

Washington University in St. Louis

## Washington University Open Scholarship

---

All Theses and Dissertations (ETDs)

---

1-1-2011

### The Gut Microbiome In Healthy And Severely Malnourished Humans

Tanya Yatsunenko

*Washington University in St. Louis*

Follow this and additional works at: <https://openscholarship.wustl.edu/etd>

---

#### Recommended Citation

Yatsunenko, Tanya, "The Gut Microbiome In Healthy And Severely Malnourished Humans" (2011). *All Theses and Dissertations (ETDs)*. 671.

<https://openscholarship.wustl.edu/etd/671>

This Dissertation is brought to you for free and open access by Washington University Open Scholarship. It has been accepted for inclusion in All Theses and Dissertations (ETDs) by an authorized administrator of Washington University Open Scholarship. For more information, please contact [digital@wumail.wustl.edu](mailto:digital@wumail.wustl.edu).

WASHINGTON UNIVERSITY

Division of Biology and Biomedical Sciences

Molecular Microbiology and Microbial Pathogenesis

Dissertation Examination Committee:

Jeffrey I. Gordon, Chair

Daniel Goldberg

Robert Heuckeroth

Scott Hultgren

Mark Manary

Clay Semenkovich

David Wang

The Gut Microbiome in Healthy and

Severely Malnourished Humans

by

Tanya Yatsunenko

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 2011

Saint Louis, Missouri

Copyright by  
Tanya Yatsunenko  
2011

## **Abstract of the Dissertation**

The gut microbiome in healthy and severely malnourished humans

by

Tanya Yatsunenko

Doctor of Philosophy in Biology and Biomedical Sciences

(Molecular Microbiology and Microbial Pathogenesis)

Washington University in St. Louis, 2011

Professor Jeffrey I. Gordon, Chairperson

Human large intestine is home to tens of trillions of microbes belonging to all three domains of life. The functions encoded by the genes in this community (microbiome) include processing and production of macro- and micronutrients. Much remains unknown about the factors that determine the assembly of the gut microbial community starting at birth, and if disruptions in the assembly of this ‘microbial metabolic organ’ early in life result in physiologic and metabolic deficits later in life. The central goal of my thesis was to characterize development of the gut microbiome early in life, with a focus on describing the relationship between the microbiome and nutritional status.

My thesis consists of three parts. Because the degree of temporal variation in the gut microbiome in children and adults in healthy and diseased state was not well described, I began by using metagenomic methods and a variety of computational and statistical tools to characterize the proportional representation of bacterial phylotypes and gene functions in the fecal communities of seven healthy adult USA monozygotic twin pairs sampled over a four-month period. I found that the fecal microbiota and microbiome are stable within each co-twin even in the face of ecologic invasion with a popular commercial fermented dairy product. I then compared the fecal microbiota and microbiome in 524 healthy infants,

children and adults living in three different countries with distinct cultural traditions (USA, Malawi and Amazon region of Venezuela). I found that interpersonal variation in babies is significantly greater than between adults, and that the microbiota evolves towards an adult configuration during the first three years of life in all three populations sampled. In addition, distinct patterns of functional maturation were observed which involved microbial genes encoding enzymes that participate in the biosynthesis of several vitamins. Finally, I characterized assembly of gut microbiomes in a cohort of Malawian twins concordant for healthy status and twins discordant for severe forms of malnutrition (kwashiorkor or marasmus). Twins were sampled during their first three years of life, including before, during and after treatment with a peanut butter-based ready-to-use therapeutic food (RUTF). In the case of the discordant twins, both co-twins were treated with RUTF. My comparative metagenomic analyses revealed notable differences in the responses to RUTF in kwashiorkor versus healthy co-twins.

## Acknowledgements

I am thankful to many colleagues and friends who provided their support during my years in graduate school. I am most of all deeply grateful to my mentor, Jeff Gordon. Jeff's wisdom, kindness, infinite energy and enthusiasm have always amazed and inspired me. During the last five years I learned not only how to be a better scientist, but also how to be a better human being. I am also thankful to him for assembling a wonderful group of people in his lab, without whom this thesis would not be possible.

Jill Manchester and Sabrina Wagoner, the "Moms" of the lab have always been extremely helpful and kind to everyone. I would like to give many thanks to Sabrina for spending hours sorting frozen tubes sent to us from Malawi in the cold "mass spec room". I would also like to acknowledge Jill's and Sabrina's contribution to the increase of my waistline due to my consumption of delicious fruit pizzas, cakes and other desserts they produced over the years at various celebrations. Stephanie Amen has been immensely patient and helpful with many administrative tasks. Laura Kyro helped with formatting figures, thesis and manuscripts.

Upon my arrival in the lab I was lucky enough to share a workspace with the loquacious Marios Giannakis. His wisdom was constantly overflowing into the air around me. When he left, a quiet void was left in our lab, which has never been completely filled. However, I must admit that my productivity increased at least two-fold after his departure. Special thanks to Brian Muegge for helping with establishing our 16S and shotgun sequencing pipelines. Thanks go to Alejandro Reyes who was unlucky to be one of the few Perl 'gurus' and therefore was subjected to being bugged constantly at the beginning of my programming experience. I'm very grateful to Federico Rey who shared his wisdom on life and science with me, and his careful and insightful edits to the manuscript. Many wonderful people in the lab contributed to my experience and evolving view on life, and I learned something from every person. There is not enough space to individually thank everyone;

but interactions with many of you have always been inspiring, and I will always remember and miss them. Thanks to Ansel Hsiao for tolerating me in the shared bench area, as well as for his insight into numerous random questions.

A number of non-Gordon lab friends made my life eventful and enjoyable during the time at the graduate school. I'm particularly thankful to Pavel Zhuravlev and Cristina Greavu for their wonderful friendship. Boris Iyutin was in large part responsible for my decision to apply for graduate school, as well as many other endeavors he inspired me to attempt.

Indi Trehan has been doing a heroic job in Malawi for the past 2 years, supervising the twin study described in the Chapter 4. Rob Knight and students and postdocs in his lab have been great collaborators. I would like to thank Dr. Bill Goldman, a former director of the Microbiology program, for making my transition from the Earth and Planetary Sciences to the DBBS smooth.

My work would not have been possible without the support of the Bill and Melinda Gates foundation. Their funding allowed 80+ 454 sequencing runs just for my few projects.

I was very lucky to meet my wonderful husband Andrew Grimm at the beginning of my graduate work at DBBS and at the end of his. He has been tremendously encouraging and supportive of all my small accomplishments throughout the years in the DBBS. His expertise and wisdom in pediatric sciences, and recently gastroenterology, have been useful in interpretation of some of my data. I hope this 'symbiosis' will continue for a long time.

And lastly I'd like to thank the creators of Dropbox and Zotero for allowing me to write my thesis from any location without ever losing it (sadly, I have no financial support from these companies to disclose that could have influence the previous statement).

## **Dedication**

To the women of Malawi

For their strong spirit, love, kindness, and heroic efforts in keeping their families alive



## Table of Contents

Abstract of the Dissertation .....	ii
Acknowledgements.....	iv
Dedication .....	vi
Table of Contents .....	vii
List of Figures .....	xi
List of Tables.....	xv

## Chapter 1

### Introduction

Methods for studying the gut microbiota.....	2
Adult gut microbiota.....	4
Gut microbiota in humans living in different geographic regions.....	6
Temporal variation of adult microbiota .....	7
Assembly of the gut microbiota in infants.....	8
Is the assembly of the gut microbiota follows the same pattern in infants living in different parts of the world?.....	11
Is the assembly and function of the gut microbiota compromised in severe childhood malnutrition?.....	12
Gut microbiota configuration in children with malnutrition.....	14
Overview of the dissertation .....	15
References.....	18

## Chapter 2

### **The impact of a consortium of fermented milk strains on the human gut microbiome: a study involving monozygotic twins and gnotobiotic mice**

Abstract.....	28
Introduction.....	29
Results .....	31
Human studies.....	31

Study design and assessment of intrapersonal and interpersonal variations in the fecal microbiota of monozygotic twin pairs over a four-month period .....	31
Effects of FMP consumption on the functional gene repertoire of the fecal microbiome .....	34
Studies in gnotobiotic mice.....	36
The repertoire of carbohydrate active enzymes (CAZymes) in members of the FMP consortium and model human gut microbial community .....	37
Introducing the FMP strain consortium produces minimal changes in the species representation of the 15-member model human gut microbiota ...	38
Microbial RNA-Seq analysis of the response of <i>B. animalis</i> subsp. <i>lactis</i> to the gut environment and members of the 15-member community to the FMP strain consortium .....	40
Identifying predictive features from the model community metatranscriptome data using a Random Forests classifier .....	44
Metabolomic analyses.....	44
Microbiome transcriptional responses to FMP strains that are shared by gnotobiotic mice and humans .....	46
Discussion .....	48
Materials and Methods.....	52
Acknowledgements.....	53
References.....	55
Figure Legends.....	60
Figures.....	65
Supplementary Material .....	72
Supplementary Materials and Methods .....	72
Microbial genome sequencing .....	72
Annotation and comparative genomic analysis .....	72
Culturing of <i>B. animalis</i> subsp. <i>lactis</i> .....	73
Human studies.....	74
Studies in gnotobiotic mice.....	77
Supplementary Results.....	87
Human studies.....	87
Studies in gnotobiotic mice.....	89
<i>In vitro</i> studies.....	90

Supplementary References.....	91
Supplementary Figure Legends .....	96
Supplementary Figures .....	100
Supplementary Table Legends .....	109
Supplementary Tables .....	112

### Chapter 3

#### **Human gut microbiome differentiation viewed across cultures, ages and families**

Abstract .....	131
Results and Discussion .....	132
Changes in the taxonomic/phylogenetic composition of fecal bacterial communities as a function of age and population .....	133
Shared functional changes in the microbiome as children mature .....	138
Population- and age-specific differences in the representation of microbiome functions .....	141
Effects of kinship on the microbiome across countries .....	145
Methods.....	147
References.....	151
Acknowledgements.....	156
Figure Legends.....	157
Figures.....	159
Supplemental Figure Legends.....	164
Supplemental Figures.....	169
Supplemental Tables .....	188

### Chapter 4

#### **Temporal variation in the gut microbiomes of healthy and twin pairs discordant for severe malnutrition.**

Introduction.....	198
Study Design.....	199
Influence of genetics, geography and gender on susceptibility to malnutrition .....	203

Sampling fecal microbiomes from twins who were concordant for healthy status and twins who were discordant for severe malnutrition .....	203
Comparison of fecal microbiomes across all children .....	205
Temporal variation of the gut microbiomes of twin pairs who remained healthy .....	206
Temporal variation in the fecal microbiomes of twin pairs discordant for kwashiorkor.....	208
Changes in KEGG ECs involved in various metabolic functions in the fecal microbiomes of co-twins discordant for kwashiorkor .....	209
Temporal variation in fecal microbiomes of twin pairs discordant for marasmus.....	212
Changes in KEGG ECs involved in various metabolic functions in the fecal microbiomes of co-twins discordant for marasmus .....	213
Methods.....	218
References .....	220
Figure Legends.....	223
Figures.....	228
Table Legends .....	248
Tables .....	250

## **Chapter 5**

### **Conclusions and Future Directions**

Enhancing the nutritional value of food via intra-familial probiotics.....	276
Filling the gaps in our understanding of the assembly of the gut microbiota.....	279

### **Appendices**

Appendix A .....	286
Appendix B .....	293
Appendix C .....	300
Appendix D.....	303
Appendix E .....	313

## List of Figures

### Chapter 2

#### **The Impact of a Consortium of Fermented Milk Strains on the Gut Microbiome of Gnotobiotic Mice and Monozygotic Twins**

Figure 1.	Experimental design for human and mouse studies.....	65
Figure 2.	Metagenomic studies of human fecal microbiomes sampled over time. ...	66
Figure 3.	Correspondence analysis of <i>B.animalis</i> subsp. <i>lactis</i> CAZyme gene expression. ....	67
Figure 4.	‘Top-down’ analysis of the effects of the FMP strain consortium on the model 15-member community’s metatranscriptome. ....	68
Figure 5.	Mouse and human communities share transcriptional responses to the FMP strain consortium involving ECs related to carbohydrate metabolism.....	69
Figure 6.	Select urinary metabolites whose levels are altered after the introduction of the FMP strain consortium into mice harboring a defined model human gut microbiota. ....	70
Figure 7.	Shared transcriptional responses to FMP strain exposure in mice and humans.....	71
Figure S1.	Levels of <i>B.animalis</i> subsp. <i>lactis</i> (CNCM I-2494) in human fecal samples collected before, during, and after consumption of an FMP.....	100
Figure S2.	KEGG pathway coverage ratios suggest that the model human gut microbiome encodes many of the functions present in more complex human fecal communities. ....	101
Figure S3.	CAZyme profiles of the 20 bacterial strains introduced into gnotobiotic mice.....	102
Figure S4.	Summary of analysis pipelines used in this study.....	103
Figure S5.	COPRO-Seq-based time series analysis of the abundance of members of the model human microbiota and of the FMP strain consortium in the feces of gnotobiotic mice. ....	104
Figure S6.	Top-down analysis of the model community’s transcriptional response to the FMP strain consortium reveals up-regulation of genes involved in interconversion of propionate and succinate.....	105
Figure S7.	A species’ contribution to the metatranscriptome is not necessarily proportional to its abundance in the 15-member community. ....	106

Figure S8.	Bottom-up analysis of genes whose expression changes significantly after introduction of the FMP strain consortium. ....	107
Figure S9.	The number of RNA-Seq reads, obtained from human fecal samples that map to genomes in the FMP strain consortium, peaks shortly after FMP consumption begins. ....	108

### Chapter 3

#### Human gut microbiome differentiation viewed across cultures, ages and families

Figure 1.	Differences in the fecal microbial communities of Malawians, Amerindians and residents of the USA at different ages. ....	159
Figure 2.	Changes in the representation of genes involved in folate and cobalamin biosynthesis and metabolism in fecal microbiomes as a function of age. ....	160
Figure 3.	Geographic differences in the bacterial functional structure of fecal microbiomes in 3 populations. ....	162
Figure 4.	Differences in the fecal microbial communities between family members across the 3 populations studied. ....	163
Figure S1.	Large interpersonal variation between children. ....	169
Figure S2.	Principal Coordinates Analysis of UniFrac distances between 524 sampled individuals. ....	170
Figure S3.	Changes in the representation of bacterial taxa in the fecal microbiota as a function of age and geographic region. ....	171
Figure S4.	Geographic differences in the bacterial phylogenetic structure of adult fecal microbiomes. ....	172
Figure S5.	Enterotype analysis. ....	173
Figure S6.	Most abundant non-bacterial members identified in the fecal microbiota. ....	174
Figure S7.	The number of ECs identified is similar in adult and infant fecal microbiomes, while the fraction of reads with assignable EC annotations declines with age in all 3 populations. ....	175
Figure S8.	Analysis of Hellinger distances between KEGG KO profiles. ....	176
Figure S9.	PCoA and Procrustes analysis of 16S rRNA and shotgun datasets annotated with KEGG ECs (a), KEGG KOs (b) and COGs (c). ....	177
Figure S10	Age-related changes in the proportional representation of genes encoding ECs involved in folate metabolism. ....	178

Figure S11.	Age-related changes in the proportional representation of genes encoding ECs involved in cobalamin biosynthesis. ....	179
Figure S12.	Spearman correlation between gut microbial species predicted to synthesize vitamins B12 and folate and their representation in fecal microbiomes at different ages and in different populations.....	180
Figure S13.	Changes in EC representation in fecal microbiomes as a function of age and population.....	181
Figure S14.	Proportional representation of 126 microbial genomes in the fecal microbiomes of breastfed Malawian twins and breast-fed and formula fed USA twins (1-5 months old). ....	182
Figure S15.	Examples of genes encoding ECs whose abundance is significantly greater in the fecal microbiomes of USA formula-fed compared to breast-fed twins (2-5 months/old).....	183
Figure S16.	Percentage of fecal microbiome gene content in sampled members of the three populations that is also represented in the METAHIT gene catalog generated from 124 adult Europeans. ....	184
Figure S17.	Principal Coordinate Analysis of Hellinger distances between the KEGG KO profiles of adult USA, Amerindian and Malawian fecal microbiomes from the present study and from 70 European microbiomes in the METAHIT dataset.....	185

## Chapter 4

### **Temporal variation in the gut microbiomes of healthy and twin pairs discordant for severe malnutrition**

Figure 1.	Geographic location of the villages where the study was conducted. ....	228
Figure 2.	Study design.....	229
Figure 3.	Anthropometric data for twin pairs whose gut microbiomes were sequenced.....	230
Figure 4.	The fraction of shotgun pyrosequencer reads that had significant annotation in the KEGG database decreases with increasing age. ....	233
Figure 5.	Large interpersonal variations are observed in the functional configurations of fecal microbial communities at early ages.....	234
Figure 6.	Principal coordinate analysis (PCoA) of Hellinger distances generated from KEGG KO profiles.....	235

Figure 7.	Analysis of Hellinger distances between and within healthy infant microbiomes.....	236
Figure 8.	Microbiomes of children within a twin pair become dissimilar with age.....	237
Figure 9.	Age and family membership explain the largest variation in the healthy microbiomes.....	238
Figure 10.	Taxonomic changes with age in healthy twin pairs.....	239
Figure 11.	Distances between microbiomes of twins discordant for kwashiorkor....	240
Figure 12.	Principal Coordinate Analysis (PCoA) of Hellinger distances generated from KEGG KO profiles.....	241
Figure 13.	Taxonomic changes with age in twins discordant for kwashiorkor.....	242
Figure 14.	Example of changes in taxonomic and functional composition in a twin pair discordant for kwashiorkor.....	243
Figure 15.	Distances between microbiomes of twins discordant for marasmus.....	244
Figure 16.	Principal Coordinate Analysis (PCoA) of Hellinger distances generated from KEGG KO profiles of twins discordant for marasmus.....	245
Figure 17.	Taxonomic changes with age in twins discordant for marasmus.....	246
Figure 18.	Example of changes in taxonomic and functional composition in a twin pair discordant for marasmus.....	247



## List of Tables

### Chapter 2

#### **The Impact of a Consortium of Fermented Milk Strains on the Gut Microbiome of Gnotobiotic Mice and Monozygotic Twins**

Table S1.	Characteristics of adult female monozygotic (MZ) twins enrolled in study. ....	112
Table S2.	Summary of human fecal metagenomic data sets. ....	113
Table S3.	Features of the microbial genomes in the 5-member FMP strain consortium and the 15-member model human gut microbiota. ....	118
Table S4.	Carbohydrate -active enzyme (CAZy) annotation data. ....	118
Table S5.	COPRO-Seq analysis of bacterial species abundance in mouse fecal samples. ....	119
Table S6.	INSeq analysis. ....	122
Table S7.	Differentially expressed <i>B. animalis</i> subsp. <i>lactis</i> (CNCM I-2494) genes. ....	122
Table S8.	Top-down function-level analysis of the impact of the FMP strain consortium on the model human gut microbiota’s metatranscriptome. ...	122
Table S9.	Model human gut microbiota membrane transport genes demonstrating more than or equal to fourfold increases or decreases in their expression after introduction of the FMP strain consortium. ....	122
Table S10.	Bottom-up (gene-level) analysis of the impact of the FMP strain consortium on the model community’s metatranscriptome. ....	122
Table S11.	Results of random forests–supervised classification analysis. ....	123
Table S12.	Urine metabolites whose levels change significantly in transitions between colonization states. ....	125
Table S13.	ShotgunFunctionalizeR analysis of EC-level changes in the metatranscriptome as a function of FMP strain introduction into mice and humans. ....	125
Table S14.	Primers and amplification conditions used for quantitative PCR assays of FMP consortium strains in fecal DNA. ....	125

Table S15.	List of 136 microbial genomes used to analyze human fecal RNAseq data.....	126
------------	--	-----

### Chapter 3

#### Human gut microbiome differentiation viewed across cultures, ages and families

Table S1.	Diet survey conducted in two Amerindian villages.....	188
Table S2.	Summary of study participants and of fecal bacterial 16S rRNA and whole community DNA sequence datasets.....	188
Table S3.	P values (Student t-test with 1000 Monte Carlo permutations) of UniFrac and Hellinger distances between fecal communities of children and adults shown in Fig. 1b,c.....	189
Table S4.	List of the 126 reference human gut microbial genomes.....	190
Table S5.	Spearman correlations of relative abundances of reads that map to microbial genomes in fecal microbiomes with age for each country.....	193
Table S6.	Results of Random Forests classifier of OTUs (species-level phylotypes) that discriminate the adult fecal microbiota of USA and non-USA residents. ....	194
Table S7.	ECs identified by Random Forests and ShotgunFunctionalizeR that exhibit significant age-associated differences.....	196
Table S8.	ECs identified by Random Forests and ShotgunFunctionalizeR that exhibit significant population-specific differences in babies (0-6 months old). ..	196
Table S9.	ECs identified by Spearman correlation analysis that exhibit age-associated changes in their proportional representation. ....	196
Table S10.	ECs identified by Random Forests and ShotgunFunctionalizeR that exhibit significant differences in their representation in the fecal microbiomes of 4 breast-fed USA twin pairs versus 4 formula-fed USA twin pairs (2-5 months old). ....	196
Table S11.	ECs identified by Random Forests and ShotgunFunctionalizeR that exhibit significant population-specific differences in the fecal microbiomes of adults.....	196

## Chapter 4

### **Temporal variation in the gut microbiomes of healthy and twin pairs discordant for severe malnutrition**

Table 1.	Number of twin pairs enrolled in study, their zygoty, gender and nutritional status.....	250
Table 2.	Gender distribution in undernourished twin pairs. ....	251
Table 3.	Characteristics of twins with malnutrition.....	252
Table 4.	Characteristics of families whose fecal microbiomes were subjected to shotgun sequencing.....	253
Table 5.	Information on whole community DNA sequence datasets.....	254
Table 6.	List of the 127 reference sequenced human gut microbial genomes. ....	261
Table 7.	ECs identified by Spearman correlation analysis that exhibit age-associated changes in their proportional representation in healthy fecal microbiomes.....	263
Table 8.	ECs whose representation is significantly different in the fecal microbiomes of healthy versus malnourished co-twins before and at the time of presentation with kwashiorkor, two weeks into RUTF treatment and 1 month after cessation of RUTF. ....	263
Table 9.	ECs whose representation is significantly different in the gut microbiomes of a healthy co-twin versus his co-twin with kwashiorkor (pair k138). ..	264
Table 10.	ECs whose representation in the fecal microbiome of a healthy co-twin was significantly different than in his/her co-twin with marasmus when sampled at the presentation with marasmus, two weeks into the RUTF treatment period and 1 month after cessation of RUTF.....	274
Table 11.	ECs whose representation was significantly different in the fecal microbiome of a healthy twin versus his co-twin who presented with marasmus (pair m229). ....	274

## **Chapter 1**

### **Introduction**

Among the diversity of life on our planet, and all the networks and patterns occurring within this diversity, few organisms exist that can survive in isolation. We humans are part of this diversity. We are connected and dependent on the other forms of life that surround us. Specifically we have formed a symbiotic relationship with microbes. Our most densely populated body habitat is the distal gut where up to one trillion microbial cells can reside in a millimeter of luminal contents. I will use the term '*microbiota*' when referring to the organismal composition of this gut community, and the term '*microbiome*' referring to its aggregate pool of microbial genes.

The gut microbiota contains representatives of all three domains of life on Earth – Bacteria, which dominate this ecosystem, Archaea and Eukarya, plus viruses. Our interaction with the gut microbiota provides many benefits, ranging from protection against invasion by pathogenic organisms, development of the innate and adaptive arms of the immune system, to harvesting energy/nutrients from otherwise indigestible components of our diet (e.g. complex plant polysaccharides) and synthesis of vitamins (K, B12, folate, riboflavin). Thus, we should consider a person's genotype and metatype as being the sum of genetic and metabolic features encoded and expressed by his/her *H. sapiens* genomes and microbiomes.

Many questions regarding the human gut microbiota are unanswered. For example, the factors that determine the assembly of this microbial 'organ' starting at birth and its subsequent development and adaptations in healthy and diseased hosts are not well understood.

### **Methods for studying the gut microbiota**

The most unbiased and thorough studies of microbial diversity in the human gut have come from sequencing the small subunit ribosomal (SSU) RNA gene. This gene has been a long-favored phylogenetic marker due to its ubiquitous presence in all microorganisms

and its large degree of sequence conservation (Woese et al. 1975, Lane et al. 1985). The composition of microbial communities can be deciphered using culture-independent methods centered around sequencing of SSU rRNA genes (16S rRNA in the case of members of Bacteria and Archaea, 18S rRNA in the case of members of Eukaryota). Since members of Bacteria dominate most microbial communities, the majority of efforts have centered on developing methods for characterizing bacterial 16S rRNA genes. PCR reactions containing primers directed against highly conserved region of this gene are used to generate amplicons that span both conserved and variable regions of the 16S rRNA gene. The amplicons can then be sequenced, as part of a pool in which each biospecimen's 16S rRNA amplicons are amplified with a unique, sample-specific, error-correcting oligonucleotide identifier ('barcode'), all amplicons from all samples in a given survey are combined, and the mixture is subjected to 'multiplex' sequencing with massively parallel DNA sequencers (initially 454 FLX pyrosequencer and most recently Illumina HiSeq2000). However, these high-throughput molecular methods, and the computational tools needed to mine the resulting datasets became available only in the last several years (Hamady et al. 2008, Walters et al. 2011, Caporaso et al. 2011).

The human microbiota have fascinated scientists since the first microscope was constructed by Antonie van Leeuwenhoek. At the beginning, microscopy was the primary tool for defining diversity in the gut microbiota. This was followed by culture-based efforts, with an explosion in information occurring with the development of robust methods for culturing anaerobic bacteria (Hungate 1950, Miller and Wolin 1974, Balch et al. 1979). This allowed metabolic and other properties expressed by various species to be delineated in pure culture. However, because complex metabolic interrelationships are established among members of a microbial community, we have little information about the functional capacities of individual species in a community occupying a given habitat, or of the community as whole.

One starting point for reconstructing the functional capacity of a community is to perform shotgun sequencing of its metagenome (microbiome). This involves random fragmentation of a community DNA preparation, and sequencing of the resulting fragments (Venter et al. 2004). The sequences ('reads') are then compared to known genes in hierarchically annotated databases, such as KEGG, and assigned to orthologous gene groups (e.g., KEGG KOs). Functional profiles of many microbiomes can be compared to one another using a variety of multivariate statistical tools, such as principal coordinate analysis.

### **Adult gut microbiota**

**Table 1** provides an abbreviated historical overview of some of the reported metagenomic studies of the human gut microbiota and microbiome. These studies have taught us a number of important lessons. *First*, within a given individual, the microbiota has a complex biogeography (i.e., its composition varies along the length as well as the width of the gut). Intrapersonal variation along the length of the distal gut, which harbors the vast majority of our gut symbionts, is smaller compared to interpersonal variation. This makes the fecal microbiota a reasonable choice for monitoring intrapersonal variations in community ecology and for conducting comparative studies of interpersonal differences [a fecal sample is easy and safe to procure and 50% of its biomass is microbial, (Eckburg et al. 2005)]. *Second*, the configuration of a persons' microbiota appears to be unique. However, genetically related adult family members share more microbial lineages than unrelated individuals. Importantly, the overall degree of phylogenetic similarity between the fecal microbiota of monozygotic (MZ) adult twins is not significantly different than the degree of similarity between dizygotic adult twins (DZ), underscoring the importance of early environmental exposures and implying that host genotype is not solely responsible for the configuration of the gut microbiota. Despite large interpersonal variation in bacterial composition of the microbiota, the gene content of the microbiome is more conserved (Turnbaugh et al. 2009, Muegge et al. 2011). This has allowed a group of shared genes to be identified

and designated as a ‘core microbiome’. *Third*, compared to the human genome, the gut microbiome contains a greater representation of genes encoding enzymes involved in the degradation of complex polysaccharides, the metabolism of amino acids and xenobiotics as well as the biosynthesis of vitamins (Gill et al., 2006, Turnbaugh et al. 2010, Qin et al. 2010) ; *Fourth*, comparisons of 60 phylogenetically diverse, carnivorous, herbivorous and omnivorous mammals with three types of gut physiologies (simple gut, hind gut fermenters, foregut fermenters) revealed that diet has been the principal factor that has shaped the configuration of the gut microbiota and microbiome during the course of mammalian evolution (Ley et al. 2008; Muegge et al., 2011). Studies in humans, and in gnotobiotic mice that harbor human gut microbiota have established the importance of diet in defining community structure and function (Turnbaugh, Ridaura, et al. 2009, Faith et al. 2011, Muegge et al. 2011, Wu et al. 2011).

**Table 1. Major sequencing-based studies describing the species and gene content of the human gut microbiome.**

<b>Findings</b>	<b>Subjects</b>	<b>16S or Shotgun</b>	<b>Sequencing coverage</b>	<b>Reference</b>	<b>Year</b>
<ul style="list-style-type: none"> <li>• Inter-individual variation is greater than intra-individual</li> <li>• Large diversity previously undescribed bacterial species; <i>M.smithii</i> is the dominant archaeon</li> </ul>	Stool and mucosa from three healthy USA adults	16S rRNA	16S rRNA Sanger sequencing (13,355 seqs)	Eckburg et al.	2005
<ul style="list-style-type: none"> <li>• Relative to the human genome, an enriched capacity for carbohydrate, amino acid, vitamin, xenobiotic metabolism</li> </ul>	Stool from 2 healthy USA adults	Both	78 Mb of shotgun (Sanger) sequences plus 2,062 16S rRNA Sanger sequences	Gill et al.	2006
<ul style="list-style-type: none"> <li>• Diet and host phylogeny are primary drivers of gut microbiota</li> </ul>	59 mammals and humans	16S rRNA	>20,000 16S rRNA Sanger sequences	Ley et al.	2008
<ul style="list-style-type: none"> <li>• Unique gut microbiota in each individual – no shared bacterial taxa across all 154 individuals (abundance cutoff 0.5%)</li> <li>• Shared gene content</li> <li>• Familial similarity; no significant difference between MZ and DZ twins</li> </ul>	50 USA healthy adult MZ and DZ twin pairs and their mothers	Both	9,920 Sanger and 1,937,461 454 16SrRNA sequences; 2.14 Gb 454 shotgun	Turnbaugh et al.	2009
<ul style="list-style-type: none"> <li>• Diet dominant shaper of functional features of the gut microbiome; carnivores have enhanced capacity for amino acid catabolism, herbivores for amino acid synthesis</li> <li>• Functional capacity can be predicted from taxonomic composition</li> </ul>	33 mammals and 18 USA healthy adults	Both	149,675 16S rRNA 454 sequences; 2,160,000 shotgun reads	Muegge et al.,	2011



The importance of diet in shaping gut microbiota had been recognized long before the advent of metagenomic methods (Cannon 1921, Porter and L. Rettger 1940, Torrey 1919). A number of experiments were conducted in humans and animals who were fed human diets from several days to several months, usually composed of a single ingredient, such as ground beef or egg yolks or potatoes (Cannon 1921). In one experiment, fecal microbiota was examined in three subjects before, during and after consumption of a diet composed exclusively of meat (Torrey and Montu 1931). Two of the three subjects were Arctic explorers who consumed meat for 13 months. During that period the frequency of fecal sample collection varied from every 2 -3 days to 6 months resulting in 16 samples from one man, and 8 from another. A third subject consumed meat for 10 days during which time three fecal samples were obtained. Bacterial diversity was determined by direct microscopic counting of diluted fecal samples spread onto a glass slide and stained with Gram's stain, as well as by culture-based methods. Only a few isolates were identified (at that time their assigned taxonomic names were *Bacillus coli*, *Proteus*, *Enterococcus*, *Staphylococcus*, *Lactobacillus acidophilus*, *B.bifidus* and *B.welchii*). One common observation was made across all three humans: direct microscopic count revealed decreased number of bacteria when meat diet was consumed, but reverted to the original density when shifted to the regular diet. The authors speculated that the decrease in isolation of certain bacteria was due to the absence of the carbohydrates required for their growth.

### **Gut microbiota in humans living in different geographic regions**

If the gut microbiota of mammals can be delineated based on the dietary habits of the host, is it possible to distinguish human gut microbiomes based on the genetic makeup of the human populations and their cultural traditions, especially their dietary habits? Much like human haplogroups, is it possible to classify human gut microbiomes based on shared taxonomic and functional features and would they overlap with the haplogroups (implying the influence of genetics)? Does the “map of the human gut microbiome” parallel

anthropologic characterizations of human cultural evolution and diversity? Several reports have described differences in the bacterial taxonomic composition between various human populations: the majority are based on analysis of a handful of subjects and use culture-based or other methods that did not allow broad sampling of bacterial diversity (e.g., Peach et al. 1974). Furthermore, there is a dearth of comprehensive comparisons to date of the gut microbiota in healthy people living in economically highly developed versus economically least developed countries: a recent report of fecal microbiota in children living in Europe (Italy) and West Africa (Burkina Faso) attributed the differences between the two populations to the differences in dietary habits (De Filippo et al. 2010). Moreover, the functional diversity of the gut microbiome in humans living in different parts of the world is largely unknown. I will address these questions in the Chapter 3 of my thesis where I compare the phylogenetic and functional composition of gut microbial communities in humans living in three different countries located on three different continents with distinct cultural traditions.

### **Temporal variation of adult microbiota**

Any comparison of the gut microbiota of healthy versus sick humans living in a given country, or humans from different geographic regions is challenging: i.e., because each individual's microbiota is unique, detection of changes across multiple unrelated people may be a daunting task. To understand the differences associated with a given host variable (age, gender, physiological phenotype, pathologic state, etc), understanding the degree of variation within a healthy person's gut microbiota is necessary: how does the composition of a fecal microbiota sampled at a given time from a given individual compare to his/her microbiota sampled a day, a month, or a year earlier? A few studies examining the gut microbiota of small number of healthy individuals over a short period of time using molecular methods such as TGGE and FISH, concluded that intra-individual variation is less than inter-individual variation (Zoetendal et al. 1998, Franks et al. 1998). Bacterial 16S rRNA-

based sequencing followed, allowing broader and less biased methods to estimate this variation (Les Dethlefsen et al. 2008, L. Dethlefsen and D. A. Relman 2010, Caporaso et al. 2011). Recently, deep Illumina-based sequencing of 16S rRNA amplicons generated from fecal specimens collected from two individuals on a daily basis for up to a year, showed that individual's microbiota is quite variable, with fewer than 10% phlotypes persisting (i.e. detectable at the level of sampling employed) over the period surveyed (Caporaso et al. 2011). So far there have not been published reports characterizing temporal variation in the gene content of the gut microbiome in a cohort of healthy related or unrelated individuals. In the chapter 2 of my thesis I address this question by studying seven healthy adult female MZ twin pairs: the degree of temporal variation of their gut microbiota and microbiomes was defined, and effect of consumption of a commercially available fermented dairy product on this variation was determined.

### **Assembly of the gut microbiota in infants**

While we have some idea of the structure and function of the healthy adult gut microbiota and microbiome, as well as the influence of diet, many questions remain unanswered. How and when do we acquire these microbes? Are their identifiable shared versus distinctive features of microbiota/microbiome assembly in infants and children living in different geographic and cultural contexts? Are we programmed to evolve a microbial community peculiar to each of us; how do events experienced early in our lives shape maturation and differentiation of our microbiota and microbiome? Descriptions of the patterns of the initial microbial colonization of a newborn human began to appear more than a century ago. The fact that babies are born 'germ-free' and then rapidly colonized by bacteria within hours of birth was recognized early in the history of microbiology (Rettger and Cheplin 1921). Theodor Escherich in 1885 was the first scientist to conduct a systematic survey of feces in breast-fed babies starting at birth and through the first few months of life. His observations were based on direct microscopy counts and morphological descriptions of cells spread

on glass slides (Escherich 1988, Hall and O'Toole 1935). When examining the meconium of two infants who died during birth he did not find any bacteria, concluding that the meconium was sterile. According to his observations, the “first settlers” in a newborn baby were “cocci or yeasts” that were also found in the air, presumably one of the sources of colonization. The number of bacterial cells increased within the first 24 hours of life. He also noted that the bacilli observed in the meconium of a newborn were also found in the feces of that same baby 8–10 days later. In 1899, Henry Tissier isolated Bifidobacterium (*Bacillus bifidus* at that time) from the feces of a breast-fed baby. In 1900, Moro isolated *Bacillus acidophilus*. Both species were claimed to be predominant in an infant gut (Rettger and Cheplin 1921).

Culture-based studies revealed that the complexity of a gut microbiota increases with age (Rettger and Cheplin 1921). With the advent of 16S rRNA-based DNA microarrays and with more limited amplicon sequencing, Palmer et al. 2007 demonstrated in a cohort of 14 children, studied from birth through 1 year of life, that each child appeared to have a unique pattern of microbial colonization. Intriguingly, they found a greater degree of similarity in the one twin pair enrolled in this study, leading them to conclude that early environmental exposures and/or host genetics play an important role.

Despite the seemingly chaotic period of initial microbial colonization, some obvious and shared features about initial succession have been defined. Colonization starts within hours after birth: in vaginally-delivered infants, members of Proteobacteria dominate in the first few days of life followed by Bifidobacteria, which prevail in the first several months of life (Favier et al. 2002). The gut microbiota is highly variable within a baby even during the period of exclusive breast-feeding. This implies that physiological maturation of the host, including the maturation of the innate and adaptive arms of the immune system (which, in turn may be influenced by the microbiome itself forming a positive feedback loop), as well as other host factors influence community assembly. Variations in the composition of mother's breast milk over the course of lactation, and the influence of this biochemical and

immunologic variation on the microbiome are not well described or understood.

During the period of weaning onto solid foods, diversity increases dramatically, with greater representation of taxa usually found in adults, such as members of Firmicutes and Bacteroidetes. Interestingly, members of Bacteroidetes and Firmicutes have been detected in the first few days of life, which poses a question “is everything there from the beginning” and are changes primarily in relative representation of different bacteria in response to diet, host physiology and various ill-defined environmental stimuli?

Many studies reported that microbes that colonize an infant’s gut were derived from mother’s skin, vagina and feces, as well as the environment in which birth occurred (Dudgeon and Jewesbury 1924, Brook et al. 1979, Mändar and Mikelsaar 1996). A recent bacterial 16S rRNA-based study of 10 breast-fed newborn babies and their mothers living in Venezuela characterized over a 24 h period reported significant differences in fecal microbiota composition between babies delivered vaginally versus those delivered by cesarean section (Dominguez-Bello et al. 2010). The fecal microbiota in vaginally-delivered babies were dominated by species present in their mother’s vagina, while the bacteria present on mother’s skin were well represented in the microbiota of infants delivered by C-section. However, given the short duration of this study, it was unclear how long these latter taxa persisted and how they affected subsequent gut community assembly. It is also unclear how many taxa are transmitted from other family members, unrelated caregivers that interact with a baby, or other environmental microbial reservoirs (including pets).

Culture-based studies have revealed that the microbiota of formula fed babies is quite different from breast-fed. The former have less Bifidobacteria and more Firmicutes and Bacteroidetes (H. J. Harmsen et al. 2000, Yoshioka et al. 1983).

Only a handful studies reported the functional changes in the gut microbiomes of growing healthy babies (Kurokawa et al. 2007, Koenig et al. 2011). In a report of one baby sampled over a year, fecal microbiomes sampled during the first three months of life

contained genes involved in utilization of lactose and galactose which are highly abundant in breast milk, as well as genes involved in the degradation of sialic acid residues present in glycans found in the mucus overlying the gut epithelium. Remarkably, both studies reported the presence of genes encoding enzymes involved in the degradation of complex polysaccharides in the microbiomes of breast-fed infants implying that the infant gut microbiome is equipped to utilize complex glycans long before encountering them.

### **Is the assembly of the gut microbiota follows the same pattern in infants living in different parts of the world?**

The vast majority of studies of gut microbial community assembly in infants have been conducted in Europe and other economically well developed countries. This raises the obvious question of what differences exist in the maturation of the gut microbiota and microbiomes in individuals living in Western societies versus countries where lifestyles and cultural traditions are quite different (e.g., where sanitation is poor, and where diets contain much less fat and protein)?

De Filippo et al. (De Filippo et al. 2010) used 16S rRNA sequencing to demonstrate differences in the fecal microbiota of 15 Italian children and 14 children living in West Africa (Burkina Faso). These children were 1-6 years old: each provided a single fecal sample. The microbiota of children living in Burkina Faso had higher representation of Bacteroidetes compared to those living in Italy. However, since these children were not subjected to serial sampling, it is unknown at what developmental stage these differences became evident: i.e. shortly after birth or during weaning on solid foods, or later.

There have been only few (mostly culture-based) studies of infants living in other economically least developed or less developed countries (Guatemala, India, Ethiopia); they have confirmed the prevalence of Enterobacteria in the first few days of life, with subsequent shifts to Bifidobacteria dominated communities (Mata and Urrutia, 1971). The majority of studies of the gut microbiota in economically developing countries have focused

on determining whether potential pathogens are present in the stools of children suffering from diarrhea and other illnesses. However, no study has been published to my knowledge comparing the gut microbiota and microbiomes of healthy infants in developed and developing countries. I address this question in the Chapter 3 of my thesis.

### **Is the assembly and function of the gut microbiota compromised in severe childhood malnutrition?**

Malnutrition is a major global health problem. It is estimated that almost half of all deaths in children under five years of age is directly or indirectly caused by malnutrition (UNICEF 2008).

A person is malnourished (undernourished) if his or her diet does not provide adequate calories and micronutrients for growth and maintenance, or if he or she is unable to fully utilize the energy and nutrients contained in the food that he or she eats (UNICEF 2006). Malnutrition is diagnosed using anthropometric measurements of height and weight, which are then compared to the median measurements in an international reference population adjusted for age and gender. These are converted into Z scores, which describe the number of standard deviations from median weight for age (WAZ, a measure of acute malnutrition), height for age (HAZ, measure of chronic malnutrition) or weight for height (WHZ, another measure of acute malnutrition, UNICEF 2008).

Severe malnutrition, which is defined if a child's WHZ score is less than -3 or if a child presents with edema (WHO 2004) can lead to three syndromes – marasmus, kwashiorkor, and marasmic kwashiorkor. Marasmus, derived from the Greek word *marasmos* or wasting, is characterized by severe wasting. The highest incidence of marasmus occurs between 6-17-months-of-age (Courtright and Canner 1995, Ahmed et al. 2009). Kwashiorkor, the name derived from the Ga language in Ghana meaning “the sickness of the weaning”, occurs in about 2% malnourished children, at older ages, usually 1-4 years; it

is more difficult to treat, and has a higher mortality rate than marasmus (Scrimshaw and Viteri 2010). It is characterized by the presence of edema, fatty liver (hepatic steatosis) and de-pigmented skin (Blackburn 2001). Children with marasmic kwashiorkor have both wasting and edema.

Kwashiorkor was first described by Cicely Williams in 1931 (Williams 1935). A number of hypotheses have been proposed on the etiology of this disease; however, none yet received a wide support. A long prevailing hypothesis held that kwashiorkor was caused by hypoalbuminemia due to a diet low in protein. However, the diets of children with kwashiorkor do not differ from those with marasmus (Golden 2002, Lin et al. 2007). Moreover, edematous malnutrition can resolve on a low-protein diet without significant accompanying changes in the levels of plasma proteins.

The cause of malnutrition does not result from food insecurity alone. Infectious diseases play a large role (Golden 2002, Prentice et al. 2008). For example, infection with enteropathogens often leads to diarrhea, one of the leading causes of childhood deaths, resulting in nutrient malabsorption and suppression of appetite (Schaible and Kaufmann 2007, Victora et al. 2008). HIV, malaria and tuberculosis lead to immune suppression. The effects of these infectious diseases on gut microbial ecology and in turn on the nutrient processing and other metabolic activities of the gut microbiome have not been described. Pathogenic organisms have evolved numerous mechanisms for nutrient sequestration from the host. For example, reliance on host-derived iron is a feature of members of *Yersinia*, *Chlamydia*, *Salmonella* and *Mycobacterium* (Schaible and Kaufmann 2004).

Severe malnutrition compromises the innate and acquired immune responses, increasing susceptibility to infection (Schaible and Kaufmann 2007). Antibody responses are reduced in malnourished children. Severe thymus atrophy, defined by death and decreased proliferation of CD4+CD8+ thymocytes, has been observed in malnourished subjects. Such abnormalities are reversible with nutritional interventions (Savino et al. 2007).



Both malnutrition and infection produce changes in gut morphology that may affect the microbiota and nutrient processing. Villus atrophy occurs in the small intestines of malnourished hosts (often in the context of a histopathologic state known as environmental enteropathy), affecting the surface area available for nutrient processing and absorption (Brunser et al. 1968, Welsh et al. 1998, Tabrez and Roberts 2001, Redmond et al. 1971). Malnutrition also results in reduced thickness of the polysaccharide-rich mucus slime layer, a microhabitat where embedded microbes can maintain a foothold in the ecosystem and avoid washout and where members of the microbiota can physically juxtapose themselves in order to effectively establish syntrophic (nutrient sharing) relationships with one another, as well as provide protection from invasion of pathogens. Thus, loss of mucus glycans could affect the composition, stability as well as the metabolic activities of the microbiota/microbiome.

Current protocols for the management of moderate and severe malnutrition still result in 30% case fatality rates for children with marasmus, and 60% for children with kwashiorkor (Collins 2007). Recently, a new dietary formula has been developed that revolutionized treatment of moderate and severe malnutrition. This ready-to-use-therapeutic food (RUTF) is a mixture of peanut butter, sugar and milk fortified with vitamins and minerals. So far, it has been extensively used in Malawi, where 90% malnourished children respond successfully (Linneman et al. 2007). The short term and longer term impact of this supplementation on gut microbial ecology and microbiome function has not been defined, nor is there information about why it has poor efficacy in some children but not others, or whether its success will be comparable in children living in other countries with distinctive cultures and diets.

### **Gut microbiota configuration in children with malnutrition.**

The gut microbiota was implicated in the pathogenesis of severe malnutrition, such as kwashiorkor several decades ago (Smythe 1958). The hypothesis was that the gut micro-

biota competes with the host for nutrients thus causing malnutrition, and therefore should be 'eliminated' using antibiotics. Very few reports exist describing the gut microbiota in children affected with severe malnutrition: only one study describes the gut microbiome in only one malnourished child: the major findings in this one case were high prevalence of enteropathogens and genes associated with them (Gupta et al. 2011). A number of studies have reported the presence of 'bacterial overgrowth' in the small intestine. Using culture-based methods (Mata et al. 1972) examined bacteria in the small intestines as well as the feces of 13 children with severe malnutrition before, during and after nutritional therapy, as well as 4 healthy controls. Bifidobacteria were dominant in the feces of healthy children, but in only one third of children who suffered from severe malnutrition. Facultative aerobes sometimes outnumbered anaerobes in children with malnutrition, but not in the healthy controls. In all cases of malnutrition, the microbiota responded to dietary intervention: total bacterial counts decreased in the jejunal aspirates, and the anaerobic bacteria prevailed over facultative anaerobes in feces. Notably, although enteropathogens were not detected in the majority of children with malnutrition, most children contained intestinal parasites. Importantly, there was no difference seen across all subjects either at the time of presentation with malnutrition or as a response to therapeutic intervention in terms of types of bacteria that were isolated.

### **Overview of the dissertation**

The central hypotheses of my thesis are as follows: (i) the gut microbiome has a definable pattern of functional maturation during postnatal life and this maturation plays a key role in myriad physiologic aspects underlying the healthy growth of infants and children; (ii) features of this maturation are shared across diverse human populations yet at the same time there are definable patterns of microbiome differentiation across ages and cultures/geography; (iii) the microbiota and microbiome are biomarkers and mediators of nutritional status in children and adults – as such, considerations of the nutritional requirements of humans

should incorporate information about the functional potential and/or activities of their microbiomes; (iv) severe forms of malnutrition are associated with and to some extent caused by disruptions in the functions normally encoded and expressed by the gut microbiome.

With these hypotheses in mind, the goal of my thesis was to describe the organismal composition and gene content of the gut microbiota/microbiomes of infants and children who were severely malnourished, prior to, during and after treatment with a peanut butter-based RUTF. These infants lived in Malawi, one of the poorest countries in the world where malnutrition is rampant. Because of the interpersonal variations that exist in gut microbial communities, my work focused on twins who were concordant for healthy status or who were discordant for marasmus or kwashiorkor. In the case of the discordant twins, and in accordance with current standards of care in Malawi, both co-twins were treated with RUTF. All twins studied (n=317) were enrolled in a Bill and Melinda Gates Foundation-sponsored project.

In order to determine if the gut microbiome is altered in malnourished children and the degree to which the microbiome responds to a nutritional resuscitation, I needed to first describe the variation of the gut microbiome in healthy individuals. In addition, to test how applicable my findings from Malawi would be for the rest of the world, I had to characterize the organismal and gene content of the microbiomes in healthy humans across multiple geographic regions. Therefore, my thesis consists of three parts. In Chapter 2, I describe the gut microbiota and microbiome in healthy adult USA MZ twins who were each sampled nine times over a four-month period. I found that the microbiota and microbiome are variable within each twin, but the variation was less between co-twins compared to unrelated individuals. This study helped to establish a baseline variation in the healthy adult microbiome. Furthermore, it allowed me to develop methods for analyzing the datasets I generated from the fecal microbiota and microbiomes of infants and children during the later phases of my thesis. In Chapter 3, I describe the variation in gut microbiota and microbiome in 524 healthy adults and children living in three different countries, with

very distinct cultural traditions, located on three continents (three metropolitan areas in the USA, two villages of Amerindians located in the Amazon region of Venezuela, plus four rural villages in Malawi). Interpersonal variation in babies was significantly greater than between adults. Nevertheless, related family members shared more features of their microbiota and microbiomes than unrelated individuals. In addition, common patterns of community assembly, and functional maturation of microbiomes were observed across all three countries, as well as distinct features associated with Western versus non-Western societies. Finally, in Chapter 4 I used metagenomic methods and a variety of computational/statistical tools to characterize assembly of the gut microbiomes of healthy and severely malnourished Malawian twins during their first three years of life, and the effects of RUTF on the configuration of their microbiomes.

## **References**

- Ahmed, T., S. Rahman, and A. Cravioto. Oedematous Malnutrition. *The Indian Journal Of Medical Research* **130**, 651-654 (2009).
- Balch, W.E., G.E. Fox, L. J. Magrum, C. R. Woese, and R. S. Wolfe. Methanogens: Reevaluation Of A Unique Biological Group. *Microbiological Reviews* **43**, 260-296 (1979).
- Blackburn, G.L. Pasteur's Quadrant And Malnutrition. *Nature* **409**, 397-401 (2001).
- Brook, I., C., T. Barrett, C. R. Brinkman, W. J. Martin, and S. M. Finegold. Aerobic And Anaerobic Bacterial Flora Of The Maternal Cervix And Newborn Gastric Fluid And Conjunctiva: A Prospective Study. *Pediatrics* **63**, 451 -455 (1979).
- Brunser, O., A. Reid, F. Monckeberg, A. Maccioni, and I. Contreras. Jejunal Mucosa In Infant Malnutrition. *The American Journal Of Clinical Nutrition* **21**, 976-983 (1968).
- Cannon, P.R. The Effects Of Diet On The Intestinal Flora. *The Journal Of Infectious Diseases* **29**, 369-385 (1921).
- Caporaso, J.G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, et al. Qiime Allows Analysis Of High-Throughput Community Sequencing Data. *Nature Methods* **7**, 335-336 (2010).
- Caporaso, J.G., C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, et al. Moving Pictures Of The Human Microbiome. *Genome Biology* **12**, R50 (2011).
- Collins, S. Treating Severe Acute Malnutrition Seriously. *Archives Of Disease In Childhood* **92**, 453-461 (2007).
- Courtright, P. and J. Canner. The Distribution Of Kwashiorkor In The Southern Region Of Malawi. *Annals Of Tropical Paediatrics* **15**, 221-226 (1995).

- De Filippo, C., D. Cavalieri, M. Di Paola, M. Ramazzotti, J. Baptiste Poullet, S. Massart, S. Collini, G. Pieraccini, and P. Lionetti. Impact Of Diet In Shaping Gut Microbiota Revealed By A Comparative Study In Children From Europe And Rural Africa. *Proceedings Of The National Academy Of Sciences* **107**, 14691 -14696 (2010).
- Dethlefsen, L. and D. A. Relman. Colloquium Paper: Incomplete Recovery And Individualized Responses Of The Human Distal Gut Microbiota To Repeated Antibiotic Perturbation. *Proceedings Of The National Academy Of Sciences* **108**, 4554-4561 (2010).
- Dethlefsen, L., S. Huse, M. L. Sogin, and D. A. Relman. The Pervasive Effects Of An Antibiotic On The Human Gut Microbiota, As Revealed By Deep 16s rRNA Sequencing. *Plos Biology* **6**, E280 (2008).
- Dominguez-Bello, M. G., E. K. Costello, M. Contreras, M. Magris, G. Hidalgo, N. Fierer, and R. Knight. Delivery Mode Shapes The Acquisition And Structure Of The Initial Microbiota Across Multiple Body Habitats In Newborns. *Proceedings Of The National Academy Of Sciences* **107**, 11971-11975 (2010).
- Dudgeon, L. and R. C. Jewesbury. The Bacteriology Of Human Milk. *The Journal Of Hygiene* **23**, 64-76 (1924).
- Eckburg, P.B., E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman. Diversity Of The Human Intestinal Microbial Flora. *Science* **308**, 1635 -1638 (2005).
- Escherich, T. The Intestinal Bacteria Of The Neonate And Breast-Fed Infant. *Reviews Of Infectious Diseases* **10**, 1220-1225 (1988).
- Faith, J., N. P. McNulty, F. E. Rey, and J. I. Gordon. Predicting A Human Gut Microbiota's Response To Diet In Gnotobiotic Mice. *Science* **333**, 101 -104 (2011).

- Favier, C. F., E. E. Vaughan, W. M. De Vos, and A. D. L. Akkermans. Molecular Monitoring Of Succession Of Bacterial Communities In Human Neonates. *Appl. Environ. Microbiol.* **68**, 219-226 (2002).
- Franks, A., H., Hermie J. M. Harmsen, G. C. Raangs, G. J. Jansen, F. Schut, and G. W. Welling. Variations Of Bacterial Populations In Human Feces Measured By Fluorescent In Situ Hybridization With Group-Specific 16s Rrna-Targeted Oligonucleotide Probes. *Appl. Environ. Microbiol.* **64**, 3336-3345 (1998).
- Gill, S. R., M. Pop, R.T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson. Metagenomic Analysis Of The Human Distal Gut Microbiome. *Science* **312**, 1355 -1359 (2006).
- Golden, M.H. The Development Of Concepts Of Malnutrition. *The Journal Of Nutrition* **132**, 2117s-2122s (2002).
- Gupta, S.S., M. H. Mohammed, T. S. Ghosh, S. Kanungo, G. B. Nair, and S. S. Mande. Metagenome Of The Gut Of A Malnourished Child **3**, 7-7 (2011).
- Hall, I.C., and E. O'Toole. Intestinal Flora In New-Born Infants: With A Description Of A New Pathogenic Anaerobe, Bacillus Difficilis. *Am J Dis Child* **49**, 390-402 (1935).
- Hamady, M., J.J. Walker, J. K.Harris, N. J. Gold, and R. Knight. Error-Correcting Barcoded Primers For Pyrosequencing Hundreds Of Samples In Multiplex. *Nat Meth* **5**, 235-237 (2008).
- Harmsen, H.J., A.C. Wildeboer-Veloo, G.C. Raangs, A.A. Wagendorp, N. Klijn, J. G. Bindels, and G.W. Welling. Analysis Of Intestinal Flora Development In Breast-Fed And Formula-Fed Infants By Using Molecular Identification And Detection Methods. *Journal Of Pediatric Gastroenterology And Nutrition* **30**, 61-67 (2000).
- Hungate, R. E. The Anaerobic Mesophilic Cellulolytic Bacteria. *Bacteriological Reviews* **14**, 1-49 (1950).

- Koenig, J. E., A. Spor, N. Scalfone, A. D. Fricker, J. Stombaugh, R. Knight, L. T. Angenent, and R. E. Ley. Succession Of Microbial Consortia In The Developing Infant Gut Microbiome. *Proceedings Of The National Academy Of Sciences* **108**, 4578-4585 (2011).
- Kurokawa, K., T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, et al. Comparative Metagenomics Revealed Commonly Enriched Gene Sets In Human Gut Microbiomes. *DNA Research* **14**, 169 -181 (2007).
- Lane, D. J., B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin, and N. R. Pace. Rapid Determination Of 16s Ribosomal RNA Sequences For Phylogenetic Analyses. *Proceedings Of The National Academy Of Sciences* **82**, 6955 -6959 (1985).
- Ley, R. E., M. Hamady, C. Lozupone, P. J. Turnbaugh, R. R. Ramey, J. S. Bircher, M. L. Schlegel, et al. Evolution Of Mammals And Their Gut Microbes. *Science* **320**, 5883 1647 -1651 (2008).
- Lin, C. A., S. Boslaugh, H. M. Ciliberto, K. Maleta, P. Ashorn, A. Briend, and M. J. Manary. A Prospective Assessment Of Food And Nutrient Intake In A Population Of Malawian Children At Risk For Kwashiorkor. *Journal Of Pediatric Gastroenterology And Nutrition* **44**, 487-493 (2007).
- Linneman, Z., D. Matilsky, M. Ndekha, M. J. Manary, K. Maleta, and M. J. Manary. A Large-Scale Operational Study Of Home-Based Therapy With Ready-To-Use Therapeutic Food In Childhood Malnutrition In Malawi. *Maternal and Child Nutrition* **3**, 206-215 (2007).
- Mändar, R., and M. Mikelsaar. Transmission Of Mother's Microflora To The Newborn At Birth. *Biology Of The Neonate* **69**, 30-35 (1996).
- Mata, L. J., F. Jiménez, M. Córdón, R. Rosales, E. Prera, R. E. Schneider, and F. Viteri. Gastrointestinal Flora Of Children With Protein--Calorie Malnutrition. *The American Journal Of Clinical Nutrition* **25**, 118-126 (1972).



- Mata, L. J., and J.J. Urrutia. Intestinal Colonization Of Breast-Fed Children In A Rural Area Of Low Socioeconomic Level. *Annals Of The New York Academy Of Sciences* **176**, 93-109 (1971).
- Miller, T. L., and M. J. Wolin. A Serum Bottle Modification Of The Hungate Technique For Cultivating Obligate Anaerobes. *Applied Microbiology* **27**, 985-987 (1974).
- Muegge, B. D., J. Kuczynski, D. Knights, J. C. Clemente, A. González, L. Fontana, B. Henrissat, R. Knight, and J. I. Gordon. Diet Drives Convergence In Gut Microbiome Functions Across Mammalian Phylogeny And Within Humans. *Science* **332**, 970-974 (2011).
- Palmer, C., E. M. Bik, D. B. Digiulio, D. A. Relman, and P. O. Brown. Development Of The Human Infant Intestinal Microbiota. *Plos Biol* **5**, E177 (2007).
- Peach, S., F. Fernandez, K. Johnson, and B. S. Drasar. The Non-Sporing Anaerobic Bacteria In Human Faeces. *Journal Of Medical Microbiology* **7**, 213 -221 (1974).
- Porter, J.R., and L. Rettger. Influence Of Diet On The Distribution Of Bacteria In The Stomach, Small Intestine And Cecum Of The White Rat. *The Journal Of Infectious Diseases* **66**, 104-110 (1940).
- Prentice, A. M., M. E. Gershwin, U. E. Schaible, G. T. Keusch, C. G. Victora, and J. I. Gordon. New Challenges In Studying Nutrition-Disease Interactions In The Developing World. *The Journal Of Clinical Investigation* **118**, 1322-1329 (2008).
- Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, et al. A Human Gut Microbial Gene Catalogue Established By Metagenomic Sequencing. *Nature* **464**, 59-65 (2010).
- Redmond, A.O., R. O. Kaschula, C. Freese, and J. D. Hansen. The Colon In Kwashiorkor. *Archives Of Disease In Childhood* **46**, 470-473 (1971).
- Rettger, L. F., and H. A. Cheplin. *A Treatise On The Transformation Of The Intestinal*

- Flora: With Special Reference To The Implantation Of Bacillus Acidophilus*. Yale University Press, (1921).
- Savino, W., M. Dardenne, L. A. Velloso, and S. D. Silva-Barbosa. The Thymus Is A Common Target In Malnutrition And Infection. *The British Journal Of Nutrition* **98**, S11-16 (2007).
- Schaible, U. E., and S. H. E. Kaufmann. Iron And Microbial Infection. *Nature Reviews. Microbiology* **2**, 946-953 (2004).
- Scrimshaw, N. S., and F. E. Viteri. Incap Studies Of Kwashiorkor And Marasmus. *Food And Nutrition Bulletin* **31**, 34-41 (2010).
- Smythe, P.M. Changes In Intestinal Bacterial Flora And Role Of Infection In Kwashiorkor. *The Lancet* **272**, 724-727 (1958).
- Tabrez, S., and I. M. Roberts. Malabsorption And Malnutrition. *Primary Care* **28**, 505-522, (2001)
- Torrey, J. C. The Regulation Of The Intestinal Flora Of Dogs Through Diet. *The Journal Of Medical Research* **39**, 415-447 (1919).
- Torrey, J. C., and E. Montu. The Influence Of An Exclusive Meat Diet On The Flora Of The Human Colon. *The Journal Of Infectious Diseases* **49**, 141-176 (1931).
- Turnbaugh, P. J., C. Quince, J. J. Faith, A. C. Mchardy, T. Yatsunenko, F. Niazi, J. Affourtit, et al. Organismal, Genetic, And Transcriptional Variation In The Deeply Sequenced Gut Microbiomes Of Identical Twins. *Proceedings Of The National Academy Of Sciences* **107**, 7503-7508 (2010).
- Turnbaugh, P. J., V. K. Ridaura, J. J. Faith, F. E. Rey, R. Knight, and J. I. Gordon. The Effect Of Diet On The Human Gut Microbiome: A Metagenomic Analysis In Humanized Gnotobiotic Mice. *Science Translational Medicine* **1**, 6ra14 (2009).

- Turnbaugh, P. J., M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M.L. Sogin, et al. A Core Gut Microbiome In Obese And Lean Twins. *Nature* **457**, 480-484 (2009).
- UNICEF. State Of The World's Children 2008 - Child Survival, (2008).
- UNICEF Progress For Children: A Report Card On Nutrition, (2006).
- Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, et al. Environmental Genome Shotgun Sequencing Of The Sargasso Sea. *Science* **304**, 66-74 (2004).
- Victora, C. G., L. Adair, C. Fall, P. C. Hallal, R. Martorell, L. Richter, and H. S. Sachdev. Maternal And Child Undernutrition: Consequences For Adult Health And Human Capital. *Lancet* **371**, 340-357 (2008).
- Walters, W. A., J. G. Caporaso, C. L. Lauber, D. Berg-Lyons, N. Fierer, and R. Knight. Primerprospector: De Novo Design And Taxonomic Analysis Of Barcoded Polymerase Chain Reaction Primers **27**, 1159-1161 (2011).
- Welsh, F. K., S. M. Farmery, K. Maclennan, M. B. Sheridan, G. R. Barclay, P. J. Guillou, and J. V. Reynolds. Gut Barrier Function In Malnourished Patients. *Gut* **42**, 396-401 (1998).
- WHO. Severe Malnutrition: Report Of A Consultation To Review Current Literature, 6-7 September 2004, (2004).
- Williams, C. Kwashiorkor. A Nutritional Disease Of Children Associated With A Maize Diet. *The Lancet* 1151 (1935).
- Woese, C. R., G. E. Fox, L. Zablen, T. Uchida, L. Bonen, K. Pechman, B. J. Lewis, and D. Stahl. Conservation Of Primary Structure In 16s Ribosomal Rna. *Nature* **254**, 83-86 (1975).

Wu, G. D., J. Chen, C. Hoffmann, K. Bittinger, Y.-Y Chen, S. A. Keilbaugh, M. Bewtra, et al. Linking Long-Term Dietary Patterns With Gut Microbial Enterotypes. *Science* **334**, 105-108 (2011).

Yoshioka, H., K. Iseki, and K. Fujita. Development And Differences Of Intestinal Flora In The Neonatal Period In Breast-Fed And Bottle-Fed Infants. *Pediatrics* **72**, 317-321 (1983).

Zoetendal, E. G., Antoon D.L. Akkermans, and W. M. De Vos. Temperature Gradient Gel Electrophoresis Analysis Of 16s rRNA From Human Fecal Samples Reveals Stable And Host-Specific Communities Of Active Bacteria. *Appl. Environ. Microbiol.* **64**, 3854-3859 (1998).

