from 0 to 6.

Figure B presents a log-log plot of reads per million over time for live VLP and heat-killed VLP treatments. The graph includes 99th, 95th, 50th, and 25th quantiles, with arrows indicating the time point of gavage.

Figure C illustrates the genome of phage lambda with the cl-cro region and cl-cl intergenic region. The cl and cro genes are indicated along with the cl-cro region. The cl-cl intergenic region is also marked.

Figure D depicts a prophage with the cl-rha region and a 600 bp intergenic region. The prophage is labeled with the OR1, OR2, and OR3 operators. The predicted operator region is also indicated.