## **Chapter 2**

**The impact of a consortium of fermented milk strains on the human gut microbiome: a study involving monozygotic twins and gnotobiotic mice**

This chapter has been published as:

Nathan P. McNulty, **Tanya Yatsunenko**, Ansel Hsiao, Jeremiah J. Faith, Brian D. Muegge, Andrew L. Goodman, Bernard Henrissat, Raish Oozeer, Stéphanie Cools-Portier, Guillaume Gobert, Christian Chervaux, Dan Knights, Catherine A. Lozupone, Rob Knight, Alexis E. Duncan, James R. Bain, Michael J. Muehlbauer, Christopher B. Newgard, Andrew C. Heath, and Jeffrey I. Gordon

“The Impact of a Consortium of Fermented Milk Strains on the Gut Microbiome of Gnotobiotic Mice and Monozygotic Twins”. *Science Translational Medicine* 26 October **2011** 3:106ra106.

## **Abstract**

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

## **Introduction**

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

A necessary starting point for assessing how the structure and functions of the human microbiome are related to our biology is to characterize the normal variations that occur in these communities, their gene pools, and their gene expression profiles both within and between individuals. This requires carefully designed studies where potentially confounding variables such as host genotype, diet and various environmental exposures can be controlled and systematically manipulated. Monozygotic (MZ) twins represent one way to constrain some of these variables, given that they have more similar genotypes and have experienced more similar dietary and other early environmental exposures than any other combination of individuals. A complementary approach is to use germ-free mice colonized at various points in their life with defined collections of microbes, with sequenced genomes, that represent major phylogenetic lineages encountered in the body habitats of human populations of interest. Gnotobiotic mice harboring ‘synthetic’ model human microbiomes, where all component organisms and microbial genes are known, can be reared under conditions where a number of the variables that confound human studies are ex-



tremely well controlled. Insights gleaned from these gnotobiotic animals can be applied back to humans (8).

Common intended or unintentional disturbances to our microbiomes include changes in our diets, consumption of antibiotics, and ingestion of live microbial strains posited to improve health. The latter include commercially available probiotics that are incorporated into fermented milk products (FMPs). With increasing regulatory pressure to validate the composition and health claims of probiotics and ‘functional’ foods, it is particularly important to develop informative translational medicine pipelines so that proof-of-concept clinical trials can be performed using validated biomarkers for quantitative phenotyping of subjects and of their responses. The present study demonstrates one type of approach. It uses adult MZ co-twins and metagenomic methods to first define temporal fluctuations in the organismal and gene content and gene expression profiles of their fecal microbial communities as a function of administration of a widely used commercial FMP. It then takes the five sequenced strains present in the FMP and introduces them as a consortium, at a dose analogous to that experienced by humans, into gnotobiotic mice containing a model human microbiome composed of 15 sequenced human gut symbionts. Quantitative analyses of temporal changes in the proportional representation of microbial species and genes, and of microbiome gene expression and metabolism before and after an ecological ‘invasion’ with the 5-member FMP microbial consortium, has provided insights into the ways that FMP strains and the indigenous model gut community respond to one another. The transcriptional responses were used as biomarkers to interrogate metatranscriptome datasets obtained from the MZ twins’ fecal specimens.

## **Results**

### **Human studies**

#### **Study design and assessment of intrapersonal and interpersonal variations in the fecal microbiota of monozygotic twin pairs over a four-month period**

Details concerning the seven adult MZ twin pairs recruited for this study are provided in **Table S1**. All had been vaginally delivered; none consumed antibiotics in the four months prior to and during their participation in the present study; none had a history of gastrointestinal diseases (including irritable bowel syndrome) or any other acute or chronic medical conditions; none were consuming dietary supplements or probiotics at the time of enrollment; and none had a history of gluten sensitivity or other food allergies, nor were any vegans or lacto-vegetarians.

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

Three fecal samples were obtained over the course of a 4-week period prior to initiation of FMP consumption ('pre-treatment phase'; see **Fig. 1A**). Each co-twin then consumed two servings of the FMP per day for 7 weeks (breakfast and dinner). Four fecal samples were collected at defined intervals during this treatment period, while two additional samples were collected during the 4 weeks following cessation of FMP consumption ('post-treatment phase'; **Fig. 1A**). Participants kept a daily log of their FMP consumption

and stool parameters including frequency. Statistical analyses of this log indicated that in this population, FMP consumption was associated with significantly softer stools but no significant changes in stool frequency (see Supplementary Material). However, based on existing regulatory criteria, our study of this small cohort was insufficiently powered to draw clinical conclusions about these stool parameters. Moreover, the MZ twin population recruited was comprised entirely of healthy individuals, so these data cannot be used to make statements about the impact of FMP consumption on stool softness in unhealthy patient populations.

All fecal samples collected during the three phases of this study were frozen at  $-20^{\circ}\text{C}$  within 30 min of their passage, and maintained at this temperature during overnight shipment to a biospecimen repository where they were subsequently stored at  $-80^{\circ}\text{C}$  prior to metagenomic analyses. To assess intra- and interpersonal variations in microbial community structure, we performed multiplex 454 FLX pyrosequencing of amplicons generated from variable region 2 (V2) of bacterial 16S rRNA genes present in fecal DNA. A total of 431,700 sequencing reads were obtained from 126 fecal samples ( $3,426 \pm 2,665$  sequences per sample, **Table S2A**). Noise due to PCR and pyrosequencing artifacts was removed from this dataset using software incorporated into the QIIME suite of 16S rRNA analysis tools (9). De-noised reads were binned into species-level phylogenetic types (phylotypes), with a species defined as isolates that share  $\geq 97\%$  identity in their 16S rRNA gene sequence. To ensure even coverage across samples, each of the 126 datasets was subsampled to 1,640 reads per fecal microbiota. A phylogenetic tree was built from one representative sequence from each phylotype using FastTree's approximately maximum-likelihood implementation (10) and communities were compared using unweighted UniFrac (11): the UniFrac metric measures community similarity based on the degree to which their members share branch length on a reference phylogenetic tree of Bacteria.

To quantify temporal variation in community composition within and between MZ twins, we generated a matrix of unweighted UniFrac distances for all pairwise compari-

sons of all 126 fecal samples obtained from the twins in our study. This matrix allowed us to compare any two fecal communities separated by all possible time intervals between sampling for each individual in each of the 7 twin pairs (**Fig. 2A**). The results indicated that no matter how far apart in time sample collection occurred (1 week to 4 months), the phylogenetic distance between communities from the same individual was less than the distance between communities between co-twins or unrelated individuals. UniFrac distances between samples harvested from a given individual increased modestly as a function of time during the 4-month period, although the changes were not statistically significant (**Fig. 2A**).

Each sample contained  $163 \pm 3$  (mean  $\pm$  SEM) observed species-level phylotypes. Four of the total 1,673 phylotypes identified in our dataset were found in all 126 samples; all belonged to the family Lachnospiraceae (order Clostridiales; phylum Firmicutes) and represented  $2.5 \pm 0.04\%$  of the 16S rRNA sequences in each sample.  $24.6 \pm 0.4\%$  of species-level phylotypes observed in a given sample were consistently represented in all 9 samples from that individual (**Fig. 2B**): the family-level taxa to which these species belong consist principally of Lachnospiraceae, Ruminococcaceae, and Veillonellaceae (phylum Firmicutes), the Bacteroidaceae and Rikenellaceae (phylum Bacteroidetes), and Coriobacteriaceae (phylum Actinobacteria).  $13.7 \pm 0.2\%$  of the observed phylotypes were represented in all samples from both co-twins (**Fig. 2B**).

### **Impact of FMP consumption on fecal bacterial community composition**

A qPCR assay disclosed that 1 week after initiation of FMP consumption, the level of representation of *B. animalis* subsp. *lactis* (CNCM I-2494) was  $10^7$  cell equivalents (CE)/g feces; this level was sustained in all 14 individuals throughout the ensuing 7 weeks of FMP consumption (i.e., there were no statistically significant differences between the 1, 2, 4 and 7 week time points as determined by Friedman test with post-hoc correction). The Spearman correlation test revealed no significant effect of human family membership on

the levels of *B. animalis* subsp. *lactis* during FMP consumption. Levels fell to below the limits of detection of the assay in all but 4 participants within 2 weeks of cessation of FMP consumption (**Fig. S1**); two of these individuals represented a twin pair, while the other two individuals were unrelated to each other.

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

### **Effects of FMP consumption on the functional gene repertoire of the fecal microbiome**

To determine the effects of FMP consumption on the representation of gene functions in the microbiome, we performed shotgun sequencing on 48 of the fecal DNA preparations generated from 4 of the MZ twin pairs (n=6 samples/individual; 2 fecal samples obtained before, 2 during, and 2 after cessation of FMP consumption; see **Fig. 1A**). Two of these twin pairs lived together, while two pairs lived 3 and 932 miles apart. A 634 Mb dataset was generated (60,863±28,775 sequences per sample, average length 238 nt; **Table S2B**). A BLASTX search against version 54 of the Kyoto Encyclopedia of Genes and Genomes (KEGG) GENES database (12-14) yielded a total of 2,205±26 (mean±SEM) KEGG Orthology identifiers (KOs) per microbiome sample: 64% of the KOs in a given sample (1,417±46) were consistently represented in all 6 samples from that individual; 55% were

consistently represented in all samples from both co-twins; 892 KOs (41% of the total KOs in a given sample) were identified in all 48 samples forming a core set of shared fecal microbiome-associated functions. **Fig. 2C** provides a visual representation of this conserved set of 892 KOs: 38% of the 892 belong to six KEGG categories — ‘membrane transport’, ‘carbohydrate metabolism’, ‘DNA replication and repair,’ ‘amino acid metabolism,’ ‘translation,’ and ‘metabolism of co-factors and vitamins.’

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

At this point in our analysis, the human studies indicated that exposure of a healthy individual’s resident gut microbiota to the FMP strains did not produce a detectable perturbation in fecal bacterial species composition, nor did it have a broad effect on the functional profile of fecal microbiome genes. To help guide further analysis of the human datasets, we turned to a simplified *in vivo* model of the human gut microbiota. We based our selection of model community members on several criteria. All members of this model community, or their close relatives, would be represented in the fecal microbiota of the MZ twins and other sampled human populations. They would encompass the three major bacterial phyla

present in this host habitat (Firmicutes, Bacteroidetes, Actinobacteria), and would have deep draft or finished genome sequences available. Gnotobiotic mice harboring such a model human microbiome would be used to characterize the impact of FMP strain introduction on the community's species and microbial gene abundances, as well as the microbiome's transcriptional profile, and to ascertain the impact of the model community on the abundance and gene expression profiles of the FMP strains whose genome sequences were also known. The knowledge gleaned would then be used to help guide further analysis of the human fecal microbiome datasets, including microbial RNA-Seq datasets generated from a subset of the human fecal samples.

### **Studies in gnotobiotic mice**

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

**Fig. 1B** presents the study protocol. Two groups of adult 6–8 week old germ-free C57Bl/6J male mice were colonized with a single gavage of the 15-member community ( $6 \times 10^6$  cfu/member, total of  $9 \times 10^7$  cfu). Each group ( $n=5$  animals) was maintained on a low fat, plant polysaccharide-rich diet. Fourteen and fifteen days after gavage with the 15-member community, both groups of mice were inoculated with a mixture of the five FMP strains. One group received a second pair of gavages of the five strains 7d and 8d later, and a third pair 21d and 22d after the first inoculation of the FMP consortium (multiple treatment group). Each gavage consisted of a community composed of  $2 \times 10^7$  cfu: 25% ( $5 \times 10^6$  cfu) *S. thermophilus*; 25% *B. animalis* subsp. *lactis*, and 25% *L. lactis* subsp. *cremoris*, with the remaining 25% split between the two *L. delbrueckii* subsp. *bulgaricus* strains

(12.5% each;  $2.5 \times 10^6$  cfu/strain). Dosing was based on the following considerations: (i) a daily dose of two cups of the FMP contains  $\sim 10^{10}$  cfu of *B. animalis* subsp. *lactis*; (ii) assuming  $\sim 10^{14}$  bacteria in the human gut, the ratio of the number of input *B. animalis* subsp. *lactis* cfu to the human gut symbiont population is approximately  $10^{-4}$ ; (iii) to maintain this ratio of  $10^{-4}$  in mice, and assuming  $10^{11}$ - $10^{12}$  organisms in the mouse gut, we administered a total of  $10^7$  *B. animalis* subsp. *lactis* cfu per gavage period (one period equals two gavages within 24h); (iv) the difference in cfu between the least and most abundant microbial species in the FMP product remains  $\leq 2$ -fold during manufacture and storage; therefore, each species in the gavage was represented at equivalent levels. By administering the strain consortium directly by gavage, rather than the corresponding commercial fermented milk product, we were able to more precisely control dosing and avoid unintended colonization of the gnotobiotic mice with other microbial species.

### **The repertoire of carbohydrate active enzymes (CAZymes) in members of the FMP consortium and model human gut microbial community**

The genomes of the five FMP strains in this study were sequenced, either completely (*B. animalis* subsp. *lactis*) or at a deep draft level (other four strains) for subsequent analyses of their representation in the model community after gavage of gnotobiotic mice and so we could define their *in vivo* patterns of gene expression (**Table S3**).

Analysis of the 48 CAZyme families (16) identified in the five FMP strains, and the 126 CAZyme families identified in the 15-member model human microbiota disclosed that 23 of the 24 CAZyme glycoside hydrolase (GH) families, 11 of the 12 glycosyltransferase (GT) families, 4 of the 4 carbohydrate esterase (CE) families, and 4 of the 8 carbohydrate binding modules (CBMs) represented in the former were also represented in the latter. The FMP consortium contains only six CAZyme families that were not represented in the model human gut community. Three of these are associated with *L. lactis* subsp. *cremoris*: of these, two are predicted to play roles in the binding and metabolism of chitin (**Fig.**



**S3; Table S4**), The other three are from *B. animalis* subsp. *lactis*: *BALAC2494\_01193* encodes a GT39 family mannose transferase involved in O-glycosylation of proteins; *BALAC2494\_01288* specifies a predicted beta-mannanase carrying a C-terminal CBM10 carbohydrate-binding module predicted to bind cellulose; *BALAC2494\_01971* gives rise to a protein with a CBM23 module predicted to bind mannan.

### **Introducing the FMP strain consortium produces minimal changes in the species representation of the 15-member model human gut microbiota**

We used COmmunity PROfiling by Sequencing (COPRO-Seq), a generally applicable method based on highly parallel DNA sequencing (17), to quantify the proportional representation of each component of the 15 member microbiota and of the FMP consortium in our gnotobiotic mice. Sequencing reads generated from fecal DNA samples collected before, during and after introduction of the FMP strains were analyzed as described in **Fig. S4A**. Briefly, “informative” tags (i.e., reads that can be mapped uniquely to a single genome) were first identified. Informative tags were then summed by species to generate digital “counts” of abundance. In cases where a read could not be assigned with certainty during COPRO-Seq analysis, it was ignored. To account for this fact, species-specific counts were normalized using their “informative genome size” (defined as the percent of all possible k-mers a genome can produce that are unique multiplied by the total genome length). Multiplex sequencing using the Illumina GA-IIx instrument yielded sufficient numbers of reads per sample so that an organism comprising  $\geq 0.003\%$  of the community could be detected: for a mouse colonized at  $10^{11-12}$  cfu/ml cecal contents or feces, this represents  $10^6$  cfu/ml.

COPRO-Seq produced several notable findings. *First*, community assembly prior to introduction of the FMP strains occurred in a highly reproducible manner, both within and between the two groups of animals (**Fig. S5A, Table S5A**). This reproducibility ensured that animals in both treatment groups harbored communities with structures compa-

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

The *B. thetaiotaomicron* component of the 15-member human gut microbiota was composed of a library of 34,544 randomly inserted transposon mutant strains covering 3,435 of the organism's 4,779 genes (72%). As noted in Supplementary Material and **Table S6**, by comparing the representation of mutants in fecal samples before and after introducing the FMP strains, we were able to determine that their presence did not affect the profile of *B. thetaiotaomicron*'s genetic determinants of fitness in the distal gut.

## **Microbial RNA-Seq analysis of the response of *B. animalis* subsp. *lactis* to the gut environment and members of the 15-member community to the FMP strain consortium**

Moving beyond COPRO-Seq based structural analysis, we performed microbial RNA-Seq analysis to determine the functional impact of exposing the established model human community to the FMP strains, and to ascertain which FMP consortium genes are most highly expressed in the intestines of these animals. *B. animalis* subsp. *lactis* attained sufficient abundance in gnotobiotic mice to allow profiling of its transcriptome at late time points (days 35, 36, and 42). When its *in vivo* patterns of gene expression were compared with those documented during mid-log and stationary-phase in MRS medium and in the commercial FMP (see Supplementary Material and **Table S7**), we noted that the *BALAC2494\_00604-BALAC2494\_00614* locus, encoding enzymes involved in the catabolism of xylo-oligosaccharides (XOS) (18), was strongly upregulated *in vivo* (average across the locus; 27-fold at the day 42 time point compared to mid-log phase in MRS monoculture; 128-fold compared to the FMP, **Table S7**). Xylose is the main building block of dietary hemicelluloses. Addition of this pentose sugar is also one of the first steps in O-glycosylation of host mucins. These results support previous observations suggesting XOSs may serve as potent ‘bifidogenic factors’, whose consumption increases the densities of Bifidobacteria in the gut (19, 20).

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

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

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

The genes that exhibited the highest fold-change in expression were heavily skewed towards the KEGG categories ‘carbohydrate metabolism’ and ‘membrane transport.’ The latter includes a number of ABC- and PTS-type carbohydrate transporters (**Table S9**). When these KEGG category-level responses were subsequently broken down into KEGG pathways (**Fig. 4B**), it was apparent that the most significant responses in the ‘carbohydrate metabolism’ category involved increases in ‘starch and sucrose metabolism’, ‘fructose and mannose metabolism’, and ‘pentose and glucuronate interconversions.’

Transcript data were subsequently binned by enzyme commission (EC) number. The levels of mRNAs encoding these ECs at each time point were compared using ShotgunFunctionalizerR, an R-based statistical and visualization tool originally designed to identify genes significantly enriched or depleted in environmental microbiomes (21, 22). Using this approach, we were able to determine that the ‘starch and sucrose metabolism’ pathway response to the FMP strains was driven by significant upregulation of genes en-

coding three enzymes involved in metabolism of dietary plant polysaccharides: (i) EC 3.2.1.65 (levanase), which cleaves 2,6-beta-D-fructofuranosidic linkages in 2,6-beta-D-fructans (levans); (ii) EC 3.1.1.11 (pectinesterase), which de-esterifies pectin to pectate and methanol; and (iii) EC 2.4.1.20 (cellobiose phosphorylase), which uses cellobiose formed from partial hydrolysis of cellulose as its substrate to generate alpha-D-glucose-1-phosphate and D-glucose. The genes encoding these ECs, which catalyze early steps in three entry points of the ‘starch and sucrose metabolism’ KEGG pathway, underwent significant increases in their expression within 1d after introduction of the FMP consortium (**Fig. 5A**). The levels of expression of these genes either increased further (levanase) or were sustained (the other two ECs) in both the single and multiple treatment groups through the remaining 4 weeks of the experiment (**Fig. 5A**). The levanase response showed remarkable species specificity: this gene is represented in 8 members of the 15-member community, yet the community’s transcriptional response is driven almost exclusively by the levanase in *Bacteroides vulgatus* (*BVU\_1663*; **Fig. 4C**). In contrast, the pectinesterase response was distributed across 6 members of the community (*B. caccae*, *B. ovatus*, *B. thetaiotaomicron*, *B. vulgatus*, *B. WH2*, *C. aerofaciens*), with changes in transcription largely due to pectinesterase genes found in *B. ovatus* (*BACOVA\_03576*, *BACOVA\_03581*, *BACOVA\_04902*), *B. thetaiotaomicron* (*BT\_4109*, *BT\_4110*), *B. vulgatus* (*BVU\_1116*), and *B. WH2* (*BACWH2\_3569*, *BACWH2\_3615*). Increases in the proportional abundance of cellobiose phosphorylase transcripts reflected the contributions of three community members: *B. uniformis*, *E. rectale*, and *R. obeum* (**Table S8**).

The KEGG ‘starch and sucrose metabolism’, ‘pentose and glucuronate interconversions’ and ‘pentose phosphate’ pathways process products generated by these three enzymes. **Fig. 5B** shows that many of the other components of these pathways that are upregulated in the 15-member community when the FMP strain consortium is introduced. ShotgunFunctionalizeR also identified significant increases in the expression of genes en-

coding five ECs that participate in the generation of propionate and succinate: the induction occurred within 1d after the FMP strains were introduced and involved acetate kinase (EC 2.7.2.1; catalyzes a bidirectional reaction between propanoyl phosphate and propionate), phosphate acetyltransferase (EC 2.3.1.8), methylmalonylCoA decarboxylase (EC 4.1.1.41), propionylCoA carboxylase (EC 6.4.1.3) and methylmalonylCoA mutase (EC 5.4.99.2, yields succinylCoA as its product) (**Fig. S6**). Only a single treatment with the FMP consortium was required to produce a sustained response involving the enzymes that can yield propionate (**Fig. S6**).

A breakdown of RNA-Seq reads by the community member genome to which they mapped revealed that the abundance of a species in the 15-member community did not necessarily correlate with its contribution to the community transcript pool. At the time point sampled immediately prior to invasion (d14), two of the most extreme outliers were *B. WH2* (comprised  $39.6 \pm 1.6\%$ ; mean $\pm$ SD) of the community but only contributed  $15.4 \pm 2\%$  of the raw reads to the total RNA-Seq read pool) and *R. obeum* ( $2.1 \pm 0.4\%$  of the community;  $18.2 \pm 4.4\%$  of the transcript pool) (**Fig. S7**). These observations indicate that community-level transcriptional responses can be driven by species representing small fractions of the microbiota.

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

## Identifying predictive features from the model community metatranscriptome data using a Random Forests classifier

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

### Metabolomic analyses

To evaluate the impact of invasion with the 5-member FMP consortium on microbial-host co-metabolism, we performed untargeted gas chromatography-mass spectrometry (GC/MS) on urine samples collected at multiple time points (days 0, 14 and 42) from members of the single- and multi-treatment groups (**Fig. 1B**). A metabolite profile was constructed for each urine sample (n=19) using the spectral abundances of all identifiable metabolites. A total of 198 metabolites met our reverse match score cutoff of 65% and were present in at least 50% of samples at one or more time points (for an explanation of the reverse match score, see (24) and **Table S12**). Comparing day 0 and 14 samples revealed 39 metabolites whose levels were significantly higher or lower following colonization with the defined 15- member community (see **Table S12A**). The changes included decreases in

the levels of oligosaccharides we would expect to be consumed by members of the microbiota [melibiose (87% decline); raffinose/maltotriose (98%); note that oligosaccharides are by their nature difficult to identify with certainty with the present, non-targeted GC/MS technique, and our annotations of these metabolites as melibiose, and raffinose/maltotriose are provisional]. The observed 3.4-fold increase in pyrogallol, a polyphenol, is consistent with the known ability of many gut microbes to cleave these molecules from polyphenols present in dietary plant material. A 4.4-fold increase in taurine following the initial colonization of mice was also noted, probably a result of microbial deconjugation of taurine from bile compounds.

**Table S12B** lists urinary metabolites that change significantly after introducing the five FMP strains (compare day 14 versus 42 in **Table S12B**). Fructose and xylose were not significantly affected by introduction of the defined 15-member community but increased significantly following introduction of the FMP strain consortium (2.3- and 2.9-fold, respectively; **Fig. 6A,B**). Increases in fructose may reflect an enhanced capacity of the community to liberate this monosaccharide from levan and other polyfructans via levanase-catalyzed reactions. Increases in xylose might be explained by the additional xylanase activity introduced by *B. animalis* subsp. *lactis* (**Fig. 3**), or alternatively by the induction of microbiome genes encoding xylan-degrading enzymes ( e.g., *BACOVA\_04387* and *BACOVA\_04390*, which were upregulated 5.2- and 11.0-fold, respectively, following introduction of the FMP strains, **Table S10**). Changes in other metabolites such as xanthosine (**Fig. 6C**), a purine metabolite, suggest that the metabolic consequences of FMP strain introduction extend beyond the processing of carbohydrates.

Collectively, our transcriptional and metabolite analyses indicated that introducing FMP strains that constitute a small fraction of a defined model human gut microbiota signals the microbiota to change its metabolic activities, including activities related to carbohydrate metabolism. With this information in hand, we returned to the human fecal samples to determine the extent to which observations made in our gnotobiotic mouse model were



applicable to humans.

### **Microbiome transcriptional responses to FMP strains that are shared by gnotobiotic mice and humans**

Microbial RNA-Seq analysis was performed on human fecal samples obtained 1 week prior to FMP consumption, 1 and 4 weeks into the consumption period, and 4 weeks following cessation (both co-twins from family 1; one co-twin from family 3; see **Table S1**). Using an analysis pipeline comparable to the one employed for the mouse data, we first aligned all RNA-Seq reads against a reference set of 127 human gut microbial genomes plus the FMP strain genomes, binned the aligned transcripts based on their EC annotations, and used ShotgunFunctionalizeR to identify ECs whose abundances were significantly changed as a function of FMP exposure (Benjamini-Hochberg adjusted p-value <0.01).

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

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

(fell to below the limits of detection) within four weeks after FMP administration was stopped (**Fig. 5A**).

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

Human fecal transcripts were detected that mapped to the *B. animalis* subsp. *lactis* genome (see Supplementary Material). The presence of these transcripts was limited to the period of FMP consumption, supporting the notion that they emanated from the FMP strain rather than from a related species present within the microbiota (**Fig. S9**). This clear linkage to FMP consumption was not evident in the case of other members of the consortium, so we could not confidently analyze their patterns of gene expression *in vivo*. The highest number of mapped reads to the *B. animalis* subsp. *lactis* genome was obtained 1 week after FMP administration began: among the 4,000 reads, we were able to detect transcripts from all but 1 of the 10 genes in the *BALAC2494\_00604-BALAC2494\_00614* locus that encodes enzymes involved in the catabolism of xylo-oligosaccharides, leading us to conclude that this locus is highly expressed in the distal human gut, just as it is in our mouse model.

## **Discussion**

Repeated sampling of seven healthy MZ adult twin pairs over a 4-month period emphasized that intrapersonal variation in bacterial community structure was less than interpersonal variation, with co-twins having significantly more similar phylogenetic and taxonomic structure in their fecal microbiota compared to those from unrelated individuals (9, 25, 26). The results also showed that (i) consumption of a fermented milk product containing 5 bacterial strains was not associated with a statistically significant change in the proportional representation of resident community members within and between individuals; (ii) the appearance and disappearance of strains comprising the FMP consortium did not exhibit familial patterns in the fecal microbiota; and (iii) *B. animalis* subsp. *lactis* CNCM I-2494 was the most prominent assayed member of the consortium represented in the microbiota during the 7-week period of FMP consumption. Analyses of the fecal gene repertoire over the course of the 16 weeks of the experiment indicated that (i) variations in the functional features of the (fecal) microbiome were less than the variations in bacterial species composition; (ii) there was no significant difference in the degree of similarity in representation of KEGG orthology group functions for a given co-twin at each time point compared to the degree of similarity that existed between co-twins, while individual and twin pair microbiomes were significantly more similar to one another than those from unrelated individuals; and (iii) there were no statistically significant changes in the representation of these functions when the FMP strain consortium was being consumed. With these findings in mind, and with each individual as well as each genetically identical co-twin serving as a control, we concluded that at least at the depth and frequency of sampling employed for this small healthy cohort, the bacterial species and gene *content* of their fecal microbiota/microbiome was not an informative biomarker for understanding whether or how this commercial fermented milk product impacted microbial community properties.

Gnotobiotic mice harboring a model 15-member gut microbial community that represented the three principal bacterial phyla present in the human gut microbiota, and whose

58,399 known or predicted protein-coding genes encompassed many of the prominent functions present in the normal adult human fecal microbiome, provided a means for characterizing the impact of the 5-member FMP strain consortium on expressed gut microbial community functions, and then applying the results to the human fecal specimens collected for this study. As with the MZ twins, introduction of the 5-member strain consortium did not significantly affect the representation of the 15 species comprising the model human microbiota. As with the MZ twins, *B. animalis* subsp. *lactis* exhibited the greatest fitness of the five FMP strains in the gut, as judged by its prominence and persistence. Unlike the human arm of the study, where all subjects consumed the FMP twice daily, the design of the mouse study, with its single versus multiple treatment regimens, allowed us to directly compare the persistence of FMP consortium members. Only *B. animalis* subsp. *lactis* and *L. lactis* subsp. *cremoris* were able to maintain a foothold in the gut ecosystem at detectable levels for the entire 4 week monitoring period after a single dose. In addition, colonization levels were not affected by the number of times the FMP strains were administered to mice.

An advantage of constructing the model human gut microbiome was that its entire predicted gene repertoire was known. This allowed us to define the impact of introducing the FMP strain consortium on the functions expressed by the overall community as well as by its individual components. A major theme emanating from our analysis was the effect of introducing the FMP consortium on carbohydrate metabolism by the community, as well as the effect of the community on a feature of carbohydrate metabolism by *B. animalis* subsp. *lactis*. The model 15-member community responded to the FMP consortium by inducing genes encoding enzymes involved in catalyzing reactions that represent the three entry points into the KEGG 'starch and sucrose metabolic pathway', as well as enzymes that catalyze fermentation of carbohydrates to propionate. The mechanism by which the FMP strains elicit this response is unclear at present, but the effect is rapid (occurring within the first 24h after invasion) and was persistent whether the consortium was introduced in a single set of gavages during a 1-day period, or with subsequent repeated gavage over a

several week period. The persistence of both the carbohydrate pathway response, and of *B. animalis* subsp. *lactis*, suggests but does not prove that the latter may be instrumental in instigating and maintaining the former.

Intriguingly, the carbohydrate response showed features of ‘differentiation.’ As noted in *Results*, the levanase response was driven almost entirely by changes in transcription in just a single species (*B. vulgatus*), the pectinesterase response by 6 community members (*B. caccae*, *B. ovatus*, *B. thetaiotaomicron*, *B. vulgatus*, *B. WH2*, *C. aerofaciens*) and the cellobiose phosphorylase response by three components of the defined model human gut microbiota (*B. uniformis*, *E. rectale*, and *R. obeum*). Of the 50 genes with predicted xylan-degrading capacity in the model microbiome (i.e., those encoding enzymes in ECs 3.2.1.37 and 3.2.1.8), only *BACOVA\_04387* and *BACOVA\_04390* (both from *B. ovatus*) were significantly upregulated after FMP strain introduction (this is ignoring xylanase genes encoded by FMP strains like *B. animalis* subsp. *lactis*). This upregulation in a limited subset of the model community coincides with an increase in urinary xylose.

The ability to attribute EC-level changes to individual genes in specific bacterial species was not possible with our RNA-Seq analysis of the human fecal samples. The differentiation of carbohydrate responses among bacterial species documented in gnotobiotic mice emphasizes a challenge and opportunity that can be addressed in these models: namely, to further delineate the niches, interactions and adaptive resource switching behaviors of community members by intentional addition, removal or substitution of taxa, and/or by their modification through genetic manipulation. Although requiring significantly more animals and loss of the ability to use an animal as its own control, future studies could be expanded to include sampling of community gene expression in different segments of the small intestine.

The increased expression of genes encoding enzymes involved in the interconversion of propionate and succinate is intriguing given the fact that this short chain fatty acid

has been linked in some reports to effects on gastrointestinal transit time. However, work in this area has yielded varying results and conclusions, perhaps because of the diversity of models and methodologic approaches used (27-30). Propionate may also link the gut microbiota and human physiology through its effects on hepatic and adipose tissue metabolism (31). Notably, another group has reported that in the *T-bet<sup>-/-</sup>Rag2<sup>-/-</sup>* mouse model of colitis, consumption of a fermented milk product containing a dairy matrix plus the same strains used in this study led to increased cecal propionate levels and a reduction in intestinal inflammation (32).

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

## **Materials and Methods**

### **Human studies**

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

*Other procedures* — Methods used for the production and distribution of the FMP to study participants, analysis of the effects of FMP consumption on stool parameters, qPCR analysis of fecal levels of FMP strains, multiplex pyrosequencing of 16S rRNA genes in fecal samples and the FMP, co-occurrence analysis, and shotgun sequencing of human fecal microbiomes are described in the *Methods* section of Supplementary Material.

### **Studies in gnotobiotic mice**

*Colonization of germ-free mice* — The justification for using mice and the protocols employed for treating them were approved by the Washington University Animal Studies Committee. Animals belonging to the C57Bl/6J inbred strain were maintained in plastic flexible film gnotobiotic isolators, and fed a standard autoclaved chow diet (B&K rat and mouse autoclavable chow #7378000, Zeigler Bros, Inc) *ad libitum*. Two groups of 6-8 week-old germ-free male animals (n=5/group) were colonized with a single gavage of 500 µl of supplemented TYG medium (TYG<sub>s</sub>; (34)) containing 15 sequenced human gut-derived bacterial symbionts (6x10<sup>6</sup> cfu/strain; total of 9x10<sup>7</sup> cfu for the community). The *B. thetaiotaomicron* component of this community was composed of a library of 34,544 transposon mutants prepared as described (34). Fourteen and fifteen days later, both groups of mice were gavaged with a mixture of the five FMP strains (each species at 5 x 10<sup>6</sup> cfu) in

300 ul of TYG. One group of mice received a second pair of gavages 7d and 8d later, and a third pair of gavages 21d and 22d after the initial FMP strain introduction.

***Other procedures*** — Methods used for sampling animals, COPRO-Seq, INSeq, microbial RNA-Seq and non-targeted metabolomics via gas chromatography/mass spectrometry (GC/MS) are described in the *Methods* section of Supplementary Material, as are methods for sequencing and annotating FMP strain genomes.

### **Acknowledgements**

We thank Jill Manchester, Jessica Hoisington-López for assistance with DNA sequencing, Maria Karlsson, David O'Donnell and Sabrina Wagoner for help with gnotobiotic mouse husbandry, Su Deng for assistance in preparing Illumina DNA libraries, Stacy Marion and Deborah Hooper for their contributions to the human study, Deanna Carlsen for coordination of FMP production and logistics, Stephan Baumann and Steven Fischer (Agilent Corp) for kindly providing the Fiehn GC/MS Metabolomics RTL library used for metabolomics analyses, members of the Gordon lab for valuable suggestions during the course of this work, and Gerard Denariáz for his continued support.

We are also grateful to Integrated Genomics for generating the draft genome sequences of *B. animalis* subsp. *lactis* (CNCM I-2494) and *S. thermophilus* (CNCM I-1630).

Funding: Supported by grants from the NIH (DK30292, DK70977) and Danone Research. Maintenance of the MOAFTS twin cohort is supported by NIH grants AA09022, AA11998, AA17915 and HD49024.

Author contributions: N.P.M., T.Y. A.C.H., and J.I.G. designed experiments; N.P.M., T.Y., A.H., J.J.F., B.D.M., A.L.G., R.O., S.C-P, G.G., J.R.B., and M.J.M. performed experiments; N.P.M., T.Y., A.H. B.D.M., A.L.G., B.H., C.C., D.K., C.A.L. R.K., A.E.D, and C.B.N. analyzed the data; N.P.M., A.H., T.Y., and J.I.G. wrote the paper

Competing interests: none declared



Accession Numbers: The genome sequences of *Bifidobacterium animalis* subsp. *lactis* (CNCM I-2494), *Lactobacillus delbrueckii* subsp. *bulgaricus* (CNCM I-1632, CNCM I-1519), *Streptococcus thermophilus* CNCM I-1630, *Lactococcus lactis* subsp. *cremoris* CNCM I-1631 are deposited in GenBank (accession numbers CP002915.1, X, X,X, and X, respectively), COPRO-Seq data in GEO (accession number GSE31943), RNA-Seq data in GEO (accession number GSE31670), 16S rRNA pyrosequencing reads in MG-RAST (accession number qiime:803), and shotgun pyrosequencing reads of human fecal community DNA in MG-RAST (accession number 4473933 to 4473980).

## References

1. J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, M. Jian, Y. Zhou, Y. Li, X. Zhang, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S. D. Ehrlich, A human gut microbial gene catalogue established by metagenomic sequencing, *Nature* **464**, 59 (2010).
2. C. F. Favier, E. E. Vaughan, W. M. De Vos, A. D. Akkermans, Molecular monitoring of succession of bacterial communities in human neonates, *Appl Environ Microbiol* **68**, 219 (2002).
3. K. Kurokawa, T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, H. Morita, V. K. Sharma, T. P. Srivastava, T. D. Taylor, H. Noguchi, H. Mori, Y. Ogura, D. S. Ehrlich, K. Itoh, T. Takagi, Y. Sakaki, T. Hayashi, M. Hattori, Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes, *DNA Res* **14**, 169 (2007).
4. M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J. M. Batto, M. Bertalan, N. Borrueil, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Dore, M. Antolin, F. Artiguenave, H. M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariáz, R. Dervyn, K. U. Foerstner, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-

- Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Merieux, R. Melo Minardi, C. M'Rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S. D. Ehrlich, P. Bork, Enterotypes of the human gut microbiome, *Nature* **473**, 174 (2011).
5. C. Palmer, E. M. Bik, D. B. DiGiulio, D. A. Relman, P. O. Brown, Development of the human infant intestinal microbiota, *PLoS Biol* **5**, e177 (2007).
  6. J. E. Koenig, A. Spor, N. Scalfone, A. D. Fricker, J. Stombaugh, R. Knight, L. T. Angenent, R. E. Ley, Succession of microbial consortia in the developing infant gut microbiome, *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4578 (2011).
  7. M. G. Dominguez-Bello, E. K. Costello, M. Contreras, M. Magris, G. Hidalgo, N. Fierer, R. Knight, Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns, *Proc Natl Acad Sci U S A* **107**, 11971 (2010).
  8. J. J. Faith, F. E. Rey, D. O'Donnell, M. Karlsson, N. P. McNulty, G. Kallstrom, A. L. Goodman, J. I. Gordon, Creating and characterizing communities of human gut microbes in gnotobiotic mice, *ISME J* **4**, 1094 (2010).
  9. J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, R. Knight, QIIME allows analysis of high-throughput community sequencing data, *Nat Methods* **7**, 335 (2010).
  10. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree: computing large minimum evolution trees with profiles instead of a distance matrix, *Mol Biol Evol* **26**, 1641 (2009).

11. C. Lozupone, M. Hamady, R. Knight, UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context, *BMC Bioinformatics* **7**, 371 (2006).
12. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res* **28**, 27 (2000).
13. M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa, From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Res* **34**, D354 (2006).
14. M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa, KEGG for representation and analysis of molecular networks involving diseases and drugs, *Nucleic Acids Res* **38**, D355 (2010).
15. P. J. Turnbaugh, C. Quince, J. J. Faith, A. C. McHardy, T. Yatsunencko, F. Niazi, J. Affourtit, M. Egholm, B. Henrissat, R. Knight, J. I. Gordon, Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins, *Proc Natl Acad Sci U S A* **107**, 7503 (2010).
16. B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, B. Henrissat, The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics, *Nucleic Acids Res* **37**, D233 (2009).
17. J. J. Faith, N. P. McNulty, F. E. Rey, J. I. Gordon, Predicting a human gut microbiota's response to diet in gnotobiotic mice, *Science* **333**, 101 (2011).
18. O. Gilad, S. Jacobsen, B. Stuer-Lauridsen, M. B. Pedersen, C. Garrigues, B. Svensson, Combined transcriptome and proteome analysis of *Bifidobacterium animalis* subsp. *lactis* BB-12 grown on xylo-oligosaccharides and a model of their utilization, *Appl Environ Microbiol* **76**, 7285 (2010).

19. C. K. Hsu, J. W. Liao, Y. C. Chung, C. P. Hsieh, Y. C. Chan, Xylooligosaccharides and fructooligosaccharides affect the intestinal microbiota and precancerous colonic lesion development in rats, *J Nutr* **134**, 1523 (2004).
20. M. Okazaki, S. Fujikawa, N. Matsumoto, Effect of xylooligosaccharide on the growth of bifidobacteria, *Bifidobacteria and Microflora* **9**, 77 (1990).
21. E. Kristiansson, P. Hugenholtz, D. Dalevi, ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes, *Bioinformatics* **25**, 2737 (2009).
22. R. D. C. Team, *R: A Language and Environment for Statistical Computing.*, (R Foundation for Statistical Computing, Vienna, Austria., 2009).
23. D. Knights, E. K. Costello, R. Knight, Supervised classification of human microbiota, *FEMS Microbiol Rev* **35**, 343 (2011).
24. S. E. Stein, An Integrated Method for Spectrum Extraction and Compound Identification from Gas Chromatography/Mass Spectrometry Data, *J. Am. Soc. Mass Spectrom.* **10**, 770 (1999).
25. R. E. Ley, P. J. Turnbaugh, S. Klein, J. I. Gordon, Microbial ecology: human gut microbes associated with obesity, *Nature* **444**, 1022 (2006).
26. P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight, J. I. Gordon, A core gut microbiome in obese and lean twins, *Nature* **457**, 480 (2009).
27. C. Cherbut, L. Ferrier, C. Roze, Y. Anini, H. Blottiere, G. Lecannu, J. P. Galmiche, Short-chain fatty acids modify colonic motility through nerves and polypeptide YY release in the rat, *Am J Physiol* **275**, G1415 (1998).
28. P. S. Kamath, M. T. Hoepfner, S. F. Phillips, Short-chain fatty acids stimulate motility of the canine ileum, *Am J Physiol* **253**, G427 (1987).

29. P. Marteau, E. Cuillerier, S. Meance, M. F. Gerhardt, A. Myara, M. Bouvier, C. Bouley, F. Tondu, G. Bommelaer, J. C. Grimaud, Bifidobacterium animalis strain DN-173 010 shortens the colonic transit time in healthy women: a double-blind, randomized, controlled study, *Aliment Pharmacol Ther* **16**, 587 (2002).
30. P. E. Squires, R. D. Rumsey, C. A. Edwards, N. W. Read, Effect of short-chain fatty acids on contractile activity and fluid flow in rat colon in vitro, *Am J Physiol* **262**, G813 (1992).
31. S. H. Al-Lahham, M. P. Peppelenbosch, H. Roelofsen, R. J. Vonk, K. Venema, Biological effects of propionic acid in humans; metabolism, potential applications and underlying mechanisms, *Biochim Biophys Acta* **1801**, 1175 (2010).
32. P. Veiga, C. A. Gallini, C. Beal, M. Michaud, M. L. Delaney, A. DuBois, A. Khlebnikov, J. E. van Hylekama Vlieg, S. Punit, J. N. Glickman, A. Onderdonk, L. H. Glimcher, W. S. Garrett, Bifidobacterium animalis subsp. lactis fermented milk product reduces inflammation by altering a niche for colitogenic microbes, *Proc Natl Acad Sci U S A* **107**, 18132 (2010).
33. A. C. Heath, W. Howells, K. K. Bucholz, A. L. Glowinski, E. C. Nelson, P. A. Madden, Ascertainment of a mid-western US female adolescent twin cohort for alcohol studies: assessment of sample representativeness using birth record data, *Twin Res* **5**, 107 (2002).
34. A. L. Goodman, N. P. McNulty, Y. Zhao, D. Leip, R. D. Mitra, C. A. Lozupone, R. Knight, J. I. Gordon, Identifying genetic determinants needed to establish a human gut symbiont in its habitat, *Cell Host Microbe* **6**, 279 (2009).

## **Figure Legends**

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

**Figure 2. Metagenomic studies of human fecal microbiomes sampled over time.** (A) 16S rRNA-based time course study of intra- and interpersonal variations in fecal bacterial community structure during the course of the four-month study. Unweighted UniFrac measurements of community distances, from pairwise comparisons of all samples obtained from a given individual, from co-twins, and from unrelated individuals are plotted as mean values  $\pm$ SEM. (B) Colored boxes represent the proportion of bacterial phylotypes that were consistently present within an individual over time (gray), between co-twins over time (orange), and in all 126 fecal samples (red). The white box represents the average number of species-level phylotypes found in a given sample. All measures of spread provided in parentheses represent  $\pm$ SEM. (C) KEGG Orthology groups (KOs) consistently present within the fecal microbiome of an individual over time (gray), between co-twins over time (orange), and in all 48 microbiomes analyzed from the four sets of MZ twins during

the four-month study (red). The white box indicates the average number of unique KOs ( $\pm$ SEM) identified in a particular sample. All measures of spread provided in parentheses represent  $\pm$ SEM. **(D)** Hellinger distance measurements of fecal microbiomes based on their KO content. Tests of statistical significance are based on 1000 permutations of a Hellinger distance matrix. Mean values ( $\pm$ SEM) are shown for the three types of comparisons (self-self; co-twin-co-twin; unrelated-unrelated individual).

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

**Figure 4. ‘Top-down’ analysis of the effects of the FMP strain consortium on the model 15-member community’s metatranscriptome.** RNA-Seq reads were mapped to the sequenced genomes of the 15 community members. Transcript counts were normalized [reads per kb of gene length per million reads (RPKM), see Supplementary Material] and



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

**Figure 5. Mouse and human communities share transcriptional responses to the FMP strain consortium involving ECs related to carbohydrate metabolism.** (A) Box plots of the proportion of all RPKM-normalized reads in mouse and human fecal metatranscriptomes represented by three ECs involved in plant biomass degradation. Individual samples are shown as black dots ( $n=2-10$ ). Boxes are also colored by fold-change, as determined by comparing mean values at a given time point to the value at the pre-treatment time point [for gnotobiotic mice pre-treatment refers to day 14; in the case of humans, pre-treatment refers to the fecal sample collected 1 week prior to initiation of FMP consumption (sample ‘Pre1’ in **Fig. 1A**)]. Statistical significance was determined using the ShotgunFunctionalizeR package in R and an adjusted *p*-value cutoff of  $< 0.01$ . Pre-treatment time points, and subsequent time points where expression levels were not significantly different from the

pre-treatment mean are colored white. **(B)** Components of KEGG ‘starch and sucrose metabolism’, ‘pentose and glucuronate interconversions’ and ‘pentose phosphate’ pathways whose expression in the 15-member model community changed compared to pre-treatment values when the 5-member FMP strain consortium was introduced. Gray indicates that the fold-change was statistically significant (adjusted  $p$ -value  $<0.01$ ). Ovals highlight the three enzymes shown in panel A. Dashed arrows indicate that multiple enzymatic reactions lead from these ECs and their indicated substrates to the products shown. These intermediate reactions have been omitted for clarity or because the omitted ECs did not manifest significant changes in their expression.

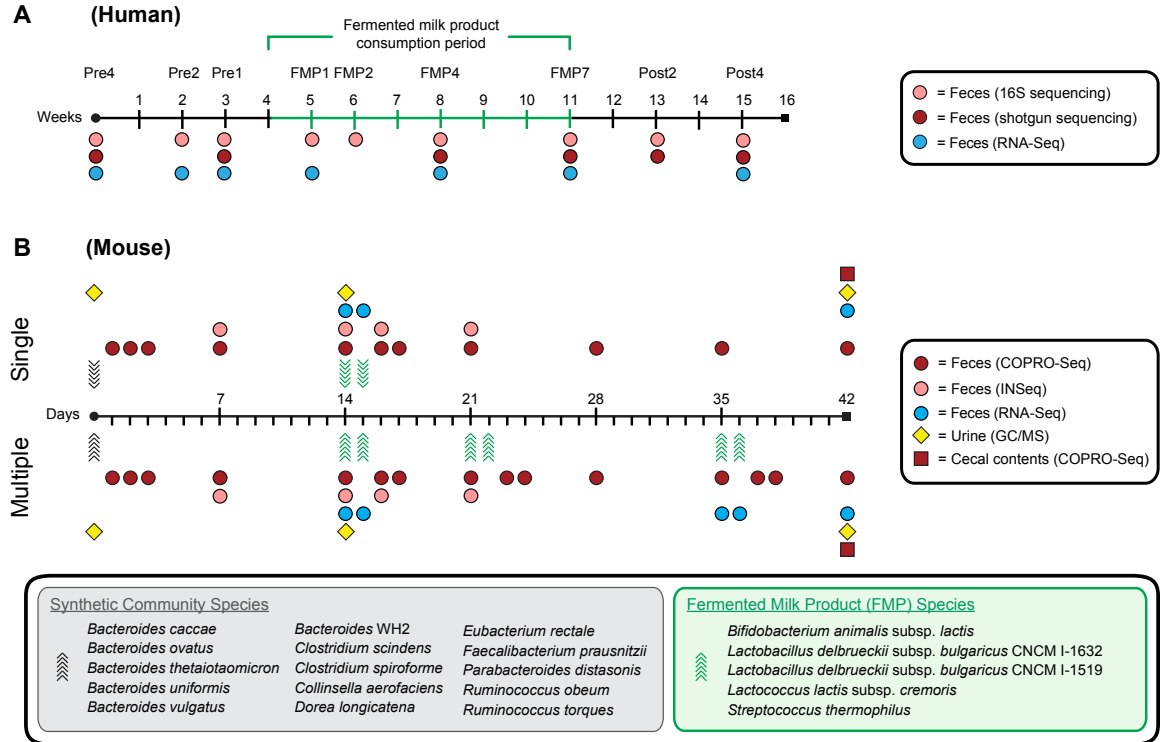
**Figure 6. Select urinary metabolites whose levels are significantly altered following the introduction of the FMP strain consortium into mice harboring a defined model human gut microbiota.** The statistical significance in pairwise comparisons shown in panels A-C was evaluated using a two-tailed Student’s  $t$ -test on the log-transformed spectral abundance of the metabolite in each sample. Values for the statistical significance of differences between time points as evaluated by one-way ANOVA, followed by FDR-correction and a post-hoc Tukey HSD test are also provided in **Table S12**. Horizontal bars represent group means, vertical bars represent  $\pm$  SEM. Abbreviations: \*,  $p < 0.05$ ; \*\*,  $p < 0.01$ ; n/s, not significant.

**Figure 7. Shared transcriptional responses to FMP strain exposure in mice and humans.** The heatmap shows ECs that exhibit a statistically significant change in their expression (ShotgunFunctionalizeR, adjusted  $p < 0.01$ ) and manifest a consistent direction of change in their expression in all four comparisons shown. Comparisons include those where the pre-treatment timepoint was compared with a timepoint shortly after FMP strains were introduced (mouse: ‘d15 vs d14’, human: ‘FMP1 vs Pre1’) and those where the pre-treatment period was compared to a timepoint several weeks after strain introduction (mouse: ‘d42multi vs d14’, human: ‘FMP4 vs Pre1’). ‘d42multi’ indicates the multiple-treatment group at day 42 of the mouse experiment. The colored boxes correspond to the KEGG cat-

egories that contain the ECs shown to the right of the heatmap. The scale refers to fold-difference in the mean of relative abundance of each EC between treatment and pre-treatment groups based on the mean number of normalized reads (RPKM) of transcripts assigned to a given EC. The 18 ECs shown at the bottom of the Figure are not associated with the five prominent KEGG categories listed. Their assigned categories are provided in **Table S13**.

# Figures

## Figure 1.



**Figure 2.**

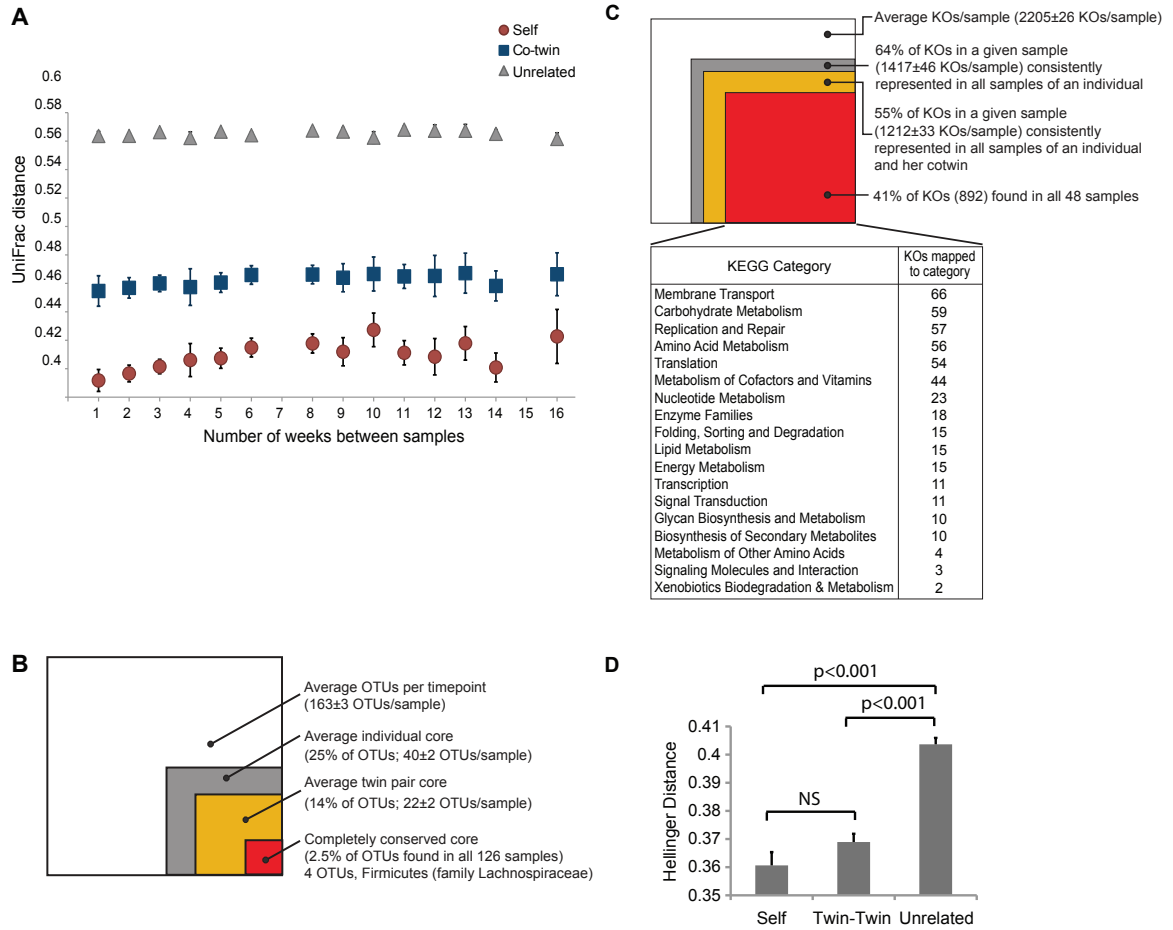
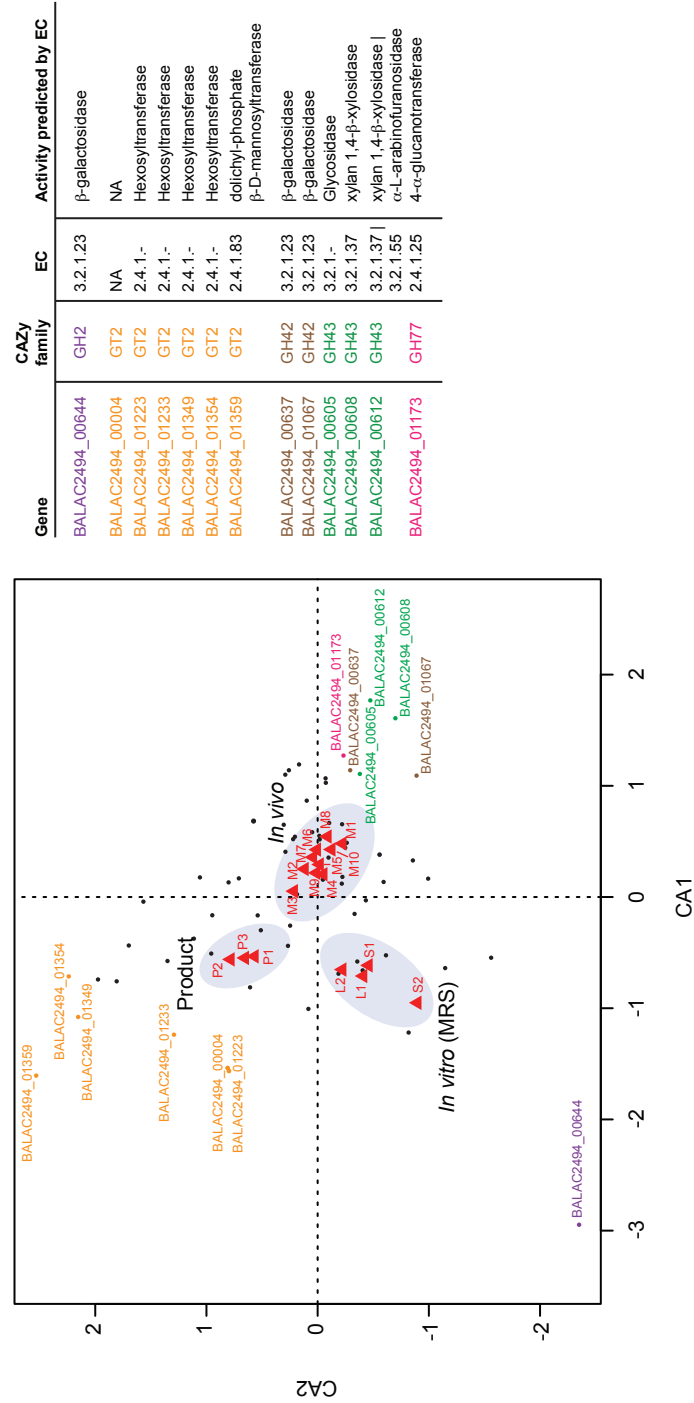
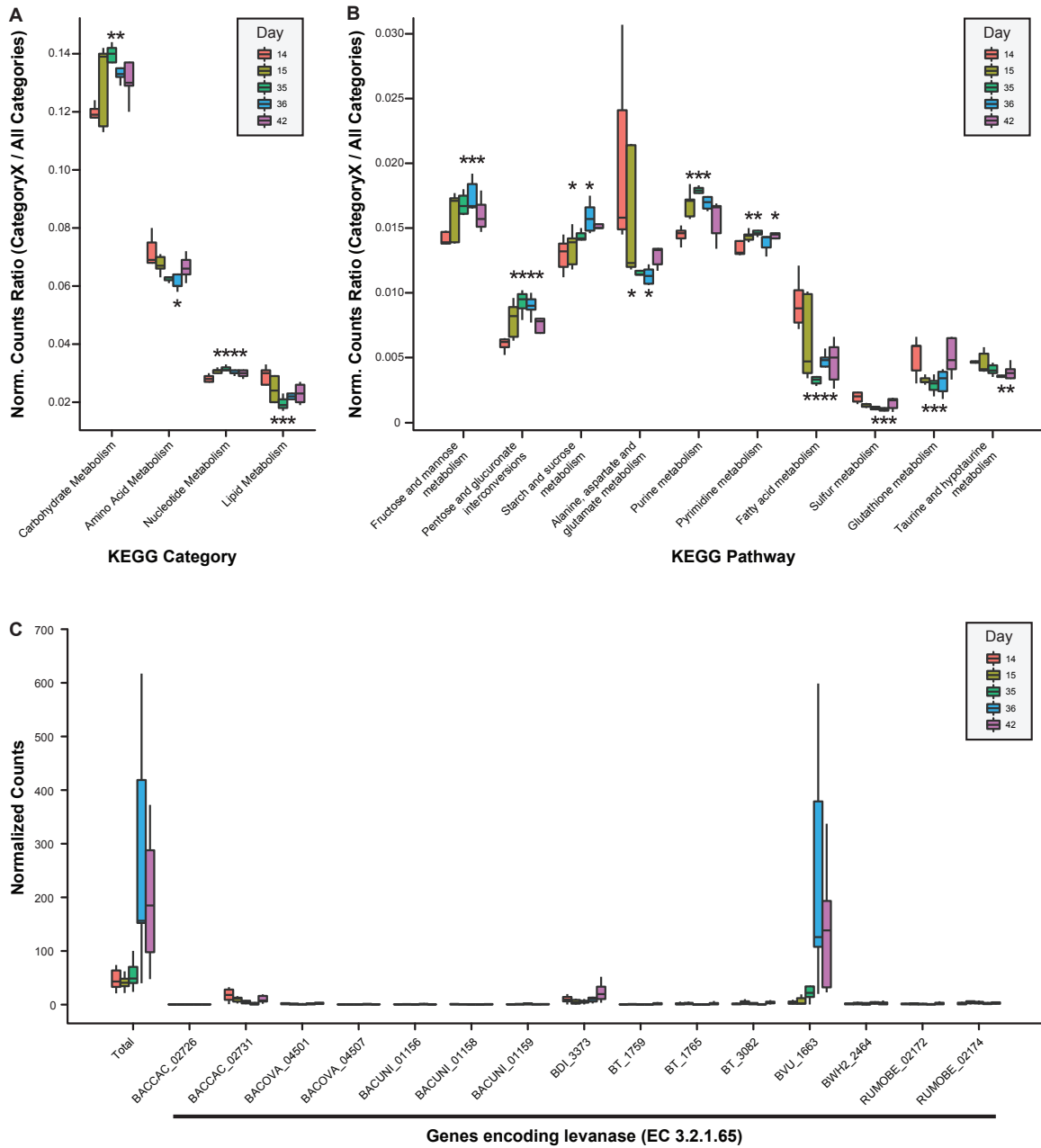


Figure 3.

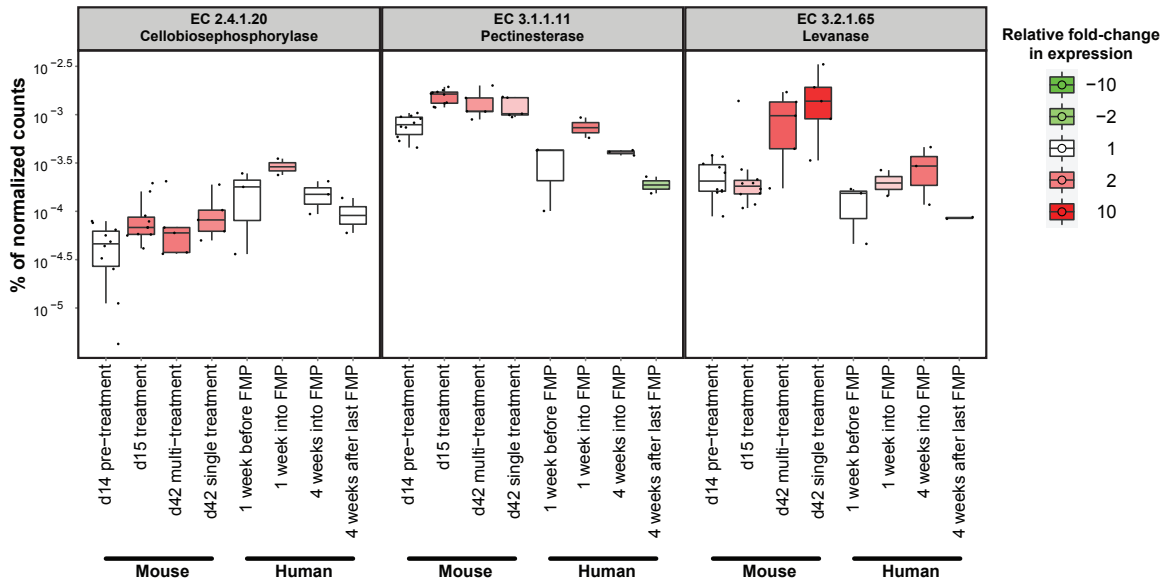


**Figure 4.**

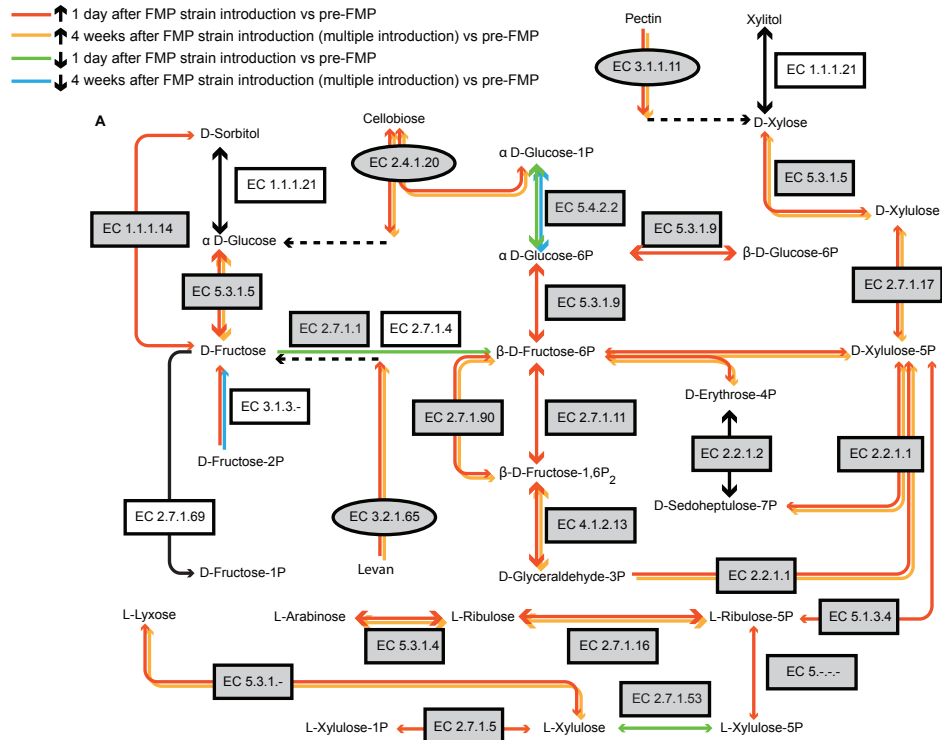


**Figure 5.**

**A**



**B**





**Figure 6.**

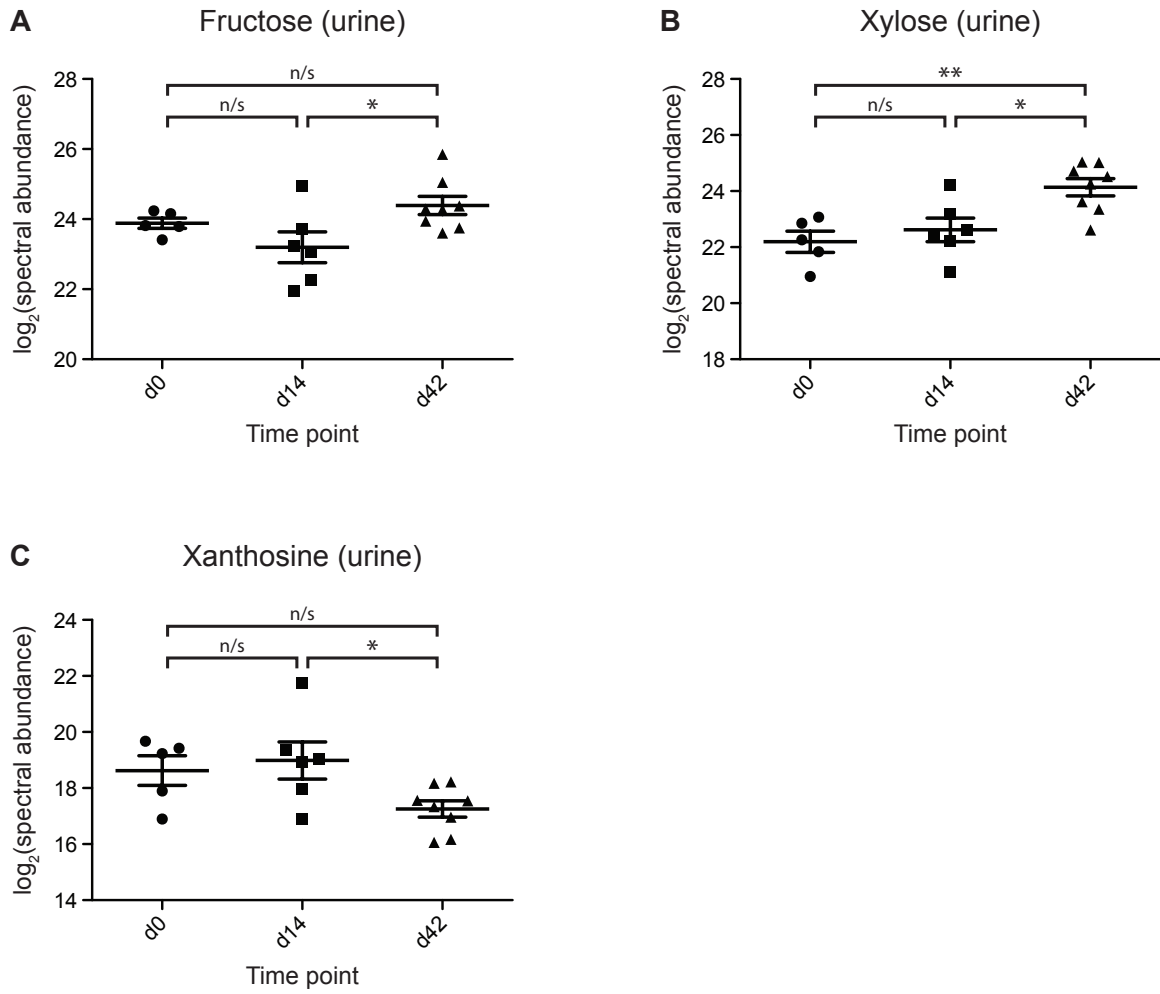
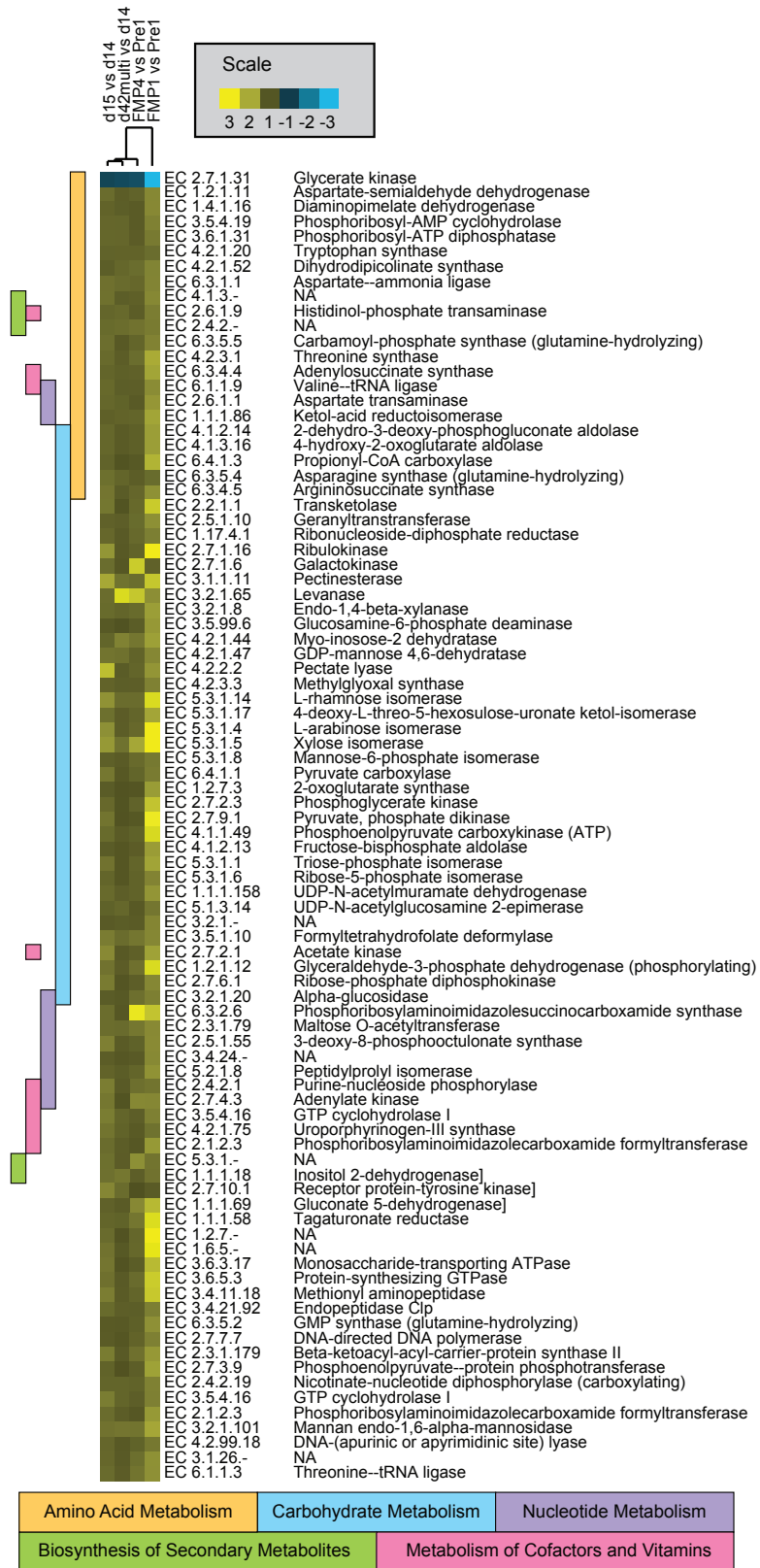


Figure 7.



## **Supplementary Material**

### **Supplementary Materials and Methods**

#### **Microbial genome sequencing**

The following bacterial strains are incorporated into the commercially available FMP: *Bifidobacterium animalis* subsp. *lactis* (strain CNCM I-2494); *Lactobacillus delbrueckii* subsp. *bulgaricus* (strains CNCM I-1632, CNCM I-1519), *Lactococcus lactis* subsp. *cremoris* (strain CNCM I-1631), and *Streptococcus thermophilus* (strain CNCM I-1630). We performed shotgun 454 FLX pyrosequencing of both *L. delbrueckii* subsp. *bulgaricus* strains, plus the *L. lactis* subsp. *cremoris* strain (39-, 41- and 51-fold coverage, respectively). Using the Newbler assembler (454 Life Sciences) and already sequenced strains of these species, we obtained draft genome assemblies with N50 contig sizes of 66,436 and 55,626 and 55,851 bp, respectively. The total sizes of the assembled *L. delbrueckii* genomes were 1,780,478 bp (CNCM I-1632) and 1,808,929 bp (CNCM I-1519), while the *L. lactis* assembly had an aggregate genome size of 2,511,332 bp. A finished genome sequence of the *B. animalis* subsp. *lactis* genome and a deep draft assembly of the *S. thermophilus* genome were previously generated by Integrated Genomics (see **Table S3** for a summary of genome metrics).

#### **Annotation and comparative genomic analysis**

The genomes of all sequenced bacterial species used in this study were annotated by BLAST searches (E-value threshold cutoff  $<10^{-5}$ ) against version 54 (v54) of the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (*S1-3*) and the Carbohydrate Active Enzyme (CAZy) database (*S4*). BLAST results were parsed into a lookup table for each genome, and each fecal microbiome, using a perl script (lookup\_KEGG\_for\_genes\_BLAST.pl) that assigns a KEGG orthology (KO) number of the top BLAST hit to each gene (minimum threshold BLAST e-value =  $10^{-5}$ ). In cases where multiple annotated da-

tabase entries shared the same lowest e-value, the gene was annotated with the KOs from each of the entries.

These lookup tables were then used to calculate the ‘coverage ratio’ of each KEGG metabolic pathway in each of the FMP consortium’s constituent bacterial genomes using the perl script `kegg_key.pl`. For each K number node of each KEGG metabolic pathway, this script searches for a gene that has been assigned that K number by our annotation pipeline in (i) a given microbial genome, (ii) defined collections of microbial genomes, or (iii) larger, incompletely sequenced microbiomes. For genes with multiple pathway assignments, the corresponding K number was placed in each of the pathways without weighting. The file of the percentage of all K number nodes present for each KEGG pathway represented in each genome or microbiome was then employed for average linkage hierarchical clustering in Cluster 3.0 (S5) using an un-centered correlation similarity metric. A heatmap visualization of this clustered data was then generated using the Java Treeview application ([jtreeview.sourceforge.net/](http://jtreeview.sourceforge.net/)).

### **Culturing of *B. animalis* subsp. *lactis***

A frozen stock of *B. animalis* subsp. *lactis* (strain CNCM I-2494) was streaked out on MRS-agar plates (BD/Difco) and transferred to a Coy chamber for overnight growth at 37°C under an atmosphere of 5% H<sub>2</sub>, 20% CO<sub>2</sub> and 75% N<sub>2</sub>. Single colonies were picked and inoculated into 10ml of pre-reduced MRS broth (BD/Difco) that had been stored for 24h in the anaerobic chamber. The medium was not supplemented with cysteine. Each culture was passaged four times to stationary phase, during which time test growth curves were used to discern growth kinetics. 100µl of an overnight culture was used to inoculate 10 ml of fresh anaerobic MRS in 27ml Balch tubes with rubber stoppers and aluminum crimp tops. The initial headspace of the tubes was composed of 5% H<sub>2</sub>, 20% CO<sub>2</sub>, and 75% N<sub>2</sub> at ambient pressure. Tubes were incubated at 37°C, and 4ml aliquots were collected at mid-log phase (9h post-inoculation, OD<sub>600</sub>=0.2) and during late stationary phase (27h post-

inoculation,  $OD_{600}=2.6$ ). Each aliquot from each culture ( $n=2$ ) was immediately combined with 8 ml of RNAprotect Bacteria Reagent (Qiagen), incubated for 5 min at room temperature, then centrifuged ( $3,220 \times g$ ; 15 min at  $25^{\circ}\text{C}$ ). The pellets were snap-frozen in liquid nitrogen and stored at  $-80^{\circ}\text{C}$ , and total cellular RNA was subsequently isolated as described previously (S6).

## **Human studies**

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

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

*Quantitative (q) PCR analysis of fecal levels of FMP strains* — qPCR was used to define the levels of selected FMP strains in fecal samples obtained from MZ co-twins and gnotobiotic mice (S8-12). The PCR primer sets targeting each strain’s 16S rRNA gene or CRISPR locus are described in **Table S14**, as are the amplification conditions. Samples were analyzed on an Applied Biosystems 7900HT instrument using SYBR green chemis-

try. Standard curves were constructed using genomic DNA prepared from a known number of bacterial cells harvested from monocultures grown to stationary phase (cells were counted by microscopy after DAPI staining);  $C_t$  values for each reaction could, therefore, be expressed in terms of cell equivalents (CE).

To calculate the concentration of a given bacterial strain in each fecal sample, three serial dilutions of extracted fecal DNA (10ng, 1ng, 0.1ng) were assayed in at least two independent qPCR reactions.  $C_t$  values falling within the linear range of the assay were referenced to the standard curves, while those outside the linear range were excluded from the analysis. For human samples, data were log-transformed and normalized to fecal mass ( $\log_{10}$  CE/g of feces). For mouse samples, data were log-transformed and normalized to mass of template DNA ( $\log_{10}$  CE/ $\mu$ g DNA).

#### ***Multiplex pyrosequencing of 16S rRNA genes in fecal samples and the FMP***

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

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

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

***Shotgun sequencing of fecal microbiomes*** — Forty-eight fecal samples from 4 twin pairs were selected for multiplex shotgun pyrosequencing of total community DNA (454 FLX chemistry). For each individual, 2 fecal samples were analyzed before initiation of FMP consumption, 2 samples during the period when FMP was being consumed, and 2 samples after consumption had ceased.

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

Datasets of reads obtained from shotgun sequencing of the twins' fecal microbiomes were used to query v54 of the KEGG GENES database (**Table S15**) (BLASTX E value  $<10^{-5}$ , bitscore  $> 50$ , and %identity  $> 50$ ). A comparable annotation was performed for published fecal microbiome gene lists that had been generated from 124 deeply sampled unrelated adult Europeans (*S18*), and from a pair of obese adult MOAFTS twins (*S19*).

### **Studies in gnotobiotic mice**

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



tants that significantly decrease in relative abundance after passage through a selective condition, INSeq allows a genome-wide map of *in vivo* fitness determinants to be created.

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

**Animal sampling** – Fecal samples were obtained from each animal at time points indicated in **Fig. 1**. Each fecal sample was collected directly as it emerged from the anus into a 1.7ml screw-cap Eppendorf tube, which was immediately deposited in a stainless steel dewer containing liquid N<sub>2</sub> (the dewer was introduced into the gnotobiotic isolator on the day of collection after it had been sterilized in the isolator's entry port with chlorine dioxide spray (Clidox-S; PRL Pharmacal)). Various subsets of samples were subjected to COPRO-Seq, INSeq and microbial RNA-Seq analyses. Blood samples were collected into lithium heparin tubes (Becton Dickinson), placed immediately on ice, and then centrifuged

(2,700 x g for 3min at 4°C). The resulting plasma supernatant was stored at -80°C until assay. Urine was collected directly into Eppendorf tubes and immediately frozen in liquid N<sub>2</sub>. Upon sacrifice, ceca were dissected and cecal contents were frozen immediately at -80°C.

***Isolation of DNA from cecal contents and feces*** – Microbial community DNA was prepared in a two-step process consisting of a crude extraction step followed by additional purification and RNase treatment.

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

*RNase treatment and further purification.* Aliquots of crude DNA were transferred to a 96-well plate. Buffer PM (Qiagen) was mixed with RNase A (Qiagen) to a final concentration of 1.3mg/ml. Three volumes of this mixture were added to each well and the reactions were allowed to incubate at room temperature for 2 min. Following RNase digestion, samples were applied to a QIAquick 96 PCR purification plate (Qiagen) and processed according to the manufacturer's instructions using a QIAvac 96 manifold. DNA

was eluted in 100µl of Buffer EB (Qiagen). DNA quality and purity were verified using a Nanodrop spectrophotometer (model ND-1000).

***Preparing DNA libraries for Illumina sequencing and COPRO-Seq analysis in a 96-well format*** – DNA libraries were prepared for sequencing using a modified version of Illumina’s sample preparation protocol for generating libraries from genomic DNA. The six steps include the following:

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

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

(iii) *Ligation*. Customized Illumina adapters containing maximally distant 4bp barcodes described elsewhere (S6) were ligated to the polyA-tailed DNA in 50 $\mu$ l reactions as follows. Thirty microliters of the inactivated A-tailing reaction described in the preceding paragraph were added to 5 $\mu$ l T4 DNA ligase buffer (10X), 5 $\mu$ l adapter mix (1 $\mu$ M final concentration per adapter), and 9 $\mu$ l water at 4°C. One microliter of T4 DNA ligase (2,000,000 U/ $\mu$ l; NEB M0202M) was subsequently added and reactions were incubated (16°C, 1h) followed by heat-inactivation (65°C, 10 min). Ligation reactions were cleaned up using the MinElute 96 UF PCR Purification Kit (Qiagen) according to the manufacturer's recommended protocol. DNA was eluted in 22 $\mu$ l Buffer EB.

(iv) *Gel Purification*. 10 $\mu$ l of each elution was separated by gel electrophoresis on 2% agarose. DNA migrating at 200bp was excised and gel slices were purified using a QIAquick 96 PCR Purification Kit (Qiagen).

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

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

A custom software pipeline was written in Perl for performing COPRO-Seq data processing in a computer cluster environment running Sun Grid Engine. These data processing steps are schematized in **Fig. S4A**. Briefly, raw Illumina GA-IIx reads from a sequencing pool were first de-plexed by barcode and trimmed to 34bp (30bp genome se-

quence + 4bp barcode). Trimmed reads were aligned to the genomes of the 20 microbial strains used in this study using Illumina's ELAND aligner. Perfect, unique alignments to the reference genomes were retained, while those mapping less than perfectly or having multiple possible alignments to the reference genomes were filtered out, ensuring that only high-quality, unambiguous reads were used. Hits to each genome were then tallied, after which the summed counts for each genome were normalized by that genome's 'informative genome size' (term defined in *Results*) to adjust for both genome size and uniqueness relative to all other genomes in the experiment. The Perl scripts supporting the COPRO-Seq analytic pipeline can be downloaded from: [http://gordonlab.wustl.edu/projects/2011-McNulty\\_et al.](http://gordonlab.wustl.edu/projects/2011-McNulty_et_al)

***Characterizing gene expression with microbial RNA-Seq*** — Following extraction of total nucleic acid with phenol-chloroform, and precipitation with isopropanol, fecal samples were subjected to DNase digestion (*S6*). Total RNA was then (i) passed through an MEGAClear column (Ambion) to deplete RNAs <200 nt (removing most 5S rRNA and tRNA species; (*S6*)); (ii) subjected to another round of DNase digestion; (iii) passed through another MEGAClear column; and (iv) subjected to a hybridization-based pull-down of 16S and 23S rRNAs using custom-designed biotinylated oligonucleotides that contain short rDNA sequences conserved across a set of 37 human gut-derived sequenced microbial genomes (*S6*). The depletion protocol, which has been adapted to 96-well format, is described elsewhere (*S6*). PCR (30 cycles) employing primers directed against the most abundant community member (typically *Bacteroides WH2*), verified the absence of detectable gDNA in the purified RNA preparations.

Doubled stranded (ds) cDNA was synthesized using random hexanucleotide primers. At the conclusion of the reaction, Illumina adapters containing sample-specific 4 nt barcodes were ligated to the dscDNA. Multiplex sequencing was performed using the Illumina GA-IIx instrument. We typically sequenced two barcoded *in vitro* samples/lane of the

8-lane flow cell; *in vivo* samples were not multiplexed (i.e., 1 sample analyzed/lane). This allowed us to identify mRNA present at levels representing  $\geq 0.001\%$  of all reads.

The pipeline for processing microbial RNA-Seq data is presented in **Fig. S4B**. The 8-20 million 36nt cDNA reads from each sequencing lane were separated by barcode, and mapped against the relevant set of genome sequences using the SSAHA2 algorithm (*S21*) to determine the raw unique-match ‘counts’ (reads) for each gene present in the relevant microbial genome or microbiome. Reads that mapped non-uniquely were added to each gene in proportion to each gene’s fraction of unique-match counts (e.g., a non-unique read that maps equally well to gene A with 18 unique reads and gene B with 2 unique reads will be scored as 0.9 of a count to gene A, and 0.1 of a count to gene B; the influence of ties is negligible for RNA-Seq given the small numbers of distinct genomes, but would become more important with more complex communities). Raw counts were then normalized to reads/kb gene length/million mapped reads (RPKM) using one or more gene position file(s) in conjunction with custom perl scripts.

Data normalization was carried out at two different levels in this study. For our ‘top-down’ analysis, data were normalized at the level of the entire community metatranscriptome (i.e., raw counts from all species were normalized simultaneously using a single gene position file that included the positions of all genes in the model community metatranscriptome). Data normalized in top-down fashion allowed us to determine, after binning gene expression values by function, how the collective operations of the model community were changing as the result of experimental perturbations. In our ‘bottom-up’ analysis, data were normalized at the level of individual species (i.e., raw counts from each individual species were normalized separately from one another, in each case using a species-specific gene position file describing the positions of only that species’ genes). Data normalized in bottom-up fashion allowed us to interrogate what statistically significant gene expression changes were occurring within a given species of interest.

Bottom-up normalized transcript data were analyzed by Cyber-T (S22) to identify mRNAs that exhibited significant differences in their levels of expression between samples. For each comparison, transcripts were then binned into a list where the magnitude of the difference in their expression was  $\geq 4$ -fold. Binned transcripts were subsequently annotated using the `kegg_counting.pl` perl script described above. Each resulting annotated dataset was used to determine the representation of individual genomes and KEGG level 2 categories within these lists.

Further functional comparisons were carried out using ShotgunFunctionalizeR, an R package designed to analyze differences between metagenomic datasets using a Poisson statistical model (S23). The `kegg_counting.pl` script was used to sum RPKM normalized reads for all transcripts annotated with an EC number obtained from BLAST to KEGG. Summed reads in each EC bin were rounded to integer format, and the data imported into ShotgunFunctionalizeR, which was then used to generate lists of transcripts encoding ECs that were differentially expressed in various samples.

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

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

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

GC/MS methods generally followed those of Fiehn (S27) and Kind (S28), and used a 6890N GC connected to a 5975 Inert single-quadrupole MS (Agilent, Santa Clara, CA). A large-volume, ProSep inlet enabled programmed-temperature vaporization and diversion of heavy contaminants away from the GC and MS, as described below, greatly reducing maintenance time (Apex Technologies, Inc., Independence, KY). The two wall-coated, open-tubular GC columns connected in series were both from J&W/Agilent (part 122-5512), DB5-MS, 15 meters in length, 0.25 mm in diameter, with an 0.25- $\mu$ m luminal film. Prior to each run, initial inlet pressures were empirically adjusted such that the resulting retention time (RT) of the TMS-D27-C14:0 standard was set at ~16.727 minutes. Under these conditions, derivatized metabolites eluted from the column and reached the electron-ionization (EI) source in the MS at known times (*e.g.*, bis-TMS-lactic acid at ~6.85 minutes, and TMS-cholesterol at ~27.38 minutes). A mid-column, microfluidic splitter (Agi-



lent) provided a means for hot back-flushing of the upstream GC column at the end of each run while the oven was held at 325°C for a terminal “bake-out” (another antifouling and anti-carryover measure analogous to that described in (S29)). During this terminal “bake-out,” the inlet was also held at 325°C, and it was purged with a large flow of the carrier gas, helium. Positive ions generated with conventional EI at 70 eV were scanned broadly from 600 to 50 m/z in the detector throughout the run.

Raw data from Agilent’s ChemStation software environment were imported into the freeware, Automatic Mass Spectral Deconvolution and Identification Software (AMDIS), developed by Drs. Steve Stein, W. Gary Mallard, and their coworkers at National Institute of Standards and Technology (S30-32); also courtesy of NIST at <http://chemdata.nist.gov/mass-spc/amdis/>). Deconvoluted spectra were identified, to the extent possible, using several commercial and public spectral libraries. Our primary source was the Fiehn GC/MS Metabolomics RTL Library (a gift from Agilent Technologies, Santa Clara, CA, part number G1676-90000). Additional spectra for comparison were gleaned from the Golm Metabolome Library (courtesy of Dr. Joachim Kopka and coworkers at the Max Planck Institute of Molecular Plant Physiology, Golm, Germany (S33); <http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html>), the commercial NIST/EPA/NIH Mass Spectral Library and our own purpose-built spectral library. Where indicated, peak alignment was performed with SpectConnect freeware (courtesy of Dr. Gregory Stephanopoulos, Massachusetts Institute of Technology, [www.spectconnect.mit.edu](http://www.spectconnect.mit.edu) (S34)). Chemometrics were performed with Mass Profiler Professional (a recent descendant of GeneSpring MS, purchased from Agilent), along with our own custom macros, written in Visual Basic for use in the Excel software environment.

The statistical significance of differences in the  $\log_2$  spectral abundances of each metabolite in samples obtained at different time points was tested using two approaches. A first-pass, highly-permissive set of pairwise comparisons was calculated between each combination of samples (d0 versus d14, d0 versus d42, d14 versus d42) using a simple

two-tailed Student's *t*-test. The resulting p-values, which were not corrected for multiple hypothesis testing, are listed in **Table S12**. Given the high ratio of hypotheses tested to samples per group (198 metabolites; 5-8 samples per group), we also guarded against false discovery by performing a more stringent set of calculations that produced a shorter list of metabolites with significant differences in abundance. This latter procedure consisted of first taking metabolite data from all three time points and subjecting them to a one-way ANOVA. The resulting p-values were then adjusted using Benjamini-Hochberg correction, generating q-values. The log<sub>2</sub> spectral abundances of all metabolites whose q-values were below 0.05 were then subjected to Tukey's HDS (Honestly Significantly Different) post-hoc test to determine which time points were significantly different from one another. All Tukey's HDS p-values that were calculated are provided in **Table S12**.

## **Supplementary Results**

### **Human studies**

*Analysis of the effects of FMP consumption on stool consistency, difficulty of passage, and frequency* — To determine whether there were differences in stool consistency and difficulty of passage of stools between pre-treatment, treatment, and post-treatment study periods, we first constructed a dataset in which the unit of analysis was 'bowel movement.' Using ordinal logistic regression, we analyzed separate models predicting stool consistency and difficulty of stool passage using 'treatment period' as the reference group. We adjusted for clustering of observations using a Huber-White robust variance estimator (STATA 2004). When data from the entire study period were included in the analyses, no significant differences were observed between study periods for either stool consistency or difficulty passing stool. Next, we conducted an alternate analysis in which data from the first two weeks of the treatment and of the post-treatment phases were omitted. We found that women had lower stool consistency scores during the last two weeks of the post-

treatment phase compared to the last two weeks of the treatment phase: i.e., stools were softer during the treatment period (OR=0.69; p=0.04). The difference in stool consistency between the pre-treatment phase and the last two weeks of the treatment phase was not significant; however, there was a significant difference between the pre-treatment versus treatment compared to the post-treatment versus treatment odds ratios (p=0.005).

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

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

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

present (the latter determined by qPCR). It is reasonable that *L. lactis* would co-occur with *B. animalis* subsp. *lactis* given the presence of both strains in the FMP. A co-occurrence analysis performed at the family level of taxonomy failed to identify any significant differences.

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

### **Studies in gnotobiotic mice**

**Measurement of adiposity** — The body weights and epididymal fat pad weights of mice from both treatment groups were measured at the time of sacrifice. We observed no significant differences between the single and multiple treatment groups in either measurement (p=0.6865, p=0.3516, respectively; two-tailed Student’s *t*-test). Furthermore, all measurements of adiposity and weight were in line with those of mice from other studies that had involved animals from the same inbred strain, who were similarly aged, the same gender,

on same diet, and who harbored comparable defined model human gut microbiota but without FMP strains.

### **In vitro studies**

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

### Supplementary References

- S1. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res* **28**, 27 (2000).
- S2. M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa, From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Res* **34**, D354 (2006).
- S3. M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa, KEGG for representation and analysis of molecular networks involving diseases and drugs, *Nucleic Acids Res* **38**, D355 (2010).
- S4. B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, B. Henrissat, The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics, *Nucleic Acids Res* **37**, D233 (2009).
- S5. M. J. de Hoon, S. Imoto, J. Nolan, S. Miyano, Open source clustering software, *Bioinformatics* **20**, 1453 (2004).
- S6. F. E. Rey, J. J. Faith, J. Bain, M. J. Muehlbauer, R. D. Stevens, C. B. Newgard, J. I. Gordon, Dissecting the in vivo metabolic potential of two human gut acetogens, *J Biol Chem* **285**, 22082 (2010).
- S7. S. J. Lewis, K. W. Heaton, Stool form scale as a useful guide to intestinal transit time, *Scand J Gastroenterol* **32**, 920 (1997).
- S8. O. Firmesse, S. Rabot, L. G. Bermudez-Humaran, G. Corthier, J. P. Furet, Consumption of Camembert cheese stimulates commensal enterococci in healthy human intestinal microbiota, *FEMS Microbiol Lett* **276**, 189 (2007).
- S9. J. P. Furet, P. Quenee, P. Tailliez, Molecular quantification of lactic acid bacteria in fermented milk products using real-time quantitative PCR, *Int J Food Microbiol* **97**, 197 (2004).

- S10. T. Matsuki, K. Watanabe, J. Fujimoto, T. Takada, R. Tanaka, Use of 16S rRNA gene-targeted group-specific primers for real-time PCR analysis of predominant bacteria in human feces, *Appl Environ Microbiol* **70**, 7220 (2004).
- S11. M. A. Nadkarni, F. E. Martin, N. A. Jacques, N. Hunter, Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set, *Microbiology* **148**, 257 (2002).
- S12. H. Sokol, B. Pigneur, L. Watterlot, O. Lakhdari, L. G. Bermudez-Humaran, J. J. Gratadoux, S. Blugeon, C. Bridonneau, J. P. Furet, G. Corthier, C. Grangette, N. Vasquez, P. Pochart, G. Trugnan, G. Thomas, H. M. Blottiere, J. Dore, P. Marteau, P. Seksik, P. Langella, Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients, *Proc Natl Acad Sci U S A* **105**, 16731 (2008).
- S13. P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight, J. I. Gordon, A core gut microbiome in obese and lean twins, *Nature* **457**, 480 (2009).
- S14. J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, R. Knight, QIIME allows analysis of high-throughput community sequencing data, *Nat Methods* **7**, 335 (2010).
- S15. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics* **22**, 1658 (2006).
- S16. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate — a Practical and Powerful Approach to Multiple Testing, *J Roy Stat Soc B Met* **57**, 289 (1995).

- S17. V. Gomez-Alvarez, T. K. Teal, T. M. Schmidt, Systematic artifacts in metagenomes from complex microbial communities, *ISME J* **3**, 1314 (2009).
- S18. J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, M. Jian, Y. Zhou, Y. Li, X. Zhang, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S. D. Ehrlich, A human gut microbial gene catalogue established by metagenomic sequencing, *Nature* **464**, 59 (2010).
- S19. P. J. Turnbaugh, C. Quince, J. J. Faith, A. C. McHardy, T. Yatsunenko, F. Niazi, J. Affourtit, M. Egholm, B. Henrissat, R. Knight, J. I. Gordon, Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins, *Proc Natl Acad Sci U S A* **107**, 7503 (2010).
- S20. A. L. Goodman, N. P. McNulty, Y. Zhao, D. Leip, R. D. Mitra, C. A. Lozupone, R. Knight, J. I. Gordon, Identifying genetic determinants needed to establish a human gut symbiont in its habitat, *Cell Host Microbe* **6**, 279 (2009).
- S21. Z. Ning, A. J. Cox, J. C. Mullikin, SSAHA: a fast search method for large DNA databases, *Genome Res* **11**, 1725 (2001).
- S22. A. D. Long, H. J. Mangalam, B. Y. Chan, L. Toller, G. W. Hatfield, P. Baldi, Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli* K12, *J Biol Chem* **276**, 19937 (2001).
- S23. E. Kristiansson, P. Hugenholtz, D. Dalevi, ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes, *Bioinformatics* **25**, 2737 (2009).



- S24. A. Liaw, M. Wiener, Classification and Regression by randomForest, *R News* **2**, 18 (2002).
- S25. D. Knights, E. K. Costello, R. Knight, Supervised classification of human microbiota, *FEMS Microbiol Rev* **35**, 343 (2011).
- S26. U. Roessner, C. Wagner, J. Kopka, R. N. Trethewey, L. Willmitzer, Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry, *Plant J* **23**, 131 (2000).
- S27. O. Fiehn, G. Wohlgemuth, M. Scholz, T. Kind, Y. Lee do, Y. Lu, S. Moon, B. Nikolau, Quality control for plant metabolomics: reporting MSI-compliant studies, *Plant J* **53**, 691 (2008).
- S28. T. Kind, G. Wohlgemuth, Y. Lee do, Y. Lu, M. Palazoglu, S. Shahbaz, O. Fiehn, FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry, *Anal Chem* **81**, 10038 (2009).
- S29. J. Chen, C. K. Meng, S. B. Narayan, W. Luan, M. J. Bennett, The use of Deconvolution Reporting Software and backflush improves the speed and accuracy of data processing for urinary organic acid analysis, *Clin Chim Acta* **405**, 53 (2009).
- S30. W. G. Mallard, J. Reed, U. D. o. C. National Institute of Standards and Technology, Ed. (Gaithersburg, MD, 1997), pp. 58.
- S31. J. M. Halket, A. Przyborowska, S. E. Stein, W. G. Mallard, S. Down, R. A. Chalmers, Deconvolution gas chromatography/mass spectrometry of urinary organic acids--potential for pattern recognition and automated identification of metabolic disorders, *Rapid Commun Mass Spectrom* **13**, 279 (1999).
- S32. S. E. Stein, An Integrated Method for Spectrum Extraction and Compound Identification from Gas Chromatography/Mass Spectrometry Data, *J Am Soc Mass*

*Spectrum* **10**, 770 (1999).

- S33. J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmuller, P. Dormann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, D. Steinhauser, GMD@CSB.DB: the Golm Metabolome Database, *Bioinformatics* **21**, 1635 (2005).
- S34. M. P. Styczynski, J. F. Moxley, L. V. Tong, J. L. Walther, K. L. Jensen, G. N. Stephanopoulos, Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery, *Anal Chem* **79**, 966 (2007).
- S35. M. A. Mahowald, F. E. Rey, H. Seedorf, P. J. Turnbaugh, R. S. Fulton, A. Wollam, N. Shah, C. Wang, V. Magrini, R. K. Wilson, B. L. Cantarel, P. M. Coutinho, B. Henrissat, L. W. Crock, A. Russell, N. C. Verberkmoes, R. L. Hettich, J. I. Gordon, Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla, *Proc Natl Acad Sci U S A* **106**, 5859 (2009).
- S36. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies, *Proc Natl Acad Sci U S A* **100**, 9440 (2003).

### Supplementary Figure Legends

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

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

**Figure S3. CAZyme profiles of the 20 bacterial strains introduced into gnotobiotic mice.** The indicated genomes were annotated for all glycoside hydrolases (GH), glycosyltransferases (GT), carbohydrate binding modules (CBM), and polysaccharide lyases (PL) using the CAZy classification scheme. The Bacteroides possess a larger and more diverse arsenal of CAZymes relative to the Firmicutes/Actinobacteria. Though most CA-

Zyme families encoded in the genomes of the FMP strains were also present in defined community members. The small number of FMP strain-specific CAZyme families (CBM5, CBM10, CBM23, CBM33, GH85, GT39) are highlighted in red. The scale refers the number of genes in a given CAZy family in a given genome.

**Figure S4. Summary of analysis pipelines utilized in this study. (A) COPRO-Seq. (B) RNA-Seq.**

**Figure S5. COPRO-Seq-based time series analysis of the abundance of members of the model human microbiota and of the FMP strain consortium in the feces of gnotobiotic mice.** Relative abundance, expressed as the  $\log_{10}$  of percent representation of all detected community members, is defined over time (d0, time of colonization with the model 15-member community; d14, time of first gavage with the FMP consortium for the single and multiple treatment groups; d21 and d35, times of subsequent gavage with the FMP consortium for the multiple treatment group). For each treatment, animals were gavaged twice over a 24h period. Mean values  $\pm$  SEM are plotted (n=5 animals/treatment group; 1 fecal sample/animal/time point; limit of detection = 0.003%). In cases where an error bar would extend below the x-axis, only the upper limit and mean are plotted. **(A)** COPRO-Seq data for 13 members of the 15 member community (*F. prausnitzii* and *C. spiroforme* were below the limits of detection throughout the study). **(B)** Data obtained from the two members of the FMP consortium that persisted at levels above the limit of detection following their introduction into mice. **(C)** Data from panel A representing the response of *C. aerofaciens* to introduction of the FMP strain consortium (see text for details).

**Figure S6. Top-down analysis of the model community's transcriptional response to the FMP strain consortium reveals upregulation of genes involved in interconversion of propionate and succinate.** Normalized RNA-Seq data were binned at the level of E.C. and comparisons were made between early responses (day 14 versus d15, representing time points just before and 1 day after gavage with the strain consortium) and late respons-

es (days 14 versus 42). Boxes and lines are colored according to the key shown above the pathway map and in the legend to panel B of Figure 5.

**Figure S7. A species' contribution to the meta-transcriptome is not necessarily proportional to its abundance in the 15-member community.** Microbial RNA-Seq data from day 14 of the mouse study were parsed by species to determine the total number of reads that each community member contributed to the total sequenced transcript pool ('meta-transcriptome') (both raw and normalized reads as defined in **Fig. S4B**). Data were further broken down into reads that could be mapped to genes with known functions (as defined by KEGG) and those with unknown functions (lacking any K number in the KEGG GENES v54 database). Mean values  $\pm$  S.D. are plotted for each of the four types of data presented. Significant differences between a species proportional abundance in the community (COPRO-Seq) and its contribution to the transcript pool are noted at the bottom of the figure next to the species name; the type of transcript data that show significant differences relative to the COPRO-Seq data are indicated by the colored box next to the species name. Note, for example, the large number of raw reads attributed to *R. obeum* despite its low proportional abundance in the community. Conversely, *Bacteroides WH2* contributes a far smaller proportion of total raw RNA-Seq reads to the pool than its relative abundance in the community might have suggested.

**Figure S8. Bottom-up analysis of genes whose expression changes significantly following introduction of the FMP strain consortium.** (A) Volcano plots of the >48,000 expressed genes detected in at least one fecal RNA sample. Colored points represent genes whose difference in expression followed introduction of the FMP strain consortium was (i)  $\geq 4$ -fold (increased or decreased) relative to the d14 pre-treatment time point and (ii) statistically significant ( $p < 0.05$ ; two-tailed Student's *t*-test). Dots are colored according to each gene's species of origin (color key shown to the right of the panel). Black dots represent genes whose change in expression is  $< 4$ -fold at the time points indicated and/or not statisti-

cally significant. See **Table S10** for a complete list of all the genes shown as colored dots. **(B)** Species-breakdown of differentially expressed genes. **(C)** KEGG category breakdown of differentially expressed genes, showing that late responses are more numerous than immediate ones, and that there is a noticeable bias towards genes involved in carbohydrate and glycan metabolism, particularly in the late response to introduction of the FMP strain consortium.

**Figure S9. The number of RNA-Seq reads, obtained from human fecal samples, that map to genomes in the FMP strain consortium peaks shortly after FMP consumption begins.** RNA-Seq reads derived from selected human samples were mapped back to the five genomes in the FMP consortium to determine which species' transcripts could be detected over time. *B. animalis* transcripts were detected only during periods of FMP consumption. Reads attributed to *L. lactis*, *S. thermophilus*, and *L. delbrueckii* at time points before consumption began may reflect 'spurious' mapping to related endogenous strains in the gut community. Facet labels located at the top of each bar chart correspond to the sample labels shown in **Figure 1A** of the main text. Numbers are presented below each bar chart correspond to the code number assigned to each de-identified co-twin (see **Table S1**). Sequence data at time points marked 'long' represent 72nt reads, while all other data represent 36nt reads.

## Supplementary Figures

Figure S1.

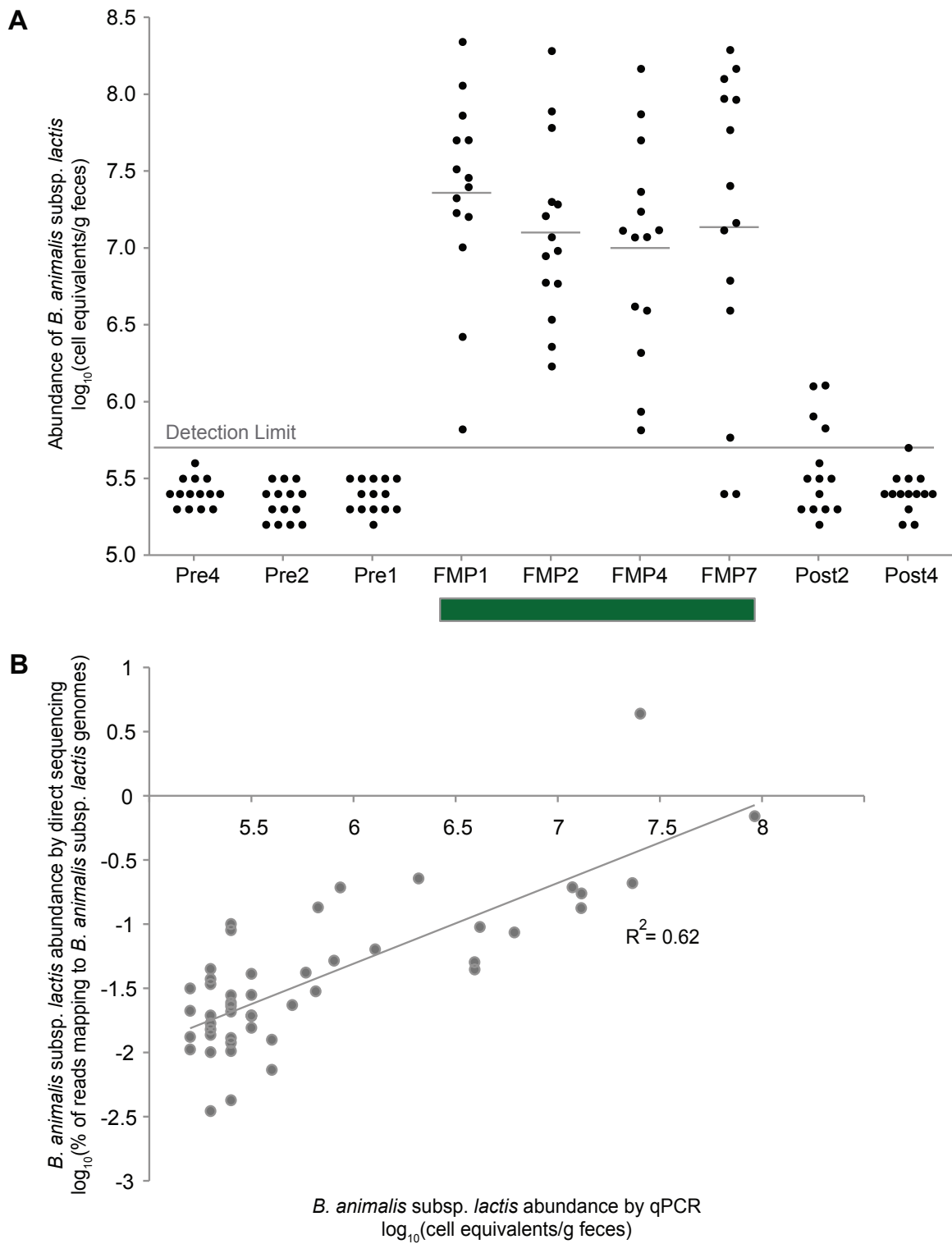


Figure S2.

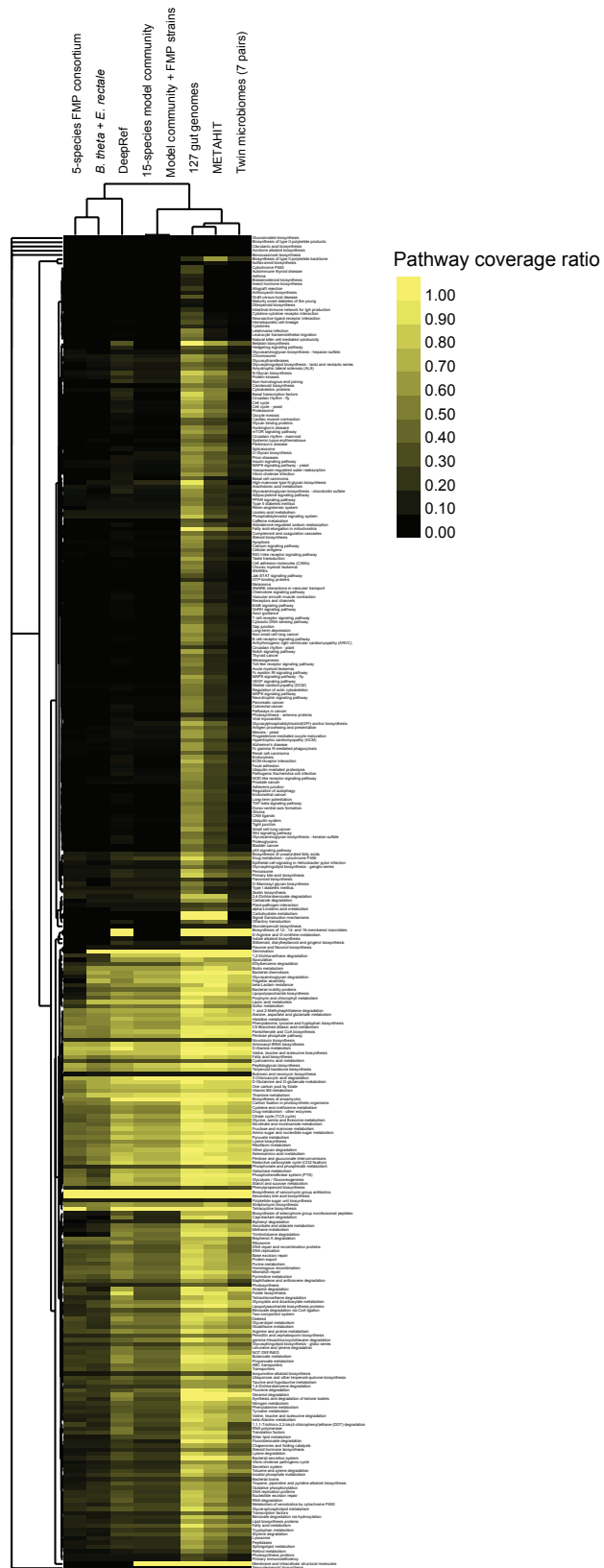




Figure S3.

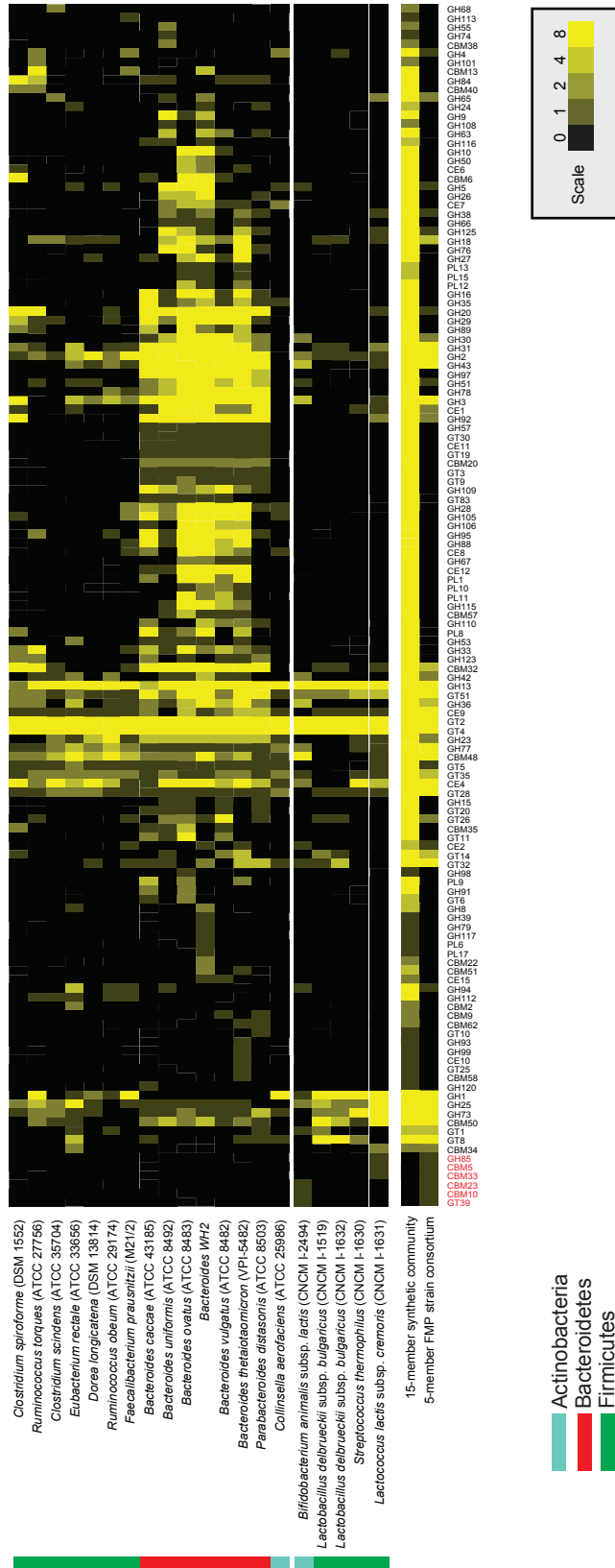
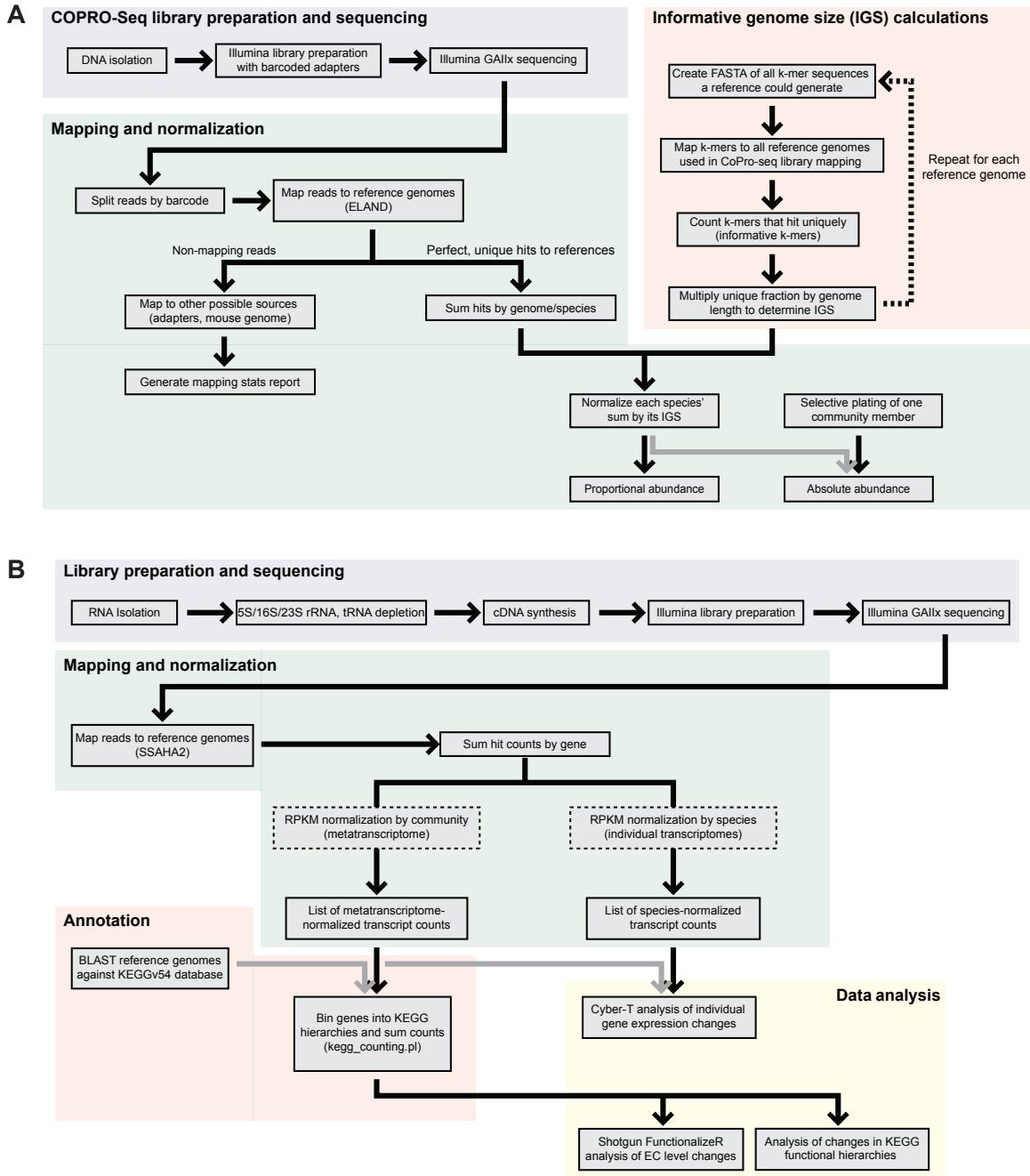
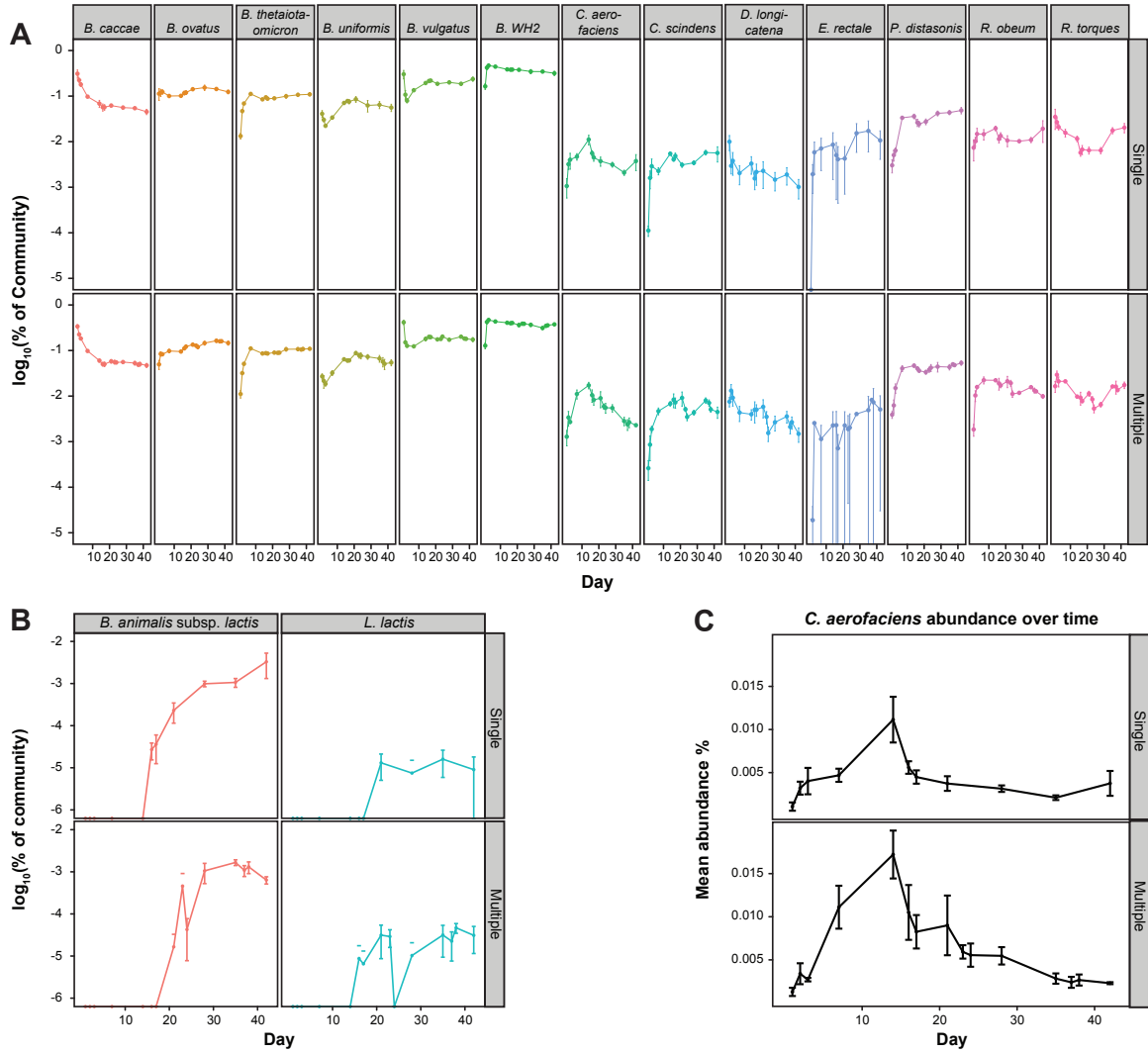


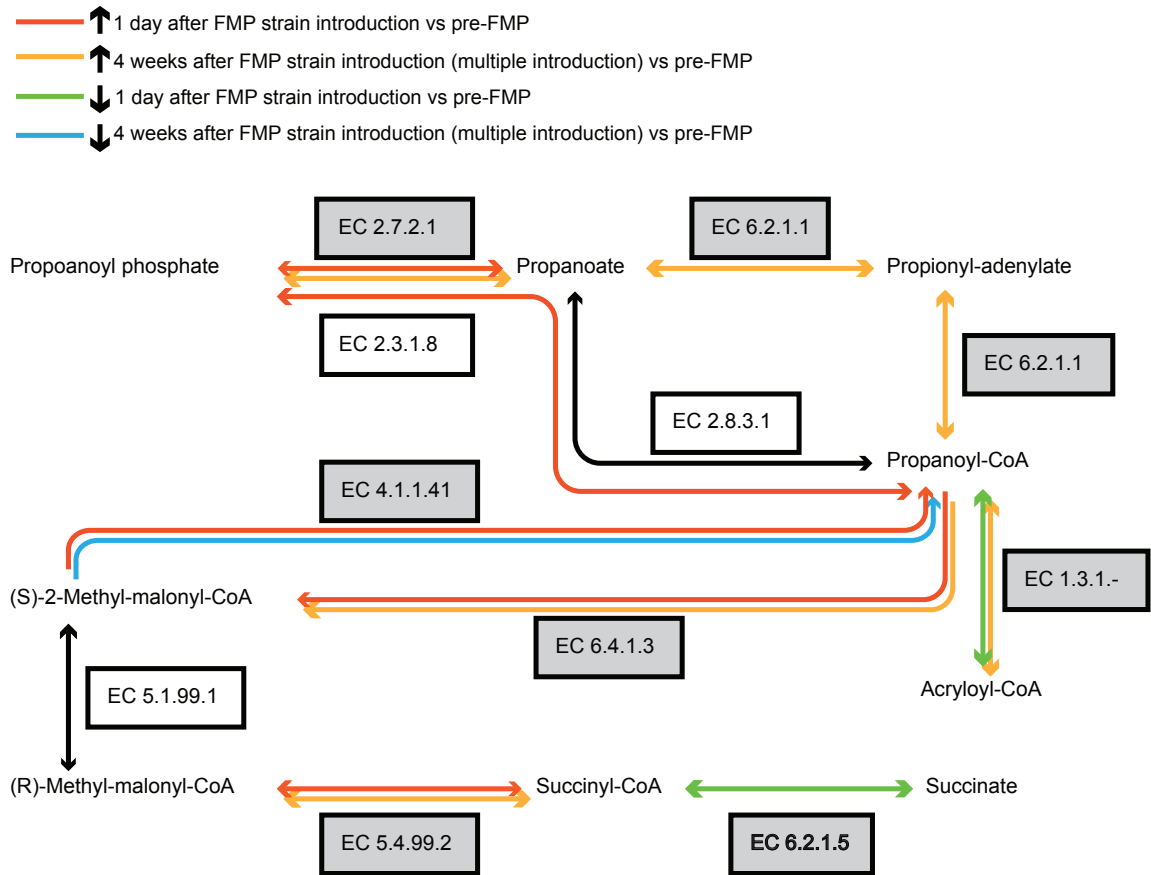
Figure S4.



**Figure S5.**



**Figure S6.**



**Figure S7.**

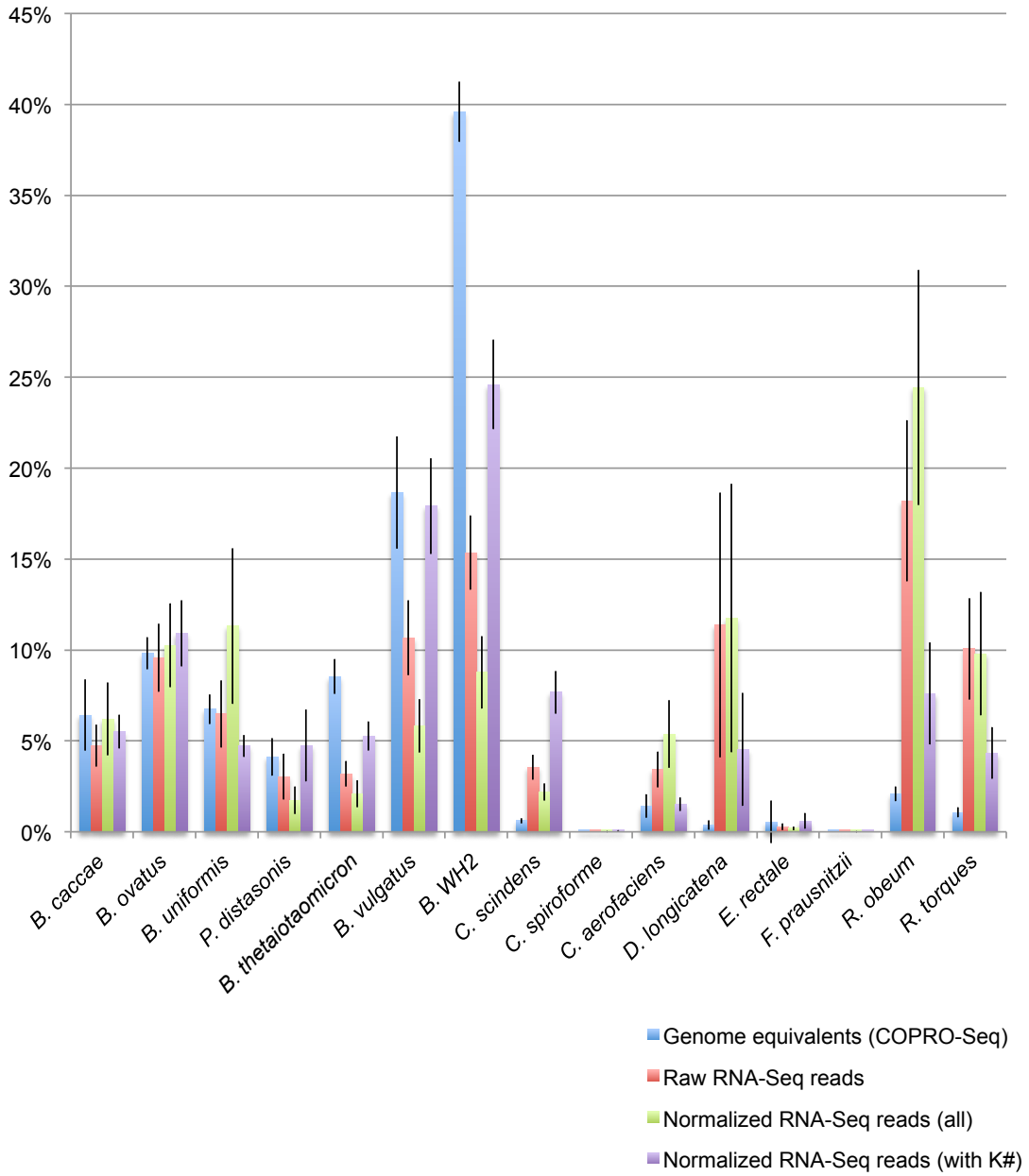
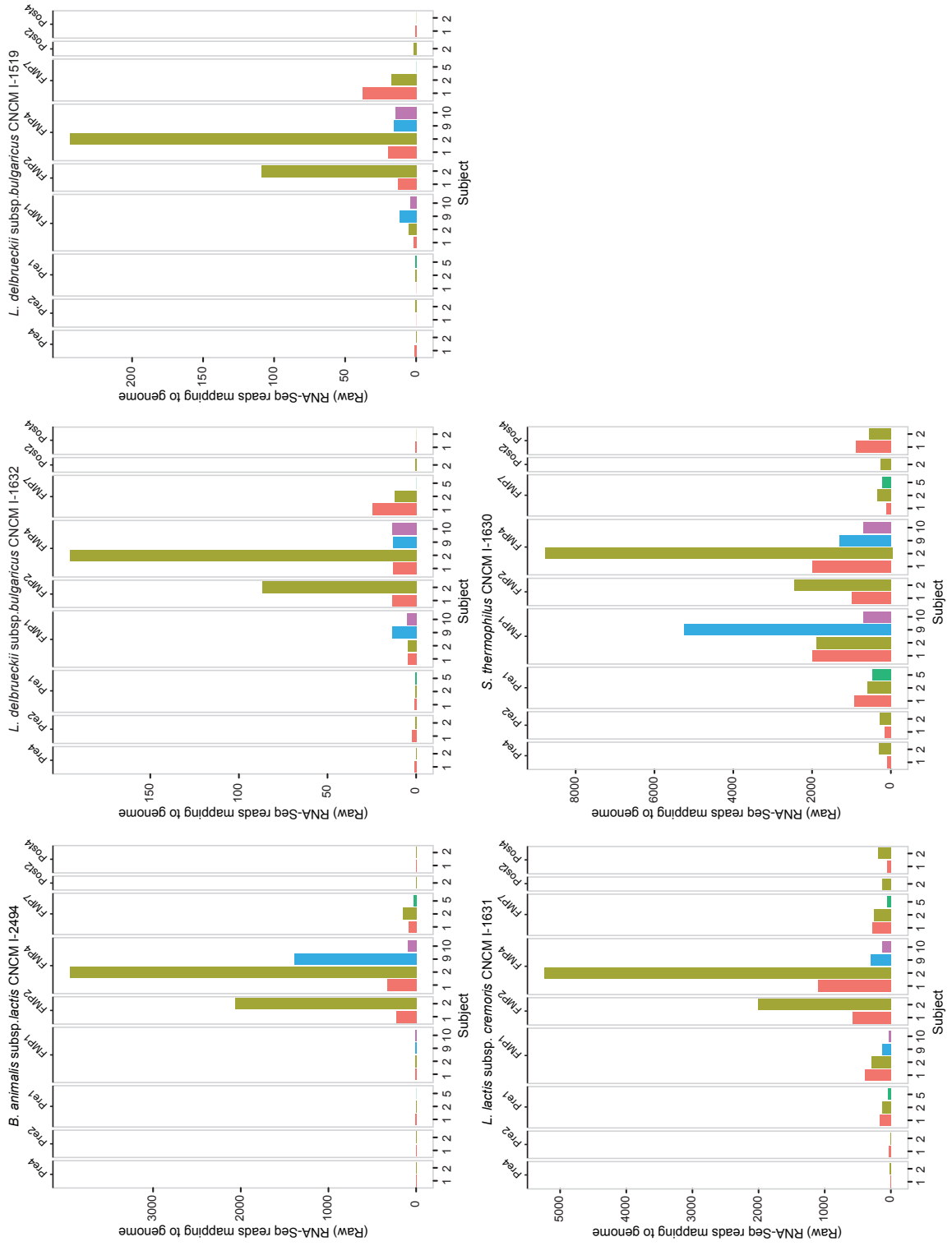
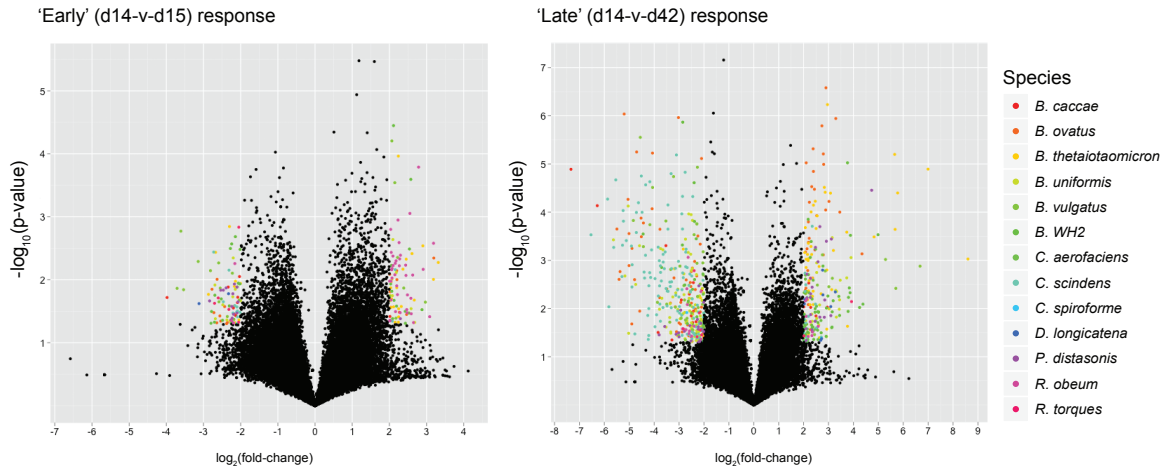


Figure S8.

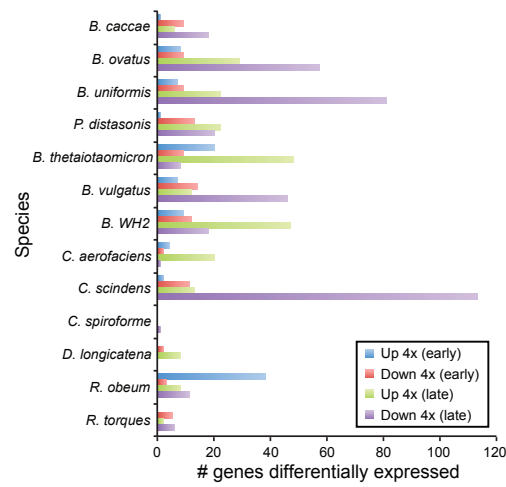


**Figure S9.**

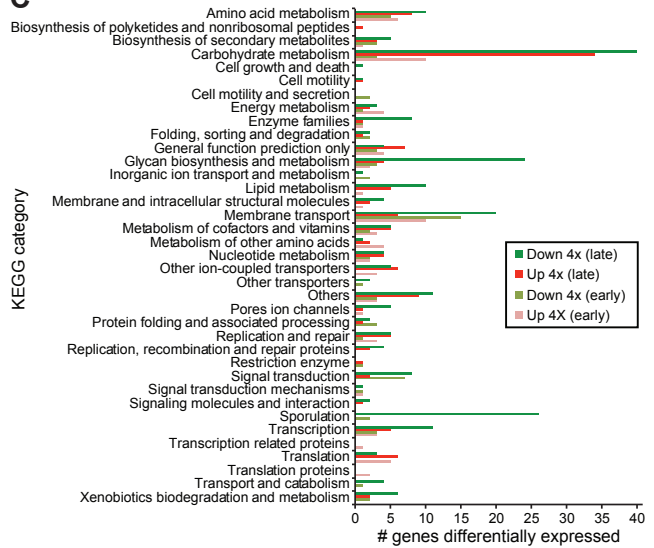
**A**



**B**



**C**



## **Supplementary Table Legends**

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

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

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

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

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

**Table S6. INSeq analysis.** (A) INSeq analysis sequencing statistics. Scale factor corresponds to counts per million normalization; underrepresented samples were re-sequenced and combined with original data so that all samples were represented by ~1 million reads. (B) Genes required by *B. thetaiotaomicron* for survival in the intestines of mice harboring the 15-member model human gut microbiota. The table describes the relative abundance of transposon insertions in each gene (rows) in the input community (average of two indepen-



dent technical replicates) and in the output communities (fecal samples collected from 10 mice two weeks after introduction of the synthetic model community, immediately prior to introduction of the FMP strain consortium). A  $z$ -test was used to identify genes whose log-transformed output to input ratios were significantly different from the overall distribution (a uniform value of 1 was added to all counts, and genes with no insertions were removed to allow ratios to be calculated). Resultant  $p$ -values were corrected for multiple hypothesis testing by  $q$ -test (S36). Genes assigned a  $q < 0.001$  are highlighted in red. Data filtering, normalization, mapping, and statistical analysis were conducted in Perl and Matlab.

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

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

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

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

**Table S11. Results of Random Forests supervised classification analysis.**

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

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

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

**Table S15. List of 127 human gut microbial genomes used to annotate human fecal microbiome datasets.**

## Supplementary Tables

**Table S1.**

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

Family ID	Twin ID	Age at enrollment (years)	BMI at enrollment (kg/m <sup>2</sup> )	Physical distance between co-twins (miles)	Breastfed, Y/N (length) <sup>1</sup>	Fermented milk product flavor(s) selected (4 shipments)
F1	F1T1	27	23	3	No	27 strawberry, 9 peach
F1	F1T2	27	23	3	No	27 strawberry, 9 vanilla
F2	F2T1	27	20	11	NA	12 strawberry, 12 peach, 12 vanilla
F2	F2T2	27	21	11	NA	12 strawberry, 12 peach, 12 vanilla
F3	F3T1	23	21	0	NA	12 strawberry, 12 peach, 12 vanilla
F3	F3T2	23	23	0	NA	12 strawberry, 12 peach, 12 vanilla
F4	F4T1	23	26	0	NA	18 strawberry, 18 peach
F4	F4T2	23	23	0	NA	18 strawberry, 18 peach
F5	F5T1	27	21	932	Yes (2 months)	36 peach
F5	F5T2	27	21	932	Yes (2 months)	18 peach, 18 strawberry
F6	F6T1	31	21	1926	NA	36 vanilla
F6	F6T2	31	20	1926	NA	12 strawberry, 12 peach, 12 vanilla
F7	F7T1	32	24	770	NA	12 strawberry, 12 peach, 12 vanilla
F7	F7T2	32	21	770	NA	18 strawberry, 18 vanilla

<sup>1</sup>NA = breastfeeding data not available

**Table S2.**

**Table S2. Summary of human fecal metagenomic datasets.**

**A. Multiplex pyrosequencing of fecal bacterial 16S rRNA V2 amplicons.**

Sample designation is as follows: F=Family ID, T=co-twin ID (1 or 2), Pre4=4 weeks prior to FMP consumption, FMP2=2 weeks into FMP consumption, Post4=4 weeks after cessation of FMP con:

Sample	Barcode Sequence	Number of 16S rRNA sequences	Barcode sequence used in the deep sequencing of 2 families	Number of 16S rRNA sequences for deep sequencing	FMP Batch*	Time Point
F1T1FMP1	AGTGTACGGTG	7141	TCGAGCGAATCT	12185	B1234	1 week into FMP consumption
F1T1FMP2	CATGTAATGCTC	2459	TCGTTACATGA	9295	B1234	2 weeks into FMP consumption
F1T1FMP4	CATATCGCAGTT	2689	TCTGCTAGATGT	10265	B1234	4 weeks into FMP consumption
F1T1FMP7	AGTTTCAGACGCT	2383	TGAGGATGATAG	10756	B1234	7 weeks into FMP consumption
F1T1Post2	AGTTCTACGTCA	2697	TGCGTGGTAGAC	12345	NA	2 weeks after cessation of FMP consumption
F1T1Post4	ATAATCTCGTCG	2669	TATCTCGAACTG	5278	NA	4 weeks after cessation of FMP consumption
F1T1Pre1	AGTGGATGCTCT	1869	TCATCGCGATAT	11458	NA	1 week before start of FMP consumption
F1T1Pre2	AGTGGATGCGGT	2838	TCAGGATTAGGG	10985	NA	2 weeks before start of FMP consumption
F1T1Pre4	AGTGAGAGAAGC	2992	TAGTTGCGAGTC	12168	NA	4 weeks before start of FMP consumption
F1T2FMP1	ATAGGCGATCTC	2888	TCGAGGACTGCA	11919	B1234	1 week into FMP consumption
F1T2FMP2	ATATCGCTACTG	2196	TCTACGGAGAGC	12271	B1234	2 weeks into FMP consumption
F1T2FMP4	CAGCTAGAACGC	2635	TCTGTGCTCTC	12257	B1234	4 weeks into FMP consumption
F1T2FMP7	ATCACGTAGCGG	3471	TGAGTCACTGGT	10817	B1234	7 weeks into FMP consumption
F1T2Post2	ATCACTAGTAC	2632	TGCTACCATGAG	14486	NA	2 weeks after cessation of FMP consumption
F1T2Post4	GATAGTGCCACT	3032	TCAGACAGACCG	10297	NA	4 weeks after cessation of FMP consumption
F1T2Pre1	CGATGCACCAGA	2318	TCATCTGACTGA	7649	NA	1 week before start of FMP consumption
F1T2Pre2	ATACTCACTCAG	2750	TCACTATGGTCA	10993	NA	2 weeks before start of FMP consumption
F1T2Pre4	ATACTATGCGC	3021	TATAGCGGCATT	11785	NA	4 weeks before start of FMP consumption
F2T1FMP1	CGCGATGTACA	2640			B1234	1 week into FMP consumption
F2T1FMP2	ATCTCTGGCATA	2379			B1234	2 weeks into FMP consumption
F2T1FMP4	CACCTCAACAGAC	3772			B1234	4 weeks into FMP consumption
F2T1FMP7	ATCTGGTGCTAT	2342			B1234	7 weeks into FMP consumption
F2T1Post2	ATCTTAGACTGC	2588			NA	2 weeks after cessation of FMP consumption
F2T1Post4	ATGACCAATCGTG	1705			NA	4 weeks after cessation of FMP consumption
F2T1Pre1	ATCGTACAACTC	2509			NA	1 week before start of FMP consumption
F2T1Pre2	ATCGCTCGAGGA	2402			NA	2 weeks before start of FMP consumption
F2T1Pre4	ATCGCGACAGAT	2352			NA	4 weeks before start of FMP consumption
F2T2FMP1	CGTGATCTCTCC	5860			B1234	1 week into FMP consumption
F2T2FMP2	ATGGATACGGCTC	2433			B1234	2 weeks into FMP consumption
F2T2FMP4	CAAGATCGACTC	3321			B1234	4 weeks into FMP consumption
F2T2FMP7	CTACTGATATCG	2954			B1234	7 weeks into FMP consumption
F2T2Post2	ATGGTCTACTAC	2678			NA	2 weeks after cessation of FMP consumption
F2T2Post4	ATGTACGGCGAC	2446			NA	4 weeks after cessation of FMP consumption
F2T2Pre1	ATGCCTGAGCAG	2495			NA	1 week before start of FMP consumption
F2T2Pre2	ATGCAGCTCAGT	2465			NA	2 weeks before start of FMP consumption

F2T2Pre4	ATGCACTGGCGA	3470	NA	4 weeks before start of FMP consumption
F3T1FMP1	CAACTATCAGCT	2575	B1234	1 week into FMP consumption
F3T1FMP2	CTCATGTACAGT	3037	B1234	2 weeks into FMP consumption
F3T1FMP4	GAGATGCCGACT	5723	B1234	4 weeks into FMP consumption
F3T1FMP7	CAAGTGAGAGAG	2624	B1234	7 weeks into FMP consumption
F3T1Post2	CACACGTGAGCA	2189	NA	2 weeks after cessation of FMP consumption
F3T1Post4	CACAGCTCGAAT	2850	NA	4 weeks after cessation of FMP consumption
F3T1Pre1	CAACAGCCACGA	2852	NA	1 week before start of FMP consumption
F3T1Pre2	CTGAGATACGCG	7378	NA	2 weeks before start of FMP consumption
F3T1Pre4	ATTATCGTGAC	2491	NA	4 weeks before start of FMP consumption
F3T2FMP1	CACGTGACATGT	2628	B1234	1 week into FMP consumption
F3T2FMP2	CATGTCTCTCCG	7791	B1234	2 weeks into FMP consumption
F3T2FMP4	ATCTGAGCTGGT	2923	B1234	4 weeks into FMP consumption
F3T2FMP7	CACCTGATTAG	2895	B1234	7 weeks into FMP consumption
F3T2Post2	CACCTGGTATATC	2842	NA	2 weeks after cessation of FMP consumption
F3T2Post4	GATGTGAGCGCT	2545	NA	4 weeks after cessation of FMP consumption
F3T2Pre1	CACGTCGATGGA	2891	NA	1 week before start of FMP consumption
F3T2Pre2	CACGGACTATAC	2819	NA	2 weeks before start of FMP consumption
F3T2Pre4	CACGACAGGCTA	2683	NA	4 weeks before start of FMP consumption
F4T1FMP1	CGACAGCTGACA	2824	B3456	1 week into FMP consumption
F4T1FMP2	CAGCGGTGACAT	2146	B3456	2 weeks into FMP consumption
F4T1FMP4	ATATGCCAGTGC	2616	B3456	4 weeks into FMP consumption
F4T1FMP7	CAGGTGCTACTA	2078	B3456	7 weeks into FMP consumption
F4T1Post2	CAGTACGATCTT	2618	NA	2 weeks after cessation of FMP consumption
F4T1Post4	CAGTCACTAACG	4127	NA	4 weeks after cessation of FMP consumption
F4T1Pre1	CAGCACTAAGCG	2460	NA	1 week before start of FMP consumption
F4T1Pre2	CGATGTCGTCAA	2650	NA	2 weeks before start of FMP consumption
F4T1Pre4	CAGATACACTTC	2486	NA	4 weeks before start of FMP consumption
F4T2FMP1	CATAGCGAGTTC	3082	B3456	1 week into FMP consumption
F4T2FMP2	CATATACTCGCA	2599	B3456	2 weeks into FMP consumption
F4T2FMP4	CGAATCGACACT	6595	B3456	4 weeks into FMP consumption
F4T2FMP7	CATCAGCGTGTA	2759	B3456	7 weeks into FMP consumption
F4T2Post2	CATCATGAGGCT	2385	NA	2 weeks after cessation of FMP consumption
F4T2Post4	CATCGTATCAAC	2419	NA	4 weeks after cessation of FMP consumption
F4T2Pre1	CGTAAGTCTACT	6754	NA	1 week before start of FMP consumption
F4T2Pre2	CGTGCAATTACA	6693	NA	2 weeks before start of FMP consumption
F4T2Pre4	GACACTCGAATC	2279	NA	4 weeks before start of FMP consumption
F5T1FMP1	ATGTCACCGTGA	2507	B3456	1 week into FMP consumption
F5T1FMP2	CACAGTGGACGT	2423	B3456	2 weeks into FMP consumption
F5T1FMP4	CAGACATTGGCGT	2659	B3456	4 weeks into FMP consumption
F5T1FMP7	CAGTCGAAGCTG	1944	B3456	7 weeks into FMP consumption
F5T1Post2	CATCTGTAGCGA	2785	NA	2 weeks after cessation of FMP consumption
	TCGATACTTGTG		11501	
	TCTACTCGTAAG		11977	
	TCTTAGACGACG		10997	
	TGAGTTCGCTAT		12226	
	TGCTAGTCATAC		11645	

F5T1Post4	CTGTTCTGAG	2241	TCCAGTGGGAGA	13324	NA	4 weeks after cessation of FMP consumption
F5T1Pre1	ATGACTCATTCG	2369	TCATGGTACACT	12299	NA	1 week before start of FMP consumption
F5T1Pre2	ATCCGATCACAG	2423	TCACTGGCAGTA	11973	NA	2 weeks before start of FMP consumption
F5T1Pre4	ATACACGTGGCG	2355	TATCAGGTGTGC	15943	NA	4 weeks before start of FMP consumption
F5T2FMP1	CTAGAACGGCACT	5888	TCGATGAACCTCG	12545	B3456	1 week into FMP consumption
F5T2FMP2	GATCCGACACTA	2697	CTAGGCGTAGTG	11339	B3456	2 weeks into FMP consumption
F5T2FMP4	GCACATCGAGCA	2750	TGATGTGTGACC	11556	B3456	4 weeks into FMP consumption
F5T2FMP7	GCATCGTCAACA	2858	TGATAGTGAGGA	12282	B3456	7 weeks into FMP consumption
F5T2Post2	GCGTTACACACA	2715	TGCTATATCTGG	11091	NA	2 weeks after cessation of FMP consumption
F5T2Post4	CTTGATGCGTAT	2734	TCGCATGAAGTC	12849	NA	4 weeks after cessation of FMP consumption
F5T2Pre1	GACTCGAATCGT	2535	TCCACGTCTCT	9673	NA	1 week before start of FMP consumption
F5T2Pre2	GACAGTTACTGC	2409	TCACCTCTCGCT	8453	NA	2 weeks before start of FMP consumption
F5T2Pre4	CTTAGCACATCA	2530	TATCGCGCGATA	12295	NA	4 weeks before start of FMP consumption
F6T1FMP1	CTCCACATGAGA	6639		B3456	B3456	1 week into FMP consumption
F6T1FMP2	CTGAGCAGAGTC	26252		B3456	B3456	2 weeks into FMP consumption
F6T1FMP4	GCATGTGCATGT	2332		B3456	B3456	4 weeks into FMP consumption
F6T1FMP7	GCTAAGAGAGTA	2941		B3456	B3456	7 weeks into FMP consumption
F6T1Post2	CTTGTCGATA	7042		NA	NA	2 weeks after cessation of FMP consumption
F6T1Post4	GACCACACGAT	2639		NA	NA	4 weeks after cessation of FMP consumption
F6T1Pre1	GAGGCTCATCAT	2948		NA	NA	1 week before start of FMP consumption
F6T1Pre2	GACTGATCAICT	2816		NA	NA	2 weeks before start of FMP consumption
F6T1Pre4	GACATCGGCTAT	2638		NA	NA	4 weeks before start of FMP consumption
F6T2FMP1	CATGGCTACACA	7935		B3456	B3456	1 week into FMP consumption
F6T2FMP2	GCATTGCGTGAG	2599		B3456	B3456	2 weeks into FMP consumption
F6T2FMP4	CGAAGACTGCTG	5210		B3456	B3456	4 weeks into FMP consumption
F6T2FMP7	ATCGATCTGTGG	2037		B3456	B3456	7 weeks into FMP consumption
F6T2Post2	ATGATCGAGAGA	2298		NA	NA	2 weeks after cessation of FMP consumption
F6T2Post4	ATGTGCGACTT	2436		NA	NA	4 weeks after cessation of FMP consumption
F6T2Pre1	CGATCGAGTGT	12510		NA	NA	1 week before start of FMP consumption
F6T2Pre2	GAGTAGCTCGTG	1948		NA	NA	2 weeks before start of FMP consumption
F6T2Pre4	GACTGCATCTTA	2652		NA	NA	4 weeks before start of FMP consumption
F7T1FMP1	ATGTGCACGACT	2087		B3456	B3456	1 week into FMP consumption
F7T1FMP2	CACATCTAACAC	1892		B3456	B3456	2 weeks into FMP consumption
F7T1FMP4	CGGAGTGTCTAT	5110		B3456	B3456	4 weeks into FMP consumption
F7T1FMP7	CATGAGTGCTAC	2333		B3456	B3456	7 weeks into FMP consumption
F7T1Post2	CAGACTGCCAGA	2066		NA	NA	2 weeks after cessation of FMP consumption
F7T1Post4	CGTGACAATGTC	7970		NA	NA	4 weeks after cessation of FMP consumption
F7T1Pre1	ATGAGACTCCAC	2595		NA	NA	1 week before start of FMP consumption
F7T1Pre2	ATCCTCAGTAGT	2462		NA	NA	2 weeks before start of FMP consumption
F7T1Pre4	ATACAGAGCTCC	2375		NA	NA	4 weeks before start of FMP consumption
F7T2FMP1	GACTCACTCAAT	2695		B3456	B3456	1 week into FMP consumption
F7T2FMP2	CTACTACAGGTG	6656		B3456	B3456	2 weeks into FMP consumption

B3456 4 weeks into FMP consumption  
 B3456 7 weeks into FMP consumption  
 NA 2 weeks after cessation of FMP consumption  
 NA 4 weeks after cessation of FMP consumption  
 NA 1 week before start of FMP consumption  
 NA 2 weeks before start of FMP consumption  
 NA 4 weeks before start of FMP consumption

F7T2FMP4 GACCGAGCTATG 3173  
 F7T2FMP7 CTCAGTATG CAG 7838  
 F7T2Post2 CAGTGCATATGC 2312  
 F7T2Post4 CATGCCAGACTGT 2406  
 F7T2Pre1 CAGAGGAGCTCT 2665  
 F7T2Pre2 CACATTGTGAGC 2620  
 F7T2Pre4 GACAGGAGATAG 2954

\* B1234=FMP batches 1,2,3,4; B3456=FMP batches 3,4,5,6

**B. Multiplex shotgun pyrosequencing of total fecal community DNA.**

Sample	Number of Shotgun Reads	Number of Filtered Reads*	Time Point
F1T1FMP4	75249	65574	4 weeks into FMP consumption
F1T1FMP7	55067	47963	7 weeks into FMP consumption
F1T1Post2	47239	41715	2 weeks after cessation of FMP consumption
F1T1Post4	42824	37301	4 weeks after cessation of FMP consumption
F1T1Pre1	45636	39910	1 week before start of FMP consumption
F1T1Pre4	48736	42450	4 weeks before start of FMP consumption
F1T2FMP4	46098	40093	4 weeks into FMP consumption
F1T2FMP7	54286	47471	7 weeks into FMP consumption
F1T2Post2	43826	38725	2 weeks after cessation of FMP consumption
F1T2Post4	57918	50959	4 weeks after cessation of FMP consumption
F1T2Pre1	59486	52254	1 week before start of FMP consumption
F1T2Pre4	64157	55886	4 weeks before start of FMP consumption
F3T1FMP4	67697	57408	4 weeks into FMP consumption
F3T1FMP7	68496	57811	7 weeks into FMP consumption
F3T1Post2	30188	26388	2 weeks after cessation of FMP consumption
F3T1Post4	63158	53287	4 weeks after cessation of FMP consumption
F3T1Pre1	49246	41765	1 week before start of FMP consumption
F3T1Pre4	63488	53593	4 weeks before start of FMP consumption
F3T2FMP4	57136	48173	4 weeks into FMP consumption
F3T2FMP7	67237	56925	7 weeks into FMP consumption
F3T2Post2	62660	52689	2 weeks after cessation of FMP consumption
F3T2Post4	64193	53538	4 weeks after cessation of FMP consumption
F3T2Pre1	52681	45020	1 week before start of FMP consumption
F3T2Pre4	59866	50473	4 weeks before start of FMP consumption
F4T1FMP4	63758	50210	4 weeks into FMP consumption
F4T1FMP7	55671	43728	7 weeks into FMP consumption
F4T1Post2	46447	35815	2 weeks after cessation of FMP consumption

F4T1Post4	44105	34134	4 weeks after cessation of FMP consumption
F4T1Pre1	46610	39617	1 week before start of FMP consumption
F4T1Pre4	61274	50261	4 weeks before start of FMP consumption
F4T2FMP4	45720	38928	4 weeks into FMP consumption
F4T2FMP7	44513	35448	7 weeks into FMP consumption
F4T2Post2	35497	27596	2 weeks after cessation of FMP consumption
F4T2Post4	64160	48809	4 weeks after cessation of FMP consumption
F4T2Pre1	32380	25396	1 week before start of FMP consumption
F4T2Pre4	57289	50679	4 weeks before start of FMP consumption
F5T1FMP4	35261	33254	4 weeks into FMP consumption
F5T1FMP7	68995	65919	7 weeks into FMP consumption
F5T1Post2	47394	44885	2 weeks after cessation of FMP consumption
F5T1Post4	58502	49499	4 weeks after cessation of FMP consumption
F5T1Pre1	62935	57038	1 week before start of FMP consumption
F5T1Pre4	56065	50262	4 weeks before start of FMP consumption
F5T2FMP4	70374	61573	4 weeks into FMP consumption
F5T2FMP7	149079	129330	7 weeks into FMP consumption
F5T2Post2	83963	71274	2 weeks after cessation of FMP consumption
F5T2Post4	212675	184097	4 weeks after cessation of FMP consumption
F5T2Pre1	46874	44073	1 week before start of FMP consumption
F5T2Pre4	85328	72400	4 weeks before start of FMP consumption

\* Sequences used after removing poor quality, duplicate and human sequences



**Tables S3 – S4.**

**Please reference provided CD for these tables.**





Day	Treatment Group	<i>B. cereus</i> (ATCC 43188)	<i>B. ovatus</i> (ATCC 8483)	<i>B. thetaiotaomicron</i> (VPI-5482)	<i>B. uniformis</i> (ATCC 8492)	<i>B. vulgatus</i> (ATCC 8482)	<i>B. WH2</i>	<i>C. aerofaciens</i> (ATCC 25868)	<i>C. scindens</i> (ATCC 3704)	<i>C. sporiforme</i> (DSM 1532)	<i>D. longicatena</i> (DSM 13814)	<i>E. rectale</i> (ATCC 33856)	<i>F. prausnitzii</i> (M21.2)	<i>P. distasonis</i> (ATCC 8505)	<i>R. obeum</i> (ATCC 23974)	<i>R. longus</i> (ATCC 27756)
16	Single	0.48873	0.22793	0.13110	0.20495	0.10806	0.6240	0.10432	0.03221	NA	0.29902	0.68600	NA	0.16074	0.02965	0.01754
16	Multiple	0.16804	0.09007	0.79358	0.48931	0.4280	0.94150	0.15101	0.61665	NA	0.17309	0.99477	NA	0.30406	0.59538	0.87070
17	Single	0.65017	0.18815	0.59192	0.60637	0.08480	0.71282	0.08438	0.51552	NA	0.52828	0.96335	NA	0.05951	0.09153	0.05729
17	Multiple	0.21077	0.01918	0.42488	0.42488	0.47140	0.69846	0.02553	0.82441	NA	0.57306	0.54324	NA	0.07838	0.06435	0.06070
21	Single	0.63669	0.00070	0.35278	0.30876	0.57826	0.33746	0.02653	0.00280	NA	0.61101	0.60619	NA	0.17029	0.04133	0.03627
21	Multiple	0.74006	0.00079	0.54566	0.02926	0.83951	0.23831	0.10292	0.50499	NA	0.52670	0.99894	NA	0.04318	0.86677	0.34378
23	Single	0.54464	0.00394	0.62177	0.24691	0.97794	0.09425	0.01309	0.11191	NA	0.75505	0.89848	NA	0.21648	0.37835	0.40710
24	Multiple	0.38535	0.06148	0.58234	0.33084	0.30410	0.14764	0.00983	0.00955	NA	0.08071	0.93842	NA	0.60339	0.00141	0.00996
28	Single	0.37289	0.06265	0.21753	0.66971	0.69239	0.33746	0.03838	0.00539	NA	0.27336	0.58699	NA	0.39302	0.01192	0.02965
28	Multiple	0.47678	0.02026	0.09967	0.38594	0.73557	0.26504	0.10042	0.03654	NA	0.37554	0.71474	NA	0.85027	0.00014	0.01264
35	Single	0.31021	0.00391	0.05838	0.62679	0.65145	0.02445	0.02445	0.68911	NA	0.40645	0.53927	NA	0.24178	0.01526	0.09903
35	Multiple	0.14026	0.00582	0.05448	0.83935	0.34994	0.01378	0.00564	0.39154	NA	0.78544	0.64365	NA	0.58053	0.00254	0.00050
37	Multiple	0.18459	0.00043	0.04011	0.60975	0.89228	0.01426	0.00494	0.84419	0.37390	0.16761	0.51879	NA	0.56125	0.00313	0.27960
38	Multiple	0.13715	0.11475	0.05484	0.37022	0.97024	0.02855	0.00528	0.12990	NA	0.34561	0.54068	NA	0.81484	0.00029	0.01959
42	Single	0.08678	0.02379	0.00883	0.30916	0.16010	0.10387	0.04868	0.92824	NA	0.17611	0.83539	NA	0.20230	0.97214	0.13903
42	Multiple					0.82101	0.03307	0.00576	0.15807	NA	0.07396	0.63106	NA	0.37634	0.00005	0.05113

Fold-changes (relative to d14)

Day	Treatment Group	<i>B. cereus</i> (ATCC 43188)	<i>B. ovatus</i> (ATCC 8483)	<i>B. thetaiotaomicron</i> (VPI-5482)	<i>B. uniformis</i> (ATCC 8492)	<i>B. vulgatus</i> (ATCC 8482)	<i>B. WH2</i>	<i>C. aerofaciens</i> (ATCC 25868)	<i>C. scindens</i> (ATCC 3704)	<i>C. sporiforme</i> (DSM 1532)	<i>D. longicatena</i> (DSM 13814)	<i>E. rectale</i> (ATCC 33856)	<i>F. prausnitzii</i> (M21.2)	<i>P. distasonis</i> (ATCC 8505)	<i>R. obeum</i> (ATCC 23974)	<i>R. longus</i> (ATCC 27756)
16	Single	-1.20041	1.15266	1.10823	1.10213	1.10810	-1.01462	-1.99315	-1.33599	NA	-2.11772	-1.70594	NA	-1.36850	-1.71191	-2.01683
16	Multiple	-1.19601	1.17481	1.01879	-1.06499	1.10763	-1.02489	-1.63938	1.22915	NA	1.25875	1.00962	NA	-1.15232	-1.16923	-1.05548
17	Single	-1.20299	1.17782	1.03195	1.06747	1.13222	-1.01080	-2.48233	-1.12103	NA	-1.53016	-2.08816	NA	-1.47380	-1.45987	-1.72719
17	Multiple	-1.10387	1.39850	-1.00082	-1.06052	1.09572	-1.01011	-2.08507	1.06725	NA	1.26079	-3.18032	NA	-1.29023	-1.29524	-1.25682
21	Single	-1.03466	1.41658	1.04614	1.20645	-1.03899	-1.02204	-2.98556	-1.75832	NA	-1.44527	-2.01661	NA	-1.31020	-1.84396	-1.79299
21	Multiple	-1.07821	1.41658	1.04614	1.36654	-1.02488	-1.12210	-1.91358	1.34567	NA	1.45808	-1.00194	NA	-1.38810	-1.05021	-1.16569
23	Multiple	-1.09670	1.34762	1.03404	1.20977	-1.00318	-1.04761	-2.90662	-1.32469	NA	-1.13564	-1.11973	NA	-1.22433	-1.15381	-1.14893
24	Multiple	-1.22314	1.25652	1.04881	1.20219	1.12609	-1.05337	-3.10265	-1.95094	NA	-2.57760	-1.11973	NA	-1.10965	-1.99345	-1.82781
28	Single	-1.07502	1.51552	1.17095	-1.13252	1.03429	-1.11521	-3.55045	-1.58638	NA	-1.49447	1.78152	NA	-1.16214	-1.91460	-1.80199
28	Multiple	-1.23134	1.51552	1.23942	1.13536	1.10872	-1.10872	-3.14934	-1.56443	NA	-1.49447	1.78152	NA	-1.05078	-1.88722	-1.49680
35	Single	-1.25933	1.41196	1.25554	-1.09655	-1.04021	-1.11632	-5.32717	1.06035	NA	-1.73955	2.00372	NA	1.21020	-1.78963	1.51788
35	Multiple	-1.13528	1.72682	1.47099	1.03910	1.11938	-1.30328	-6.08542	1.17964	NA	-1.11431	2.16482	NA	-1.10778	-1.41851	1.66878
37	Multiple	-1.25223	1.67565	1.21956	-1.21987	1.01678	-1.17050	-7.18569	1.04146	Inf	-1.88781	3.62518	NA	1.07214	-1.66746	1.65706
38	Multiple	-1.19342	1.68647	1.25351	-1.25774	1.00431	-1.13781	-6.46681	-1.33347	NA	-1.49154	3.22106	NA	1.02112	-1.75902	1.39273
42	Single	-1.51118	1.22045	1.28732	-1.26221	1.22268	-1.21726	-2.97022	1.03645	NA	-3.25388	1.24232	NA	1.34770	-1.01899	1.75489
42	Multiple	-1.27154	1.53847	1.27067	-1.18070	-1.03378	-1.08498	-7.49634	-1.51485	NA	-2.68706	2.23953	NA	1.14681	-2.24881	1.79392


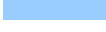

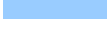
**Tables S6 – S10.**

**Please reference provided CD for these tables.**


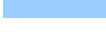

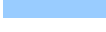
## Table S11.

Table S11. Gene expression profile features with predictive value for classifying mouse fecal samples by treatment status (i.e. pre- versus post-treatment with the FMP consortium).

### A. 'Predictive' and 'highly predictive' KEGG categories.

KEGG category	Importance score	
Germination	0.053574154	 Highly predictive  Predictive
Sporulation	0.048158403	
Cell motility and secretion	0.030330699	
Function unknown	0.020830591	
Amino acid metabolism	0.014304414	
Lipid metabolism	0.007210555	 Highly predictive  Predictive
Metabolism of other amino acids	0.006722455	
Carbohydrate metabolism	0.006280581	
Environmental adaptation	0.003092801	
Cell growth and death	0.002750286	
Pores ion channels	0.002355375	
Translation	0.002096304	
Transcription related proteins	0.002061799	
Transport and catabolism	0.001905376	
Signaling Molecules and interaction	0.001858913	
Transcription	0.001481509	

### B. 'Predictive' and 'highly predictive' KEGG pathways.

KEGG pathway	Importance score	
Atrazine degradation	0.0231	 Highly predictive  Predictive
Novobiocin biosynthesis	0.0197	
Isoquinoline alkaloid biosynthesis	0.0161	
Pentose and glucuronate interconversions	0.0119	
Fatty acid metabolism	0.0080	 Highly predictive  Predictive
Fructose and mannose metabolism	0.0072	
Alanine aspartate and glutamate metabolism	0.0072	
beta-Alanine metabolism	0.0062	
Nitrogen metabolism	0.0060	
Inositol phosphate metabolism	0.0059	
Arachidonic acid metabolism	0.0057	
Carbazole degradation	0.0054	
Taurine and hypotaurine metabolism	0.0053	
Steroid hormone biosynthesis	0.0045	
Sphingolipid metabolism	0.0044	
Phenylalanine metabolism	0.0037	
Pyruvate metabolism	0.0036	
Prion diseases	0.0034	
Carbon fixation in photosynthetic organisms	0.0032	
Streptomycin biosynthesis	0.0029	
Arginine and proline metabolism	0.0024	
Starch and sucrose metabolism	0.0023	
Other glycan degradation	0.0022	
Flavone and flavonol biosynthesis	0.0020	
Alzheimers disease	0.0015	
Selenoamino acid metabolism	0.0014	
Styrene degradation	0.0013	
Translation factors	0.0013	
Sulfur metabolism	0.0013	
Lysosome	0.0012	
Benzoate degradation via hydroxylation	0.0011	

C. 'Predictive' and 'highly predictive' enzyme commission (EC) numbers.

Enzyme commission (EC) number	Importance score		
tRNA-specific adenosine deaminase [EC:3.5.4.-]	0.009528252		Predictive
malate dehydrogenase (oxaloacetate-decarboxylating)(NADP+) [EC:1.1.1.40]	0.00848779		
fructoselysine 6-phosphate deglycase [EC:3.5.-.-]	0.008074793		
aspartate aminotransferase [EC:2.6.1.1]	0.007070721		
L-rhamnose isomerase [EC:5.3.1.14]	0.005562481		
xylose isomerase [EC:5.3.1.5]	0.004887727		
endoribonuclease Dicer [EC:3.1.26.-]	0.003280151		
pectinesterase [EC:3.1.1.11]	0.003260817		
feruloyl-CoA synthase [EC:6.2.1.34]	0.003150745		
inosose dehydratase [EC:4.2.1.44]	0.002515511		
GDPmannose 4,6-dehydratase [EC:4.2.1.47]	0.002275375		
diaminopimelate dehydrogenase [EC:1.4.1.16]	0.002246208		
foldase protein PrsA [EC:5.2.1.8]	0.002009365		
sialidase-1 [EC:3.2.1.18]	0.001700689		
arabinogalactan endo-1,4-beta-galactosidase [EC:3.2.1.89]	0.001672472		
alanine dehydrogenase [EC:1.4.1.1]	0.001672122		
2',3'-cyclic-nucleotide 2'-phosphodiesterase [EC:3.1.4.16]	0.00164777		
threonyl-tRNA synthetase [EC:6.1.1.3]	0.001639447		
transcriptional repressor NF-X1 [EC:6.3.2.-]	0.001557477		
histidinol-phosphate aminotransferase [EC:2.6.1.9]	0.001556015		
4-hydroxyphenylacetate-3-hydroxylase large chain [EC:1.14.13.3]	0.001490383		
GTP cyclohydrolase II [EC:3.5.4.25]	0.001453787		
site-specific DNA-methyltransferase (cytosine-N4-specific) [EC:2.1.1.113]	0.001353193		
GTP cyclohydrolase II [EC:3.5.4.25]	0.001337456		
leucyl-tRNA synthetase [EC:6.1.1.4]	0.001319817		
peptide-methionine (S)-S-oxide reductase [EC:1.8.4.11]	0.001315217		
rubredoxin-NAD+ reductase [EC:1.18.1.1]	0.001287784		
3-oxo-5-alpha-steroid 4-dehydrogenase 3 [EC:1.3.99.5]	0.001245141		
4-hydroxy 2-oxovalerate aldolase [EC:4.1.3.39]	0.001240895		
transcriptional activator TenA [EC:3.5.99.2]	0.001227339		
choline-sulfatase [EC:3.1.6.6]	0.001184942		
threonine 3-dehydrogenase [EC:1.1.1.103]	0.001162346		
lysostaphin [EC:3.4.24.75]	0.001128836		
formyltetrahydrofolate deformylase [EC:3.5.1.10]	0.001064236		
cyclase HisF [EC:4.1.3.-]	0.001021597		

**Tables S12 – S14.**

**Please reference provided CD for these tables.**



**Table S15.****Table S15. List of 127 human gut microbial genomes used to annotate human fecal microbiome datasets.**

Microbial strain name	Genome ID
<i>Actinomyces odontolyticus</i> ATCC 17982	NZ_AAYI00000000
<i>Akkermansia muciniphila</i> ATCC BAA-835	NC_010655
<i>Alistipes putredinis</i> DSM 17216	NZ_ABFK00000000
<i>Anaerococcus hydrogenalis</i> DSM 7454	NZ_ABXA00000000
<i>Anaerofustis stercorihominis</i> DSM 17244	NZ_ABIL00000000
<i>Anaerostipes caccae</i> DSM 14662	NZ_ABAX00000000
<i>Anaerotruncus colihominis</i> DSM 17241	NZ_ABGD00000000
<i>Bacteroides caccae</i> ATCC 43185	NZ_AAVM00000000
<i>Bacteroides capillosus</i> ATCC 29799	NZ_AAXG00000000
<i>Bacteroides cellulosilyticus</i> DSM 14838	NZ_ACCH00000000
<i>Bacteroides coprocola</i> DSM 17136	NZ_ABIY00000000
<i>Bacteroides coprophilus</i> DSM 18228	NZ_ACBW00000000
<i>Bacteroides dorei</i> DSM 17855	NZ_ABWZ00000000
<i>Bacteroides eggerthii</i> DSM 20697	NZ_ABVO00000000
<i>Bacteroides finegoldii</i> DSM 17565	NZ_ABXI00000000
<i>Bacteroides fragilis</i> 3_1_12	NZ_ABZX00000000
<i>Bacteroides fragilis</i> NCTC 9343	NC_003228
<i>Bacteroides fragilis</i> YCH46	NC_006347
<i>Bacteroides intestinalis</i> DSM 17393	NZ_ABJL00000000
<i>Bacteroides ovatus</i> ATCC 8483	NZ_AAXF00000000
<i>Bacteroides plebeius</i> DSM 17135	NZ_ABQC00000000
<i>Bacteroides</i> sp. 1_1_6	NZ_ACIC00000000
<i>Bacteroides</i> sp. D1	NZ_ACAB00000000
<i>Bacteroides</i> sp. D2	NZ_ACGA00000000
<i>Bacteroides stercoris</i> ATCC 43183	NZ_ABFZ00000000
<i>Bacteroides thetaiotaomicron</i> 3731	NC_Bthetaiotaomicron3731
<i>Bacteroides thetaiotaomicron</i> 7330	NC_Bthetaiotaomicron7330
<i>Bacteroides thetaiotaomicron</i> VPI-5482	NC_004663
<i>Bacteroides uniformis</i> ATCC 8492	NZ_AAYH00000000
<i>Bacteroides vulgatus</i> ATCC 8482	NC_009614
<i>Bacteroides</i> WH2	NC_BWH2
<i>Bacteroides xyloxylicus</i> XB1A	NC_BxyloxylicusXB1A
<i>Bifidobacterium adolescentis</i> ATCC 15703	NC_008618
<i>Bifidobacterium adolescentis</i> L2-32	NZ_AAXD00000000
<i>Bifidobacterium angulatum</i> DSM 20098	NZ_ABYS00000000
<i>Bifidobacterium animalis</i> subsp. <i>lactis</i> CNCM I-2494	NC_BanimalisDN1730010
<i>Bifidobacterium animalis</i> subsp. <i>lactis</i> AD011	NC_011835
<i>Bifidobacterium animalis</i> subsp. <i>lactis</i> HN019	NZ_ABOT00000000
<i>Bifidobacterium breve</i> DSM 20213	NZ_ACCG00000000
<i>Bifidobacterium catenulatum</i> DSM 16992	NZ_ABXY00000000
<i>Bifidobacterium dentium</i>	NC_Bdentium
<i>Bifidobacterium gallicum</i> DSM 20093	NZ_ABXB00000000
<i>Bifidobacterium longum</i> DJO10A	NC_010816
<i>Bifidobacterium longum</i> NCC2705	NC_004307
<i>Bifidobacterium pseudocatenulatum</i> DSM 20438	NZ_ABXX00000000

<i>Blautia hansenii</i> DSM 20583	NZ_ABYU00000000
<i>Blautia hydrogenotrophica</i> DSM 10507	NZ_ACBZ00000000
<i>Bryantella formatexigens</i> DSM 14469	NZ_ACCL00000000
<i>Butyrivibrio crossotus</i> DSM 2876	NZ_ABWN00000000
<i>Catenibacterium mitsuokai</i> DSM 15897	NZ_ACCK00000000
<i>Citrobacter youngae</i> ATCC 29220	NZ_ABWL00000000
<i>Clostridium asparagiforme</i> DSM 15981	NZ_ACCJ00000000
<i>Clostridium bartlettii</i> DSM 16795	NZ_ABEZ00000000
<i>Clostridium bolteae</i> ATCC BAA-613	NZ_ABCC00000000
<i>Clostridium hiranonis</i> DSM 13275	NZ_ABWP00000000
<i>Clostridium hylemonae</i> DSM 15053	NZ_ABYI00000000
<i>Clostridium leptum</i> DSM 753	NZ_ABCB00000000
<i>Clostridium methylpentosum</i> DSM 5476	NZ_ACEC00000000
<i>Clostridium nexile</i> DSM 1787	NZ_ABWO00000000
<i>Clostridium ramosum</i> DSM 1402	NZ_ABFX00000000
<i>Clostridium scindens</i> ATCC 35704	NZ_ABFY00000000
<i>Clostridium</i> sp. L2-50	NZ_AAYW00000000
<i>Clostridium</i> sp. M62/1	NZ_ACFX00000000
<i>Clostridium</i> sp. SS2/1	NZ_ABGC00000000
<i>Clostridium spiroforme</i> DSM 1552	NZ_ABIK00000000
<i>Clostridium sporogenes</i> ATCC 15579	NZ_ABKW00000000
<i>Clostridium symbiosum</i>	NC_Csymbiosum
<i>Collinsella aerofaciens</i> ATCC 25986	NZ_AAVN00000000
<i>Collinsella intestinalis</i> DSM 13280	NZ_ABXH00000000
<i>Collinsella stercoris</i> DSM 13279	NZ_ABXJ00000000
<i>Coprococcus comes</i> ATCC 27758	NZ_ABVR00000000
<i>Coprococcus eutactus</i> ATCC 27759	NZ_ABEY00000000
<i>Desulfovibrio piger</i> ATCC 29098	NZ_ABXU00000000
<i>Desulfovibrio piger</i> GOR1	NC_DpigerGOR1
<i>Dorea formicigenerans</i> ATCC 27755	NZ_AAXA00000000
<i>Dorea longicatena</i> DSM 13814	NZ_AAXB00000000
<i>Enterobacter cancerogenus</i>	NC_Ecancerogenus
<i>Escherichia coli</i> str. K-12 substr. MG1655	NC_000913
<i>Escherichia fergusonii</i> ATCC 35469	NC_011740
<i>Eubacterium bifforme</i> DSM 3989	NZ_ABYT00000000
<i>Eubacterium dolichum</i> DSM 3991	NZ_ABAW00000000
<i>Eubacterium eligens</i> ATCC 27750	NC_012778
<i>Eubacterium hallii</i> DSM 3353	NZ_ACEP00000000
<i>Eubacterium rectale</i> ATCC 33656	NC_012781
<i>Eubacterium rectale</i> DSM17629	NC_Erectale_DSM17629
<i>Eubacterium ventriosum</i> ATCC 27560	NZ_AAVL00000000
<i>Faecalibacterium prausnitzii</i> A2-165	NZ_ACOP00000000
<i>Faecalibacterium prausnitzii</i> M21/2	NZ_ABED00000000
<i>Fusobacterium</i> sp. 4_1_13	NZ_ACDE00000000
<i>Fusobacterium varium</i> ATCC 27725	NZ_ACIE00000000
<i>Helicobacter pylori</i> HPAG1	NC_008086
<i>Holdemania filiformis</i> DSM 12042	NZ_ACCF00000000
<i>Lactobacillus casei</i> ATCC 334	NC_008526

Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	NC_008054
Lactobacillus reuteri DSM 20016	NC_009513
Lactococcus lactis subsp. cremoris MG1363	NC_009004
Lactococcus lactis subsp. cremoris SK11	NC_008527
Lactococcus lactis subsp. lactis II1403	NC_002662
M23A	NC_M23A
Methanobrevibacter smithii ATCC 35061	NC_009515
Methanobrevibacter smithii DSM 2374	NZ_ABYV000000000
Methanobrevibacter smithii DSM 2375	NZ_ABYW000000000
Methanosphaera stadtmanae DSM 3091	NC_007681
Mitsuokella multacida DSM 20544	NZ_ABWK000000000
Parabacteroides distasonis ATCC 8503	NC_009615
Parabacteroides johnsonii DSM 18315	NZ_ABYH000000000
Parabacteroides merdae ATCC 43184	NZ_AAxE000000000
Parvimonas micra ATCC 33270	NZ_ABEE000000000
Prevotella copri DSM 18205	NZ_ACBX000000000
Proteus penneri ATCC 35198	NZ_ABVP000000000
Providencia alcalifaciens DSM 30120	NZ_ABXW000000000
Providencia rettgeri DSM 1131	NZ_ACCI000000000
Providencia rustigianii DSM 4541	NZ_ABXV000000000
Providencia stuartii ATCC 25827	NZ_ABJD000000000
Roseburia intestinalis L1-82	NZ_ABYJ000000000
Ruminococcus bromii L263	NC_RbromiiL263
Ruminococcus gnavus ATCC 29149	NZ_AAYG000000000
Ruminococcus lactaris ATCC 29176	NZ_ABOU000000000
Ruminococcus obeum ATCC 29174	NZ_AAVO000000000
Ruminococcus torques ATCC 27756	NZ_AAVP000000000
Shigella sp. D9	NZ_ACDL000000000
Streptococcus infantarius subsp. infantarius ATCC BAA-102	NZ_ABJK000000000
Streptococcus thermophilus CNRZ1066	NC_006449
Streptococcus thermophilus LMD-9	NC_008532
Streptococcus thermophilus LMG 18311	NC_006448
Subdoligranulum variabile DSM 15176	NZ_ACBY000000000
Victivallis vadensis ATCC BAA-548	NZ_ABDE000000000

## **Chapter 3**

### **Human gut microbiome differentiation viewed across cultures, ages and families**

## **Human gut microbiome differentiation viewed across cultures, ages and families**

Tanya Yatsunenکو<sup>1</sup>, Federico E. Rey<sup>1</sup>, Mark J. Manary<sup>2</sup>, Indi Trehan<sup>2</sup>, Maria Gloria Dominguez-Bello<sup>4</sup>, Monica Contreras<sup>5</sup>, Magda Magris<sup>6</sup>, Glida Hidalgo<sup>6</sup>, Robert N. Baldassano<sup>7</sup>, Andrey P. Anokhin<sup>3</sup>, Andrew C. Heath<sup>3</sup>, Barbara Warner<sup>2</sup>, Jens Reeder<sup>9</sup>, Justin Kuczynski<sup>9</sup>, Catherine A. Lozupone<sup>9</sup>, Christian Lauber<sup>9</sup>, Jose Carlos Clemente<sup>9</sup>, Dan Knights<sup>9</sup>, Rob Knight<sup>8,9</sup>, and Jeffrey I. Gordon<sup>1</sup>

<sup>1</sup>Center for Genome Sciences and Systems Biology and Departments of <sup>2</sup>Pediatrics and <sup>3</sup>Psychiatry, Washington University in St. Louis, <sup>4</sup>Department of Biology, University of Puerto Rico Rio Piedras, Puerto Rico <sup>5</sup>Venezuelan Institute of Scientific Research (IVIC), Caracas, <sup>6</sup>Venezuela, Amazonian Center for Research and Control of Tropical Diseases (CAICET), Puerto Ayacucho, Amazonas, Venezuela, <sup>7</sup>Department of Pediatrics, University of Pennsylvania, <sup>8</sup>Howard Hughes Medical Institute and <sup>9</sup>Department of Chemistry and Biochemistry, University of Colorado, Boulder

Address correspondence to: [jgordon@wustl.edu](mailto:jgordon@wustl.edu)

## **Abstract**

Genotypic differences between human populations are typically viewed as consisting of differences in the frequencies of shared *Homo sapiens* alleles. Another source of genetic diversity resides in differences in representation of the millions of genes and myriad gene functions within our gut microbial communities. To address the question of how gut microbiomes differ between human populations, when viewed from the combined perspective of their component microbial lineages<sup>1,2</sup>, encoded metabolic functions<sup>3,4</sup>, stage of host postnatal development<sup>4,8</sup> and environmental exposures, we have conducted a demonstration project characterizing the bacterial species present in fecal samples obtained from 524 infants, children and adults from 147 families, and the community gene content of 110 of their microbiomes. These individuals are from three different countries located on three different continents (Malawi in Africa, Venezuelan Amerindians in South America, and residents of the USA in North America) and exemplify distinctive living environments and cultural traditions. Shared features of the functional maturation of the gut microbiome were identified during the first three years of life in all three populations including, for example, age-associated changes in the representation of genes involved in the biosynthesis and metabolism of vitamins. Pronounced differences in bacterial species assemblages and functional gene repertoires were also noted between individuals residing in the USA compared to the other two countries. These distinctive features, which include differences in vitamin metabolism, are evident in early infancy as well as during adulthood. In addition, the similarity of fecal microbiomes noted among family members extends across cultures. Together, these findings emphasize the importance of considering the microbiome when evaluating the nutritional needs of humans who live in different parts of the world. Moreover, genographic studies of humans should include a longitudinal analysis of the gut microbiome component, in part to understand how the pressures of westernization are changing the microbial components of our genetic, metabolic and developmental landscapes<sup>3,9</sup>.

## **Results and Discussion**

Fecal samples were obtained from healthy individuals in families of Guahibo Amerindians residing in two villages, separated by 10 miles, located near Puerto Ayacucho in the Amazonas State of Venezuela: in Platanillal, diets are dominated by cassava, corn, fruits, fish and sporadically, meats obtained by hunting; in Coromoto, increased consumption of Western-type foods has begun to occur during the past 10 years (**Table S1**). Samples were also procured from members of families living in four rural communities of Malawi located within 10-70 miles of one another (Chamba, Makwhira, Mayaka, Mbiza). Lifestyles in these villages are very similar, and diets are relatively monotonous, dominated by maize<sup>10</sup>. In addition, we sampled families distributed across the USA including the greater metropolitan areas of St. Louis, Philadelphia and Boulder. The sampled populations included parents and siblings, and, in the USA and Malawi, monozygotic (MZ) and dizygotic (DZ) twin pairs. A total of 524 individuals and 147 families were studied: 114 individuals (34 families) from Malawi; 98 individuals (19 families) in Venezuela; and 312 individuals (94 families) from the USA (see **Table S2** for subject characteristics; note that all except 31 adults and one child from the USA were newly recruited for this study).

DNA was prepared from a single fecal sample donated by each person. Variable region 2 (V2) of bacterial 16S rRNA genes present in each fecal community was amplified by PCR and the resulting amplicons were subjected to multiplex pyrosequencing to define the phylogenetic types (phylotypes) present [ $n=3,195\pm 3,002$  (mean $\pm$ SD) pyrosequencer reads/fecal sample; total of 1,679,598 reads]. Species-level bacterial phylotypes were defined as organisms sharing  $\geq 97\%$  nucleotide sequence identity in the V2 region of their 16S rRNA genes<sup>11</sup>. In addition, we characterized functions encoded in community DNA by performing multiplex shotgun pyrosequencing of fecal DNA from a subset of 110 fecal samples, encompassing 43 families with members matched as closely as possible for age [9 Malawian; 6 Amerindian and 28 USA;  $157,036\pm 86,512$  reads/sample; total size of dataset, 5.9 Gb]. The resulting shotgun reads were annotated with KEGG Orthology group (KO)

assignments and with enzyme commission (EC) numbers (KEGG version 58). **Table S2** summarizes all of the datasets incorporated into the present study.

### **Changes in the taxonomic/phylogenetic composition of fecal bacterial communities as a function of age and population**

Many reports have examined the bacterial species content of infants and children within one population using culture-based methods. Far fewer studies have attempted to compare the gut communities of humans living under markedly different socio-economic and cultural traditions<sup>10,11</sup>. Culture-based approaches, although informative, capture only a limited portion of the species diversity present in microbial communities, and, unlike metagenomic analyses, cannot directly provide information about the functional features encoded by gut microbiomes. Culture-independent techniques have been used to define the gut microbiota at various points in postnatal development<sup>6,12</sup>, but have been limited either by the analytic methods used (FISH and DGGE can typically see only the most dominant organisms; DGGE and broad-taxon DNA microarrays cannot specifically identify organisms), by the limited number of subjects examined, or by the scope of the populations surveyed. These studies have nonetheless provided important insights. Using 16S rRNA gene-based microarrays, Palmer et al.<sup>13</sup> observed considerable intra- and interpersonal variation in fecal bacterial community structures during the first year of life in 12 unrelated children and 1 twin pair. Interpersonal variation was less within the twin pair and intrapersonal variation decreased as a function of age.

In order to determine whether there is a consistent pattern of human gut community development, we used a combination of next-generation sequencing approaches to assess microbiota species composition and gene content across ages and as a function of different cultural traditions/geography. Our analytical approach was to use the 16S rRNA data to compare populations in terms of patterns of succession of the microbiota compared across individuals, and to employ shotgun metagenomics to understand the patterns of change in



the representation of gene functions. The resulting datasets were analyzed using a variety of statistical as well as machine learning approaches to identify shared as well as discriminatory species and genes, and to relate change in microbiota to change in microbiome.

To assess the microbiota, we collected bacterial 16S rRNA data from 323 individuals aged 0-17 years (83 Malawian, 64 Amerindian and 176 residents of the USA) plus 201 adults aged 18-70 years (31 Malawians, 34 Amerindians, and 136 residents of the USA). 16S rRNA datasets were first analyzed using UniFrac, a tool that measures similarity between microbial communities based on the degree to which their component taxa share branch length on a bacterial tree of life<sup>14</sup>. There were several notable findings. *First*, the phylogenetic composition of the bacterial community evolved towards an adult-like configuration within the 3-year period following birth in all three populations (**Fig. 1a**, **Fig. S1**). *Second*, interpersonal variation is significantly greater between children than between adults<sup>4</sup>; this finding is robust to geography (**Fig. 1b**). *Third*, there were significant differences in the phylogenetic composition of fecal microbiota between individuals living in the different countries, with especially pronounced separation occurring between USA compared to Malawian and Amerindian microbiota; this was true for individuals aged 0-3 years, 3-17 years, and for adults (**Fig. 1b**, **Table S3**). We also performed unsupervised clustering using Principal Coordinates Analysis (PCoA) of UniFrac distance matrices. The results indicate that age and geography primarily explain the variation in our dataset (**Fig. S2**).

We next used the nonparametric Spearman rank correlation to determine which bacterial taxa change monotonically with increasing age within and between the three sampled populations. We only considered children who were breastfed. Because of biases exhibited by PCR primers used to amplify 16S rRNA genes of certain taxa (e.g., members of the Actinobacteria are not well represented among the V2 region of the 16S rRNA gene amplicons<sup>5</sup>), we mapped shotgun pyrosequencing reads from the fecal microbiomes of the 110 sampled individuals [24 babies (0.6-5 months old), 60 children and adolescents

(6 months to 17 years old) and 26 adults] to 126 sequenced human gut-derived microbial species (listed in **Table S4**). The results are summarized in **Table S5** and **Fig. S3a,b**. The majority ( $75\pm 20\%$ ) of all shotgun sequences in all babies mapped to members of the genus *Bifidobacterium*. Bifidobacteria continued to dominate fecal communities throughout the first year of life although their proportional representation diminished during this period, in agreement with the results of several studies of small numbers of children<sup>4,6,7</sup> (**Fig. S3a**). The advantage of using the 126 gut microbes as a reference database is that spurious hits of shotgun microbiome reads to taxa that are not present in the gut are minimized. Nonetheless, we repeated the entire analysis, blasting against 1280 genomes in KEGG. The results were similar to those obtained using just the 126 gut-derived microbial genomes (**Fig. S3b**). **Table S5a,b** lists the species-level bacterial taxa whose representation increases significantly with age in all three populations, as well as species that change in a population-specific manner as defined from analysis of the shotgun sequencing data that were available from 110 of the 524 individuals.

We used Random Forests, a supervised machine learning technique<sup>15</sup>, and the 16S rRNA datasets obtained from all 524 individuals to identify bacterial species-level operational taxonomic units (OTUs) that identify differences in fecal community composition in children and adults within and between the three populations. The purpose of a classifier such as Random Forests is to learn a function that maps a set of input values or predictors (here, relative OTU abundances in a community) to a discrete output value (here, USA versus non-USA microbiota). Random Forests is a particularly powerful classifier that can exploit non-linear relationships and complex dependencies between OTUs. The measure of the method's success is its ability to correctly classify unseen samples, estimated by training it on a subset of samples, and using it to classify the remaining samples (cross-validation). The cross-validation error is compared to the baseline error that would be achieved by always guessing the most common category. As an added benefit, Random Forests assigns an importance score to each OTU by estimating the increase in error caused by removing

that OTU from the set of predictors. In our analysis, we considered an OTU to be highly predictive if its importance score was at least 0.001; all error estimates and OTU importance scores were averaged over 100 even rarefactions of the sample communities in order to control for sequencing effort. For adults, Random Forests revealed distinct community signatures for Western (USA) and non-Western individuals (baseline error=0.286, cross-validation error=0.020  $\pm$  0.004, 64 highly predictive OTUs). Of the 64 highly predictive OTUs shown in **Fig. S4** and **Table S6**, 58 were over-represented in non-USA adults, and 44 of the 58 were assigned to the genus *Prevotella* or family *Prevotellaceae*. Malawians and Amerindians could also be distinguished from each other, although the difference was less extreme than the USA versus non-USA comparison (baseline error=0.407, cross-validation error=0.089  $\pm$  0.027, 27 highly predictive OTUs). There were only small discernable differences between infants in the above comparisons, and between adults living in the two Amerindian villages (cross-validation error greater than or equal to half of baseline error in all cases). Thus, a Western (USA) lifestyle appears to affect the bacterial component of the gut microbiota significantly, although this influence is not detectable against the high degree of variability observed in infants and children. Although the *Prevotella* were the most discriminatory lineages, removing the entire family of *Prevotellaceae* increased the classification error only slightly, all 20 of the non-Prevotellaceae OTUs are still predictive, and the average decrease in predictive accuracy when they are removed is <0.1%. Thus, as in the case of the Bifidobacteria, the Prevotellaceae provide a major component of the effect we report, but by no means all of the effect.

Confirming the importance of *Prevotella* as a discriminatory taxon, a recent study also showed that abundance of this genus was present at higher levels in the fecal microbiota of children living in West Africa (Burkina Faso) compared to children living in Europe (Italy)<sup>10</sup>. Additionally, a member of this genus is one of three bacterial species that, in European adults, distinguishes strongly among three clusters, or enterotypes, of gut microbiota configurations that are claimed to be reproducible across Western adult

populations<sup>16</sup>. Therefore, we asked whether the fecal microbiota of infants and adults in each of our three geographically and distinct populations fell into natural discrete clusters. We did not find evidence for discrete clustering (see *Methods*), but rather for continuous variation driven in adults by a trade-off between *Prevotella* and *Bacteroides*, as previously observed<sup>17</sup>. Although Western and non-Western populations tended to occupy the *Bacteroides*-rich and *Prevotella*-rich ends of the gradient, respectively, truncated sections of the gradient were reproduced in each of the three sub-populations we studied (**Fig. S5a-c**). Including infants introduces a new, strongly supported gradient driven by *Bifidobacteria*, generally orthogonal to the *Bacteroides/Prevotella* gradient. Clustering of sub-populations of increasing minimum age indicates that adult cluster membership is generally consistent, but that children between 0.6 years and 1 year of age may be clustered with adults or with younger children, depending on whether the younger children are included in the analysis (**Fig. S5d-e**).

*Non-bacterial members of the fecal microbiota* — Shotgun sequences were used to query the NCBI non-redundant nucleotide database (Blastn threshold E-value $<10^{-5}$ ) to identify the representation of organisms that belong to domains other than Bacteria in the 110 fecal microbiomes. Across all samples,  $7\pm 8\%$  of reads mapped to non-bacterial sequences. The majority of these sequences belonged to Archaea and Fungi. **Fig. S6** shows that the proportional representation of archaeal sequences is significantly higher in adults compared to children  $\leq 3$  years of age in Malawi and the USA (Mann-Whitney test;  $p < 0.05$ ; note that the differences between age groups were not statistically significant among Amerindians). More than 99% of these archaeal sequences mapped to the methanogen *Methanobrevibacter smithii*, previously shown to be the dominant archaeon in USA population<sup>18,19</sup>. The representation of fungi was significantly higher in adults compared to children in all populations; among adults, fungal sequences were significantly higher in Malawian and Amerindian versus USA microbiomes (see **Fig. S6** for the most abundantly represented taxa out of all non-bacterial sequences). As the databases of gut-associated genomes ex-

pand, it is likely that additional sequences may map to other archaea and eukaryotes.

### **Shared functional changes in the microbiome as children mature**

Very few studies have described changes in the gene content of the gut microbiome as a function of age: the largest study reported to date was carried out in 13 healthy Japanese individuals (5 children, the youngest 3-months-old, and 8 adults)<sup>4</sup>. Our dataset of 110 individuals allowed us to characterize the representation of functional gene groups [KEGG Orthology (KO) annotations and Enzyme Commission numbers (ECs)] in the microbiomes representing broader age groups (youngest 3 weeks), and several distinct geographic locations/cultural traditions. We used Hellinger distance measurements to show that just as children are significantly more different from one another than are adults in terms of their fecal bacterial community phylogenetic structure, they are also more different in terms of their repertoires of microbiome-encoded functions, as defined by the proportional representation of EC and KO assignments (**Fig. 1b,c, Fig S7, Table S3**). Moreover, as with UniFrac distances, Hellinger distances were greater between the USA and the other two populations at all ages sampled (**Fig. 1b,c, Fig S8, Table S3**). Of interest is the concordance between the two data types: accordingly, we used Procrustes analysis<sup>20</sup>, which is a method of comparing the goodness of fit between two point clouds, scaling and rotating the first point cloud to align with the second in order to test whether the relative orientation of each point is preserved in the two datasets. The goodness of fit was highly significant result ( $P < 0.001$  with 1,000 iterations) whether UniFrac (the most appropriate metric for 16S rRNA data) or Hellinger distance (for consistency with the method used on the KEGG EC and KO data) was used to reduce the OTU table (**Fig. S9** plus data not shown). COG annotations also produced similar concordance with 16S rRNA datasets (**Fig. S9**).

When examining EC profiles across 110 fecal microbiomes, we obtained the remarkable result that of the 1,349 ECs identified in the sampled populations, none was uniquely present in all adults ( $n=26$ ) but not in babies ( $n=24$ ), or in all babies but not adults.

Moreover, the total number of ECs found in adults was not significantly different compared to the total number of ECs scored in babies (sampling normalized to coverage in **Fig. S7a**). This finding was robust to culture/geography. The *fraction* of sequences with assignable KEGG EC annotations declined with increasing age in all three populations (**Fig. S7b**). This may be due to the increased complexity of the adult microbiome, with fewer representative genomes sequenced.

We used ShotgunFunctionalizerR<sup>21</sup>, a software tool designed for metagenomic analysis and based on a Poisson model, to identify 1008 ECs whose proportional representation in fecal microbiomes differed significantly between all sampled babies and all adults irrespective of their geographic location; 530 of the 1008 ECs were significantly higher in adults ( $p < 0.0001$ , **Table S7**). A prominent example of these shared age-related changes involves vitamins B12 (cobalamin) and folate metabolisms. In contrast to folate, which is synthesized by both microbes and plants, vitamin B12 is produced primarily by microbes<sup>19</sup>. The gut microbiomes of breast-fed babies are enriched in genes involved in the *de novo* biosynthesis of folate, while those of adults have a significantly higher representation of genes that metabolize dietary folate and its reduced form tetrahydrofolate (THF, **Fig. 2a**, **Fig. S10**, **Table S7**). Unlike *de novo* folate biosynthetic pathway components, which decrease with age, the proportional representation of genes encoding the majority of enzymes involved in cobalamin biosynthesis increase with age (**Fig 2b**, **Fig S11**, **Table S7**). The folate and cobalamin pathways are linked functionally: methionine synthase (EC 2.1.1.13) catalyzes formation of THF from 5-Methyl-THF and L-homocysteine, and requires cobalamin as a cofactor (**Fig. 2a**, **Fig S10**). Methionine synthase gene representation in the microbiome also increases with age (**Fig. S10**).

The low relative abundance of ECs involved in cobalamin biosynthesis in the microbiomes of babies correlates with the lower representation of members of Bacteroidetes, Firmicutes, and the archaeon *M. smithii* in their fecal microbiota (see **Fig. S12** for Spearman correlation coefficients). While the biosynthetic pathway for cobalamin is well rep-

resented in the genomes of these organisms (**Fig. S12**), *Bifidobacterium*, *Streptococcus*, *Lactococcus*, *Lactobacillus* which dominate the baby gut microbiota (**Table S5, Fig. S3**), are deficient in these genes (**Fig. S12**). In contrast, a number of these early gut colonizers contain ECs involved in folate biosynthesis/metabolism (**Fig. S12**). The traditional view of the developing infant gut is that the principal change is in the representation of Bifidobacteria. Our analysis indicated that although differences in representation of Bifidobacteria contribute to this effect, differences in vitamin metabolism among the rest of the bacteria remain even when all Bifidobacteria sequences are excluded (**Table S7b**). The changes in vitamin biosynthetic pathway representation in the microbiome correlate with published reports indicating that blood levels of folate decrease and vitamin B12 increase as babies age<sup>22,23</sup>.

Besides cobalamin and folate, the relative abundance of ECs involved in the biosynthesis of vitamins B7 (biotin) (biotin synthase, EC2.8.1.6) and thiamine (thiamine-phosphate diphosphorylase, EC2.5.1.3) are significantly higher in adult microbiomes compared to the microbiomes of babies (**Fig. 2c, Table S7**). Together, these findings suggest that the microbiota should be considered when assessing the nutritional needs of humans at various stages of development.

Random Forests asks a somewhat different statistical question from ShotgunFunctionalizeR: i.e., which genes or species are most discriminatory among different class labels, rather than which are most over/underrepresented, and tends to identify fewer features than does ShotgunFunctionalizeR when applied to the same data. Nevertheless, this complementary approach identified exactly the same major biological patterns. Random Forests analysis yielded 107 ECs that best discriminate the adult and baby microbiomes (**Table S7**, see description of the method above); these predictive ECs were among the most significantly different ECs determined by ShotgunFunctionalizeR and included ECs involved in the metabolisms of vitamin B12 and folate (**Fig. 2c, Table S7**). Random Forests revealed that ECs involved in fermentation, methanogenesis and metabolism of ar-

ginine, glutamate, aspartate and lysine were higher in the adult microbiomes, while ECs involved in the metabolism of cysteine and a fermentation pathway found in lactic acid bacteria [acetolactate decarboxylase (EC4.1.1.5) and 6-phosphogluconate dehydrogenase (EC1.1.1.4)] were represented primarily in the baby microbiomes.

When we compared representation of KOs (instead of ECs) between babies and adult microbiomes, we obtained essentially same results as reported with ECs. The only novel finding was the overrepresentation of KOs assigned to a wide variety of ABC transporters in baby microbiomes (**Table S7c**).

### **Population- and age-specific differences in the representation of microbiome functions**

ShotgunFunctionalizeR and Spearman rank correlation analyses were both used to compare EC representation in fecal microbiomes as a function of predefined categories of geographic location and age. 476 ECs were identified as being significantly different in the USA versus Malawian and Amerindian babies ( $p < 0.0001$ , ShotgunFunctionalizeR; **Table S8**). The most prominent differences involved pathways related to vitamin biosynthesis and carbohydrate metabolism. Malawian and Amerindian babies had higher representation of ECs that were components of vitamin B2 (riboflavin) biosynthetic pathway (**Fig. 3a,b**). These differences were not evident in adults (**Table S7**). Riboflavin is found in human milk, meat and dairy products. We did not measure the levels of these vitamins in mothers and in their breast milk in the sampled populations, although it is tempting to speculate the observed differences in baby microbiomes may represent an adaptive response to vitamin availability.

Studies in gnotobiotic mouse models indicate that the ability of members of the microbiota to access host-derived glycans plays a key role in establishing a gut microbial community<sup>24,25</sup>. As expected<sup>4,5</sup>, compared to adults, the baby microbiomes were enriched in ECs involved in foraging of glycans represented in mother's milk and the intestinal mucosa (mannans, sialylated glycans, galactose, fucosyloligosaccharides; **Table S7**). A number of



genes involved in utilizing these host glycans are significantly overrepresented in Amerindian and Malawian compared to USA baby microbiomes, most notably exo-alpha-sialidase and alpha-L-fucosidase (**Fig. 3a, Table S8**). These population-specific biomarkers may reflect differences in the glycan content of breast milk. In fact, the representation of these glycoside hydrolases decreases as Malawian and Amerindian babies mature and transition to a diet dominated by maize-, cassava- and other plant-derived polysaccharides. In contrast, alpha-fucosidase gene representation increases as USA infants age and are exposed to diets rich in readily absorbed sugars (**Fig. S13, Table S9**).

Another biomarker that distinguishes microbiomes based on age and geography is urease (EC3.5.1.5). Urease gene representation is significantly higher in Malawian and Amerindian baby microbiomes and decreases with age in these two populations, unlike in the USA where it remains low from infancy through adulthood (**Fig. 3a, Fig. S13**). Urea comprises up to 15% of the nitrogen present in human breast milk<sup>26</sup>. Urease releases ammonia that can be used for microbial biosynthesis of essential and nonessential amino acids<sup>27,28</sup>. Furthermore, urease plays a major role in nitrogen recycling, particularly when diets are deficient in protein<sup>29,30</sup>. Under conditions where dietary nitrogen is limiting, the ability of the microbiome to utilize urea would presumably be advantageous to both the microbial community and host. Urease activity has been characterized previously in *Streptococcus thermophilus*<sup>31</sup>. While most attribute urease to *Helicobacter* and *Proteus spp.*, the relative abundance of members of these two genera was low (<0.05%) and not significantly different between the three populations. Our analysis of metagenomic reads that matched to the 126 gut genomes revealed that the representation of five species that possess EC3.5.1.5 (*Bacteroides WH2*, *Coprococcus comes*, *Roseburia intestinalis*, *Streptococcus infantarius* and *S. thermophilus*) was significantly higher in Malawian and Amerindian compared to USA baby microbiomes (**Table S5**).

Random Forests analysis again confirmed these results, showing that the best predictors of USA vs Malawian/Amerindian baby microbiomes (**Table S8**) were among the

most significant ECs determined by ShotgunFunctionalizeR.

*Effects of breast milk versus formula feeding in USA twins* — In the analyses described above, we only considered infants who were breastfed. Epidemiologic studies have shown that formula feeding is more common in the USA than in a number of developing countries<sup>32,33</sup>. Therefore, we compared shotgun sequences generated from the fecal microbiomes of 4 USA twin pairs where both co-twins were breast-fed, and 4 USA age-matched (2-5 month old) twin pairs that were formula-fed. Formula-fed babies contained significantly fewer sequences that mapped to Bifidobacteria genomes, and more taxa belonging to the Firmicutes and Bacteroidetes compared to their breast-fed counterparts ( $p < 0.0001$ ; Mann-Whitney test; **Fig S14**). We identified 244 ECs whose proportional representation differentiated formula- and breast-fed microbiomes ( $p < 0.0001$ , ShotgunFunctionalizeR; **Fig. S15, Table S10**). The majority of the 170 genes that were overrepresented in formula-fed fecal microbiomes were involved in various aspects of carbohydrate metabolism (e.g., fructose, mannose) as well as nitrogen and amino acid metabolism (e.g., lysine biosynthesis). The representation of genes involved in biosynthesis of cobalamin and folate in formula-fed babies phenocopies what is observed in adults, i.e., the proportion of genes involved in the generation of cobalamin is higher and the representation of genes that participate in *de novo* folate synthesis is significantly lower than in breast-fed infants (**Fig. 2, Fig. S11**). These findings highlight the need to use the types of biomarkers we have identified to conduct longitudinal metagenomic studies comparing the development of the microbiomes of formula- versus breast-fed individuals. The goal would be to determine whether differences between formula- and breast-fed children persist through adulthood, and the extent to which early exposure to formula heralds microbiome-encoded metabolic programs that confer human physiologic phenotypes distinct from those of breast-fed children (e.g., ref. 33).

*Differences in adult fecal microbiomes associated with geography* — Annotation of the shotgun sequencing datasets yielded a total of 1,349 ECs in the 26 adults surveyed:

ShotgunFunctionalizeR revealed that the representation of genes encoding 893 of these ECs were significantly different in USA versus Malawian/Amerindian fecal microbiomes ( $p < 0.005$  after multiple comparison correction; 433 overrepresented in USA samples). By contrast, at this threshold only 445 ECs were identified as different between Malawian and Amerindian adults (see **Table S11** for a complete list). A USA diet is rich in protein, while diet in Malawi and Amerindian populations are dominated by corn and cassava (see **Table S1** for the results of dietary surveys of Amerindians and Malawians). The differences between USA versus Malawian/Amerindian microbiomes can be related to these differences in their diets. Genes encoding ECs whose representation are most significantly enriched in USA fecal microbiomes parallel differences observed in carnivorous versus herbivorous mammals<sup>34</sup>: ECs encoding glutamate synthase and glutamine synthase are higher in proportional representation in Malawian and Amerindian adult microbiomes and are also higher in herbivorous mammalian microbiomes<sup>34</sup> (**Fig. 3c**), while degradation of glutamine was overrepresented in USA as well as carnivorous mammalian microbiomes. Several ECs involved in the degradation of other amino acids were overrepresented in adult USA fecal microbiomes: aspartate (EC4.1.1.12), proline (EC1.5.99.8), ornithine (EC2.6.1.13) and lysine (EC5.4.3.2) (**Fig. 3c**), as were ECs involved in catabolism of simple sugars (glucose-6-phosphate dehydrogenase, 6-phosphofructokinase), sugar substitutes (L-iditol 2-dehydrogenase, which degrades sorbitol), as well as host glycans (alpha-mannosidase, beta-mannosidase, alpha-fucosidase, **Fig. 3c**). In contrast, alpha-amylase (EC 3.2.1.1), which participates in the degradation of starch, was overrepresented in the Malawian and Amerindian microbiomes, reflecting their corn-rich diet.

USA microbiomes also had significant overrepresentation of ECs involved in the biosynthesis of vitamins B12 (**Fig. 2,3c**), lipoic acid and biotin (**Fig. 3c**), the metabolism of xenobiotics [phenylacetate CoA ligase (EC 6.2.1.30) which participates in the metabolism of aromatic compounds) and mercury reductase (EC1.16.1.1)], and choloyglycine hydrolase (EC3.5.1.24) which metabolizes bile salts (the latter may reflect diets that are higher

in fat) (**Fig. 3c**).

The Random Forests classifier again confirmed these results, revealing that all 52 ECs that were the best at discriminating USA versus non-USA adult microbiomes were among the most significantly different identified by ShotgunFunctionalizeR (**Table S12**).

### **Effects of kinship on the microbiome across countries**

The definition of the family is a key parameter that differs among societies, and differences in social structures may influence the extent of vertical transmission of the microbiota and the flow of microbes and microbial genes among members of a household. Differences in cultural tradition also affect food, exposure to pets and livestock, and many other factors that could influence how and from where a gut microbiota/microbiome is acquired. We previously observed that MZ twins are no more similar to one another in gut bacterial community structure than DZ twins for adults living apart in the USA<sup>35</sup>. This result suggests that the *overall* heritability of the microbiome is low. We confirmed that the phylogenetic architecture of the fecal microbiota of MZ Malawian co-twins  $\leq 3$  years of age is not more similar than the microbiota of similarly aged DZ co-twins (n=15 MZ and 6 DZ twin pairs). We found that this is also true for MZ and DZ twin pairs aged 1-12 months of age (n=16 twin pairs), as well as teenaged twins (13-17 years-old; n=50 pairs) living together in the USA (**Fig. 4**). Although biological mothers are in a unique position to transmit an initial inoculum of microbes to their infant during and following birth, our analysis of mothers of teenage USA twins showed that their fecal microbiota were no more similar to their children than were biological fathers and that genetically unrelated but co-habiting mothers and fathers were significantly more similar to one another microbially than were members of different families (**Fig. 4**; note that no fathers were sampled in Malawi and only 4 fathers in the Amerindian cohort). These latter observations emphasize the importance of a history of numerous common environmental exposures in shaping gut microbial ecology. Moreover, the similarity in overall pattern of the effects of kinship on microbial community

structure suggests that despite the large influence of cultural factors on which microbes are present in both children and adults in each population, the bases for the degree of similarity among members of a family are consistent across the three populations studied.

The high similarity of Malawians and Amerindians microbiomes is remarkable, given the large geographic (and genetic) distances between these populations, and also implies a major influence of environment (diet) on the structure of gut microbiome. One question is the extent to which the recent deep sampling of 124 adult European microbiomes by the MetaHIT consortium (2-7.3 Gbp of shotgun sequence/fecal sample; ref. 2) represents the gene content present in the microbiomes of all modern humans of all ages. Accordingly, we tested the extent to which this gene catalog recruited reads from each of our subjects, using the 90% nucleotide identity criterion that Meta-HIT employed to identify reads as belonging to the same gene in the same microbial species<sup>2</sup>. On average, 91% of reads from the fecal microbiomes of adults living in the USA, 81% from Amerindian adult microbiomes, and 76% from Malawian adult microbiomes mapped to Meta-HIT; the corresponding numbers for children below three years of age were 79%, 72% and 78% respectively (**Fig. S15**); additionally, individuals from the MetaHIT European cohort cluster with the USA population we studied (**Fig. S16**).

Together, our results emphasize that it is essential to sample a broad population of healthy humans over time, both in terms of their age, geography and cultural traditions, in order to discover features of our microbiomes that are unique to different living circumstances. The continuous pattern of variation we observed with enterotype analysis suggests that while some features of normal variation in the human gut microbiota, such as the *Prevotella/Bacteroides* gradient, are highly reproducible even in human population subsets of reduced variability, a full accounting of the directions in which the human gut microbiota can vary will require a substantially broader cross-cultural and cross-age sampling. In addition, we need to understand how the pressures of westernization are changing the microbial parts of our genetic landscape — changes that potentially mediate the suite

of pathophysiological states (obesity, diabetes, etc.) correlated with Westernization, and changes that may influence which populations are chosen for clinical trials that test various pharmacologic agents<sup>36</sup>. In the same way that extensive efforts exist to preserve the cultural, linguistic and genetic heritage of threatened and/or assimilating populations, we must preserve humanity's microbiological heritage<sup>9</sup>: this diversity may provide fertile grounds for bio-prospecting for microbial genes and species lost through antibiotics or the Western diet that could, if restored, counterbalance some disturbing trends in global human health. Finally, given the need for global policies about sustainable agriculture and improved nutrition, it will be important to understand how we can match these policies not only to our varying cultural conditions but also to our varied gut microbiomes.

## **Methods**

**Subjects** — Subjects were recruited for the present study using procedures approved by Human Studies Committees from Washington University, the University of Pennsylvania, the University of Colorado, Boulder, the University of Malawi, the University of Puerto Rico, and the Venezuelan Institute for Scientific Research (IVIC). Subject characteristics are summarized in **Table S2**.

**Isolation of fecal DNA and multiplex pyrosequencing** — Each participant provided a fecal specimen that was frozen within 30 min. All samples were stored at -80°C prior to metagenomic analyses. Moreover, all fecal samples were subjected to a common protocol for DNA extraction. Fecal samples were pulverized with a mortar and pestle at -80°C. Genomic DNA was extracted from 400 mg aliquots of frozen pulverized samples. Amplification of amplicons from the V2 region was carried as described<sup>10</sup>.

For multiplex shotgun 454 Titanium FLX pyrosequencing, each fecal community DNA sample was randomly fragmented by nebulization 500-800 bp and then labeled with a distinct MID (**M**ultiplex **I**dentifier; Roche) according the manufacturer's protocol (Rapid

Library preparation for FLX Titanium). Equivalent amounts of 12 MID-labeled samples were pooled prior to each pyrosequencer run.

### **Data analysis**

Pyrosequencing reads were demultiplexed and binned by sample according to their barcodes. Reads with a low quality region [i.e. 50 or more consecutive nucleotides with a quality score below 25] were truncated at the beginning of the window, and reads shorter than 150 nucleotides were discarded. Reads were classified into OTUs based on a 97% sequence identity threshold using uclust<sup>36</sup>, and each OTU was assigned taxonomic information using the RDP naïve Bayesian classifier<sup>37</sup>. The set of representative sequences for the OTUs were aligned using PyNAST, and a *de novo* taxonomic tree of the sequences was constructed from the alignment based on the degree of sequence similarity.

16S rRNA amplicon sequences were processed using the QIIME (v2.1) suite of software tools<sup>38</sup>. A table of OTU counts per sample was generated and used in combination with the tree to calculate alpha and beta diversity. To generate unweighted UniFrac distance matrices, all communities were rarefied to 500 16S rRNA reads/sample. Unweighted UniFrac rather than weighted UniFrac was used for analyses due to the large differences in taxonomic representation among the samples. Nonetheless, the patterns were similar with weighted UniFrac (data not shown).

*Enterotype analysis* — Enterotype testing was performed on the rarefied versions of the 16S rRNA OTU relative abundance tables. OTU counts were binned into genus-level taxonomic groups according to the taxonomic assignments discussed above. Several distance measures were considered, including Jensen-Shannon divergence, Bray-Curtis, and weighted/unweighted UniFrac distances. Clustering was performed via partitioning around medoids in the R package “cluster”<sup>39</sup>. The choice of number of clusters and quality of the resulting clusters were assessed by maximizing the silhouette index<sup>40</sup>. Traditionally, silhouette indices of 0.5 or above have been considered evidence of reasonable cluster-

ing structure. Although some silhouette scores above 0.5 were found in this data set (e.g. for two clusters when clustering all adult populations with Jensen-Shannon divergence), re-clustering within different subpopulations (e.g. individual countries) introduced new cluster boundaries with silhouette scores still near or above 0.5, indicating that silhouette index scores may need to be substantially above 0.5 to claim clustering structure for microbial enterotype testing.

*Shotgun sequences from fecal microbiomes* – Shotgun reads were filtered using custom Perl scripts and publicly available software to remove (i) all reads <60 nt, (ii) Titanium reads with two continuous and/or three total degenerate bases (N), (iii) all duplicates (a known artifact of pyrosequencing), defined as sequences whose initial 20 nucleotides are identical and that share an overall identity of >97% throughout the length of the shortest read<sup>41</sup> and (iv) all sequences with significant similarity to human reference genomes (BLASTN with e-value  $\leq 10^{-5}$ , bitscore  $\geq 50$ , percent identity  $\geq 75\%$ ) to ensure the continued de-identification of samples.

Searches against the database of 126 human gut bacterial genomes were conducted with Blastn. A sequence read was annotated as the best hit in the database if the E-value was  $\leq 10^{-5}$ , the bit score was  $\geq 50$ , and the alignment was at least 95% identical between query and subject. Relative abundances of reads mapped to each of the 126 genomes were adjusted to genome sizes. Searches against protein-coding component of the KEGG database (v58) and against COG (v8.3) were conducted with BLASTX. (Note that when we performed searches against a separate KEGG database of intergenic regions alone, very few hits were observed). Counts were normalized to the mapped reads.  $40\pm 8\%$  reads were mapped to KEGG KOs and  $56\pm 11\%$  to COG.  $44\pm 16\%$  of the reads mapped to the 126 gut genomes using 95% sequence similarity cut-off. Unmapped reads were excluded from the analyses shown in the main text, although repeating analyses including these reads had little effect on the results. To quantify the differences in KEGG EC profiles among fecal microbiomes, evenly rarefied matrices of EC counts were created with all samples, and



Hellinger distances were calculated using QIIME.

Spearman rank correlations were carried out using the R statistical software<sup>42</sup>. To identify bacterial taxa that change with increasing age in each population, the proportion of reads that map to each of the 126 reference sequenced human gut genomes in each fecal microbiome was identified. The relative abundance of reads from each genome was then correlated with age (years) for each geographic region. To identify genes encoding ECs that change with age, the proportion of reads annotated with each EC in each fecal microbiome was identified. The relative abundance of each EC was subsequently correlated with age (years) for each geographic region.

### **Random Forests Analysis**

Random Forests analysis was applied as described in<sup>8</sup>, using the randomForest package in R<sup>43</sup>, with 500 trees and all default settings. Generalization error was estimated using 5-fold cross-validation for all comparisons involving adults from the 16S rRNA data; leave-one-out cross-validation was used for all other comparisons. For each comparison, the relevant subset of samples was extracted from the table of OTU or EC counts, and all singleton OTUs/ECs (or all OTUs/ECs present in fewer than 5 samples for the 16S rRNA comparisons involving adults) were subsequently removed. Random Forests analysis was performed for each comparison on 100 rarefied versions of the data, and the average cross-validation error estimates and OTU/EC importance estimates were reported. Rarefaction depths were chosen manually to include all samples without exceptionally low total sequences. The chosen depth for each comparison and the resulting number of samples are shown in **Table S6** and **Table S12**.

## References

- 1 Mueller, S. *et al.* Differences in fecal microbiota in different European study populations in relation to age, gender, and country: a cross-sectional study. *Appl Environ Microbiol* **72**, 1027-1033, doi:72/2/1027 [pii]10.1128/AEM.72.2.1027-1033.2006 (2006).
- 2 Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65, doi:nature08821 [pii]10.1038/nature08821.
- 3 Li, M. *et al.* Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci USA* **105**, 2117-2122, doi:0712038105 [pii]10.1073/pnas.0712038105 (2008).
- 4 Kurokawa, K. *et al.* Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* **14**, 169-181, doi:10.1093/dnares/dsm018 (2007).
- 5 Koenig, J. E. *et al.* Microbes and Health Sackler Colloquium: Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A*, doi:1000081107 [pii]10.1073/pnas.1000081107.
- 6 Favier, C. F., Vaughan, E. E., De Vos, W. M. & Akkermans, A. D. Molecular monitoring of succession of bacterial communities in human neonates. *Appl Environ Microbiol* **68**, 219-226 (2002).
- 7 Tannock, G. W. What immunologists should know about bacterial communities of the human bowel. *Semin Immunol* **19**, 94-105, doi:10.1016/j.smim.2006.09.001 (2007).
- 8 Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci U S A* **107**, 11971-11975, doi:1002601107 [pii]10.1073/pnas.1002601107.

- 9 Blaser, M. J. & Falkow, S. What are the consequences of the disappearing human microbiota? *Nat Rev Microbiol* **7**, 887-894, doi:nrmicro2245 [pii] 10.1038/nrmicro2245 (2009).
- 10 De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci U S A* **107**, 14691-14696, doi:1005963107 [pii]10.1073/pnas.1005963107.
- 11 Peach, S., Fernandez, F., Johnson, K. & Drasar, B. S. The non-sporing anaerobic bacteria in human faeces. *J Med Microbiol* **7**, 213-221 (1974).
- 12 Mackie, R. I., Sghir, A. & Gaskins, H. R. Developmental microbial ecology of the neonatal gastrointestinal tract. *The American journal of clinical nutrition* **69**, 1035S-1045S (1999).
- 13 Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol* **5**, e177, doi:07-PLBI-RA-0129 [pii]10.1371/journal.pbio.0050177 (2007).
- 14 Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**, 8228-8235, doi:71/12/8228 [pii]10.1128/AEM.71.12.8228-8235.2005 (2005).
- 15 Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. *FEMS Microbiol Rev* **35**, 343-359, doi:10.1111/j.1574-6976.2010.00251.x (2011).
- 16 Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature*, doi:10.1038/nature09944 (2011).
- 17 Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105-108, doi:10.1126/science.1208344 (2011).

- 18 Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308**, 1635-1638, doi:1110591 [pii]10.1126/science.1110591 (2005).
- 19 Hansen, E. E. *et al.* Microbes and Health Sackler Colloquium: Pan-genome of the dominant human gut-associated archaeon, *Methanobrevibacter smithii*, studied in twins. *Proc Natl Acad Sci U S A*, doi:1000071108 [pii]10.1073/pnas.1000071108 (2011).
- 20 Gower, J. C. Generalized Procrustes Analysis. *Psychometrika* **40**, 33-51 (1975).
- 21 Kristiansson, E., Hugenholtz, P. & Dalevi, D. ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* **25**, 2737-2738, doi:btp508 [pii]10.1093/bioinformatics/btp508 (2009).
- 22 Krautler, B. Vitamin B12: chemistry and biochemistry. *Biochem Soc Trans* **33**, 806-810, doi:BST0330806 [pii]10.1042/BST0330806 (2005).
- 23 Monsen, A. L., Refsum, H., Markestad, T. & Ueland, P. M. Cobalamin status and its biochemical markers methylmalonic acid and homocysteine in different age groups from 4 days to 19 years. *Clin Chem* **49**, 2067-2075, doi:10.1373/clinchem.2003.01986949/12/2067 [pii] (2003).
- 24 Hooper, L. V., Xu, J., Falk, P. G., Midtvedt, T. & Gordon, J. I. A molecular sensor that allows a gut commensal to control its nutrient foundation in a competitive ecosystem. *Proc Natl Acad Sci U S A* **96**, 9833-9838 (1999).
- 25 Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe* **4**, 447-457, doi:S1931-3128(08)00303-X [pii] 10.1016/j.chom.2008.09.007 (2008).
- 26 Harzer, G., Franzke, V. & Bindels, J. G. Human milk nonprotein nitrogen components: changing patterns of free amino acids and urea in the course of early lactation. *Am J Clin Nutr* **40**, 303-309 (1984).

- 27 Millward, D. J. *et al.* The transfer of <sup>15</sup>N from urea to lysine in the human infant. *Br J Nutr* **83**, 505-512, doi:S0007114500000647 [pii] (2000).
- 28 Metges, C. C. *et al.* Incorporation of urea and ammonia nitrogen into ileal and fecal microbial proteins and plasma free amino acids in normal men and ileostomates. *Am J Clin Nutr* **70**, 1046-1058 (1999).
- 29 Langran, M., Moran, B. J., Murphy, J. L. & Jackson, A. A. Adaptation to a diet low in protein: effect of complex carbohydrate upon urea kinetics in normal man. *Clin Sci (Lond)* **82**, 191-198 (1992).
- 30 Meakins, T. S. & Jackson, A. A. Salvage of exogenous urea nitrogen enhances nitrogen balance in normal men consuming marginally inadequate protein diets. *Clin Sci (Lond)* **90**, 215-225 (1996).
- 31 Mora, D. *et al.* Characterization of urease genes cluster of *Streptococcus thermophilus*. *J Appl Microbiol* **96**, 209-219, doi:2148 [pii] (2004).
- 32 Li, R., Darling, N., Maurice, E., Barker, L. & Grummer-Strawn, L. M. Breastfeeding rates in the United States by characteristics of the child, mother, or family: the 2002 National Immunization Survey. *Pediatrics* **115**, e31-37, doi:10.1542/peds.2004-0481 (2005).
- 33 WHO. WHO Global Data Bank on Infant and Young Child Feeding (IYCF). (2006).
- 34 Muegge, B. D. *et al.* Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332**, 970-974, doi:10.1126/science.1198719 (2011).
- 35 Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484, doi:nature07540 [pii]10.1038/nature07540 (2009).
- 36 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461, doi:btq461 [pii]10.1093/bioinformatics/btq461 (2010).

- 37 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261-5267, doi:AEM.00062-07 [pii]10.1128/AEM.00062-07 (2007).
- 38 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335-336, doi:nmeth.f.303 [pii] 10.1038/nmeth.f.303.
- 39 Kaufman, L. & Rousseeuw, P. J. *Finding groups in data : an introduction to cluster analysis*. (Wiley, 1990).
- 40 Rousseeuw, P. J. Silhouettes — a Graphical Aid to the Interpretation and Validation of Cluster-Analysis. *J Comput Appl Math* **20**, 53-65 (1987).
- 41 Teal, T. K. & Schmidt, T. M. Identifying and removing artificial replicates from 454 pyrosequencing data. *Cold Spring Harb Protoc* **2010**, pdb prot5409, doi:2010/4/pdb.prot5409 [pii]10.1101/pdb.prot5409.
- 42 Team, R. D. C. R Foundation for Statistical Computing. (2010).
- 43 Wiener, A. L. a. M. Classification and Regression by random forest. *R News* **2**, 18-22 (2002).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

### **Acknowledgements**

We thank Sabrina Wagoner, Jill Manchester for superb technical assistance, plus Brian Muegge, Andrew Grimm, Ansel Hsiao, Nicholas Griffin, and Philip Tarr for helpful suggestions, and Malick Ndao, Tara Tinnin and R. Mkakosya for patient recruitment and/or technical assistance.

This work was supported in part by grants from the NIH (DK078669, T32-HD049338), St. Louis Children's Discovery Institute (MD112009-201), the Howard Hughes Medical Institute, the Crohn's and Colitis Foundation of America, and the Bill and Melinda Gates Foundation.

### **Author contributions**

T.Y., R.K. and J.I.G designed the experiments, M.M., I.T., M.G. D-B., M.C., M.M., G.H., A.C.H., A.P.A, R.K., R.N.B., C.A.L., C.L. and B.W. participated in patient recruitment, T.Y. generated the data, T.Y., F.R., J.R., J.K., J.C.C., D.K. , R.K. and J.I.G. analyzed the results, T.Y., R.K. and J.I.G. wrote the paper.

### **Datasets**

16S rRNA and fecal microbiome datasets have been deposited in MG-RAST for release on publication.

## **Figure Legends**

**Fig 1. Differences in the fecal microbial communities of Malawians, Amerindians and residents of the USA at different ages.** (a) UniFrac distances between children and adults decrease with increasing age of children in each population. Each point shows an average distance between a child and all adults unrelated to that child but from the same country. (b,c) Large interpersonal variations are observed in the phylogenetic and functional configurations of fecal microbial communities at early ages. Malawian and Amerindian children and adults are more similar to one another than to USA children and adults. In panel b, UniFrac distances were defined from bacterial 16S rRNA data generated from the microbiota of 184 unrelated adults ( $\geq 18$  years old) and 206 unrelated children (n=32 Malawians 0.03-3 years old, 21 3-17 years old; 30 Amerindians 0.08-3 years old, 29 3-17 years old; 32 residents of the USA 0.08-3 years old, 62 sampled at 3-17 years of age). In panel c, Hellinger distances derived from EC profiles are shown for unrelated children  $\leq 3$  years of age and unrelated adults (n=9 children and 5 adults from Malawi; 11 children and 5 adults from Venezuela; 10 children and 8 adults from USA). Mean values  $\pm$  SEM are plotted. Abbreviations: \*  $p < 0.05$ ; \*\* $p < 0.005$  (Student's t-test with 1000 Monte Carlo simulations). See **Table S3** for a complete description of the statistical significance of all possible comparisons shown in the Figure.

**Fig. 2. Changes in the representation of genes involved in folate and cobalamin biosynthesis and metabolism in fecal microbiomes as a function of age.** (a) Diagram of KEGG folate metabolic pathway indicating ECs involved in the *de novo* biosynthesis of folate whose proportional representation was higher in the fecal microbiomes of breast-fed babies (0.6-5 months old, colored in yellow) compared to adults (gray). Note that the representation of genes encoding ECs involved in folate metabolism is higher in adult fecal microbiomes and in formula-fed USA baby microbiomes compared to the microbiomes of breast-fed babies in all populations. (b) Diagram of KEGG pathway for cobalamin biosynthesis, indicating ECs whose proportional representation was higher in the fecal microbi-



omes of all adults and formula-fed USA infants (gray) compared to the fecal microbiomes of breast-fed babies in all populations. However, among adults, USA fecal microbiomes have higher relative representation of ECs in this pathway compared to adult Malawian/Amerindian microbiomes. P-values for the highlighted ECs can be found in **Table S7**. (c) Age-related changes in the proportional representation of genes encoding ECs best discriminating baby and adult microbiomes. UPGMA clustering (average linkage method) of fecal microbiomes, based on the relative abundances of ECs (normalized by Z-score across all datasets). The bars on the top indicate geographic location of each human that was sampled.

**Fig. 3. Geographic differences in the bacterial functional structure of fecal microbiomes in three populations.** (a) Examples of ECs that exhibited the largest differences in proportional representation between USA and Malawian/Amerindian baby fecal microbiomes. UPGMA clustering of 10 USA, 10 Malawian and 6 Amerindian fecal microbiomes, based on the relative abundances of genes encoding ECs (normalized by Z-score across all datasets). (b) Diagram of KEGG riboflavin biosynthetic pathway indicating ECs whose proportional representation was higher in the fecal microbiomes of Malawian and Amerindian compared to USA babies; (c) Examples of ECs that exhibited the largest differences in proportional representation between USA and Malawian/Amerindian adult fecal microbiomes. UPGMA clustering of 16 adult USA, 5 Malawian and 5 Amerindian fecal microbiomes, based on the relative abundances of genes encoding ECs (normalized by Z-score across all datasets).

**Fig. 4. Differences in the fecal microbial communities between family members across the three populations studied.** UniFrac distances between the fecal bacterial communities of family members were calculated (n=19 Amerindian families, 34 Malawian families, 54 USA families with teenage twins). Mean  $\pm$  SEM values are plotted. The UniFrac matrix was permuted 1000 times; p values represent the fraction of times permuted differences were greater than real differences: ns (not significant;  $p > 0.05$ ), \*  $p < 0.05$ , \*\* $p < 0.005$ .

**Figures**

**Figure 1.**

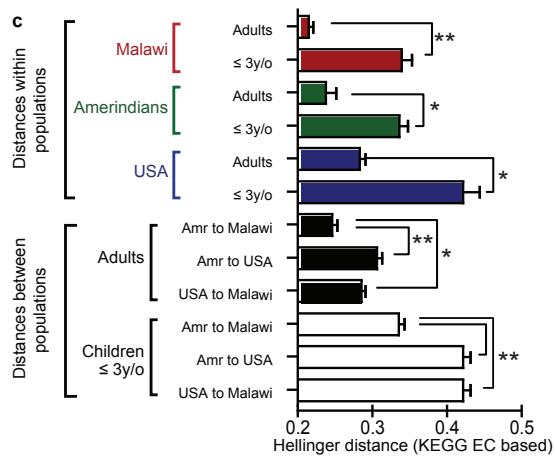
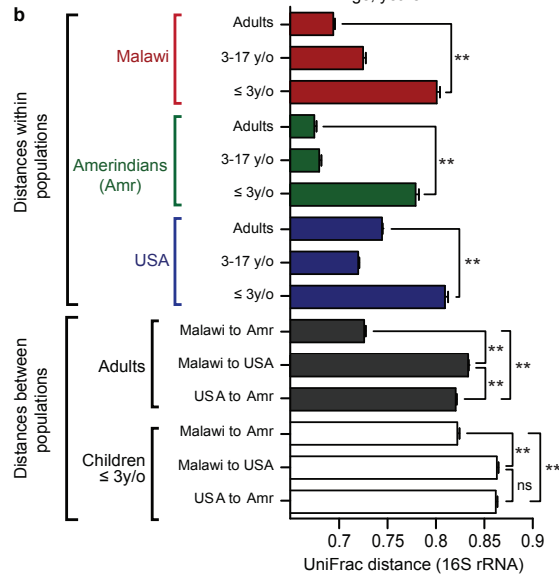
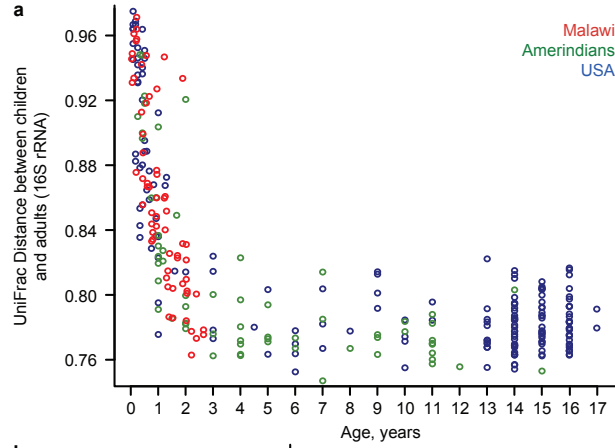


Figure 2a,b.

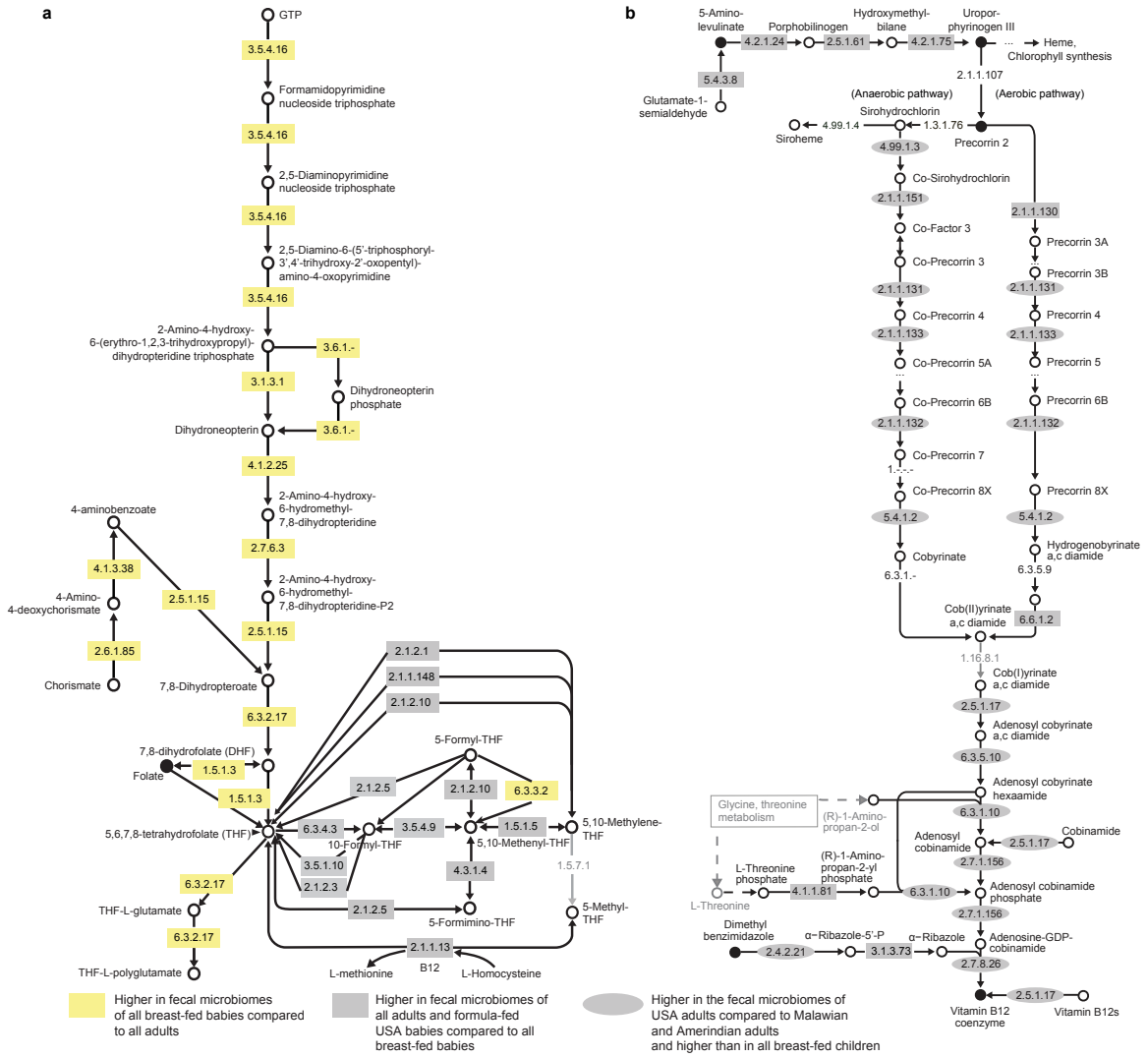
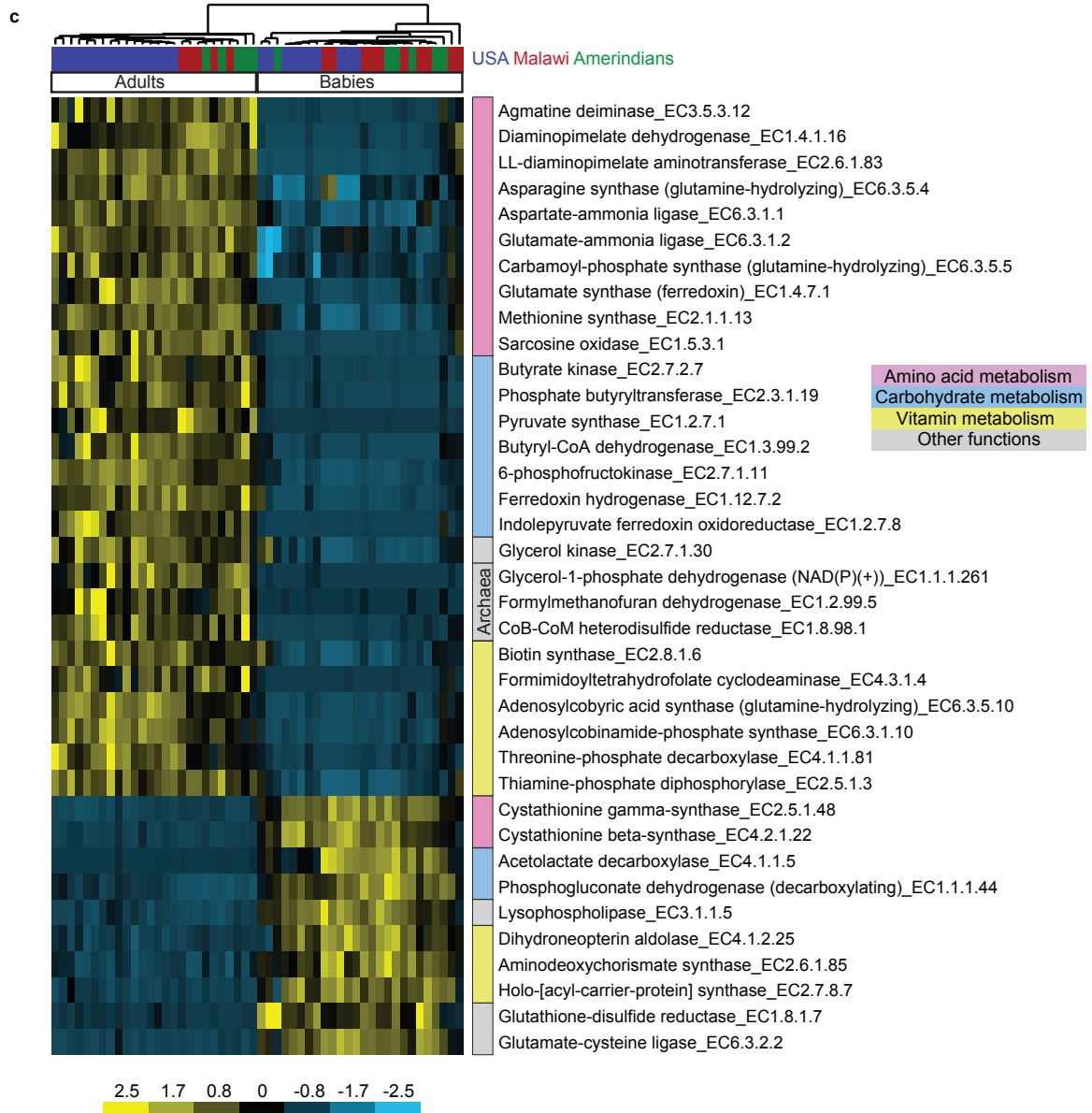


Figure 2c.



**Figure 3.**

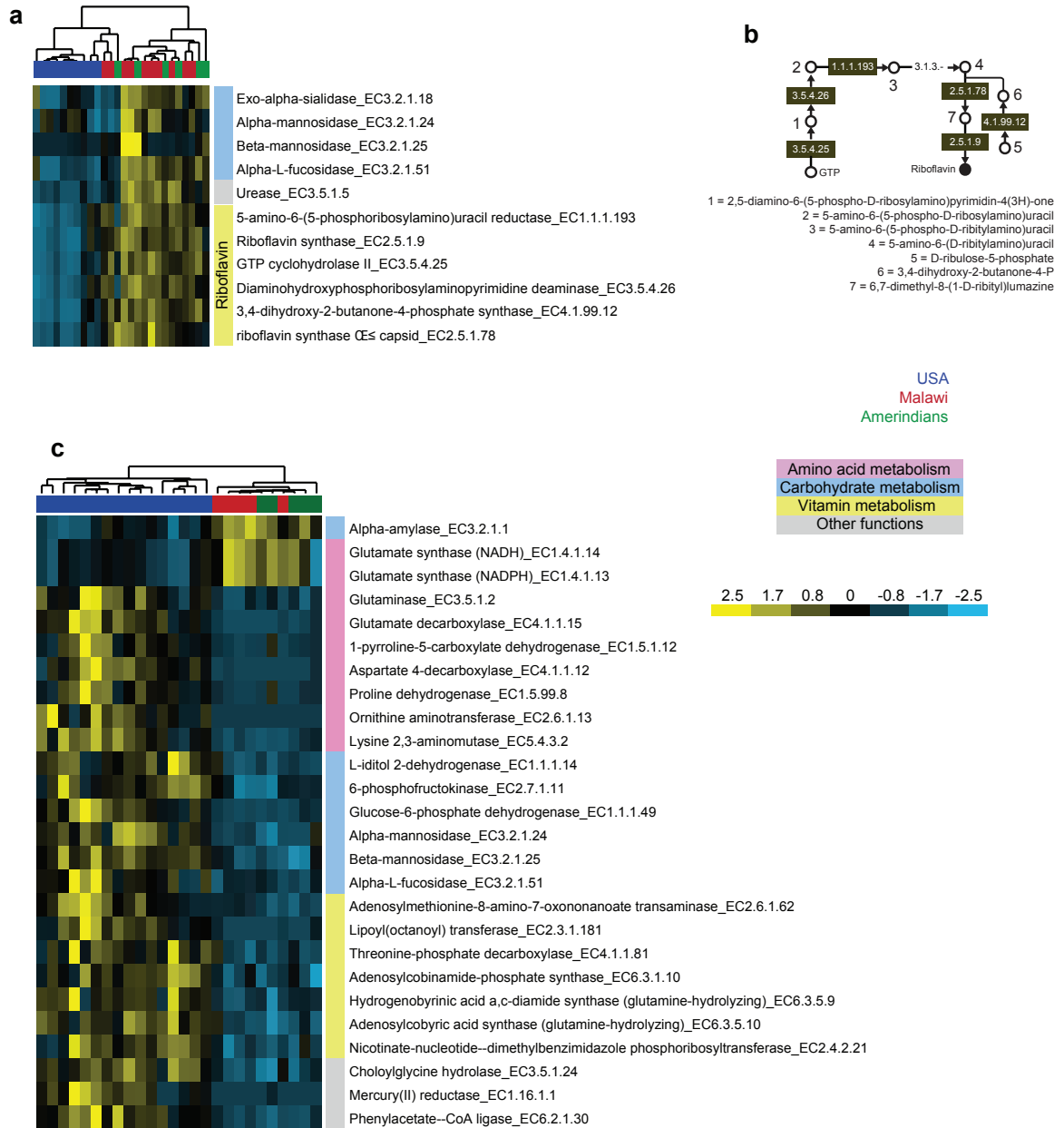
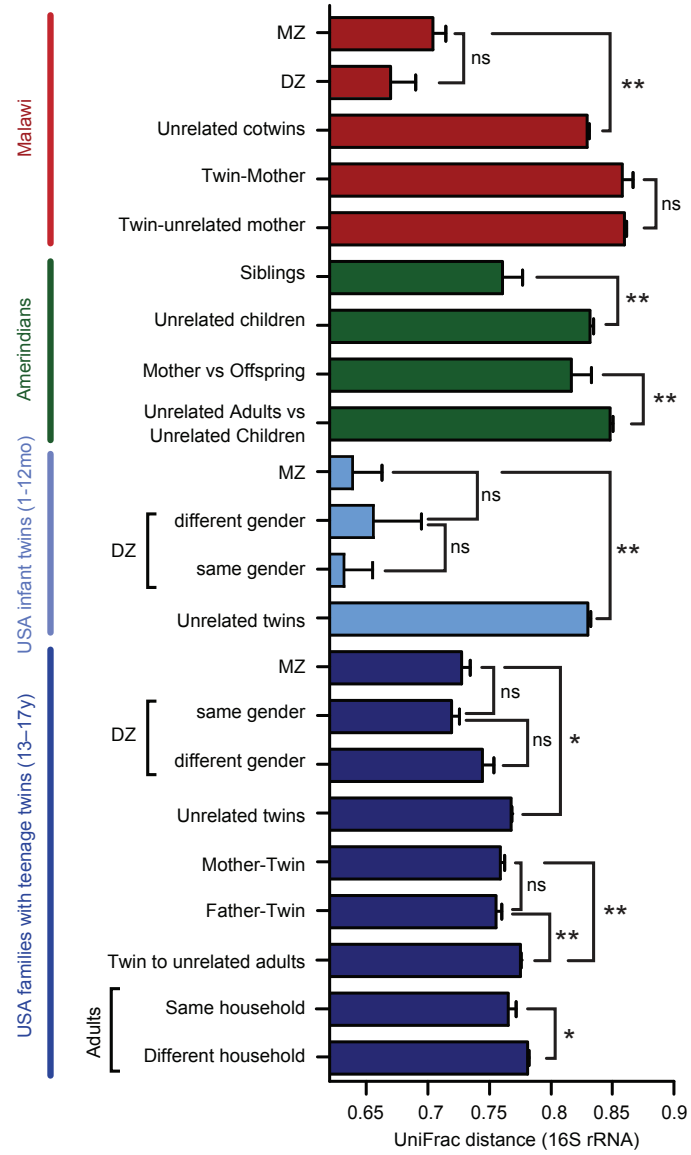


Figure 4.



### Supplemental Figure Legends

**Fig. S1. Large interpersonal variation between children.** UniFrac distances between children  $\leq 3$  years old and adults from the same population compared to adults from the other two populations. Mean values  $\pm$  SEM are plotted.  $**p < 0.001$ , Student's t-test with 1000 Monte Carlo simulations.

**Fig. S2. Principal Coordinates Analysis of UniFrac distances between 524 sampled individuals.** When adults are considered alone (a), or when babies are considered alone (b), there is clear separation among populations using unweighted UniFrac. When adults, babies and children older than 6 months are considered together (c,d), the main axis separates by age, and the differences among adult populations are more apparent than differences among infant populations, in part because of the high inter-individual variability in infants.

**Fig. S3. Changes in the representation of bacterial taxa in the fecal microbiota as a function of age and geographic region.** (a) Shotgun reads were mapped to 126 reference sequenced human gut microbial genomes. Spearman correlations (Rho values) were subsequently calculated for the representation of reads that match to each genome against age for each population (a Rho value of  $\pm 1$  indicates maximum correlation with age, a zero indicates minimum correlation). Rho values for Malawians are plotted against Rho values for Amerindians (black points) or residents of the USA (yellow points). Each point represents a genome; coordinates are correlations for the relative abundance of that genome with age in Malawians (x-axis) and Amerindians or USA residents (y-axis). Spearman correlations relating populations: Malawi vs USA,  $Rho = 0.65$ ,  $p < 10^{-15}$ ; Amerindians vs USA  $Rho = 0.78$ ,  $p < 10^{-15}$ ; Malawi vs Amerindians  $Rho = 0.66$ ,  $p < 10^{-15}$ . Lower panel presents examples of the largest changes with age in all three populations (*Bifidobacterium longum* and *Clostridium sp*), and changes that are most pronounced in Malawi and Amerindians (*Providencia rettgeri*). (b) An analysis similar to that shown in panel a, but using 1280 microbial genomes present in KEGG.

**Fig. S4. Geographic differences in the bacterial phylogenetic structure of adult fecal microbiomes.** Random Forests analysis disclosed groups of species-level phylotypes whose representation is significantly different between the three populations. Shown are relative abundances (log<sub>10</sub>) for the 64 OTUs whose removal increases estimated error by 0.001.

**Fig. S5. Enterotype analysis.** (a) Stacked bar plot of *Bacteroides/Prevotella* gradient. Each column shows relative abundances of *Bacteroides* (red), *Prevotella* (blue), and other genera (green) for a single gut community; communities are ordered according to increasing *Bacteroides* relative abundance. Box plots below show the distribution of samples from each country. (b) Classical multidimensional scaling (also known as principal coordinates analysis) of Jensen-Shannon distances between all adult ( $\geq 20$  years of age) gut communities; samples are colored by host country, and lines connect samples to their putative enterotype cluster centroids (silhouette index = 0.58). The inset shows a scatter plot of the relative abundance of *Bacteroides* and *Prevotella* along the first principal coordinate axis (PC1). (c) Clustering results for several adult sub-populations (silhouette indices: Malawi = 0.37; Amazon 0.51; non-USA = 0.50; USA = 0.68) showing new putative enterotype boundaries. (d) Enterotype clustering algorithm applied to samples from all countries and all ages; samples are colored by age. (e) Enterotype membership for partitioned subpopulations of increasing minimum age. Samples are sorted vertically first by putative cluster number, then by age within each cluster. Lines indicate samples that switched cluster membership after a partitioning step.

**Fig. S6. Most abundant non-bacterial members identified in the fecal microbiota.** Shotgun sequences were used to query the NCBI nr database (Blastn e-value threshold cutoff,  $10^{-5}$ ). The proportion of sequences that mapped to non-bacterial sequences was calculated for each age- and geographic group. The most abundant fungal sequences belong to the NCBI nr family level taxa *Ascomycota* and *Microsporidia* and were found in all three populations. In NCBI nr, 'other eukaryota' refers to sequences that do not map to fungi,



plants, arthropoda, mammals, and ‘other metazoa’. In USA microbiomes ‘other eukaryota’ was most prominently represented by *Hexamitidae*, *Trichomonadidae* families and genus *Entamoeba*, while in Malawian and Amerindian microbiomes the most abundant group was “uncultured compost protozoan”, with *Codonosigidae* and *Hexamitidae* families represented less frequently. \*\*\*  $p < 0.0005$ , \*\* $p < 0.005$ , \* $p < 0.05$  (Mann-Whitney test).

**Fig. S7. The number of ECs identified is similar in adult and infant fecal microbiomes, while the fraction of reads with assignable EC annotations declines with age in all three populations.** (a) EC matrix was rarefied to 3,650 sequences per sample, and number of ECs plotted against log (Age) for each sample. (b) Percent of sequences with KEGG annotation plotted against log (Age).

**Fig. S8 –Analysis of Hellinger distances between KEGG KO profiles.** (a,b) PCoA plots. (c) Hellinger distances derived from KO profiles are shown for unrelated children  $\leq 3$  years of age and unrelated adults (n=9 children and 5 adults from Malawi; 11 children and 5 adults from Venezuela; 10 children and 8 adults from USA). Mean values  $\pm$  SEM are plotted. Abbreviations: \*  $p < 0.05$ ; \*\* $p < 0.005$  (Student’s t-test with 1000 Monte Carlo simulations). For the analyses shown in (a-c), counts were normalized to the total number of reads for each fecal microbiome sample, thus accounting for sequences unassigned to KEGG.

**Fig. S9. PCoA and Procrustes analysis of 16S rRNA and shotgun datasets annotated with KEGG ECs (a), KEGG KOs (b) and COGs (c).** Two spheres connected by a line represent two different data types from the same fecal sample. The colors of the lines indicate the type of data; in all cases, the grey component of the line is connected to the sphere representing 16S rRNA data, while the red component of the line is connected to the sphere corresponding to that sample’s functional annotation data (EC, KO, or COG). The overall goodness of fit for the different data types ( $M^2$ ) is noted in each panel (three dimensions were used to calculate this  $M^2$  value).

**Fig. S10. Age-related changes in the proportional representation of genes encoding ECs involved in folate metabolism.** (a) KEGG pathway for folate metabolism. (b) UPGMA clustering (average linkage method) of fecal microbiomes of 24 babies and 26 adults based on the relative abundances of genes encoding ECs shown in panel A, normalized by Z-score across all datasets.

**Fig. S11. Age-related changes in the proportional representation of genes encoding ECs involved in cobalamin biosynthesis.** UPGMA clustering (average linkage method) of all 110 characterized fecal microbiomes, based on the relative abundances of ECs involved in cobalamin biosynthesis (normalized by Z-score across all datasets). The bars on the top indicate the age, breastfeeding status and geographic location of each human that was sampled.

**Fig. S12. Spearman correlation between gut microbial species predicted to synthesize vitamins B12 and folate and their representation in fecal microbiomes at different ages and in different populations.** UPGMA clustering of 126 sequenced gut genomes (average linkage method) based on the presence of the ECs involved in folate and cobalamin biosynthesis and metabolism (black squares). Spearman correlation coefficients of the proportional representation of these genomes with increasing age are shown on the right for each geographic location; a negative value indicates a decrease in the proportion of a taxon with increasing age.

**Fig. S13. Changes in EC representation in fecal microbiomes as a function of age and population.** Spearman correlation coefficients (Rho values) were calculated for the proportional representation of each EC against age for each human population. Plotted are Rho values for Malawians (X-axis) against Rho values for Amerindians (black points) or USA residents (yellow points). Each point represents an EC and coordinates are Rho values for that EC in Malawians (X-axis) and Amerindians or USA residents (Y-axis). Spearman correlation: Malawi vs USA,  $Rho=0.76$ ,  $p<10^{-15}$ ; Amerindians vs USA  $Rho=0.66$ ,

$p < 10^{-15}$ ; Malawi vs Amerindians  $Rho = 0.78$ ,  $p < 10^{-15}$ . Panels a-f show examples of ECs with similar or distinct Rho values for the three populations. The calculated Spearman correlation coefficient and the corresponding p-value for these examples are provided at the bottom of the figure.

**Fig. S14. Proportional representation of 126 microbial genomes in the fecal microbiomes of breastfed Malawian twins and breast-fed and formula fed USA twins (1-5 months old).**

**Fig. S15. Examples of genes encoding ECs whose abundance is significantly greater in the fecal microbiomes of USA formula-fed compared to breast-fed twins (2-5 months/old).** Criteria for inclusion: p-value  $< 10^{-10}$  (ShotgunFunctionalizeR) and consistent representation in a KEGG pathway in one or the other feeding group. The relative abundances of genes encoding ECs are normalized by Z-score across all datasets.

**Fig. S16. Percentage of fecal microbiome gene content in sampled members of the three populations that is also represented in the METAHIT gene catalog generated from 124 adult Europeans.** Percentage of shotgun pyrosequencing reads in each population that could be assigned to the METAHIT gene catalog using the following Blastn parameters:  $\geq 90\%$  nucleotide sequence identity between the read and a member of the gene catalog, E-value  $< 10^{-5}$ , bitscore  $\geq 50$ .

**Fig. S17. Principal Coordinate Analysis of Hellinger distances between the KEGG KO profiles of adult USA, Amerindian and Malawian fecal microbiomes from the present study and from 70 European microbiomes in the METAHIT dataset<sup>2</sup>.** First three principal coordinates are shown.

Supplemental Figures

Figure S1.

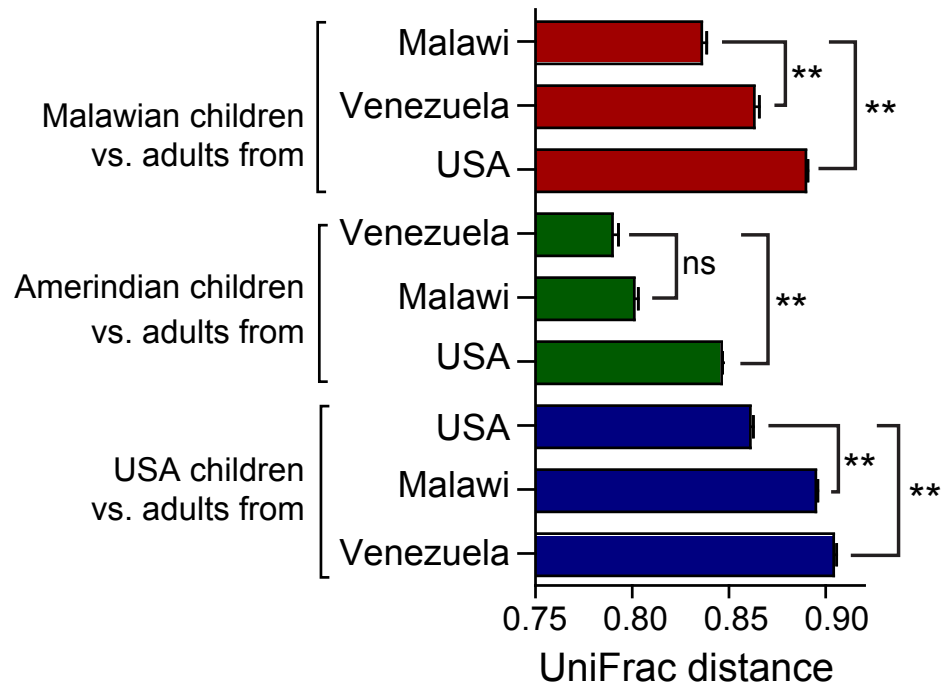


Figure S2.

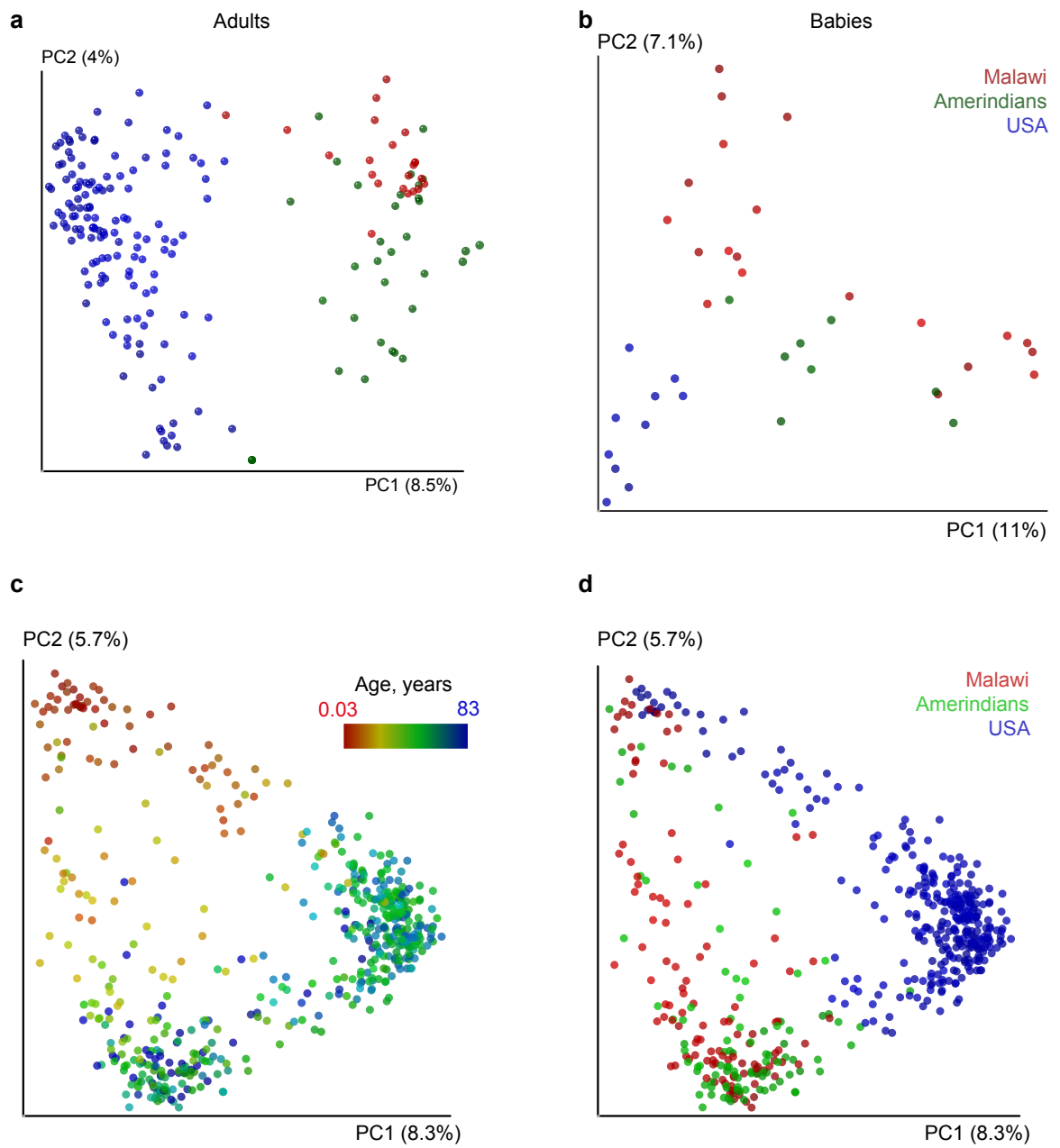


Figure S3.

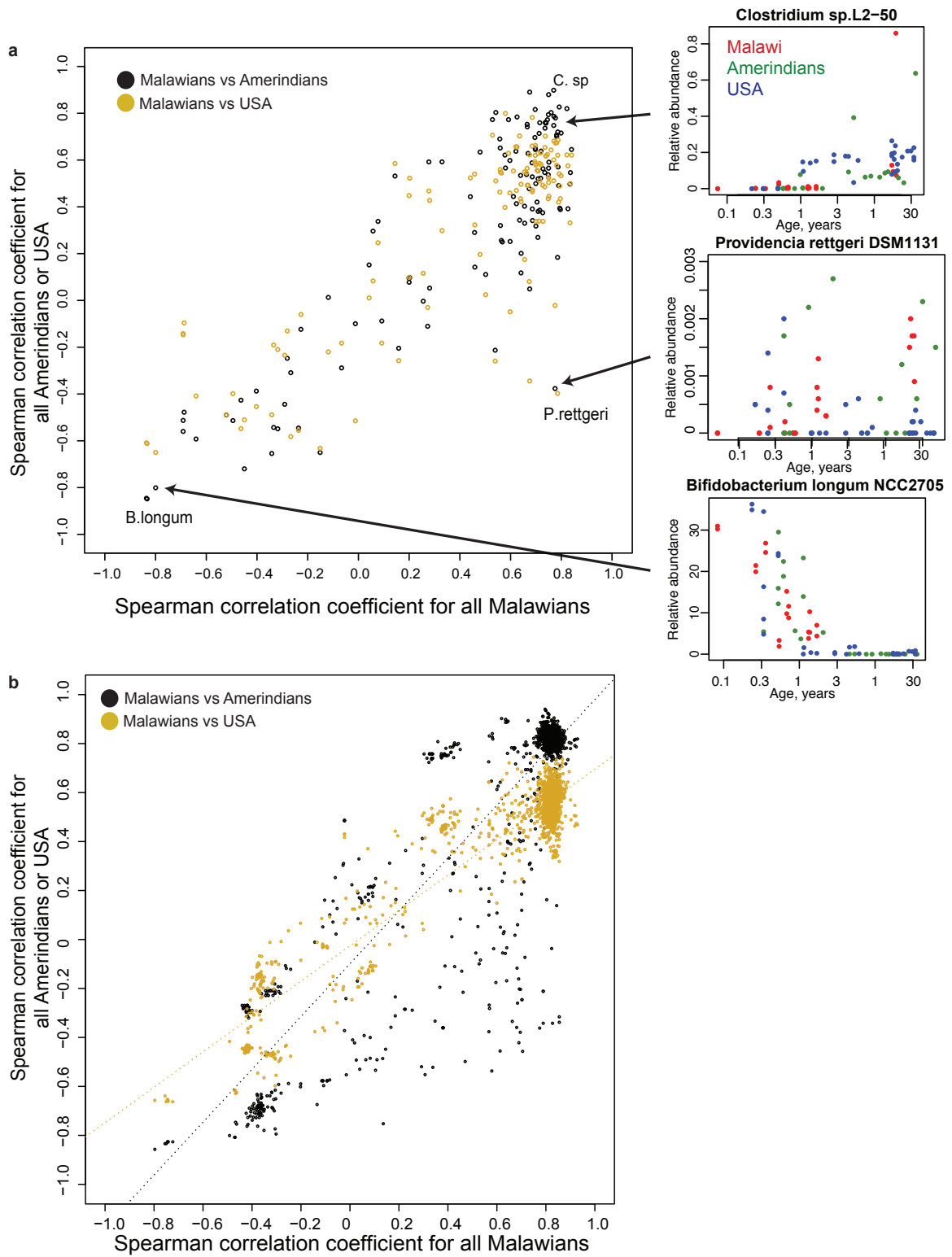


Figure S4.

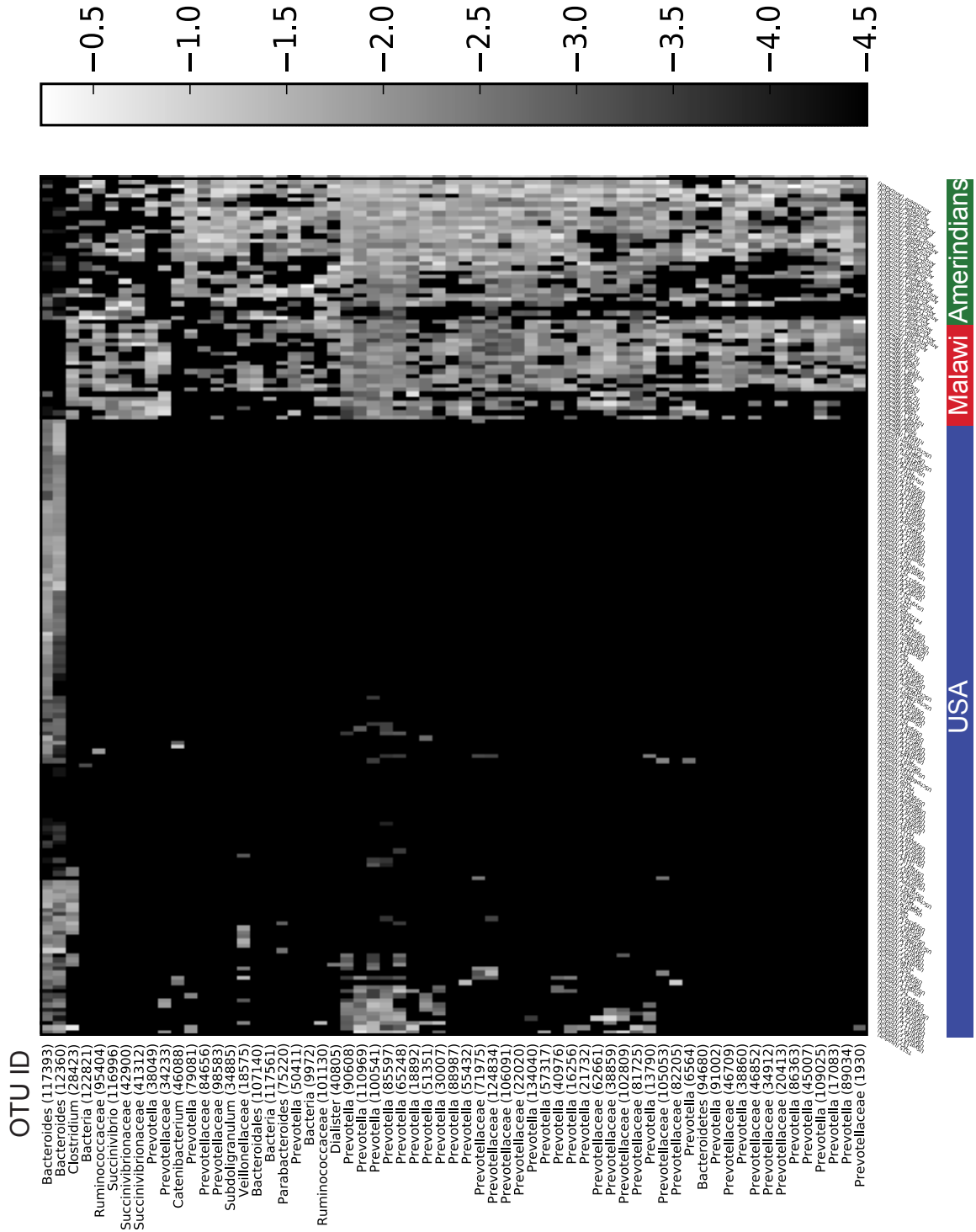


Figure S5.

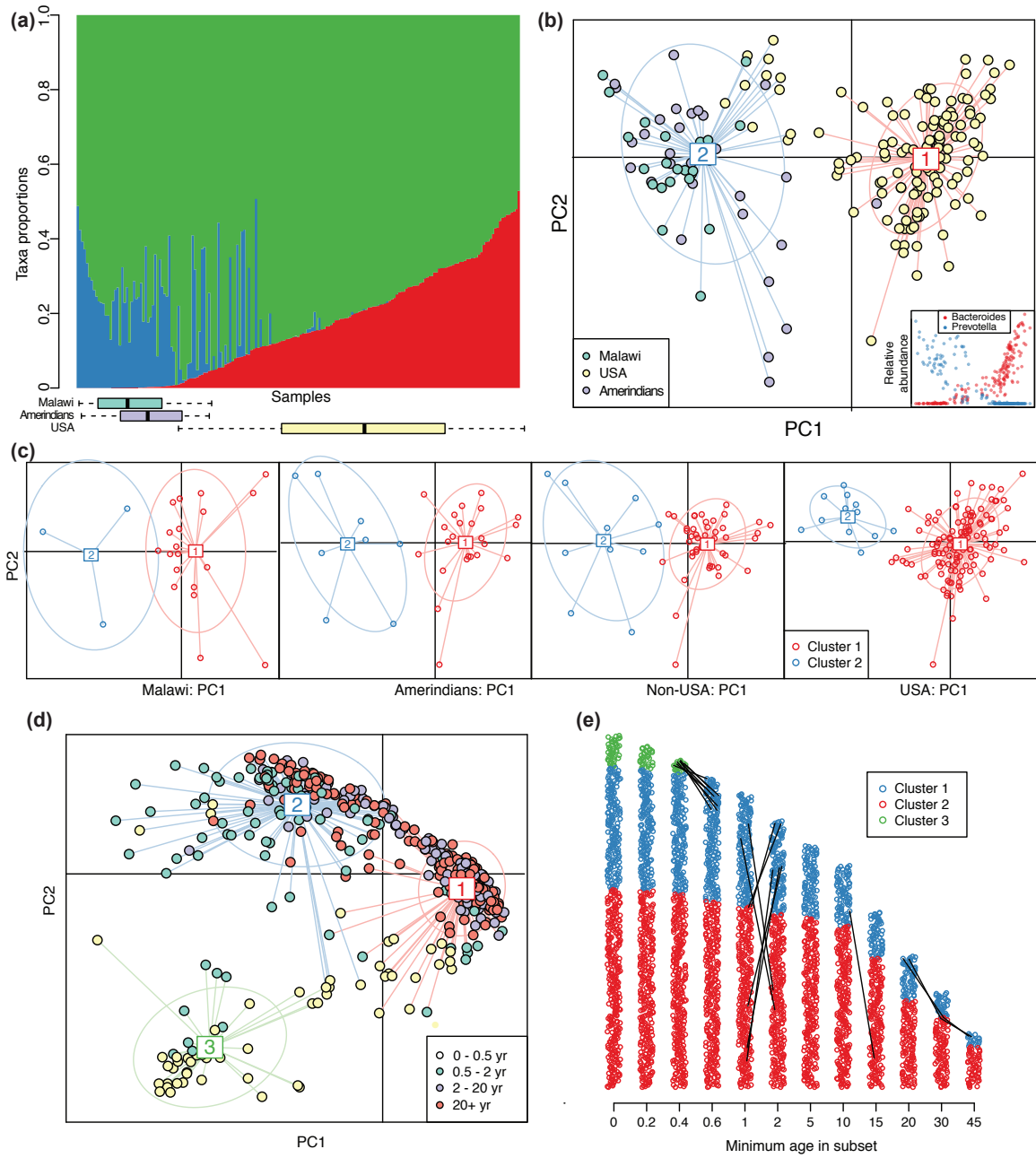




Figure S6.

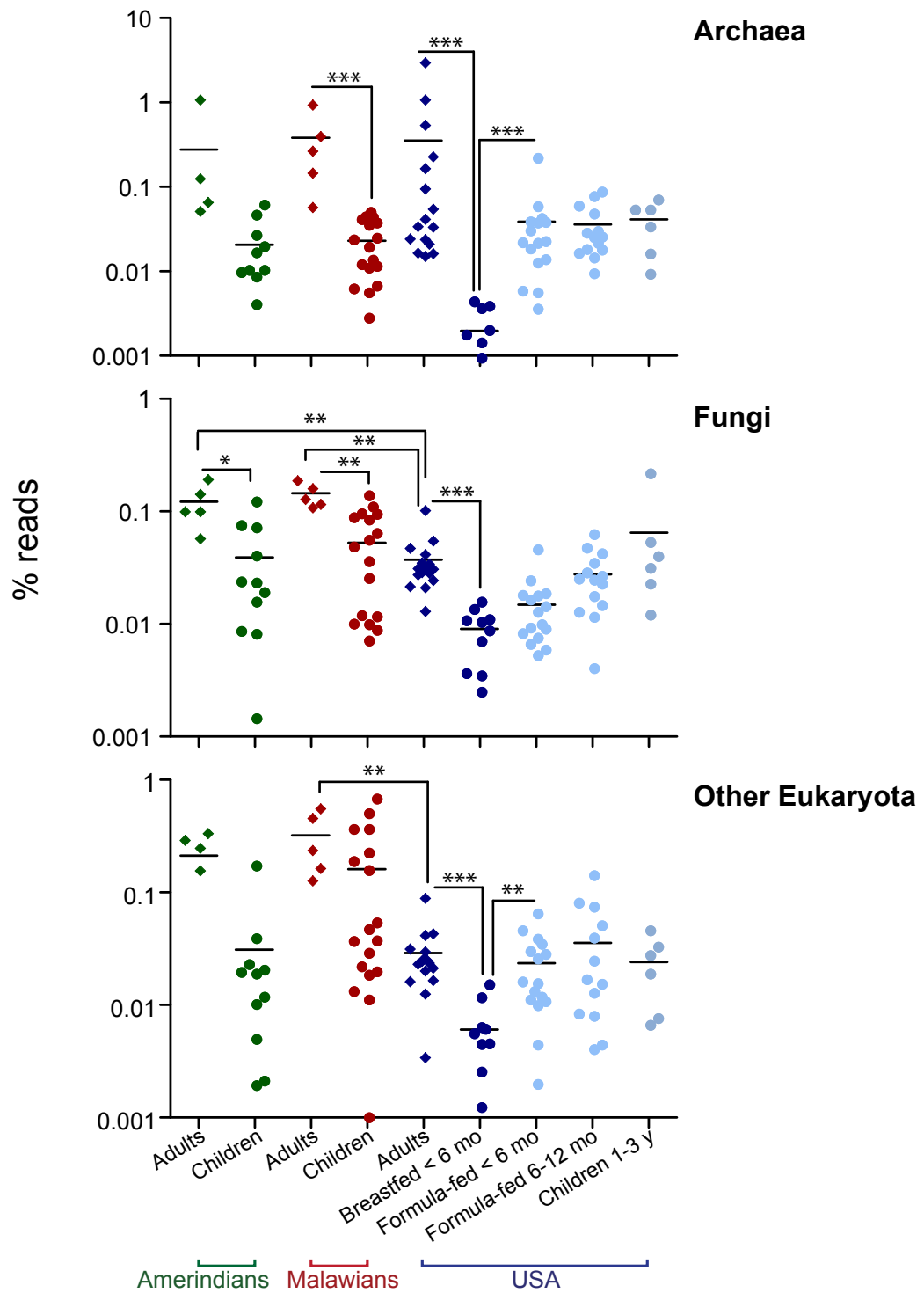


Figure S7.

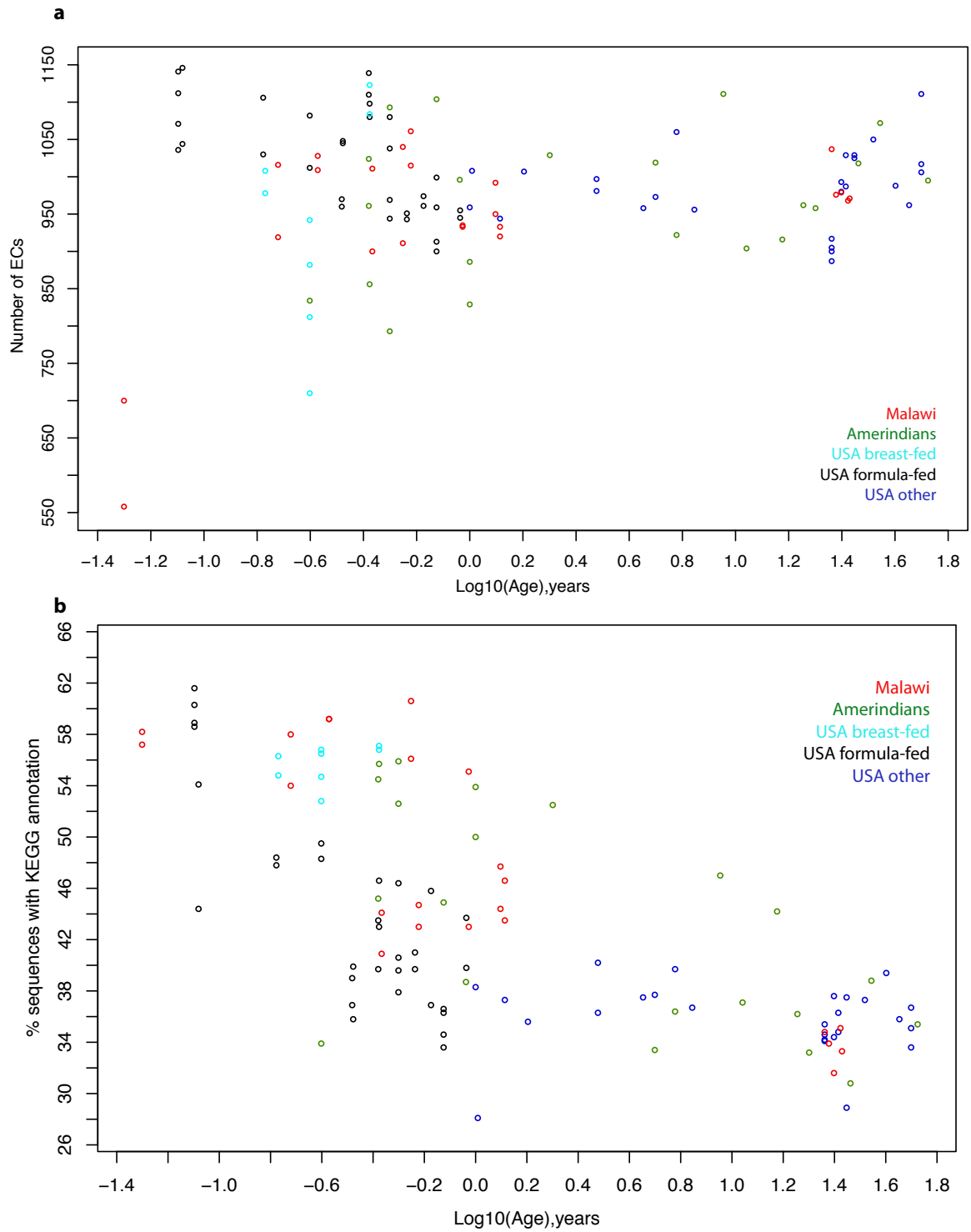


Figure S8.

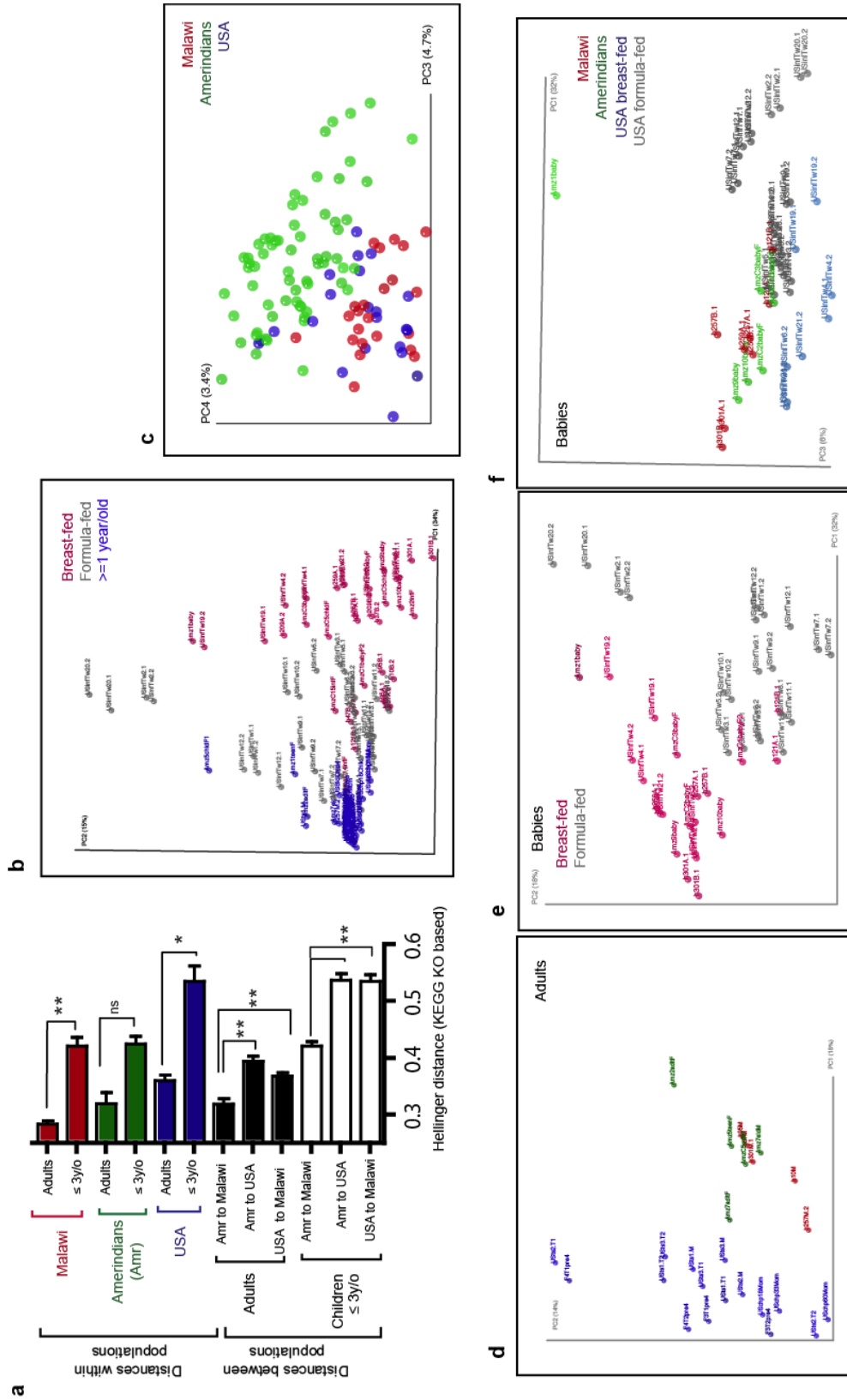


Figure S9.

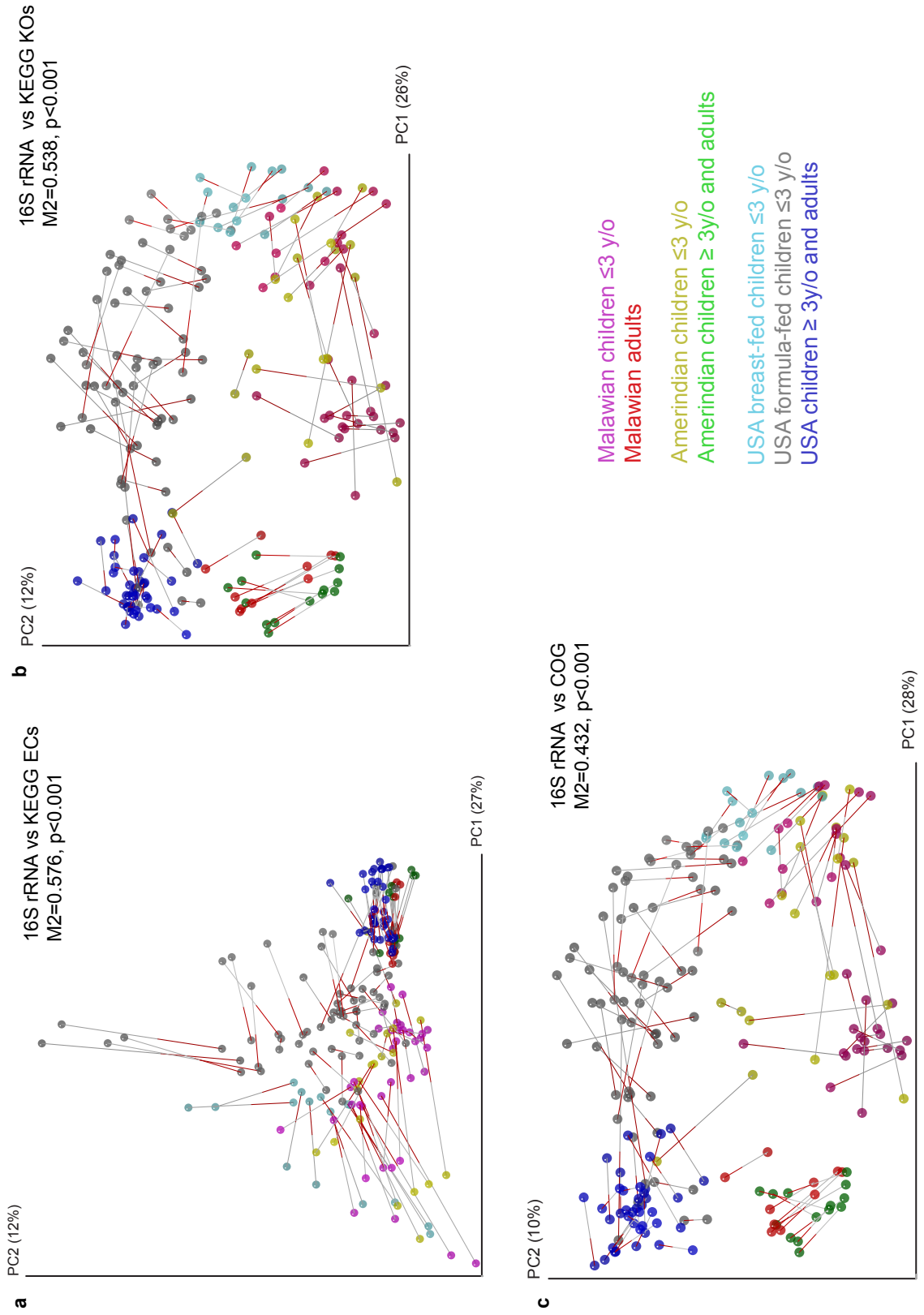


Figure S10.

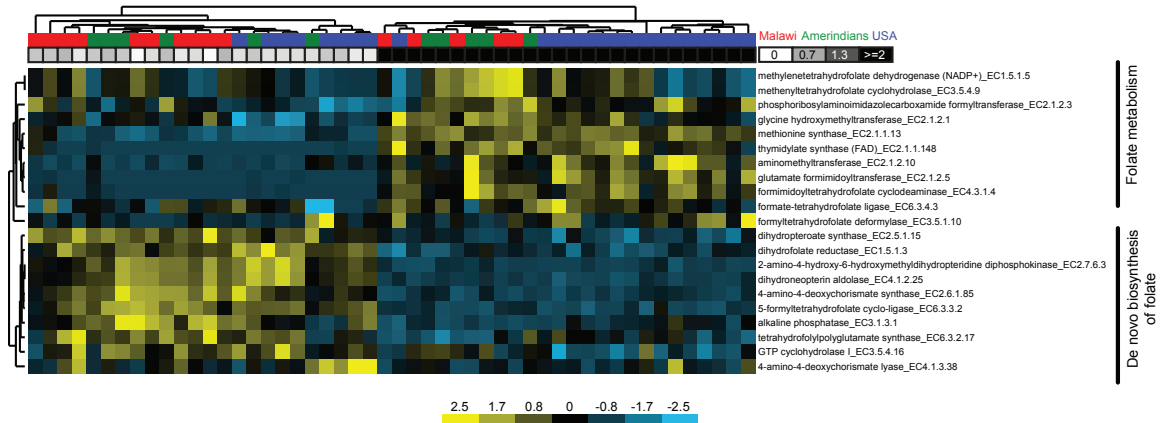
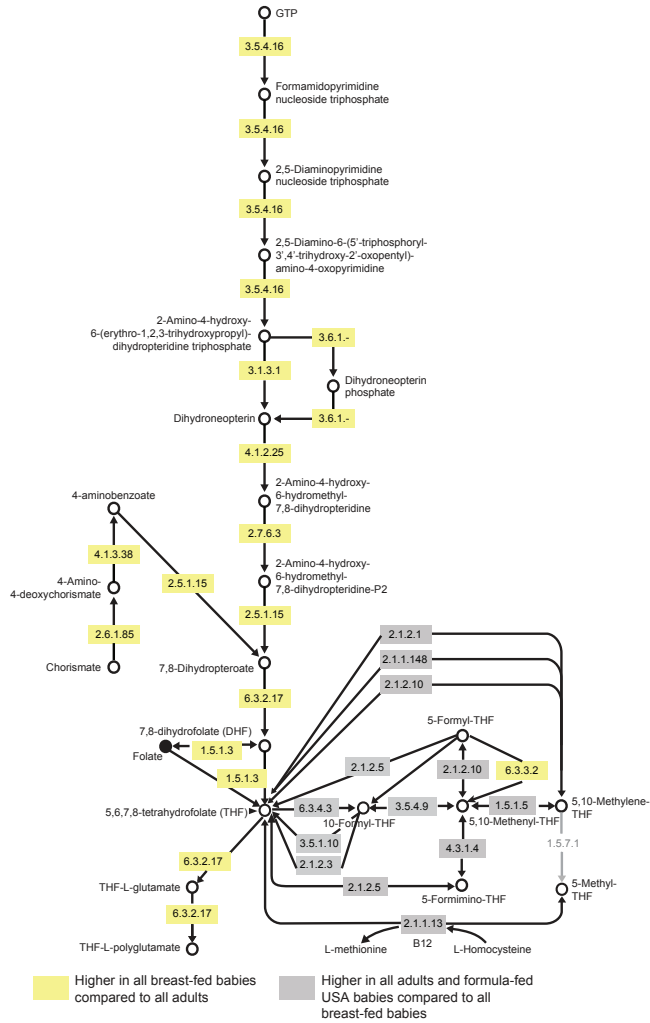


Figure S11.

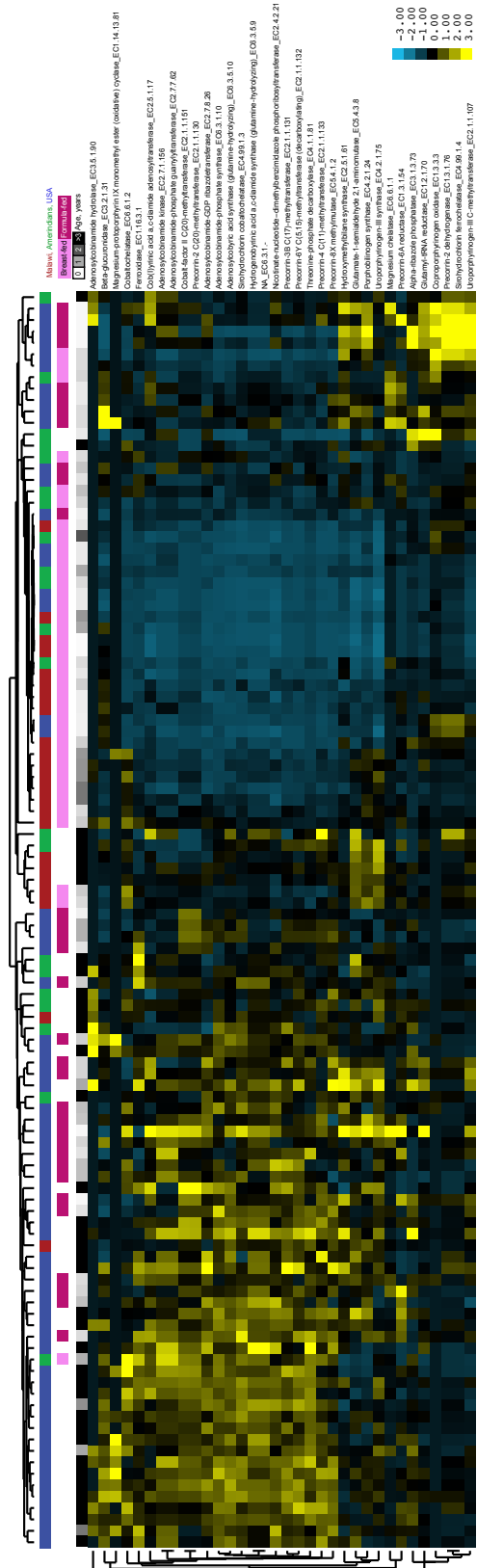


Figure S12.

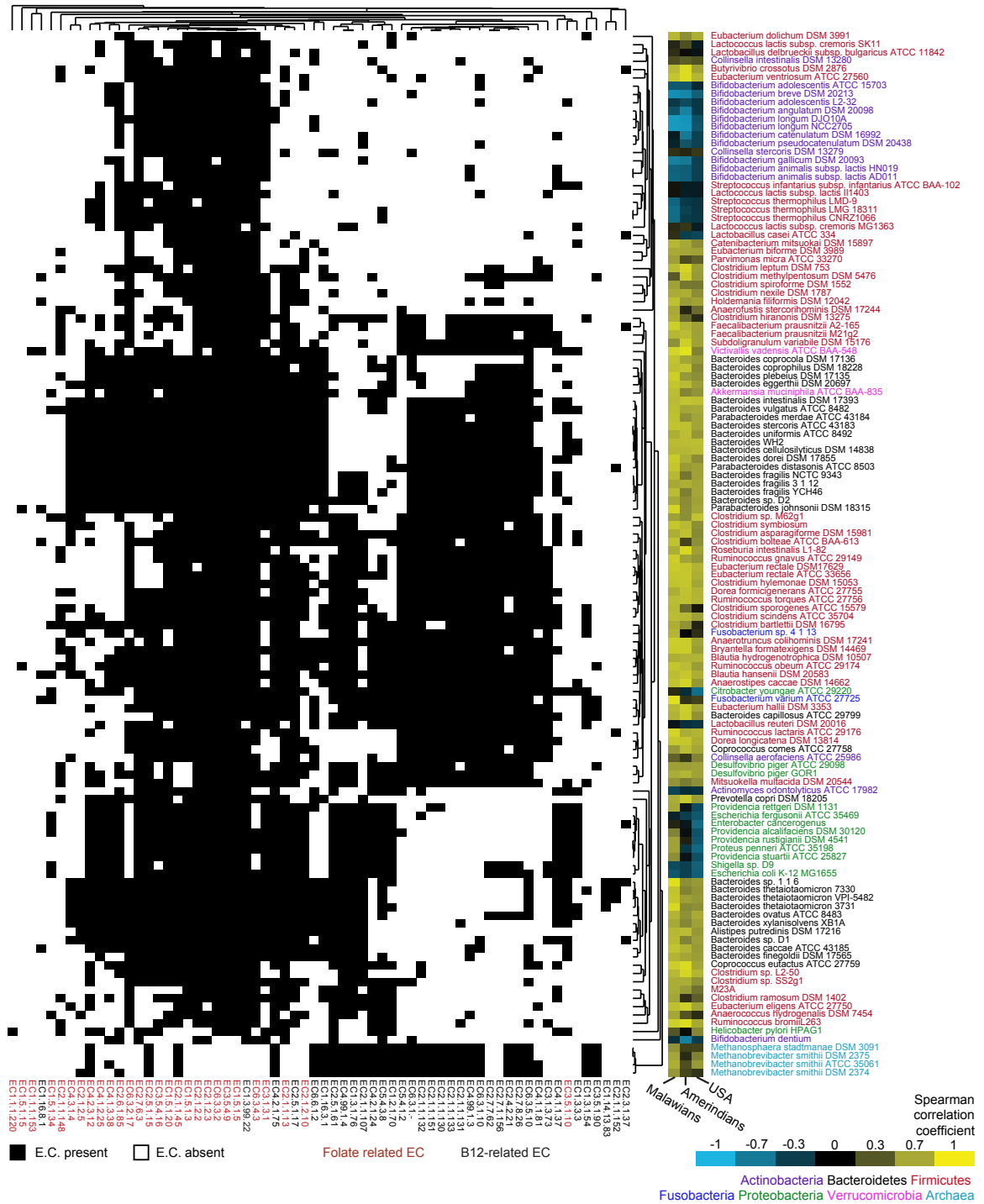


Figure S13.

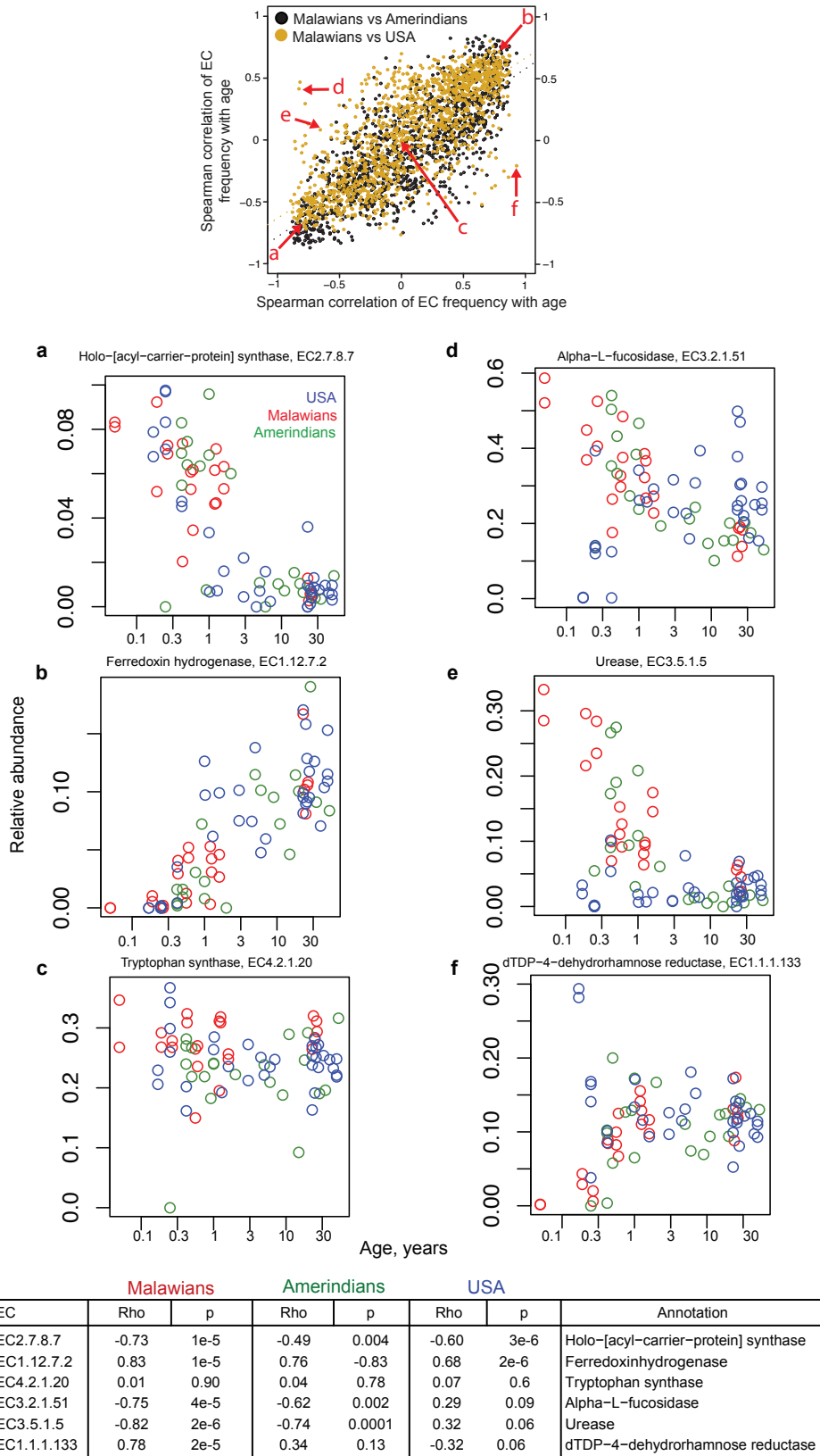




Figure S14.

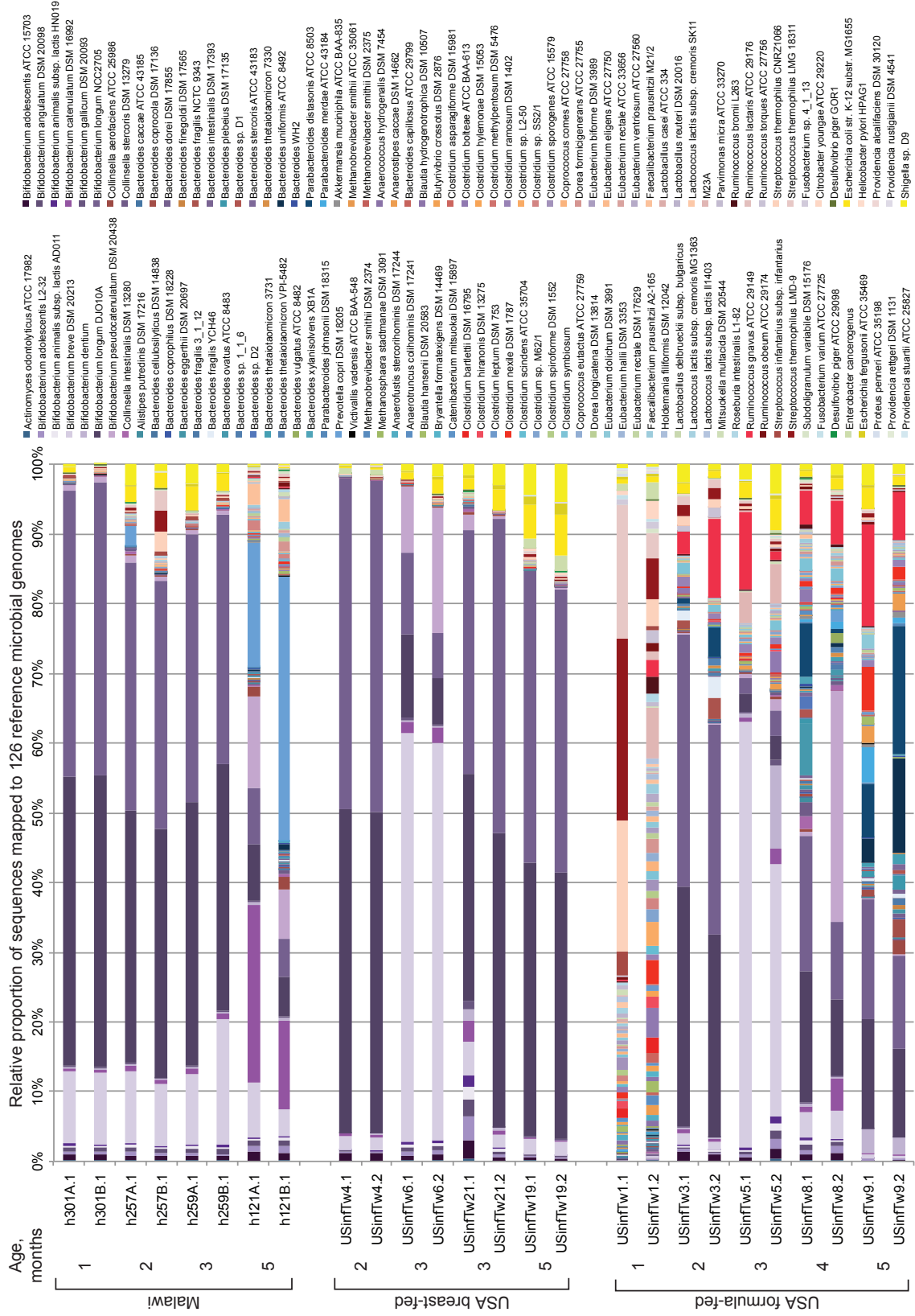


Figure S15.

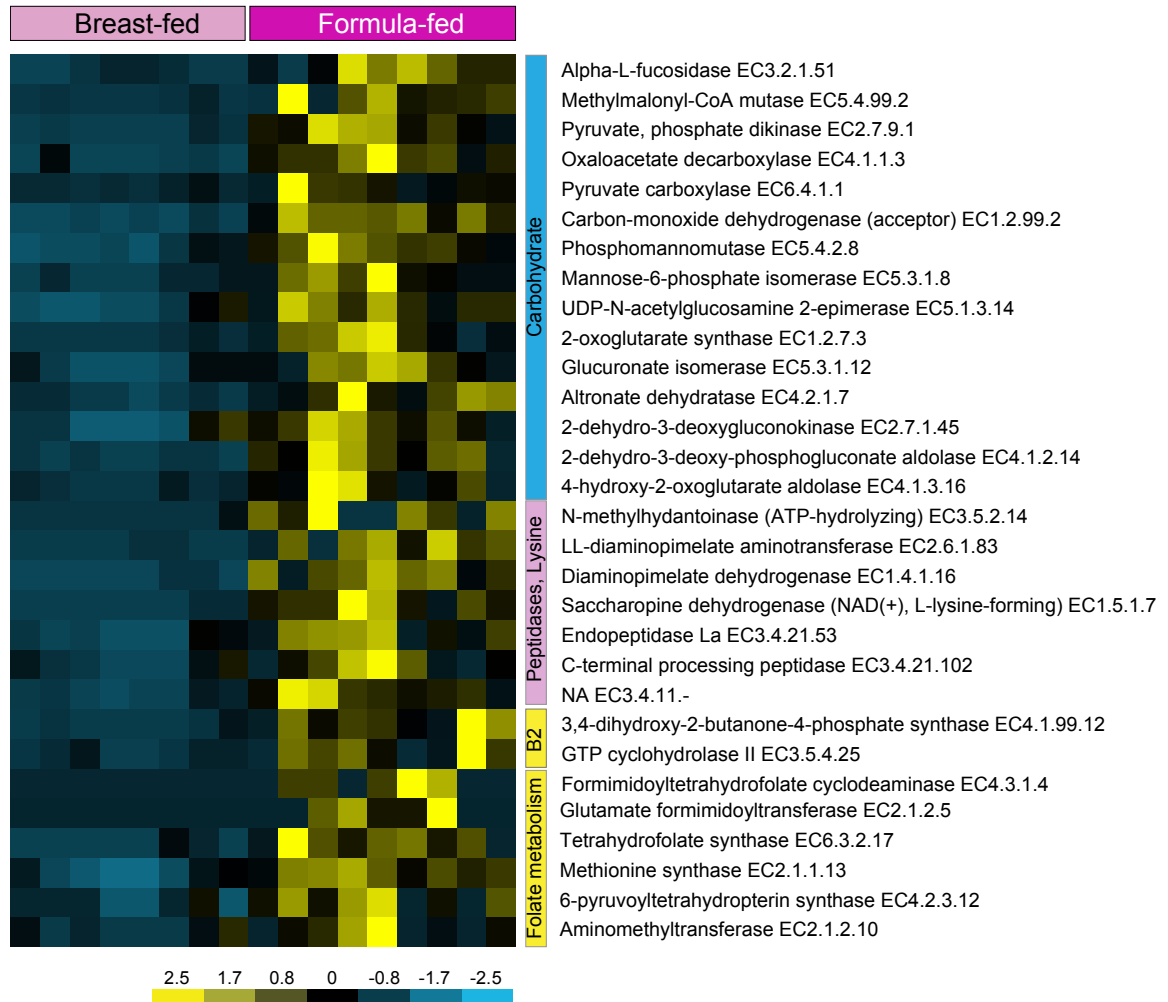


Figure S16.

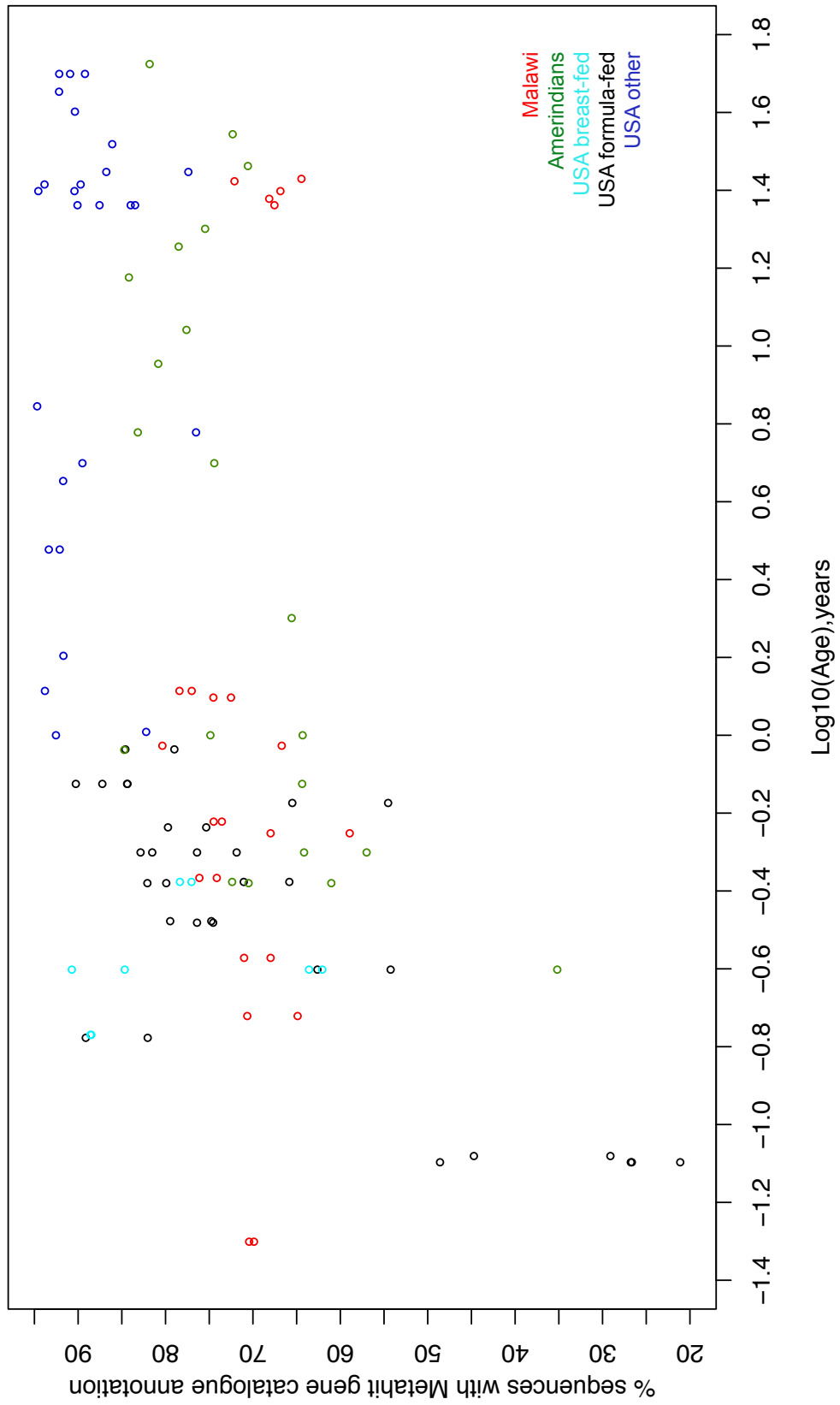
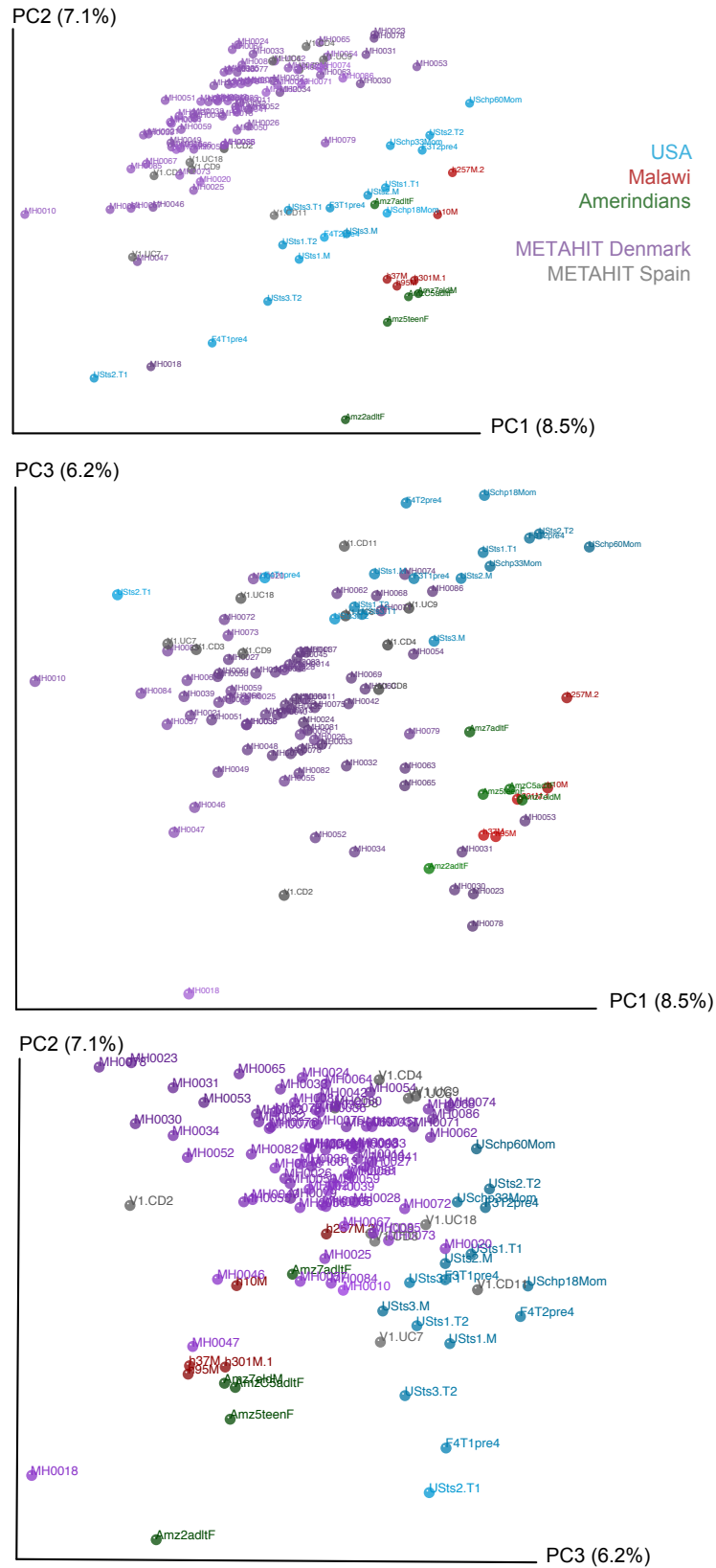


Figure S17.



## Supplemental Table Legends

**Table S1. Diet survey conducted in two Amerindian villages. (a) Platanillal (b) Coromoto.** Data are based on 24 h recall.

**Table S2. Summary of study participants and of fecal bacterial 16S rRNA and whole community DNA sequence datasets.** Our analyses also included (i) V2-derived 16S rRNA data from 30 individuals representing 10 USA families, each comprised of lean adult female twins and their mother, who had been characterized in one of our earlier publications, together with shotgun pyrosequencing data from a subset of three of these families<sup>35</sup>, plus (ii) 16S rRNA and shotgun data from two fecal samples obtained from a single USA mother and child who had been the subject of report describing the assembly of that child's gut microbiota/microbiome<sup>5</sup>.

**Table S3. P values (Student t-test with 1000 Monte Carlo permutations) of UniFrac and Hellinger distances between fecal communities of children and adults shown in Fig. 1b,c.**

**Table S4. List of the 126 reference human gut microbial genomes.**

**Table S5. Spearman correlations of relative abundances of reads that map to microbial genomes in fecal microbiomes with age for each country using (a) 126 genomes and (b) 1,280 genomes from KEGG database.**

**Table S6. Results of Random Forests classifier of OTUs (species-level phylotypes) that discriminate the adult fecal microbiota of USA and non-USA residents (performed over 100 even rarefactions of sampled communities).** The rarefaction depth for was 718 sequences/sample. One hundred even rarefactions were performed for the comparison.

**Table S7. ECs identified by Random Forests and ShotgunFunctionalizeR that exhibit significant age-associated differences. Shown are ECs with p-values < 0.0001 (adjusted for multiple comparison using Benjamini-Hochberg False Discovery Rate).**

**Table S8. ECs identified by Random Forests and ShotgunFunctionalizeR that exhibit significant population-specific differences in babies (0-6 months old). Shown are ECs with p values <**

**0.0001 (adjusted for multiple comparisons using Benjamini-Hochberg False Discovery Rate).**

**Table S9. ECs identified by Spearman correlation analysis that exhibit age-associated changes in their proportional representation.**

**Table S10. ECs identified by Random Forests and ShotgunFunctionalizeR that exhibit significant differences in their representation in the fecal microbiomes of 4 breast-fed USA twin pairs versus 4 formula-fed USA twin pairs (2-5 months old).**

**Table S11. ECs identified by Random Forests and ShotgunFunctionalizeR that exhibit significant population-specific differences in the fecal microbiomes of adults. Shown are ECs with p values < 0.0001 (adjusted for multiple comparisons using Benjamini-Hochberg False Discovery Rate).**

**Supplemental Tables**

**Tables S1 – S2.**

**Please reference provided CD for these tables.**

Table S3

Table S3. P values (Student t-test with 1000 Monte Carlo permutations) of UniFrac distances between fecal communities of children and adults shown in Fig. 1b.

	Within population comparisons				Between populations comparisons			
	Malawi		USA		Adults		Children under 3y/o	
	Children 3-17 y/o vs Children under 3y/o	Children 3-17 y/o vs Children under 3y/o	Children 3-17 y/o vs Children under 3y/o	Children 3-17 y/o vs Children under 3y/o	Amr vs Malawi	Amr vs USA	USA vs Malawi	USA vs Malawi
Within population comparisons	Adult vs Adult	0.039	0.123	<0.001	<0.001	<0.001	<0.001	<0.001
	Children 3-17 y/o vs Children under 3y/o	<0.001	<0.001	0.002	0.342	0.462	<0.001	<0.001
	Adult vs Adult	<0.001	<0.001	0.122	<0.001	<0.001	0.105	0.034
	Children 3-17 y/o vs Children under 3y/o	<0.001	0.327	<0.001	<0.001	<0.001	<0.001	<0.001
Between populations comparisons	Amr vs Malawi	<0.001	<0.001	<0.001	<0.001	0.083	<0.001	<0.001
	USA vs Malawi	<0.001	<0.001	<0.001	<0.001	0.195	<0.001	<0.001
	Amr vs USA	<0.001	<0.001	<0.001	<0.001	<0.001	0.083	<0.001
	USA vs Malawi	<0.001	<0.001	<0.001	<0.001	<0.001	0.237	0.226
Children Under 3y/o	Amr vs Malawi	<0.001	<0.001	<0.001	<0.001	<0.001	0.387	<0.001
	USA vs Malawi	<0.001	<0.001	<0.001	<0.001	<0.001	0.004	<0.001
	Amr vs USA	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	USA vs Malawi	<0.001	<0.001	<0.001	<0.001	<0.001	0.05	<0.001

P values (Student t-test with 1000 Monte Carlo permutations) of Hellinger distances between fecal communities of children and adults shown in Fig. 1c.

	Within population comparisons				Between populations comparisons			
	Malawi		USA		Adults		Children under 3y/o	
	Children 3-17 y/o vs Children under 3y/o	Children 3-17 y/o vs Children under 3y/o	Children 3-17 y/o vs Children under 3y/o	Children 3-17 y/o vs Children under 3y/o	Amr vs Malawi	Amr vs USA	USA vs Malawi	USA vs Malawi
Within population comparisons	Adult vs Adult	0.121	0.067	0.004	0.086	0.007	0.012	<0.001
	Children 3-17 y/o vs Children under 3y/o	0.059	0.02	0.074	0.006	0.09	0.012	0.367
	Adult vs Adult	0.079	0.357	0.056	0.555	0.004	0.197	0.039
	Children 3-17 y/o vs Children under 3y/o	0.003	0.021	0.045	0.004	0.157	0.013	<0.001
Between populations comparisons	Amr vs Malawi	<0.001	<0.001	<0.001	0.272	0.064	0.166	0.027
	USA vs Malawi	<0.001	<0.001	<0.001	0.003	0.021	0.023	0.023
	Amr vs USA	<0.001	<0.001	<0.001	0.001	0.001	0.572	0.385
	USA vs Malawi	<0.001	<0.001	<0.001	0.084	0.001	<0.001	<0.001
Children under 3y/o	Amr vs Malawi	<0.001	<0.001	<0.001	0.003	0.001	<0.001	0.003
	USA vs Malawi	<0.001	<0.001	<0.001	0.442	0.001	<0.001	0.442

p<0.005 p<0.05



Table S4

Table S4. List of the 126 reference human gut microbial genomes

Genome name	Genbank ID	Genome size
<i>Actinomyces odontolyticus</i> ATCC 17982	NZ_AAYI00000000	2393758
<i>Akkermansia muciniphila</i> ATCC BAA-835	NC_010655	2664102
<i>Alistipes putredinis</i> DSM 17216	NZ_ABFK00000000	2549878
<i>Anaerococcus hydrogenalis</i> DSM 7454	NZ_ABXA00000000	1889366
<i>Anaerofustis stercorihominis</i> DSM 17244	NZ_ABIL00000000	2284603
<i>Anaerostipes caccae</i> DSM 14662	NZ_ABAX00000000	3605636
<i>Anaerotruncus colihominis</i> DSM 17241	NZ_ABGD00000000	3718888
<i>Bacteroides caccae</i> ATCC 43185	NZ_AAVM00000000	4564814
<i>Bacteroides capillosus</i> ATCC 29799	NZ_AAXG00000000	4241076
<i>Bacteroides cellulosilyticus</i> DSM 14838	NZ_ACCH00000000	6726268
<i>Bacteroides coprocola</i> DSM 17136	NZ_ABIY00000000	4295617
<i>Bacteroides coprophilus</i> DSM 18228	NZ_ACBW00000000	3855443
<i>Bacteroides dorei</i> DSM 17855	NZ_ABWZ00000000	5487768
<i>Bacteroides eggerthii</i> DSM 20697	NZ_ABVO00000000	4157980
<i>Bacteroides finegoldii</i> DSM 17565	NZ_ABXI00000000	4881901
<i>Bacteroides fragilis</i> 3_1_12	NZ_ABZX00000000	5486240
<i>Bacteroides fragilis</i> NCTC 9343	NC_003228	5205140
<i>Bacteroides fragilis</i> YCH46	NC_006347	5277274
<i>Bacteroides intestinalis</i> DSM 17393	NZ_ABJL00000000	6052596
<i>Bacteroides ovatus</i> ATCC 8483	NZ_AAXF00000000	6463169
<i>Bacteroides plebeius</i> DSM 17135	NZ_ABQC00000000	4421324
<i>Bacteroides</i> sp. 1_1_6	NZ_ACIC00000000	6855195
<i>Bacteroides</i> sp. D1	NZ_ACAB00000000	5986762
<i>Bacteroides</i> sp. D2	NZ_ACGA00000000	6901960
<i>Bacteroides stercoris</i> ATCC 43183	NZ_ABFZ00000000	4009229
<i>Bacteroides thetaiotaomicron</i> 3731	NC_Bthetaiotaomicron3731	7098445
<i>Bacteroides thetaiotaomicron</i> 7330	NC_Bthetaiotaomicron7330	6894436
<i>Bacteroides thetaiotaomicron</i> VPI-5482	NC_004663	6260361
<i>Bacteroides uniformis</i> ATCC 8492	NZ_AAYH00000000	4717497
<i>Bacteroides vulgatus</i> ATCC 8482	NC_009614	5163189
<i>Bacteroides</i> WH2	NC_BWH2	7129681
<i>Bacteroides xylanisolvens</i> XB1A	NC_BxylanisolvensXB1A	5861392
<i>Bifidobacterium adolescentis</i> ATCC 15703	NC_008618	2089645
<i>Bifidobacterium adolescentis</i> L2-32	NZ_AAXD00000000	2385710
<i>Bifidobacterium angulatum</i> DSM 20098	NZ_ABYS00000000	2007108
<i>Bifidobacterium animalis</i> subsp. <i>lactis</i> AD011	NC_011835	1933695
<i>Bifidobacterium animalis</i> subsp. <i>lactis</i> HN019	NZ_ABOT00000000	1915892
<i>Bifidobacterium breve</i> DSM 20213	NZ_ACCG00000000	2297799
<i>Bifidobacterium catenulatum</i> DSM 16992	NZ_ABXY00000000	2058429
<i>Bifidobacterium dentium</i>	NC_Bdentium	2642189
<i>Bifidobacterium gallicum</i> DSM 20093	NZ_ABXB00000000	2019802
<i>Bifidobacterium longum</i> DJO10A	NC_010816	2375792
<i>Bifidobacterium longum</i> NCC2705	NC_004307	2256640
<i>Bifidobacterium pseudocatenulatum</i> DSM 20438	NZ_ABXX00000000	2304808
<i>Blautia hansenii</i> DSM 20583	NZ_ABYU00000000	3053221
<i>Blautia hydrogenotrophica</i> DSM 10507	NZ_ACBZ00000000	3565428
<i>Bryantella formatexigens</i> DSM 14469	NZ_ACCL00000000	4548960
<i>Butyrivibrio crossotus</i> DSM 2876	NZ_ABWN00000000	2482791
<i>Catenibacterium mitsuokai</i> DSM 15897	NZ_ACCK00000000	2671313
<i>Citrobacter youngae</i> ATCC 29220	NZ_ABWL00000000	5143204
<i>Clostridium asparagiforme</i> DSM 15981	NZ_ACCJ00000000	6224391

<i>Clostridium bartlettii</i> DSM 16795	NZ_ABEZ00000000	2971856
<i>Clostridium bolteae</i> ATCC BAA-613	NZ_ABCC00000000	6556988
<i>Clostridium hiranonis</i> DSM 13275	NZ_ABWP00000000	2423348
<i>Clostridium hylemonae</i> DSM 15053	NZ_ABYI00000000	3885459
<i>Clostridium leptum</i> DSM 753	NZ_ABCB00000000	3270109
<i>Clostridium methylpentosum</i> DSM 5476	NZ_ACEC00000000	3406326
<i>Clostridium nexile</i> DSM 1787	NZ_ABWO00000000	3861016
<i>Clostridium ramosum</i> DSM 1402	NZ_ABFX00000000	3234795
<i>Clostridium scindens</i> ATCC 35704	NZ_ABFY00000000	3619905
<i>Clostridium</i> sp. L2-50	NZ_AAYW00000000	2954116
<i>Clostridium</i> sp. M62/1	NZ_ACFX00000000	3836694
<i>Clostridium</i> sp. SS2/1	NZ_ABGC00000000	3141381
<i>Clostridium spiroforme</i> DSM 1552	NZ_ABIK00000000	2507485
<i>Clostridium sporogenes</i> ATCC 15579	NZ_ABKW00000000	4102125
<i>Clostridium symbiosum</i>	NC_Csymbiosum	4954054
<i>Collinsella aerofaciens</i> ATCC 25986	NZ_AAVN00000000	2439869
<i>Collinsella intestinalis</i> DSM 13280	NZ_ABXH00000000	1804297
<i>Collinsella stercoris</i> DSM 13279	NZ_ABXJ00000000	2399821
<i>Coprococcus comes</i> ATCC 27758	NZ_ABVR00000000	3238915
<i>Coprococcus eutactus</i> ATCC 27759	NZ_ABEY00000000	3102087
<i>Desulfovibrio piger</i> ATCC 29098	NZ_ABXU00000000	2826240
<i>Desulfovibrio piger</i> GOR1	AF192152	2597386
<i>Dorea formicigenerans</i> ATCC 27755	NZ_AAXA00000000	3186031
<i>Dorea longicatena</i> DSM 13814	NZ_AAXB00000000	2913833
<i>Enterobacter cancerogenus</i>	NC_Ecancerogenus	4605129
<i>Escherichia coli</i> str. K-12 substr. MG1655	NC_000913	4639675
<i>Escherichia fergusonii</i> ATCC 35469	NC_011740	4588711
<i>Eubacterium bifforme</i> DSM 3989	NZ_ABYT00000000	2415920
<i>Eubacterium dolichum</i> DSM 3991	NZ_ABAW00000000	2190453
<i>Eubacterium eligens</i> ATCC 27750	NC_012778	2144190
<i>Eubacterium hallii</i> DSM 3353	NZ_ACEP00000000	3290996
<i>Eubacterium rectale</i> ATCC 33656	NC_012781	3449685
<i>Eubacterium rectale</i> DSM 17629	NC_Erectale_DSM17629	3255606
<i>Eubacterium ventriosum</i> ATCC 27560	NZ_AAVL00000000	2869695
<i>Faecalibacterium prausnitzii</i> A2-165	NZ_ACOP00000000	3080849
<i>Faecalibacterium prausnitzii</i> M21/2	NZ_ABED00000000	3126983
<i>Fusobacterium</i> sp. 4_1_13	NZ_ACDE00000000	2268505
<i>Fusobacterium varium</i> ATCC 27725	NZ_ACIE00000000	3321664
<i>Helicobacter pylori</i> HPAG1	NC_008086	1596366
<i>Holdemania filiformis</i> DSM 12042	NZ_ACCF00000000	3803745
<i>Lactobacillus casei</i> ATCC 334	NC_008526	2895264
<i>Lactobacillus delbrueckii</i> subsp. <i>bulgaricus</i> ATCC 1	NC_008054	1864998
<i>Lactobacillus reuteri</i> DSM 20016	NC_009513	1999618
<i>Lactococcus lactis</i> subsp. <i>cremoris</i> MG1363	NC_009004	2529478
<i>Lactococcus lactis</i> subsp. <i>cremoris</i> SK11	NC_008527	2438589
<i>Lactococcus lactis</i> subsp. <i>lactis</i> II1403	NC_002662	2365589
M23A	NC_M23A	4338875
<i>Methanobrevibacter smithii</i> ATCC 35061	CP000678.1	1853160
<i>Methanobrevibacter smithii</i> DSM 2374	NZ_ABYV00000000	1727775
<i>Methanobrevibacter smithii</i> DSM 2375	NZ_ABYW00000000	1704865
<i>Methanosphaera stadtmanae</i> DSM 3091	NC_007681	1767403
<i>Mitsuokella multacida</i> DSM 20544	NZ_ABWK00000000	2574556
<i>Parabacteroides distasonis</i> ATCC 8503	NC_009615	4811379
<i>Parabacteroides johnsonii</i> DSM 18315	NZ_ABYH00000000	4612238

Parabacteroides merdae ATCC 43184	NZ_AAXE00000000	4431877
Parvimonas micra ATCC 33270	NZ_ABEE00000000	1703772
Prevotella copri DSM 18205	NZ_ACBX00000000	3507873
Proteus penneri ATCC 35198	NZ_ABVP00000000	3747729
Providencia alcalifaciens DSM 30120	NZ_ABXW00000000	4029346
Providencia rettgeri DSM 1131	NZ_ACCI00000000	4749568
Providencia rustigianii DSM 4541	NZ_ABXV00000000	3965844
Providencia stuartii ATCC 25827	NZ_ABJD00000000	4603561
Roseburia intestinalis L1-82	NZ_ABYJ00000000	4380675
Ruminococcus bromii L263	NC_RbromiiL263	2240019
Ruminococcus gnavus ATCC 29149	NZ_AAYG00000000	3501911
Ruminococcus lactaris ATCC 29176	NZ_ABOU00000000	2729735
Ruminococcus obeum ATCC 29174	NZ_AAVO00000000	3624708
Ruminococcus torques ATCC 27756	NZ_AAVP00000000	2739406
Shigella sp. D9	NZ_ACDL00000000	4717340
Streptococcus infantarius subsp. infantarius ATCC	NZ_ABJK00000000	1925087
Streptococcus thermophilus CNRZ1066	NC_006449	1796226
Streptococcus thermophilus LMD-9	NC_008532	1856368
Streptococcus thermophilus LMG 18311	NC_006448	1796846
Subdoligranulum variabile DSM 15176	NZ_ACBY00000000	3237471
Victivallis vadensis ATCC BAA-548	NZ_ABDE00000000	5294868

**Table S5**

**Please reference provided CD for this table.**

Tables S6

**Table S6. Results of Random Forests classifier of OTUs (species-level phylotypes) that discriminate adult fecal microbiota of USA and non-USA residents.** The rarefaction depth for was 718 sequences/sample. One hundred even rarefactions were performed for the comparison.

OTU-ID	Taxonomic assignment	Importance Score		Relative abundance of an OTU in each adult population (mean ± sd)				
		Average	Standard Error	Malawi	Amerindians	USA		
134040	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.01009	0.00012424	0.0172 ± 0.01766	0.0114 ± 0.011934	0.000002 ± 3E-05		
124834	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae	0.00696	0.00012612	0.00517 ± 0.00608	0.0101 ± 0.010416	0.000206 ± 0.0016		
90608	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00685	9.91E-05	0.01498 ± 0.01602	0.0155 ± 0.016171	0.000286 ± 0.0012		
55432	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00678	0.00010322	0.01106 ± 0.02112	0.0081 ± 0.011712	0.000227 ± 0.0026		
85597	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00676	8.56E-05	0.03394 ± 0.02579	0.033 ± 0.029436	0.002633 ± 0.0114		
71975	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae	0.00660	8.82E-05	0.0064 ± 0.00684	0.0177 ± 0.023211	0.000365 ± 0.0024		
106091	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae	0.00624	0.00011498	0.00284 ± 0.00335	0.0057 ± 0.007412	0 ± 0		
46709	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae	0.00536	9.17E-05	0.0087 ± 0.00874	0.0078 ± 0.014577	0 ± 0		
42900	Root;Bacteria;Proteobacteria;Gammaproteobacteria;Aeromonadales;Succinimicrobiota	0.00530	9.10E-05	0.0086 ± 0.01379	0.0071 ± 0.020178	0 ± 0		
22020	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae	0.00523	0.00011548	0.00183 ± 0.00277	0.004 ± 0.005169	0 ± 0		
109025	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae	0.00520	9.38E-05	0.00981 ± 0.01118	0.0029 ± 0.004513	0 ± 0		
17083	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00448	0.00010839	0.00454 ± 0.00549	0.007 ± 0.010248	0.000073 ± 0.0008		
65248	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00436	9.82E-05	0.00432 ± 0.00534	0.0038 ± 0.005857	0 ± 0		
89034	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00410	7.49E-05	0.00895 ± 0.00895	0.0117 ± 0.013095	0.001242 ± 0.0059		
40805	Root;Bacteria;Firmicutes;Clostridia;Clostridiales;Verrucomicrobiales;Dialister	0.00395	9.74E-05	0.00301 ± 0.00399	0.0039 ± 0.00786	0.000011 ± 0.0001		
36049	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00395	7.82E-05	0.00841 ± 0.0088	0.0007 ± 0.002106	0 ± 0		
100541	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00355	7.28E-05	0.01172 ± 0.01134	0.0097 ± 0.010827	0.001401 ± 0.0064		
98583	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae	0.00340	0.00010314	0.00063 ± 0.00157	0.0043 ± 0.005684	0 ± 0		
21732	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00333	6.47E-05	0.00286 ± 0.00727	0.0043 ± 0.006524	0 ± 0		
38860	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00301	7.31E-05	0.00537 ± 0.0094	0.0047 ± 0.008697	0 ± 0		
16256	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00300	8.32E-05	0.00259 ± 0.00369	0.0033 ± 0.006924	0.000014 ± 9E-05		
30007	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00289	8.49E-05	0.00267 ± 0.00443	0.0051 ± 0.008684	0.000193 ± 0.0018		
46852	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae	0.00286	5.20E-05	0.00718 ± 0.00891	0.0063 ± 0.013526	0 ± 0		
18892	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00269	8.55E-05	0.0038 ± 0.01146	0.0013 ± 0.001197	0.00004 ± 0.0003		
38859	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae	0.00266	6.82E-05	0.00624 ± 0.00998	0.0004 ± 0.000677	0.000194 ± 0.0019		
102809	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00238	8.11E-05	0.00297 ± 0.00513	0.0024 ± 0.004531	0.000106 ± 0.0008		
40976	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00231	8.07E-05	0.00183 ± 0.00288	0.0024 ± 0.00464	0.000008 ± 6E-05		
95404	Root;Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae	0.00228	6.76E-05	0.00226 ± 0.00194	0.0006 ± 0.002154	0.000011 ± 0.0001		
57317	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00220	7.22E-05	0.00102 ± 0.00142	0.0012 ± 0.001279	0.000003 ± 4E-05		
45007	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00204	5.60E-05	0.00498 ± 0.00825	0.0026 ± 0.005408	0 ± 0		
13790	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00203	7.42E-05	0.00626 ± 0.00991	0.001 ± 0.001726	0.000033 ± 0.0002		
62661	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae	0.00202	5.82E-05	0.00276 ± 0.00351	0.0006 ± 0.001058	0.000043 ± 0.0005		
94680	Root;Bacteria;Bacteroidetes	0.00201	4.86E-05	0.00394 ± 0.00645	0.0039 ± 0.007223	0 ± 0		
6564	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00192	5.32E-05	0.00216 ± 0.0053	0.0017 ± 0.003152	0 ± 0		
20413	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae	0.00189	5.26E-05	0.00465 ± 0.00967	0.0005 ± 0.000723	0 ± 0		
117561	Root;Bacteria	0.00188	5.86E-05	0.00066 ± 0.00161	0.0022 ± 0.003446	0 ± 0		
51351	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00187	7.36E-05	0.00096 ± 0.00156	0.0024 ± 0.003298	0.000057 ± 0.0003		
110969	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Prevotellia	0.00182	7.47E-05	0.00149 ± 0.00135	0.0022 ± 0.002164	0.000266 ± 0.0014		
117393	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae;Bacteroides	0.00181	3.27E-05	0.00082 ± 0.00356	0.0006 ± 0.001217	0.036813 ± 0.0438		
82205	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae	0.00170	5.18E-05	0.00391 ± 0.00624	0.0009 ± 0.001963	0.000106 ± 0.0012		
101130	Root;Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae	0.00165	6.33E-05	0.0013 ± 0.00222	0.0011 ± 0.001961	0.000044 ± 0.0005		
84656	Root;Bacteria;Bacteroidetes;Bacteroidales;Bacteroidales;Prevotellaceae	0.00153	5.34E-05	0.00017 ± 0.0004	0.0036 ± 0.006497	0 ± 0		

116996	Root;Bacteria;Proteobacteria;Gammaproteobacteria;Aeromonadales;Succinivibrionaceae;Succinivibrio	0.00152	5.86E-05	0.00248 ± 0.0046	0.002 ± 0.005316	0	± 0
88987	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Prevotellaceae;Prevotella	0.00147	5.99E-05	0.00119 ± 0.00127	0.0005 ± 0.000558	0	± 0
34912	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Prevotellaceae	0.00144	4.15E-05	0.00216 ± 0.004	0.0042 ± 0.012565	0	± 0
1930	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Prevotellaceae	0.00143	5.00E-05	0.00943 ± 0.02897	0.0008 ± 0.001948	0.000001	± 8E-06
46088	Root;Bacteria;Firmicutes;Erysipelotrichales;Erysipelotrichaceae;Catenibacterium	0.00141	6.70E-05	0 ± 0	0.0028 ± 0.004394	0.000183	± 0.0019
86363	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Prevotellaceae;Prevotella	0.00134	4.85E-05	0.00554 ± 0.0123	0.0028 ± 0.009092	0	± 0
75220	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Prevotellaceae;Prevotella	0.00133	5.09E-05	0.0009 ± 0.00165	0.0024 ± 0.005323	0.000005	± 5E-05
91002	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Parabacteroides	0.00126	4.18E-05	0.00291 ± 0.00419	0.0012 ± 0.002882	0	± 0
18575	Root;Bacteria;Firmicutes;Clostridia;Clostridiales;Prevotellaceae;Prevotella	0.00126	5.24E-05	0.0019 ± 0.00261	0.0037 ± 0.004578	0.000288	± 0.0015
122821	Root;Bacteria;Firmicutes;Clostridia;Clostridiales;Veillonellaceae	0.00122	4.64E-05	0.00237 ± 0.00492	0.006 ± 0.012972	0.000006	± 7E-05
9972	Root;Bacteria	0.00118	3.55E-05	0.00198 ± 0.00348	0.0039 ± 0.01721	0	± 0
105053	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Prevotellaceae	0.00114	5.51E-05	0.00134 ± 0.00233	0.0012 ± 0.001915	0.000011	± 1E-04
79081	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Prevotellaceae;Prevotella	0.00114	6.39E-05	0.00027 ± 0.00092	0.0009 ± 0.000829	0.000007	± 5E-05
41312	Root;Bacteria;Proteobacteria;Gammaproteobacteria;Aeromonadales;Succinivibrionaceae	0.00114	4.71E-05	0.00212 ± 0.00565	0.001 ± 0.002844	0	± 0
12360	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Succinivibrionaceae	0.00110	3.21E-05	0.00014 ± 0.00064	0.0007 ± 0.002117	0.03027	± 0.0457
34233	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidales;Bacteroides	0.00110	3.05E-05	0.00488 ± 0.00747	0.0001 ± 0.000349	0.000015	± 0.0001
107140	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Prevotellaceae	0.00110	3.98E-05	0.00099 ± 0.00228	0.0011 ± 0.001783	0	± 0
34885	Root;Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Subdoligranulum	0.00109	6.40E-05	0.00014 ± 0.00033	0.0008 ± 0.000773	0	± 0
28423	Root;Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Clostridiaceae 1;Clostridium	0.00108	4.16E-05	0.00266 ± 0.00288	8E-05 ± 0.000137	0.000238	± 0.001

**Tables S7 – S11.**

**Please reference provided CD for these tables.**

## **Chapter 4**

**Temporal variation in the gut microbiomes of healthy and twin pairs discordant for severe malnutrition.**



## **Introduction**

Undernutrition is a leading cause of childhood deaths in the world (Bryce, 2005). Those who suffer from severe forms of malnutrition, such as kwashiorkor or marasmus, are especially at high risk. There are a number of long-term consequences of undernutrition, including stunting and neurodevelopmental disorders. As noted in Chapter 1 of this thesis, kwashiorkor was first described by Williams in 1931 (Williams 1973). Prominent phenotypic features include generalized edema, hepatic steatosis and depigmentation of the skin (Blackburn 2001). Its etiology remains obscure. An early hypothesis held that it was caused by a low protein diet. However, epidemiologic studies subsequently showed that the diets of children with kwashiorkor do not differ significantly from those with marasmus (Golden 2002, Lin et al. 2007). Moreover, the edematous malnutrition that characterizes kwashiorkor can resolve on a low-protein diet without accompanying changes in plasma protein levels. Marasmus, which has a higher mortality rate (Scrimshaw and Viteri, 2010), is characterized by severe wasting: its onset typically occurs earlier than kwashiorkor [6-17 months versus up to 4 years (Courtright & Canner 1995, Ahmed et al. 2009)]. A ready-to-use-therapeutic food (RUTF), composed of peanut butter, sugar, vegetable oil and milk fortified with vitamins and minerals, has been developed recently: clinical studies indicate that it has increased efficacy in the treatment of severe forms of malnutrition compared to standard nutritional rehabilitation protocols (Ciliberto et al. 2005).

The role of the gut microbiome in the pathogenesis of kwashiorkor and marasmus is unclear. However, there is mounting evidence that the nutritional value of food is influenced in part by a consumer's gut microbial community (microbiota), and that food in turn shapes the composition and operations of microbiota and its vast collection of microbial genes (the 'gut microbiome') (Muegge et al., 2011; Wu et al., 2011). Information is also rapidly accumulating that many features of our metabolic phenotypes (metabotypes) are a reflection of enzymatic activities encoded in our human genomes and gut microbiomes. Chapter 3 describes a developmental program where the proportional representation of

genes encoding metabolic functions related to micro- and macronutrient processing and biosynthesis changes as healthy infants and children develop. The link between infections that occur within and outside the gastrointestinal tract, and the development of nutritional deficiencies has been emphasized for many years (Golden 2002, Prentice et al., 2008). We are learning how our gut microbial communities and immune systems co-develop, how microbiota influences mucosal barrier function and impedes invasion with enteropathogens. This barrier function can be disrupted by malnutrition as well as by perturbations in immune function. Poor nutrition in turn, increases the risk for infection.

Together, these observations give rise to the following testable hypotheses: (i) the gut microbiome is a microbial metabolic organ that provides key functions needed for healthy postnatal growth and development; (ii) disturbances in microbiome assembly and function, including those prompted by enteropathogen infection, affect the risk for kwashiorkor or marasmus; (iii) in a self-reinforcing pathogenic cascade, undernutrition in infants affects gut microbiome functions involved in determining host nutritional status, thus further worsening health status; (iv) there may be a number of gut microbiome configurations associated with kwashiorkor or marasmus among different hosts and even within a given host over time; (v) microbiome configurations associated with kwashiorkor or marasmus may be differentially affected by RUTF and features that are reconfigured during the treatment may not persist after its withdrawal, indicating a need for longer term nutritional support to repair microbiome-associated metabolic lesions that may underlie lingering host phenotypes associated with a history of kwashiorkor or marasmus.

### **Study Design**

Prospective, longitudinal, comparative metagenomic studies of functional features present in the developing gut microbiomes of healthy infants and children, and those who develop these two severe forms of malnutrition, have not been reported. Therefore, we have conducted a pilot study characterizing the configuration of fecal microbiota and microbiomes

in mono- and dizygotic twin pairs who were born in Malawi and became discordant for kwashiorkor and marasmus. Malawi, located in southeastern Africa, has one of the highest infant mortality rates in the world (81 deaths/1,000 live births), with 42% child's deaths associated with malnutrition (Pelletier et al., 1994). In designing the study, we had to take into account that there is considerable intra- as well as interpersonal variation in the organismal content of the gut microbiota during infancy (Palmer et al. 2007). We reasoned that a healthy co-twin in a twin pair discordant for kwashiorkor or marasmus represented a very desirable control given his/her genetic relatedness to the affected co-twin, and their similar exposures to diet, the microbiota of family members including the mother, and to other microbial reservoirs that exist in their shared early environment. Following each co-twin in a twin pair prospectively would allow each individual to serve as his/her own control (e.g. comparing their microbiome structure before the onset of disease, plus during and after treatment); if there were many different routes to disrupted microbiome structure/function, then each discordant twin pair could provide a 'vignette' of the pathologic process. Being a twin not only increases the risk of malnutrition but comparing the incidence of discordance for severe malnutrition among mono- versus dizygotic twins would allow us to assess the contributions of host genotype to kwashiorkor or marasmus. Moreover, in Malawi, the standard of care for twins discordant for kwashiorkor or marasmus is to treat both the healthy and malnourished co-twin with RUTF, allowing us to compare and contrast the responses of their microbiomes prior to, during and after treatment. Finally, as an additional set of controls, we defined the gut microbiome assembly and encoded functions in twin pairs who remained healthy and lived in the same geographic location as discordant pairs.

Twin pairs who were less than 3 years old were enrolled regardless of their health status from five villages (Makhwira, Mitondo, M'biza, Chamba, Mayaka) located in the southern region of Malawi (**Fig. 1**). There are two seasons in this region: a rainy season that lasts from November to May and coincides with the highest incidence of malnutrition, and dry season (CIA 2011). In the southern region, 58.9% infants are exclusively breastfed

during the first 6 months of life; 70.3% children continue to be breastfed in the first 2 years of life (WHO 2006).

317 twin pairs and 3 families of triplets were recruited. The average age at enrollment was  $9\pm 6$  months; all but 19 twin pairs were being breastfed (average age at the cessation of breastfeeding,  $18\pm 5$  months). Children were followed until they were 36 months old. Zygosity testing, using buccal DNA and a custom array containing 48 autosomal SNPs with high heterozygosity in the Yoruban HapMap sample (see *Methods*), revealed that 46 of the twin pairs, 1 set of triplets, and 2 children in another triplet were monozygotic (MZ). The percentage of same gender twin pairs in the recruited cohort was higher than opposite gender pairs ( $p=7.53e-9$ , Binomial test, probability of success = 66.24%, CI =(0.61,0.71), **Table 1a**).

Our study design is depicted in **Fig.2**. Assessments of nutritional status and anthropometric data were conducted once a month; a fecal sample was collected every 2 months up until 1 year of age, and every 3 months thereafter. Kwashiorkor was diagnosed based on the presence of a pitting edema. Marasmus was diagnosed when an infant or child had a weight-for-height Z score (WHZ) that was less than -3. Children with WHZ scores between -2 and -3 were classified as moderately undernourished. Twin pairs who developed kwashiorkor or marasmus immediately received a peanut butter-based, ready-to-use therapeutic food (RUTF), while those with moderate malnutrition received a soy-based diet (Matilsky et al., 2009). After the diagnosis with malnutrition, health status was monitored and fecal sample collected every 2 weeks until the child recovered (in the case of marasmus, when WHZ scores of children with marasmus became greater than -2; in the case of kwashiorkor, with resolution of edema). On average, twin pairs where one or both children developed kwashiorkor received 4 weeks of RUTF treatment, while those with marasmus were treated for 7 weeks.

50% of the enrolled twin pairs remained healthy throughout the study, 43% twin pairs became discordant for undernutrition, and 7% manifested concordance for under-

nutrition (**Table 1a**). The prevalence of discordant compared to concordant phenotypes was significant ( $p < 2e-16$ , Binomial and chi-squared tests). Moderate malnutrition was significantly more frequent than severe malnutrition, affecting 60% of the discordant twin pairs ( $p = 0.02$ , Chi-square test). 7% of the children in this cohort experienced episodes of kwashiorkor, 2.5% developed marasmus, and 14% exhibited moderate malnutrition. 10% of children experienced multiple episodes of these forms of undernutrition (i.e., kwashiorkor, marasmus and/or moderate malnutrition), with the most frequent combination being marasmus and moderate malnutrition (5.4% children, **Table 2**). **Table 3** indicates the ages, anthropometric measurements and incidence of symptoms of diarrhea, fever, cough and vomiting at the time a child presented with undernutrition. The average age at the presentation with marasmus was  $11 \pm 4$  months; the corresponding values for kwashiorkor and moderate malnutrition were  $16 \pm 7$  and  $14 \pm 7$  months, respectively. Children with marasmus had the lowest anthropometric values: weight for height Z score (WHZ) was  $-3.6 \pm 0.6$ , height-for-age Z score (HAZ) was  $-3.7 \pm 1.3$  and weight-for-age Z score (WAZ) was  $-4.6 \pm 1$ . In addition, children with marasmus suffered significantly more episodes of diarrhea than did those with kwashiorkor or moderate malnutrition ( $p < 0.05$ ; see **Table 3**). Moreover, the highest proportion of deaths was found in co-twins that were members of twin pairs discordant for marasmus, compared to those who were discordant for kwashiorkor or moderate malnutrition ( $p = 0.0027$ , Chi-square test): one co-twin died during the study in 43% of the twin pairs who were discordant for marasmus versus 9% in twin pairs discordant for kwashiorkor and 14% pairs discordant for moderate malnutrition (**Table 1a**). The majority of deaths were caused by diarrhea, malaria or pneumonia. Unfortunately, death of a child in a family was not a rare event: in 120 out of 320 (37.5%) sampled families at least 1 child (not twin) died in a family at some point before the study began. HIV prevalence was low among twins: only 1 twin pair was HIV positive and 11 of 266 sampled mothers were reported as positive.

### **Influence of genetics, geography and gender on susceptibility to malnutrition**

Recruitment of both, MZ and dizygotic (DZ) twin pairs allowed us to assess the role of host genetics in the susceptibility to malnutrition. There was no significant relationship between the concordance for malnutrition and zygosity: the number of MZ twin pairs concordant for moderate or severe malnutrition was not significantly different from the number of DZ twins concordant for any of these diseases (**Table 1b**). We did not find significant differences in the number of MZ versus DZ twin pairs affected with any type of malnutrition in our twin cohort (Chi-squared and Fisher's exact tests). Among same gender twin pairs discordant for severe malnutrition, only 1 MZ twin pair was discordant for marasmus, while 7 MZ twin pairs were discordant for kwashiorkor; however, this difference was not significant, perhaps due to low number of subjects. Taking all 135 discordant pairs into account, we did not find a statistically significant difference in the incidence of discordance for marasmus, kwashiorkor, or moderate malnutrition in MZ versus DZ twins (**Table 1a**). We did not find any association between gender and the type of malnutrition, or the frequency of malnutrition, or discordance for malnutrition among twin pairs (Chi-squared and Fisher's exact tests, **Table 2**).

We found a significant association between where a village was located and the frequency of kwashiorkor ( $p=0.0044$ , Chi-square test, **Table 3**): the incidence of kwashiorkor was lowest in the 2 southernmost villages surveyed (**Fig. 1**); these villages were located in the Shire River valley, where soil is typically more fertile. Cultivation of cotton and sugar cane occurs in this area unlike in the 3 other sites. A lower prevalence of kwashiorkor in this area had been reported before (Courtright and Canner, 1995).

### **Sampling fecal microbiomes from twins who were concordant for healthy status and twins who were discordant for severe malnutrition**

For our pilot survey of functional changes in the fecal microbiomes in these children, we focused on 9 same gender twin pairs who never developed malnutrition through-

out the course of the study (“healthy” group), 13 same gender twin pairs who became discordant for kwashiorkor, and 10 same gender twin pairs discordant for marasmus. **Table 4** provides a summary of the characteristics of these families, and **Table 5** provides detailed information about each fecal sample collected from these twins. Anthropometric data in the form of Z scores collected over the course of the study for these twin pairs is shown in **Fig. 3**. On average,  $5 \pm 1$  fecal samples were collected every 2-3 months from twins who remained healthy. For twin pairs discordant for severe malnutrition, fecal samples were collected every 2-3 months before diagnosis of malnutrition, every 2 weeks during the RUTF treatment and 1 month immediately after cessation of RUTF therapy (**Fig. 2**). This resulted in  $8 \pm 3$  samples from twins discordant for kwashiorkor and  $8 \pm 5$  samples from twin pairs discordant for marasmus (**Table 4**). Fecal samples were immediately frozen in liquid nitrogen at the site where they were produced by the children, transported to the laboratory, and subsequently stored at  $-80^{\circ}\text{C}$ .

Total genomic DNA was isolated from each pulverized fecal sample, and subjected to multiplex shotgun pyrosequencing (454 FLX Titanium chemistry;  $82,138 \pm 43,730$  high quality sequences per sample; total size of the dataset is 14.5 Gb). Human sequences ( $6 \pm 14\%$  of reads) were identified by BLASTn searches of the human genome, and discarded. The remaining reads were functionally annotated by comparison to the KEGG database (version 58). Reads were assigned to KEGG orthologous groups (KOs) and enzyme commission numbers (ECs) (BLAST e-value cutoff  $<10^{-5}$ );  $40 \pm 8\%$  of sequences received annotations using this procedure (**Fig.4**). In addition to sequencing whole community DNA, we sequenced amplicons generated from variable region 2 of bacterial 16S rRNA genes represented in 1,041 fecal samples in order to survey the phylogenetic composition of these communities as a function of age, health status, and treatment. On average,  $2,055 \pm 2,254$  reads were obtained from each fecal sample, resulting in 2,139,576 total reads (average read length 250 nt).

## Comparison of fecal microbiomes across all children

To compare functional gene profiles across all 453 fecal samples that had been subjected to shotgun sequencing, we calculated Hellinger distances from KEGG KO assignments. We confirmed our previous finding (Chapter 3 of this thesis) of greater inter-individual variation at younger ages: the average distance between microbiomes sampled during the first few months of life was much greater than the distance between microbiomes sampled at older ages (**Fig. 5**). We then used principal coordinates analysis (PCoA) of the Hellinger distances to visualize variation in this dataset (**Fig. 6**). Principal coordinate 1, which explained the largest amount of variation (18%), was strongly associated with age and family membership (linear mixed model, **Fig. 6a,b**): i.e., within each family the functional composition of fecal microbiomes was changing at a similar rate as children matured. Microbiomes of twins affected with kwashiorkor or marasmus did not cluster in a fashion that was significantly differently from their healthy co-twins along PC1 (**Fig. 6c,d**). This lack of strong clustering is not surprising given the high inter-individual variation observed during the first 3 years of life even in the microbiomes of healthy children from Malawi as well as other countries, plus the strong familial similarity (Chapter 3 of this thesis).

When we examined the distribution of samples (sequenced and annotated fecal microbiomes) along other principal coordinates, we noticed that variation along PC2 was driven by several samples primarily from twin pairs discordant for marasmus (**Fig. 6e**), although the association between position along PC2 and health status was not significant according to Spearman correlation and mixed model regression analyses. However, the position of these samples along PC2 was positively correlated with the representation of bacterial taxa belonging to the family *Enterobacteriaceae*, and negatively correlated with the representation of the genera *Bifidobacterium* and *Colinsella* (**Fig 6f**). Although a large number of common enteropathogens are found in the family *Enterobacteriaceae*, gut microbiomes with the highest proportional representation of this family did not come from children with the most severe cases of malnutrition.



Because of the high degree of intra- and inter-personal variation due to age, and because of the significant familial similarity, we did not expect to observe common responses across all families to malnutrition or to RUTF treatment. Nonetheless, our study design allowed us to control for age and family effects by comparing a healthy co-twin to his or her malnourished co-twin. But first we proceeded with the analysis of twin pairs who remained healthy throughout the study, in order to define baseline variation of the gut microbiome in healthy children in this region.

### **Temporal variation of the gut microbiomes of twin pairs who remained healthy**

The adult human gut microbiome is relatively stable over time: a recent study of healthy adult USA female twins who were sampled over 4 months showed that intra-individual variation was smaller than variation between twins from the same family; in other words, microbiomes from the same person were most similar to one another, and then to a co-twin (McNulty et al. 2011). However, the extent of variation in infant microbiomes within an individual and family (in this case twin pair) has not been described in Malawi or other populations. Therefore, we first asked if intrapersonal as well as ‘familial’ similarity appears early in life. Hellinger distances were calculated between 93 KEGG KO profiles obtained from 9 healthy twin pairs sampled between 3 weeks and 24.5 months of age (**Table 5**). When Hellinger distances between microbiomes sampled from the same child were compared to distances between microbiomes of related children, we found that unlike in adults, on average the degree of intrapersonal variation was not smaller than variation between co-twins. This was true for phylogenetic composition (defined by UniFrac distances obtained from comparison of 16S rRNA sequences), as well as functional gene content, and was independent of distance metrics (**Fig. 7**). In other words, familial similarity was evident in the first 2 years of life while the individual makeup of the gut microbial community was not observed during the age range of 1-24 months. However, as co-twins mature, the distance between their microbiomes increases; this was most evident in terms

of phylogenetic composition (**Fig. 8**, linear mixed model regression). As co-twins grow they become more autonomous, and they are able to interact with other people, animals and other microbial reservoirs in their environment, which may explain the observed increasing dissimilarity in the phylogenetic content of their microbiomes.

Despite this large variation early on, common patterns of ‘maturation’ of the gut microbiome could be described. The rate of ‘maturation’ of the gut microbiome with age did not differ significantly between families, when positions of microbiomes along PC1 were regressed against age (**Fig. 9**). Just as we noted in Chapter 3 of this thesis, fecal microbiomes sampled during the first 6 months were dominated by members of Bifidobacteria; as children age, the proportional representation of Bifidobacteria diminished while the representation of Bacteroidetes and Firmicutes increased (**Fig. 10a**). Interestingly, in this infant population, the representation of different Bifidobacteria species varied with age: *B. longum* dominated during the first 7 months; this phylotype was superseded by *B. catenulatum* and *B. pseudocatenulatum* (**Fig. 10b**).

To characterize age-associated changes in the functional repertoire of microbiomes, we used Spearman correlation analysis of the representation of ECs encoded by the fecal microbial communities of infants aged 1-24 months and their mothers (**Table 7**). The results were consistent with those reported in the Chapter 3 of this thesis. Intriguingly, we detected ECs involved in the degradation of complex polysaccharides in the samples as young as 3 weeks of age (e.g. starches) despite the fact that the diet was exclusively breast milk, indicating that the healthy infant gut microbiome is already equipped to process more complex dietary components even though they will be encountered later as supplemental feeding begins.

### **Temporal variation in the fecal microbiomes of twin pairs discordant for kwashiorkor**

Having characterized the variation in healthy twin pairs, we turned to 13 twin pairs in which one of the co-twins developed kwashiorkor (see **Tables 4** and **5** for subject characteristics). We started first by comparing intra- and inter-personal variations in phylogenetic and KEGG KO profiles. Much like in twins who remained healthy, the temporal variation within a child was equal to the variation between co-twins, but still smaller than in unrelated children (**Fig. 11a**). We then asked if the increased dissimilarity with age that we observed in healthy twins (**Fig. 7**) was also evident in the microbiomes of co-twins discordant for kwashiorkor. When distances between co-twins in a given twin pair were compared over time, there was no significant relationship with age, unlike twin pairs who remained concordant for healthy status during the first three years of life (**Figs. 7, 11b,c**). This result indicates greater degree of variation in the microbiomes of co-twins discordant for kwashiorkor compared to twin pairs who remained healthy. Nevertheless, there were no significant correlations between Hellinger distances between co-twin microbiomes at each of three key time points: at the time of presentation with kwashiorkor, 2 weeks into RUTF treatment, and 1 month after cessation of RUTF (**Fig. 11d,e**).

We then used PCoA of KO profiles to further compare microbiomes. The results revealed that much like in the healthy twin pairs, age and family membership were significantly correlated with the PC1 (31% variation, mixed model linear regression, **Fig. 12a,b**). Because the “age” variable implies not only physiological maturation of a growing child, but also behavioral and dietary changes, exposure to pathogens, as well as changes in the nutritional status of a host, we wanted to know if the fecal microbiomes of children who had kwashiorkor matured differently from their healthy siblings. When positions along PC1 were compared using only 3 samples – at the time one of the twins presented with kwashiorkor, 2 weeks into RUTF and 1 month after cessation of RUTF, we found that microbiomes of healthy children progressed steadily towards more ‘mature’ configuration: microbiomes 1 month after cessation of RUTF were significantly ‘older’ compared

to before RUTF ( $p < 0.05$ , Friedman test with Dunn's post-hoc comparison, **Fig. 12c,d**). However, this was not the case for their siblings with kwashiorkor: the difference between before and after RUTF was not significant, but on average, fecal microbiomes sampled at 2 weeks into RUTF treatment were significantly more 'mature' compared to those sampled at kwashiorkor ( $p < 0.05$  after Dunn's post-hoc test). These findings indicate that the fecal microbiomes of kwashiorkor co-twins were more responsive to RUTF than the microbiomes of their healthy co-twins. But unlike the steady 'maturation' observed in the healthy co-twins, the average position along PC1 shifted towards a 'younger' state once RUTF treatment stopped, implying that the kwashiorkor-associated microbiome, or the environment in the gut was not able to sustain the 'mature' configuration induced by RUTF (**Fig. 12c,d**).

Similar to twin pairs who remained concordant for healthy status throughout the study, Bifidobacteria dominated in the first 10 months of life twin pairs discordant for kwashiorkor, with subsequent dominance by members of the Bacteroidetes and Firmicutes (**Fig. 13a**). Using the 127 human gut genomes database for phylogenetic assignments of shotgun sequences generated from fecal microbiomes, we compared representation of major bacterial phyla in discordant co-twins at the time that the co-twin first presented with kwashiorkor, 2 weeks after initiation of RUTF treatment, and 1 month after cessation of RUTF. We found that representation of Actinobacteria decreased significantly with RUTF only in co-twins afflicted with kwashiorkor, but not their healthy siblings ( $p < 0.05$ , **Fig. 13b**).

### **Changes in KEGG ECs involved in various metabolic functions in the fecal microbiomes of co-twins discordant for kwashiorkor**

We next determined the functional gene changes associated with kwashiorkor and RUTF treatment. Because age and familial similarity largely determined the composition of fecal microbiomes across all twin pairs, we examined the functional differences in each family individually. We used Fisher's exact test to determine the differences in representation of

genes encoding KEGG ECs between a healthy and malnourished co-twin within a family at the time of presentation with kwashiorkor, 2 weeks into RUTF treatment and 1 month after termination of RUTF. As expected each family had a unique collection of ECs that were significantly different at each comparison (see **Table 8** for all ECs whose proportional representation was significantly different within each family). We provide an example of family k138, where functional differences in the fecal microbiomes of the twin pair were most pronounced compared to other twin pairs (**Figs. 12d** and **14**). These MZ male co-twins were enrolled in the study at 12.4 months of age when both were healthy; their last visit took place when they were 36.6 months old. WHZ and WAZ scores were declining in both twins prior to the development of kwashiorkor, but remained within the ‘healthy’ range (**Fig. 14a**). All Z scores were slightly lower in the co-twin who developed the disease. Weight improved immediately following RUTF treatment (**Fig. 14a**), and remained stable until both boys received anti-malarial treatment for 2 months when their weights dropped again temporarily. There were no increased symptoms of cough, fever, diarrhea or vomiting at the presentation of kwashiorkor (**Fig. 14b**), suggesting that acute infection may not have been the cause of kwashiorkor. We compared the representation of major bacterial phyla in both children (**Fig. 14c**): the co-twin who developed kwashiorkor had a significantly higher representation of Actinobacteria in his fecal microbiome prior to and at the time of diagnosis (Fisher’s exact test,  $p < 0.05$  after FDR correction), and a significantly lower representation of Bacteroidetes and Firmicutes. Actinobacteria decreased following RUTF treatment, but then increased again once the co-twins returned to their usual diet. When we compared KEGG KO profiles using PCoA, examination of microbiomes along PC1, which explains the most variation in the data, revealed that the trajectories of maturation of microbiomes in both co-twins paralleled one another with the malnourished twin falling slightly behind (see **Fig. 14d**). Comparing the representation of KEGG ECs before, during and after RUTF treatment, revealed an overrepresentation of ECs in the healthy child at the time his co-twin developed kwashiorkor that are involved in the car-

bohydrate metabolism (fructose/mannose, galactose metabolism), as well as glycerolipid metabolism. Among ECs overrepresented in the kwashiorkor co-twin's microbiome were those involved in the cysteine/methionine metabolism, selenocompound and glutathione metabolism, as well as degradation of host glycans (**Table 9**). Four of these ECs were also overrepresented in this co-twin 2 months prior to the development of the disease (alpha-mannosidase, beta-glucosidase, purine nucleosidase, and an EC with a transferase activity).

We used the results of the Spearman correlation analysis obtained from comparisons of healthy twins and their mothers (**Table 7**) to assess the degree of 'maturation' of healthy and kwashiorkor co-twin's microbiomes. A negative value for this coefficient implies that representation of an EC declines with increasing age and is usually present in infants. The average Spearman coefficient of ECs overrepresented in the kwashiorkor microbiome was -0.635, while average coefficient of ECs overrepresented in the healthy twin was 0.5, indicating that the kwashiorkor microbiome was underdeveloped relative to his healthy sibling (**Fig. 14e**). This trend reversed once both co-twins in this discordant pair were treated with RUTF for 2 weeks: the kwashiorkor co-twin's microbiome now looked more 'mature' compared to his healthy sibling, which correlates with our finding of more vigorous response to RUTF in kwashiorkor co-twins compared to their healthy siblings (**Fig. 12 c,d; Fig. 14e**). Among the 78 ECs whose proportional representation was higher during RUTF treatment in the child with kwashiorkor were those involved in the biosynthesis of vitamin B12, nitrogen metabolism, amino acid metabolism (cysteine/methionine, lysine, aspartate), as well as glucuronate interconversion (**Table 9**). Upon withdrawal of RUTF, the kwashiorkor co-twin's microbiome appeared less 'mature' relative to that of his healthy sibling, as judged by the Spearman coefficients of the significantly different ECs (**Fig. 14e**). The representation of EC2.7.1.69 (protein-*N*-phosphohistidine-sugar phosphotransferase), which is involved in the translocation of phosphorylated sugars into bacteria and forms part of the phosphotransferase system, was significantly lower in the kwashiorkor co-twin's fecal microbiome at all time points surveyed: i.e., before and at the

time of diagnosis, as well as during RUTF treatment.

In two other families (k46 and k268, **Table 8**), the microbiomes of kwashiorkor co-twins appeared to be less ‘mature’ compared to their healthy co-twins as judged by the values of Spearman coefficients of the ECs significantly underrepresented in the malnourished microbiomes. In both families, these ECs were involved in the carbohydrate metabolism (for example, pullulanase, alpha-amylase, beta-glucosidase). In addition, in these 2 families, the underrepresentation of genes encoding enzymes involved in carbohydrate metabolism was observed a month before they were diagnosed with kwashiorkor.

Taken together, these results suggest that RUTF induces a ‘temporal maturation’ of the microbiomes of children with kwashiorkor but not their healthy siblings. ‘Regression’ back to ‘immature’ microbiome appears when children go back on their regular diet, suggesting that their microbiomes (or the environment in the intestine) are not able to sustain the re-configuration induced by short-term nutritional therapy.

### **Temporal variation in fecal microbiomes of twin pairs discordant for marasmus**

Similar to healthy and twins discordant for kwashiorkor, familial similarity was greater than intra-individual similarity in the fecal microbiomes of twins discordant for marasmus (**Fig.15a**). Over time, the distances between twins increased, similar to twins who remained healthy, but the distances were even greater than those of healthy twins, and considerably higher than those with kwashiorkor whose discordance was lower than in pairs concordant for healthy status (**Fig.15b,c**). This finding was surprising given that these twin pairs on average were younger than twins discordant for kwashiorkor, and given that in healthy pairs the distance between microbiomes is lower at younger ages (**Fig. 8**).

As in the case with kwashiorkor twins, we focused on three fecal samples from each discordant twin pair: samples obtained at the presentation with marasmus, the first sample collected after 2 weeks of RUTF treatment, and first sample collected 1 month

after the cessation of RUTF to compare the distances between co-twins at each time point. Although microbiomes in each family (twin pair) changed differently, on average distances between co-twins at 1 month after cessation of RUTF were smaller than at the time of presentation with marasmus (**Fig. 15d,e**). We used PCoA to compare KEGG KO profiles of microbiomes. The results revealed that much like in the healthy twin pairs, age and family membership were significantly correlated with PC1 (33% variation explained). When positions along PC1 were compared using only 3 samples, we found that the fecal microbiomes of both healthy and marasmus co-twins progressed steadily towards a more functionally mature microbiome configuration, and in contrast to co-twins who had kwashiorkor, the fecal microbiomes of children with marasmus did not respond to RUTF with significantly greater functional changes than their healthy siblings (**Fig. 16**). Similar to co-twins who remained healthy, Bifidobacteria dominated the fecal microbiomes of marasmus co-twins during the first 10 months, and were gradually ‘replaced’ by Bacteroidetes and Firmicutes (**Fig. 17**). However, a greater representation of Proteobacteria was noted in the microbiomes of these families compared to twins who remained healthy or those discordant for kwashiorkor (**Fig. 10 and 13**), although the difference was not significant across all subjects.

### **Changes in KEGG ECs involved in various metabolic functions in the fecal microbiomes of co-twins discordant for marasmus**

Due to the strong effects of age and family membership, we used Fisher’s exact test to compare the representation of KEGG ECs in the fecal microbiomes of each of the discordant twin pairs before, during and after RUTF treatment (**Table 10**). As in the case of kwashiorkor, each family had a unique response to the disease and RUTF. We used Spearman correlation analysis to assess the degree of ‘maturation’ of the microbiomes before, during and after RUTF. Prior to RUTF treatment, in 4 of 10 discordant pairs, the microbiomes of the co-twins with marasmus appeared less mature. With RUTF treatment, the fecal



microbiome of the marasmus co-twin appeared less mature in 6 of 10 pairs (**Fig. 16**). In this comparison, 5 twins with marasmus had an increased representation of urease, as well as several ECs involved in host glycan degradation (e.g., alpha-mannosidase and exo-alpha-sialidase). As noted in Chapter 3, urease releases ammonia that can be used for microbial biosynthesis of essential and nonessential amino acids. In addition, urease plays a major role in nitrogen recycling. Under conditions where dietary nitrogen is limiting, the ability of the microbiome to utilize urea should be advantageous to both the microbial community and host. Increased representation of genes involved in foraging of host glycans would be an adaptive response if glycans in the diet were limiting or if a microbiome lacked the glycoside hydrolases needed to degrade classes of polysaccharides that were in their diets.

We provide example of family m229 where changes in the representation of ECs were most pronounced. One of the co-twin's in this male dizygotic pair remained healthy throughout the study period. The other co-twin presented with marasmus at 11 weeks of age (WHZ -3.3, versus -1.7 for his co-twin), and was treated with RUTF for 8 weeks. He then developed kwashiorkor twice later in the study; the first time 4 weeks after cessation of RUTF; 14 months after a second 4 week-long round of RUTF treatment, he was again diagnosed with kwashiorkor and again treated with RUTF (see **Fig. 18a** for anthropometric measurements; note that throughout the study anthropometric measurements for the undernourished twin remained lower than of his healthy sibling). Moreover, the number of days with fever, diarrhea and vomiting was significantly higher for the sick co-twin ( $p < 0.05$ , Poisson test, **Fig. 18b**).

When we compared the representation of major bacterial phyla across microbiomes, the co-twin with marasmus had lower representation of Actinobacteria ( $p < 0.05$ ) compared to his healthy sibling (**Fig. 18c**). This phylum increased in relative abundance following RUTF treatment and continued to dominate, especially after the withdrawal of RUTF when that child developed kwashiorkor. Remarkably, the response to RUTF after the appearance of kwashiorkor was similar to the response observed in twins discordant for this disease

that we described earlier (**Fig. 13b**): i.e., the representation of Actinobacteria decreased with RUTF in the severely undernourished child, but not his healthy sibling, along with the significant increase of Bacteroides (mainly Prevotella, **Fig. 18c**). When we examined the functional configuration (KEGG KO profile) using PCoA analysis (**Fig. 18d**), we observed that following presentation with marasmus, the first round of RUTF treatment and subsequent development of kwashiorkor, the microbiome of the malnourished co-twin remained in “younger” coordinate space compared to his healthy sibling. However, following the second RUTF intervention, his microbiome ‘matured’.

To further characterize functional differences between the microbiomes of this twin pair, we used Fisher’s exact test on the relative abundances of KEGG ECs. At the time of presentation with marasmus, 50 ECs had a significantly higher proportional representation in the severely undernourished co-twin compared to his healthy sibling (**Table 11**). They included ECs involved in butanoate, glyoxylate and sulfur metabolism as well as cephalosporin-C deacetylase, which degrades beta-lactam group antibiotics. Following 2 weeks of RUTF treatment, ECs that were significantly overrepresented in the malnourished microbiome included those involved in host glycan and urea degradation (alpha-mannosidase, exo-alpha-sialidase, alpha-l-fucosidase, urease) (note that these ECs were also overrepresented in the microbiomes of co-twins with marasmus in five other twin pairs at the time of RUTF treatment; **Table 11**). Importantly, all ECs overrepresented in the microbiome of the co-twin had lower Spearman correlation coefficients (**Fig. 18e**), suggesting that microbiome was ‘underdeveloped’ relative to the microbiome of his healthy twin. This trend became more obvious upon the withdrawal of RUTF at which point the malnourished twin developed kwashiorkor: i.e., ECs significantly overrepresented in his microbiome had lower Spearman coefficients compared to the healthy twin; they included alpha-mannosidase, ECs involved in glutathione metabolism, as well as myo-inositol degradation (**Table 11**). Remarkably, with the second RUTF administration for treatment of kwashiorkor, ECs that had a significantly higher representation in the fecal microbiome of the undernourished

co-twin appeared to be more ‘mature’ compared to his healthy twin, as judged by their Spearman coefficients (**Fig. 18e**). This result is similar to what we described in the previous section about the responses of co-twins discordant for kwashiorkor.

These results suggest that the fecal microbiomes of twins discordant for marasmus are less similar to one another compared to twins who remained healthy throughout the study or who were discordant for kwashiorkor. Unlike the more pronounced response to RUTF in children with kwashiorkor, no significant difference in the degree of the response was found across all twin pairs discordant for marasmus. Nonetheless, when microbiomes were compared in each twin pair at the time of first treatment with RUTF, a greater degree of functional ‘maturation’ of the microbiome was noted among healthy co-twins compared to their siblings with marasmus. The prevalence of ECs involved in the degradation of host glycans and urea may indicate a paucity of bioavailable nutrients in the distal guts of children with marasmus despite RUTF treatment – either because most nutrients in RUTF are absorbed in their small intestines or their microbiomes may not already have, fail to adopt configurations that have a level functional maturity or capacity to utilize key nutrient represented in this therapeutic diet.

Taken together, our results show the value of twin studies in deciphering differences in microbiome configurations in healthy versus malnourished infants and children in the face of the significant intra- and interpersonal, and familial variation, in microbial community structure that is normally manifest during this stage of development. Our data indicate that the microbiome is not only a biomarker but a potential causal factor in the development of severe undernutrition. It is also reporter of the response and efficacy of a given therapeutic food intervention. Our results underscore that there may be many specific routes by which the microbiome contributes to malnutrition; a general trend, more obvious in the case of kwashiorkor, is functional immaturity of this microbial metabolic organ as it relates to nutrient metabolism/biosynthesis. The fact that in the subset of kwashiorkor microbiome characterized, we see a robust response to RUTF followed by a regression in

state functional maturation, raises the possibility that despite current clinical parameters indicating successful treatment, longer term nutritional support may be required to ameliorate persistent and irreparable metabolic dysfunction, or to permit time for functional repair/regeneration. Using the microbiome to judge these parameters should be very useful in selecting the type and duration of nutritional support, especially in the case of disorders such as marasmus. Finally, it will be very important to take these observations made in twins and perform direct functional assays of the microbiome. One approach, described in a separate study from our lab, involves transplantation of previously frozen fecal microbial communities, obtained from the untreated discordant twin pairs characterized above into germ-free mice. This allows a fecal community from a single donor to be replicated with a high degree of accuracy in multiple recipient mice who are given the same diets as those consumed by the donor: i.e., a macro- and micronutrient deficient Malawi diet, followed by RUTF, followed by a Malawi diet. This approach allows assessment of the degree to which donor phenotypes can be transmitted via their microbiomes, a determination of the contributions of diet and microbiome to these phenotypes, a detailed assessment of community structure along the length of the gut, characterization of myriad features microbial community metabolism and host-microbial co-metabolism as well as immune function under highly controlled conditions, direct assessment of the contributions of RUTF components and/or duration of administration, definition of multiple other features of physiology and pathology inside and outside of the GI tracts of these humanized gnotobiotic animals, and use of the resulting datasets to further characterize and understand the human donors that made these models possible.

## **Methods**

**Subjects recruitment and sample collection** – The current study was approved by Human Studies Committees from Washington University and University of Malawi. Twin pairs were recruited through the health centers located in the five sites surveyed. A team of one USA pediatrician and a minimum of two trained local personnel visited each site every week. During each visit, the weight and height of each infant or child was measured in 3 replicates; during each visit mothers of twins were interviewed about whether her children had symptoms of cough, fever, diarrhea and vomiting in the preceding days. During each visit for a scheduled fecal sample collection, each child wore a commercial diaper until a sample was deposited. An aliquot of sample was taken with a clean spatula and collected into sterile plastic 2 ml tube, which was immediately deposited into an aluminum tank filled with liquid nitrogen. Upon arrival to the laboratory, tubes with samples were transferred and stored at – 80°C until further processing. Buccal smears were collected once for zygosity test using Oragene kits. RUTF and soy-based diet were produced locally.

**Isolation of fecal DNA and multiplex pyrosequencing** – Fecal samples were pulverized with a mortar and pestle at -80°C. Genomic DNA was extracted from 200 mg of frozen sample as described in (Muegge et al. 2011). For multiplex shotgun 454 Titanium FLX pyrosequencing, each fecal community DNA sample was processed according to the manufacturer's protocol (Rapid Library preparation for FLX Titanium, Roche). Equivalent amounts of 12 samples each labeled with a unique barcode sequence were pooled prior to each pyrosequencer run with FLX Titanium chemistry.

## **Data analysis**

**Zygosity tests** – We selected 48 autosomal SNPs with high heterozygosity in the Yoruban HapMap sample (N=90) from the Illumina DNA Test Panel of 360 SNP loci that have been optimized for the BeadExpress (Catalog # GT-17-221). The SNP minor allele frequencies in Yorubans ranged from 0.28-0.43, and each autosome was sampled by at least one

marker with the exception of 7,19, and 21, with minimum marker separation of 3Mb. We ran PREST (Pedigree Relationship Statistical Test, McPeck & Sun 2000) to estimate the identity-by-descent (IBD) sharing and kinship between the relative pairs, and compared reported relationships versus the inferred relationships from the genetic data.

**16S rRNA data processing and analysis** was carried out using QIIME as previously described (Muegge et al, 2011)

### **Shotgun sequences from fecal microbiomes**

Raw sequences were processed as described previously in Muegge et al (2011). Functional annotation against version 58 of the KEGG database, where sequences without an assigned KO number were removed, was carried out using BLASTX with the following parameters: e-value cut-off less than  $1e^{-5}$ , bitscore  $> 50$ , and percent identity  $> 50\%$ , and option  $-z$  2214788408. In the cases where a shotgun sequence had a significantly equal match to more than one (n) KOs, all KOs were accounted for and each assigned  $1/n$  counts.

### **Taxonomic composition of the shotgun sequence data**

Shotgun sequences were mapped to the database of 127 sequenced bacterial and archaeal genomes, listed in the **Table 6**. BLASTN was used with the following cut-offs: e-value  $< 1e^{-20}$ , bitscore  $> 50$ , percent identity  $> 50$ , percent alignment  $> 80\%$  (Arumugam et al., 2011). Relative abundance of each genome was adjusted to a genome length.

### **Statistical analyses**

Hellinger distances and principal coordinates analyses were carried out using QIIME software. Spearman correlation and Fisher's exact tests were conducted using R statistical software. Linear mixed model regression was conducted using R package NLME.

## **References**

- Ahmed T., S. Rahman , A. Cravioto Oedematous malnutrition. *Indian J. Med. Res.* **130** :651–4 (2009).
- Arumugam M., J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D.R. Mende, et al. Enterotypes of the human gut microbiome. *Nature* **473**, 174–80 (2011).
- Blackburn, G.L. Pasteur’s Quadrant and malnutrition. *Nature* **409**, 397–401 (2001).
- Bryce J., Boschi-Pinto C., Shibuya K., Black R.E. WHO estimates of the causes of death in children. *Lancet* **365**, 1147–1152 (2005).
- CIA. CIA - The World Factbook 2011; Available from: <https://www.cia.gov/library/publications/the-world-factbook/geos/mi.html> (2011)
- Ciliberto M.A., H. Sandige H., M.J. Ndekha M.J., P. Ashorn, A. Briend A., H.M. Ciliberto, et al. Comparison of home-based therapy with ready-to-use therapeutic food with standard therapy in the treatment of malnourished Malawian children: a controlled, clinical effectiveness trial. *The American Journal of Clinical Nutrition* **81**, 864–70 (2005).
- Courtright P., J. Canner The distribution of kwashiorkor in the southern region of Malawi. *Ann Trop Paediatr.* **15**, 221–6 (1995).
- Golden M.H.N. The development of concepts of malnutrition. *J. Nutr.* **132**, 2117S–2122S (2002).
- Lin C.A., S. Boslaugh, H.M. Ciliberto, K. Maleta, P. Ashorn, A. Briend, et al. A prospective assessment of food and nutrient intake in a population of Malawian children at risk for kwashiorkor. *J. Pediatr. Gastroenterol. Nutr.* **44**, 487–93 (2007).
- Matilsky D.K., K. Maleta, T. Castleman, M.J. Manary Supplementary feeding with fortified spreads results in higher recovery rates than with a corn/soy blend in moderately wasted children. *J. Nutr* **139**, 773–8 (2009).

- McNulty N. P., T. Yatsunenکو, A. Hsiao, J.J. Faith, B. D. Muegge, A. L. Goodman, B. Henrissat, R. Oozeer, S. Cools-Portier, G. Gobert, C. Chervaux, D. Knights, C.A. Lozupone, R. Knight, A.E. Duncan, J.R. Bain, Muehlbauer M. J., Newgard C., Heath A. C., Gordon J. I., The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. *Sci. Transl. Med.* **3**, 106ra106 (2011).
- McPeck M.S., L. Sun. Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet.* **66**,1076–94 (2000).
- Muegge B.D., J. Kuczynski, D. Knights, J.C. Clemente, A. González, L. Fontana, et al. Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans. *Science* **332**, 970–4 (2011).
- Palmer C., E.M. Bik, D.B. DiGiulio, D.A. Relman, P.O. Brown. Development of the Human Infant Intestinal Microbiota. *PLoS Biol.* **5**, :e177 (2007).
- Pelletier D.L., E.A. Frongillo EA Jr, D.G. Schroeder, J.P. Habicht. A methodology for estimating the contribution of malnutrition to child mortality in developing countries. *J. Nutr.* **124**, 2106S-2122S (1994).
- Prentice A.M., M.E. Gershwin, U.E. Schaible, G.T. Keusch, C.G. Victora, J.I. Gordon. New challenges in studying nutrition-disease interactions in the developing world. *J Clin Invest* **118**, 1322–9 (2008).
- Scrimshaw N.S., F.E. Viteri. INCAP studies of kwashiorkor and marasmus. *Food Nutr Bull.* **31**, 34–41 (2010).
- WHO. WHO | Infant and young child feeding data by country, Available from: <http://www.who.int/nutrition/databases/infantfeeding/countries/en/index.html#M> (2006)
- Williams C.D. Deficiency diseases in infants, a report by Miss C. D. Williams from Gold Coast Colony Annual Medical Report, 1931-1932, p. 93. *Nutr. Rev.* **31**, (1973).



Wu G.D., J. Chen, C. Hoffmann, K. Bittinger, Y-Y. Chen, S.A. Keilbaugh, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105–8 (2011).

## **Figure Legends**

**Figure 1. Geographic location of the villages where the study was conducted.** A health center at each of the five highlighted villages served as a meeting place for the twin visits.

**Figure 2. Study design.** Twin pairs who were less than three-years-of-age were enrolled for the study. Every month, anthropometric data was collected from each twin pair and their nutritional status monitored. Fecal samples were collected every two months from twins who were less than 1 year old, and thereafter every three months from twins between one and three-years-of-age. If one or both children in a twin pair developed kwashiorkor or marasmus, both were given RUTF and fecal samples were collected every two weeks until nutritional status improved as judged by anthropometric measurements.

**Figure 3. Anthropometric data for twin pairs whose gut microbiomes were sequenced.** Each column shows data for each twin pair: weight-for-height Z-scores, (WHZ), height-for-age Z-scores (HAZ), and weight-for-age Z scores (WAZ) are plotted against their age over the course of the study. For twins discordant for undernutrition, the time of presentation with kwashiorkor or marasmus is indicated with a yellow or cyan dot, respectively. (a) twins concordant for healthy status; (b) twin pairs discordant for kwashiorkor; (c) twin pairs discordant for marasmus.

**Figure 4. The fraction of shotgun pyrosequencer reads that had significant annotation in the KEGG database decreases with increasing age.** Percent of reads with significant BLAST hits to genes in the KEGG database is plotted against age for each category of twin pairs: both healthy; pairs discordant for kwashiorkor; pairs discordant for marasmus.

**Figure 5. Large interpersonal variations are observed in the functional configurations of fecal microbial communities at early ages.** Each point represents an average Hellinger distance between unrelated children at each age range identified on the Y-axis.

**Figure 6. Principal coordinate analysis (PCoA) of Hellinger distances generated from KEGG KO profiles.** (a) PC1 coordinates of all 453 sequenced twin fecal microbiomes are plotted against age. Each sphere represents a microbiome, colored by age; (b) same as (a) colored by a health status of a twin pair; (c) PC1 coordinate of twin pairs discordant for kwashiorkor (207 microbiomes) is plotted against age, colored by the health status of each twin (those who developed kwashiorkor over the course of the study are colored in red before and after onset of the disease); (d) same as (c) for twin pairs discordant for marasmus (169 microbiomes); (e) PC2 coordinates of 453 sequenced microbiomes are plotted against age and colored by the health status of a twin pair. (f) Spearman correlation between position of microbiomes along PC2 and relative abundance of sequences with representation in 127 reference human gut microbial genomes. Microbiomes spread along PC2 have increased representation of *Enterobacteriaceae* and decreased representation of Bifidobacteria.

**Figure 7. Analysis of Hellinger distances between and within healthy infant microbiomes.** Distances between all microbiomes that originate from a child, and distances between microbiomes sampled from a twin pair, are plotted next to an average distance between microbiomes of unrelated children: (a) Unweighted Unifrac measurements of phylogenetic distances between microbial communities; (b) Weighted UniFrac measurement; (c) Hellinger distance derived from 16S rRNA OTUs; (d) Hellinger distance derived from KEGG KO profiles.

**Figure 8. Microbiomes of children within a twin pair become dissimilar with age.** Distances between twins are calculated at each age using (a) unweighted UniFrac distances generated from 16S rRNA datasets, and (b) Hellinger distances generated from KEGG KO profiles. Each line represents a family.

**Figure 9. Age and family membership explain the largest variation in the healthy microbiomes.** PC1 coordinate is plotted against age for 97 fecal microbiomes from twin

pairs who remained concordant for healthy status. Microbiomes are colored by family membership.

**Figure 10. Taxonomic changes with age in healthy twin pairs.** (a) Relative abundance of bacterial phyla and Archaea are plotted against age for all sampled microbiomes; (b) The representation of members of Bifidobacteria (phylum Actinobacteria) are plotted against age for all sampled microbiomes.

**Figure 11. Distances between microbiomes of twins discordant for kwashiorkor.** (a) Distances between all microbiomes originated from a child, and distances between microbiomes sampled from a twin pair are plotted next to an average distance between microbiomes of unrelated children for Unweighted UniFrac and Hellinger distances derived from KEGG KO profiles; Unweighted UniFrac distances (b) and Hellinger distances generated from KEGG KO profiles (c) between twins at each age sampled; (d) same as (c) for only 3 time points – before, during and after RUTF, average  $\pm$  SEM; (e) same as (d) but individual family is shown.

**Figure 12. Principal Coordinate Analysis (PCoA) of Hellinger distances generated from KEGG KO profiles.** (a) PC1 coordinates of 207 microbiomes are plotted against age; (b) same as (a) colored by a family membership; (c) Average  $\pm$  SEM PC1 coordinate before, during and after RUTF for co-twins with kwashiorkor and their healthy siblings; (d) same as (c) but with each family shown.

**Figure 13. Taxonomic changes with age in twins discordant for kwashiorkor.** (a) Relative abundance of bacterial phyla and Archaea are plotted against age for all sampled microbiomes; (b) The representation of Actinobacteria decreases with RUTF in the microbiomes of kwashiorkor co-twins, but not in the microbiomes of their healthy siblings, Friedman test  $p=0.0054$  with Dunn's multiple comparison correction (\*  $p<0.05$ ); average values  $\pm$  SEM are plotted.

**Figure 14. Example of changes in taxonomic and functional composition in a twin pair discordant for kwashiorkor.** (a) Weight-for-height Z scores (WHZ), Height-for-age Z-scores (HAZ), and Weight-for-age Z scores (WAZ) collected over the course of the study. (b) Number of days with fever, cough, diarrhea and vomiting preceding each visit to a health center. (c) Relative abundance of bacterial phyla and archaeal sequences in the microbiomes of the co-twins in this twin pair. (d) PC1 derived from Hellinger distances obtained from KEGG KO counts plotted against age. (e) ECs shown as segments, identified by Fisher's exact test to be significantly different between the fecal microbiomes of the healthy versus kwashiorkor co-twins before (upper panel), during (middle) and after (lower panel) RUTF intervention. Corresponding Spearman correlation values, obtained from healthy twin pairs and their mothers is shown next to each EC. For descriptions of ECs shown see **Table 9**.

**Figure 15. Distances between microbiomes of twins discordant for marasmus.** (a) Distances between all microbiomes originated from a child, and distances between microbiomes sampled from a twin pair are plotted next to an average distance between microbiomes of unrelated children for Unweighted UniFrac and Hellinger distance derived from KEGG KO profiles; Unweighted UniFrac distances (b) and Hellinger distances generated from KEGG KO profiles (c) between twins at each age; (d) same as (c) for only 3 time points – before, during and after RUTF, average  $\pm$  SEM; (e) same as (d) but individual family is shown.

**Figure 16. Principal Coordinate Analysis (PCoA) of Hellinger distances generated from KEGG KO profiles of twins discordant for marasmus.** (a) Average $\pm$ SEM PC1 coordinate before, during and after RUTF treatment for co-twins with marasmus and their healthy siblings; (b) same as in (a) but with each family shown.

**Figure 17. Taxonomic changes with age in twins discordant for marasmus.** Relative abundance of bacterial phyla and Archaea are plotted against age for all sampled fecal microbiomes.

**Figure 18. Example of changes in taxonomic and functional composition in a twin pair discordant for marasmus.** (a) Weight-for-height Z scores (WHZ), Height-for-age Z scores (HAZ), and Weight-for-age Z scores (WAZ) collected over the course of the study. (b) Number of days with fever, cough, diarrhea and vomiting preceding each visit to a health center. (c) Relative abundance of bacterial phyla and archaeal sequences in the microbiomes of two twins. (d) PC1 derived from Hellinger distances obtained from KEGG KO counts plotted against age. (e) ECs shown as segments, identified by Fisher's exact test to have significantly different representation in the fecal microbiomes of the healthy versus marasmus co-twin before (upper panel), during and after (middle panels) RUTF intervention, as well as two weeks after second RUTF treatment (lower panel). The Spearman correlation value calculated from healthy twin pairs and their mothers is shown next to each EC. For descriptions of ECs shown see **Table 11**.

**Figures**

**Figure 1.**



Figure 2.

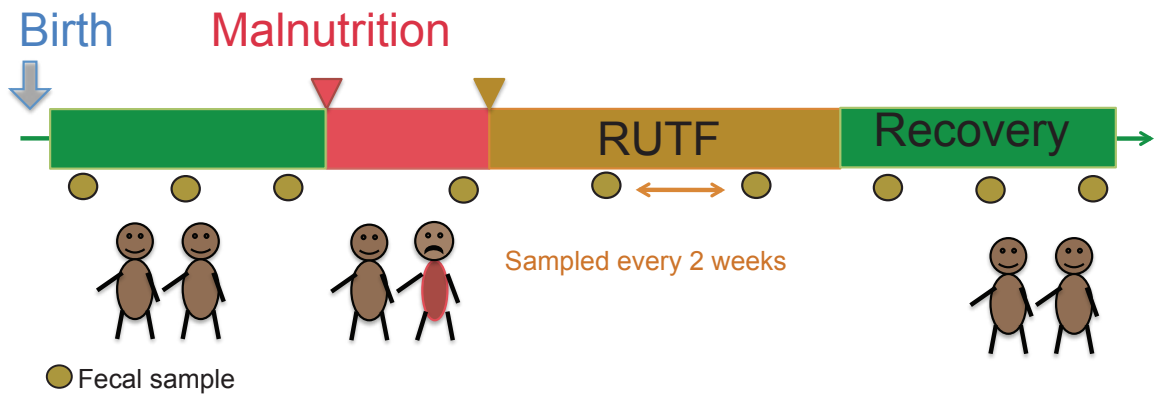




Figure 3a.

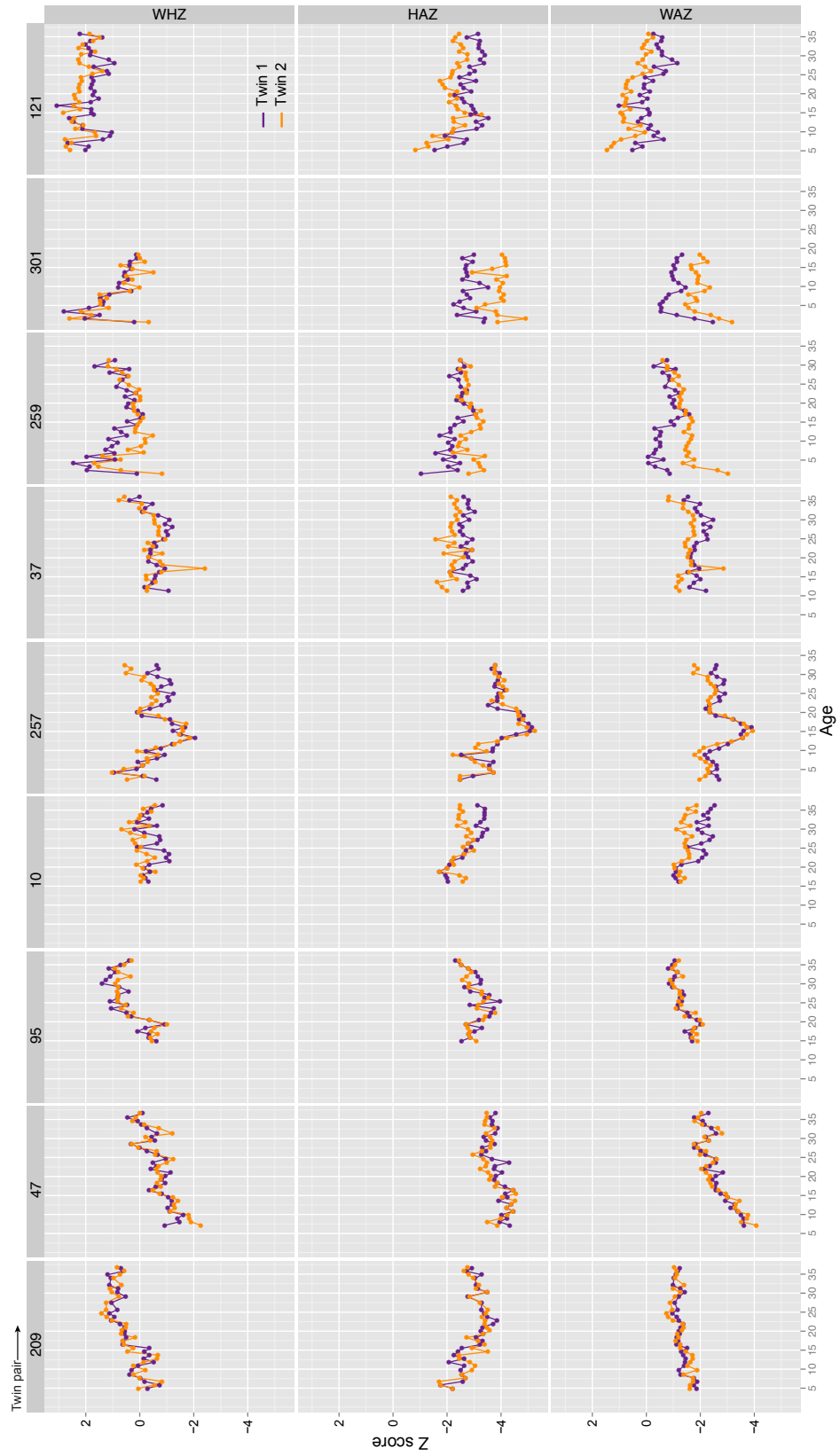


Figure 3b.

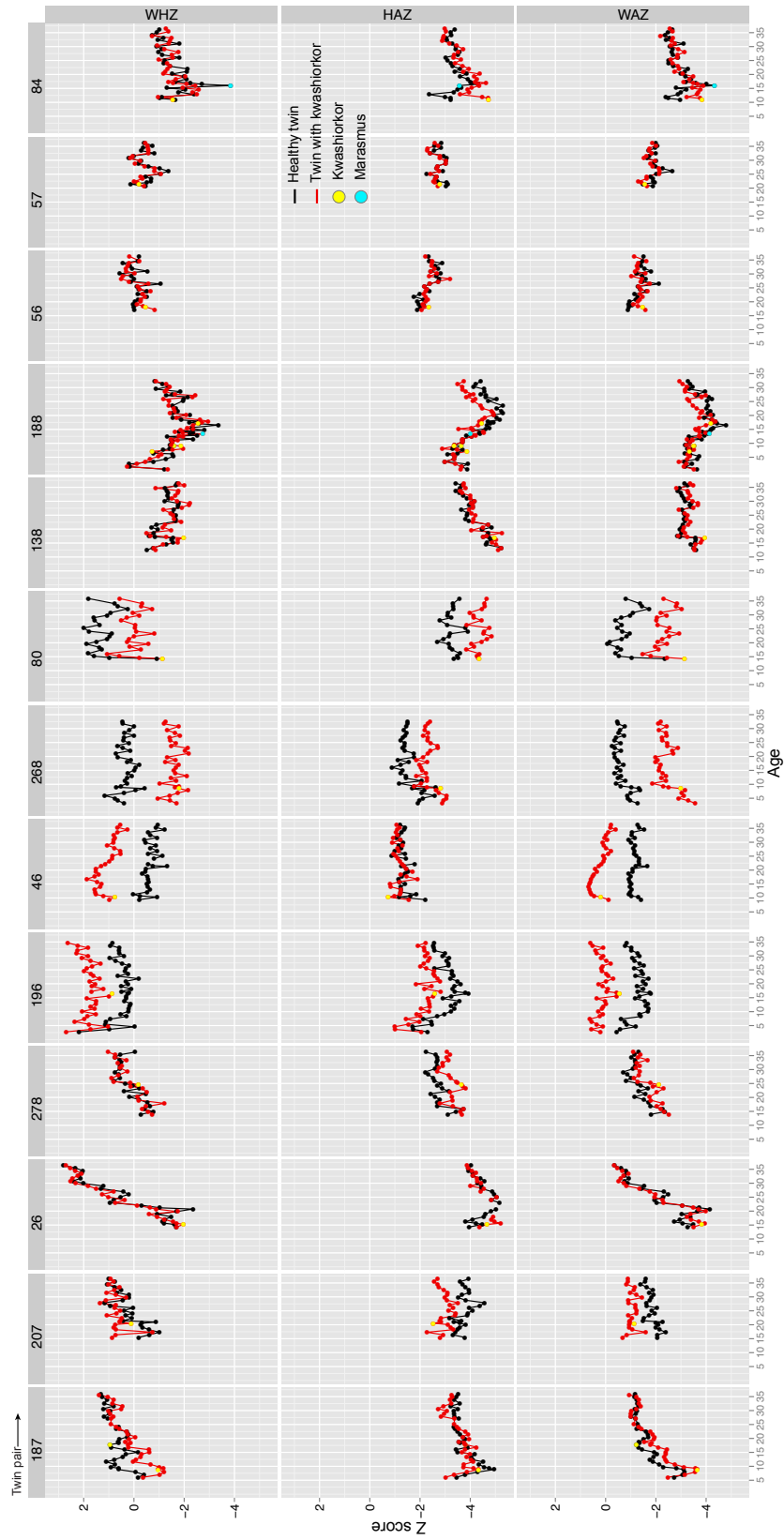
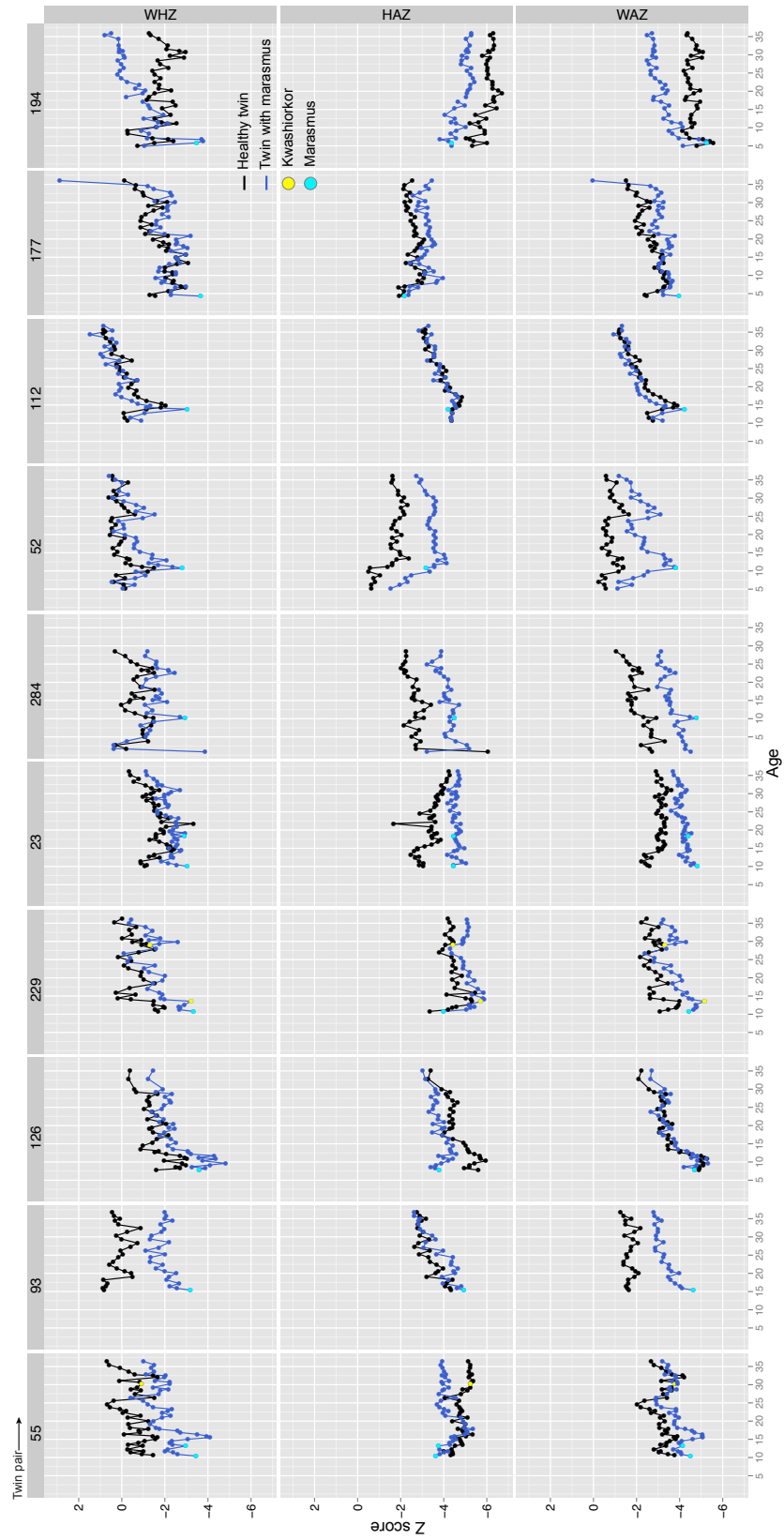


Figure 3c.



**Figure 4.**

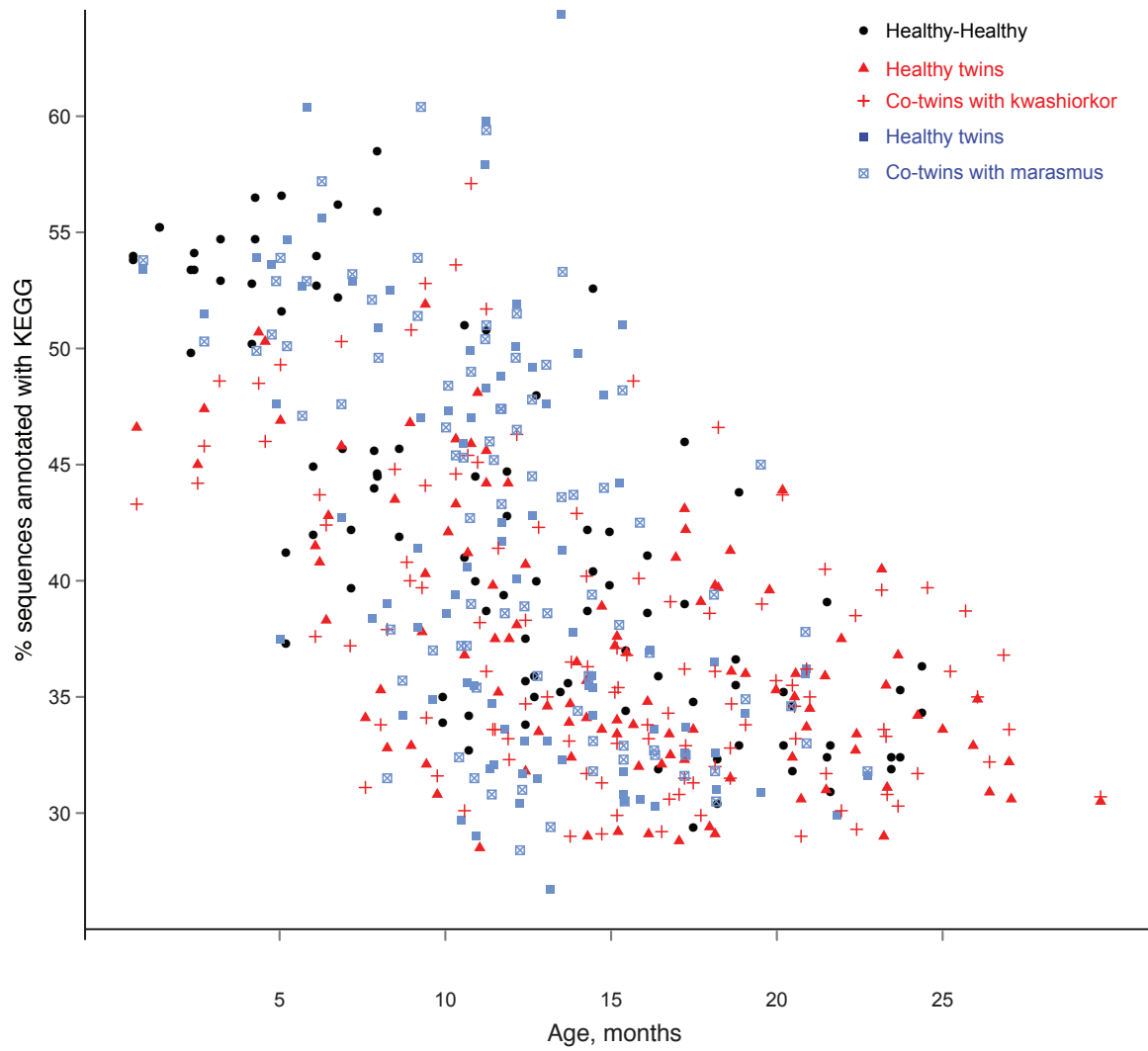
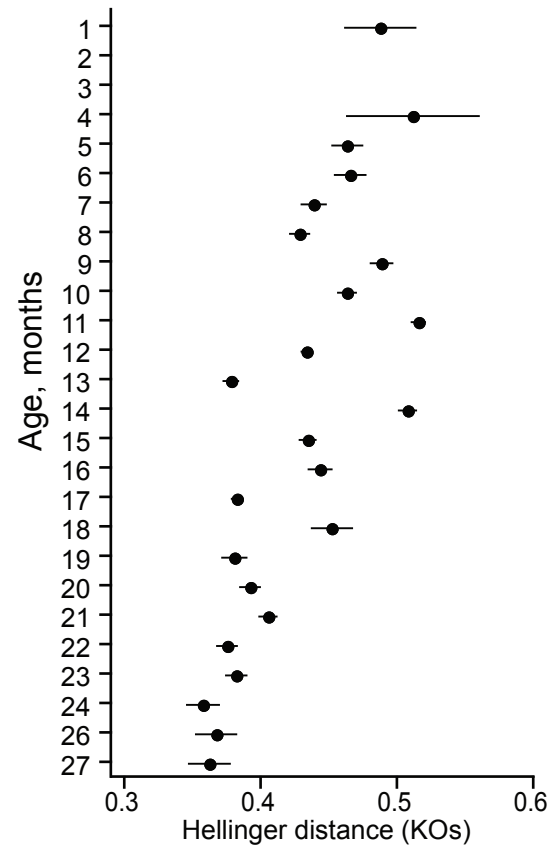
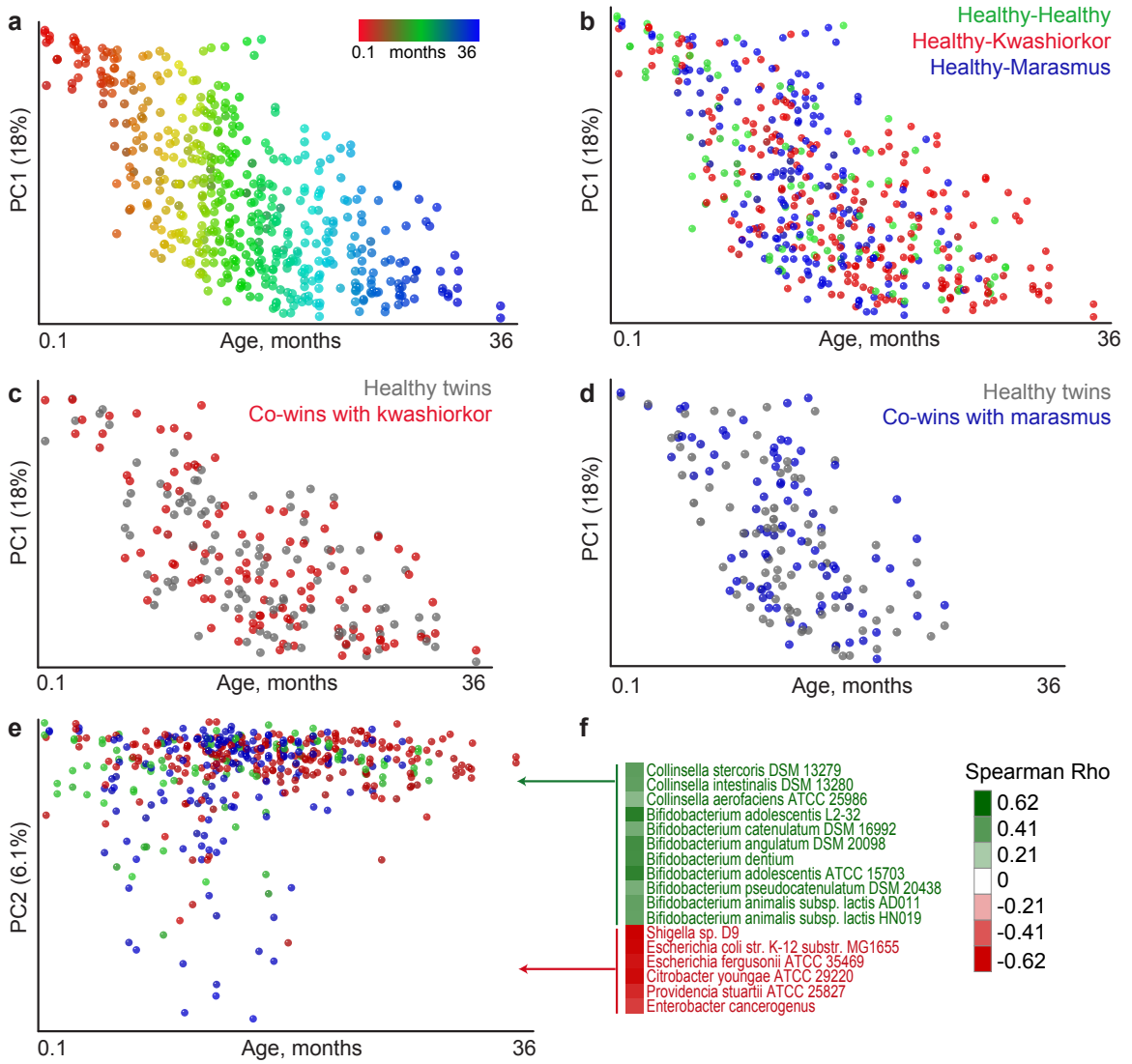


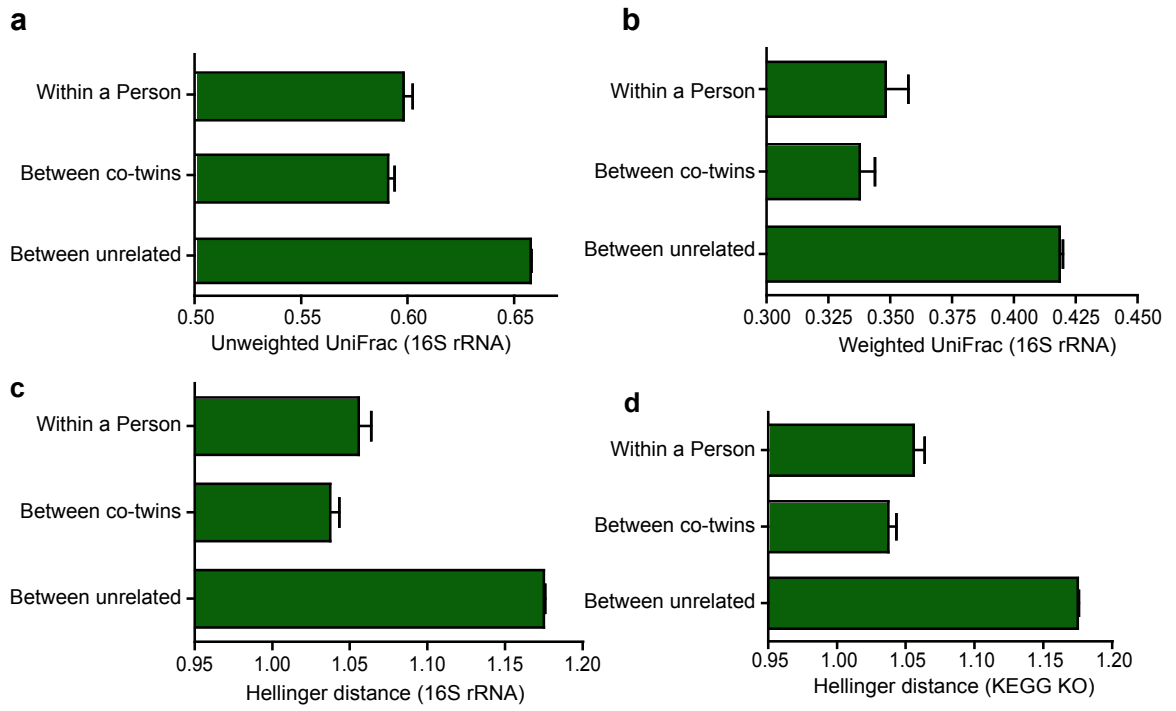
Figure 5.



**Figure 6.**



**Figure 7.**



**Figure 8.**

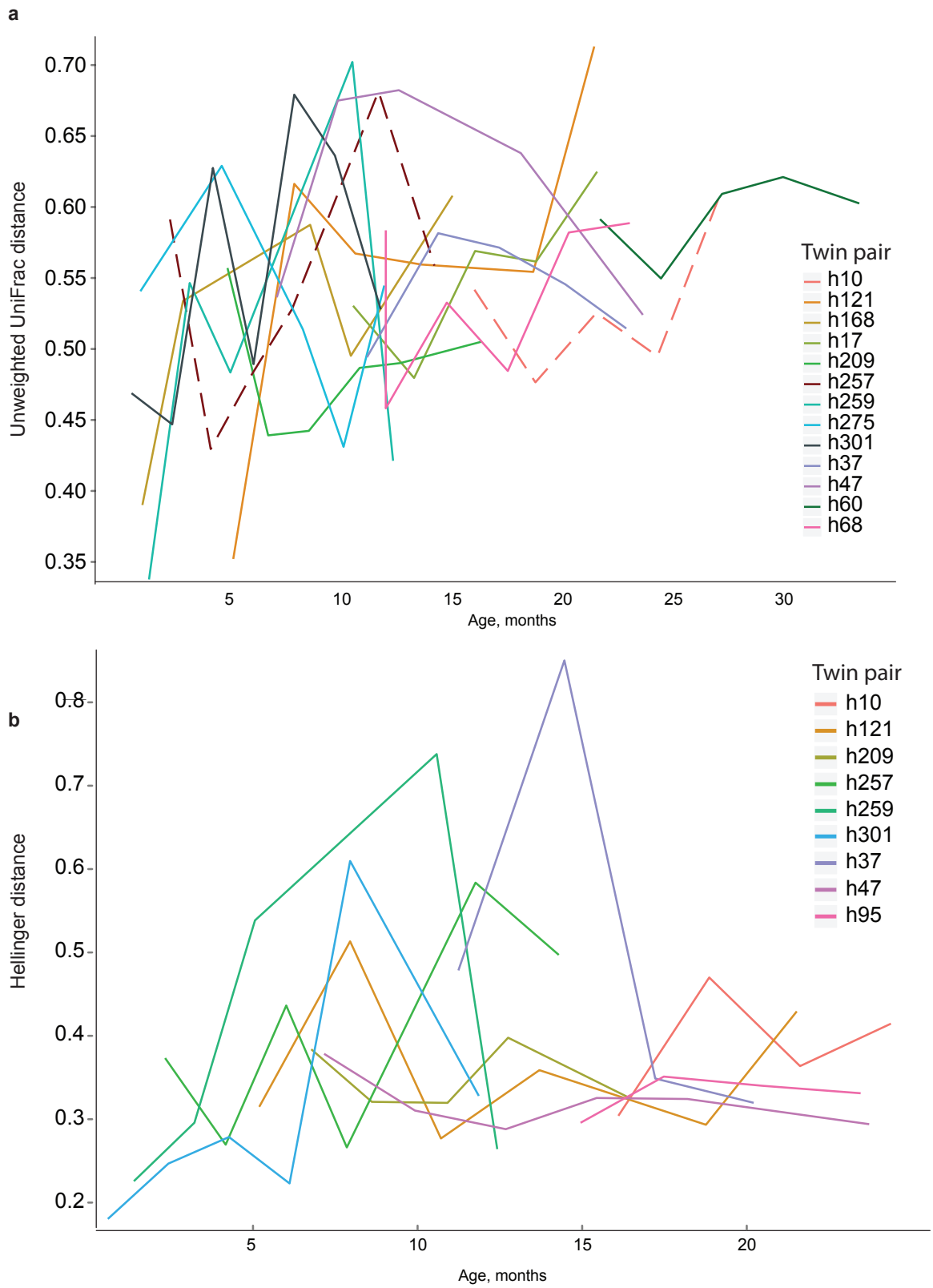
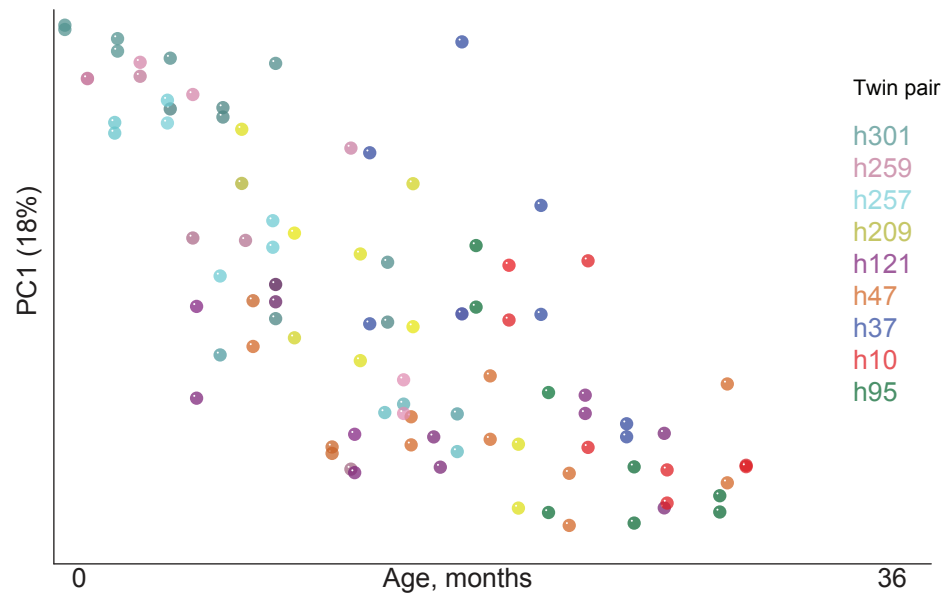
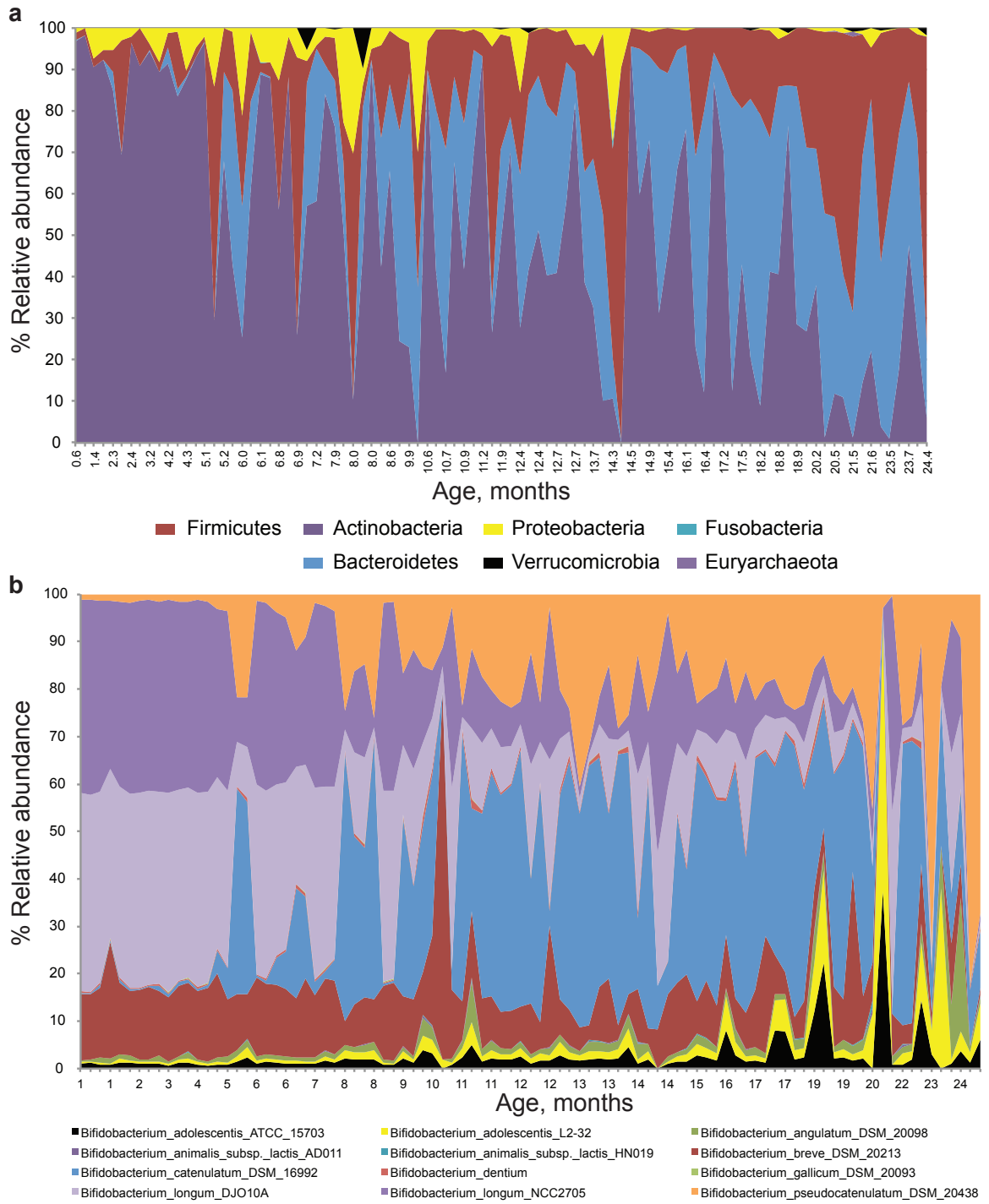




Figure 9.



**Figure 10.**



**Figure 11.**

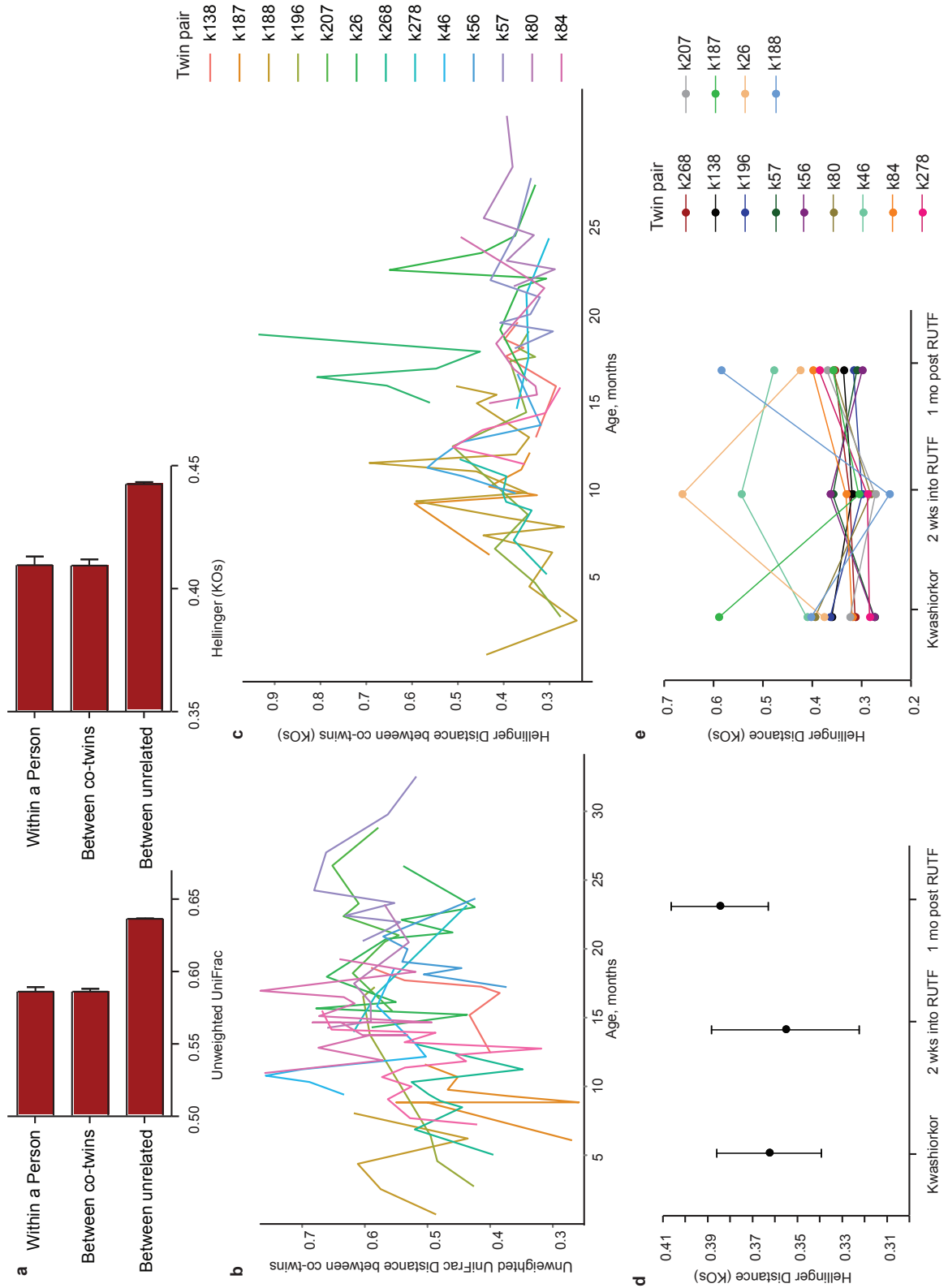
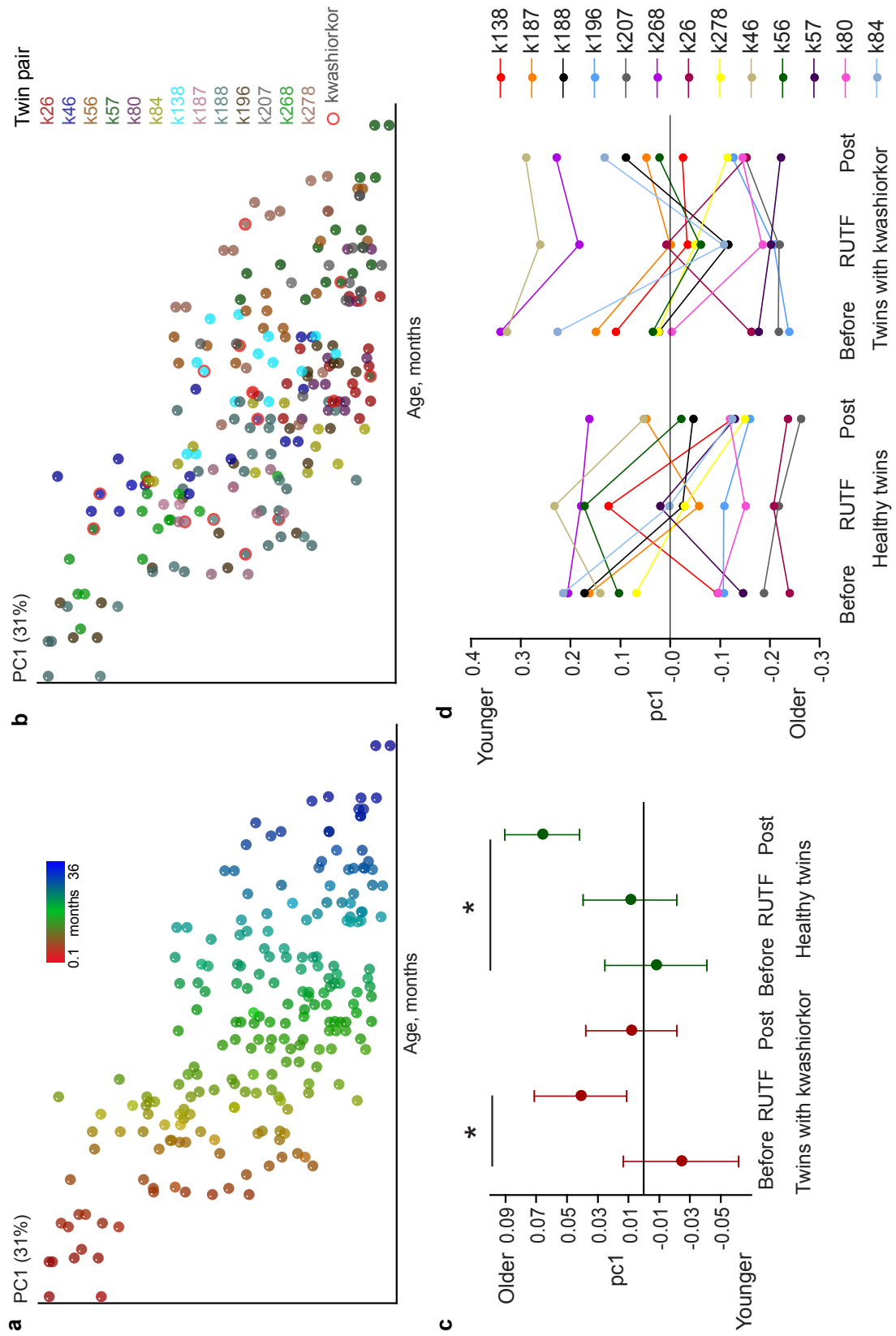


Figure 12.



**Figure 13.**

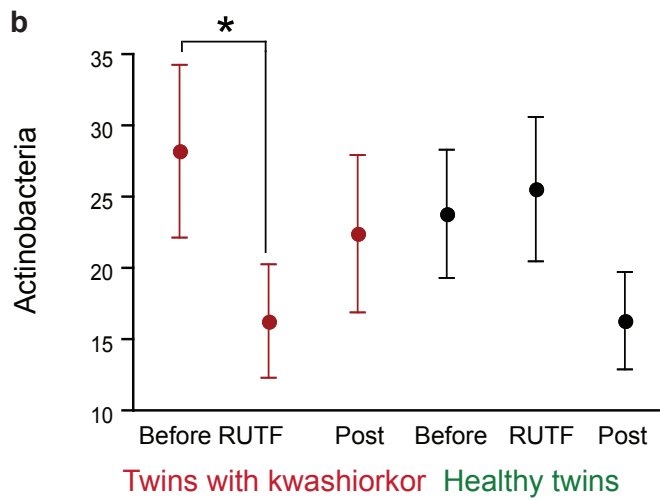
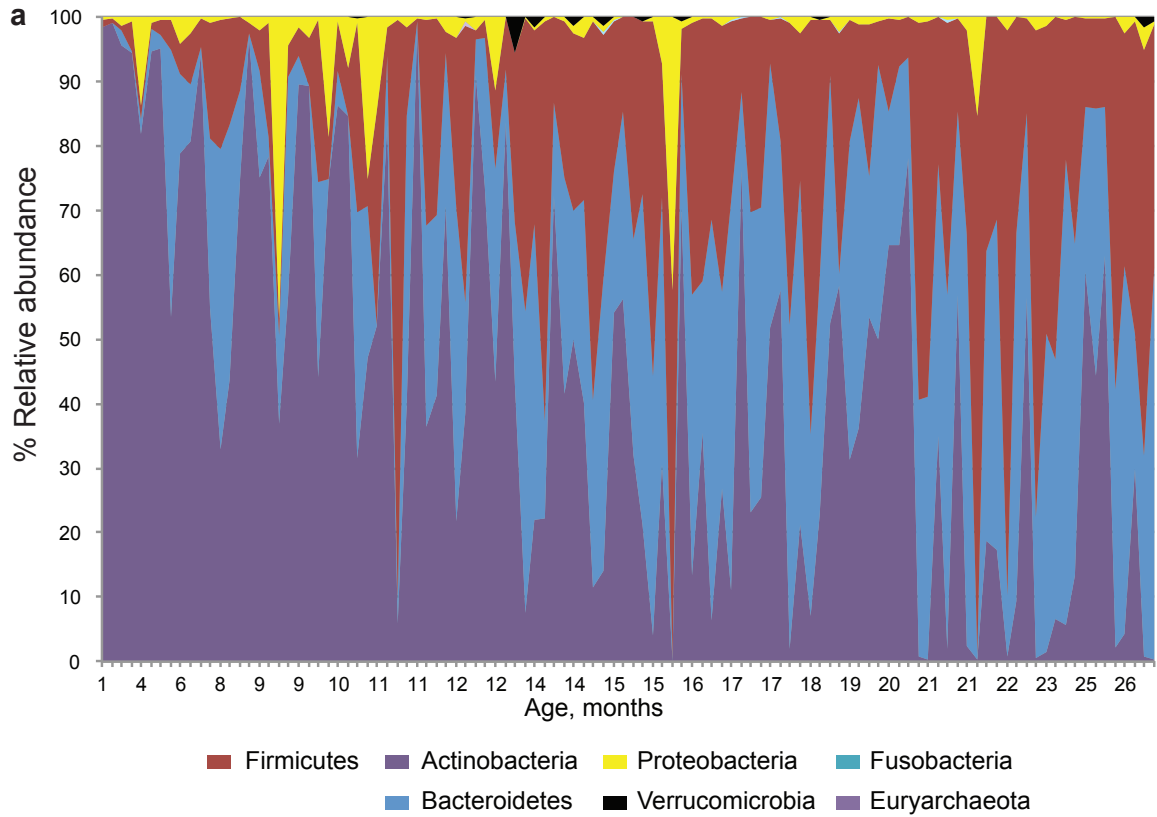
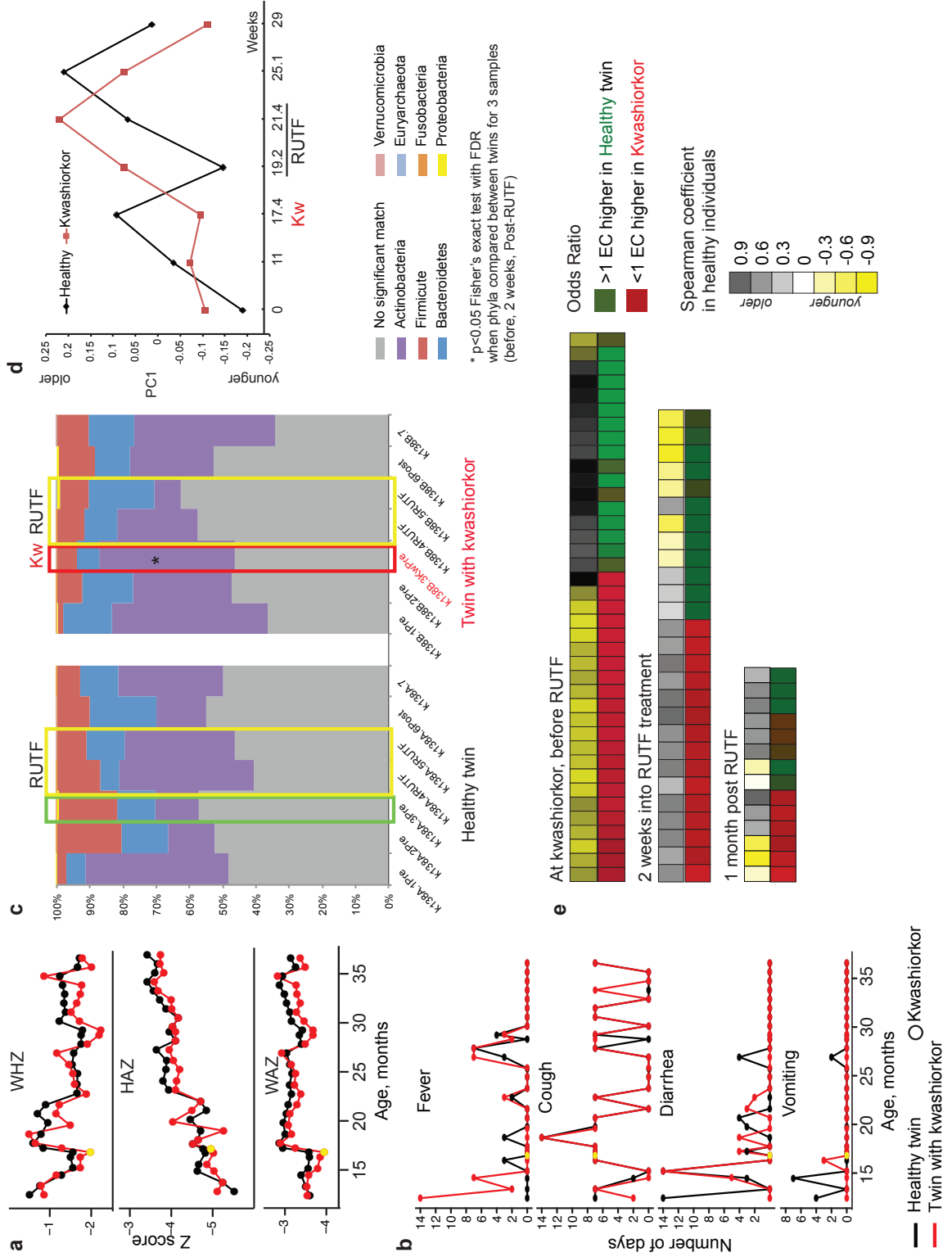
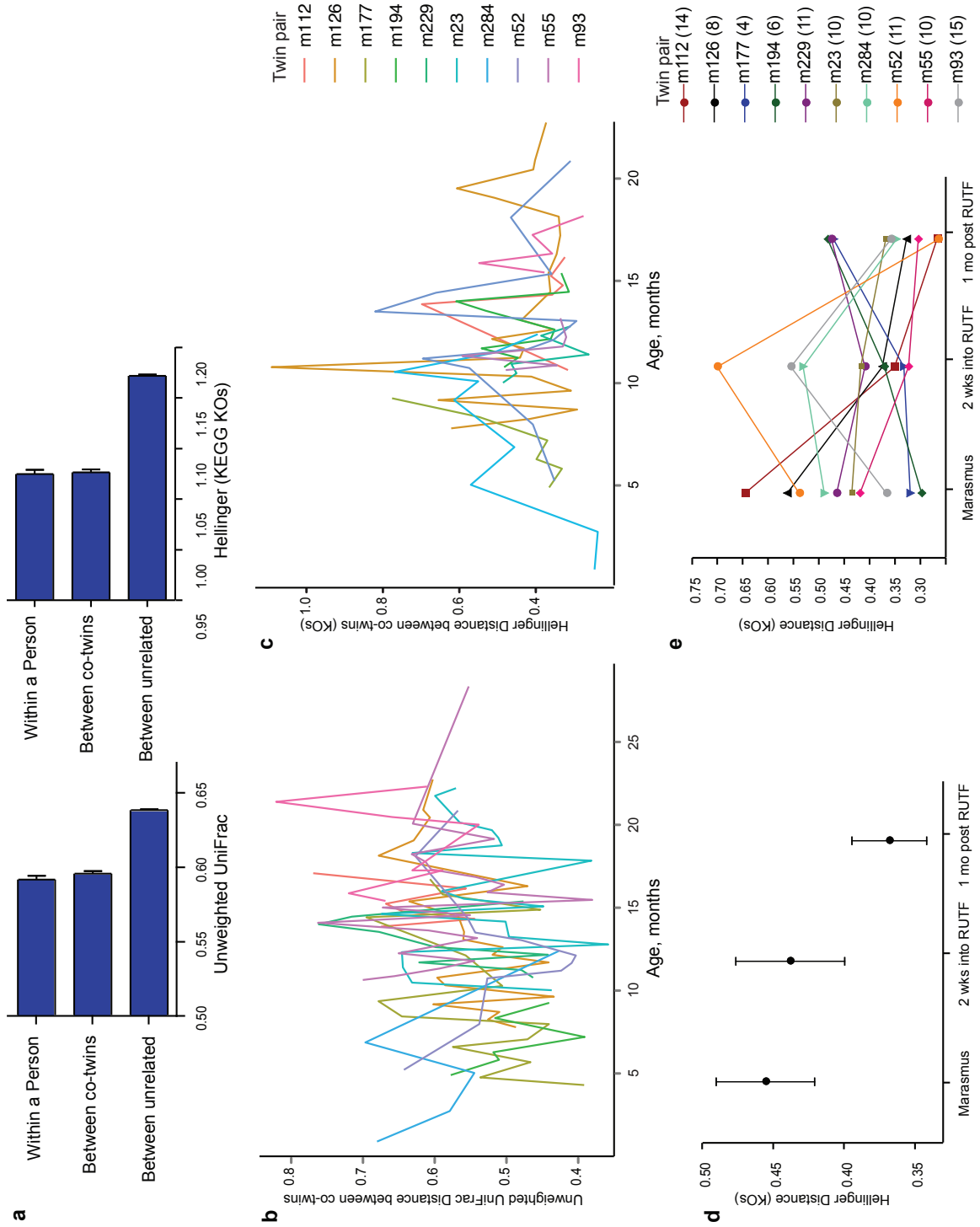


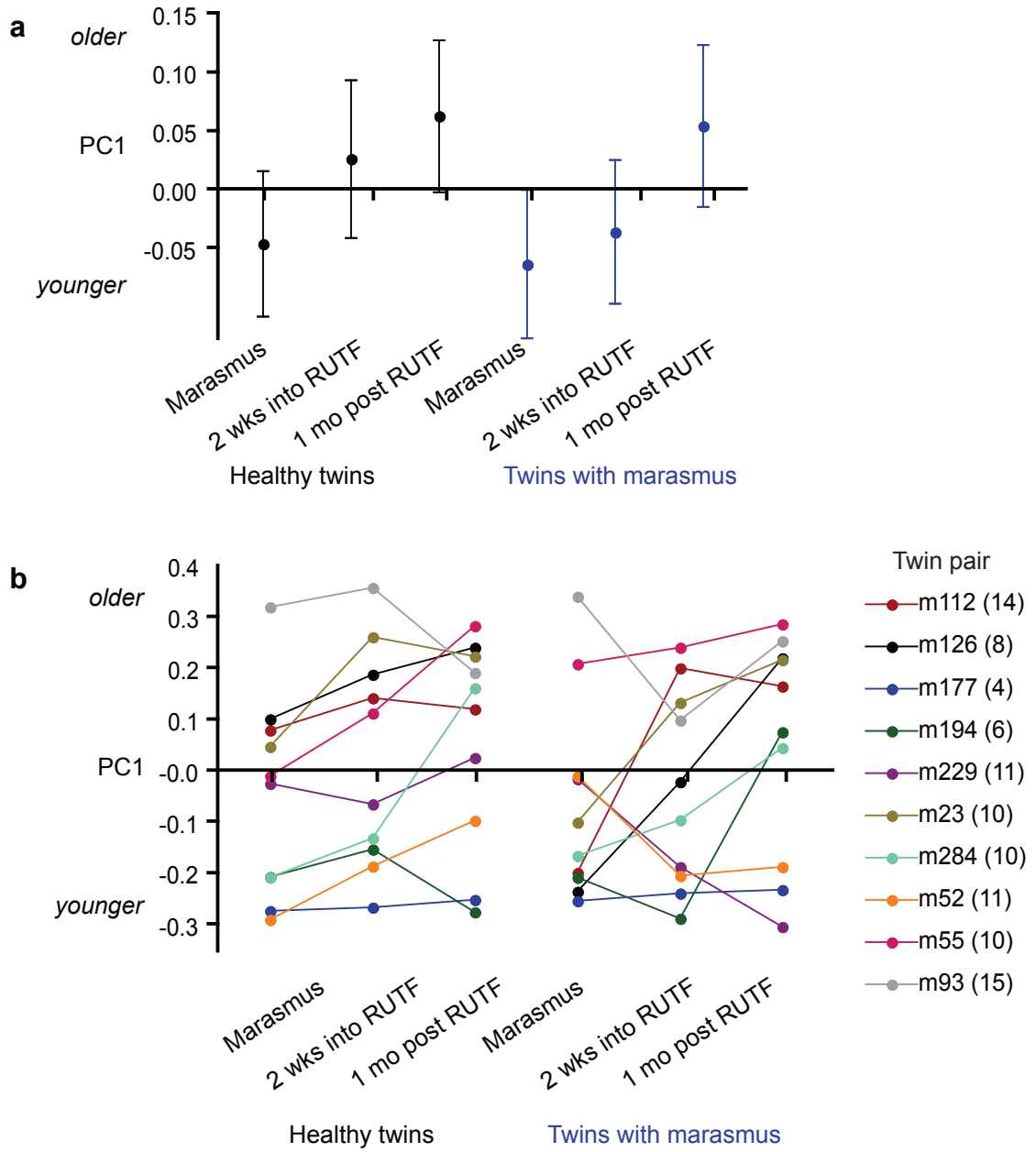
Figure 14.



**Figure 15.**



**Figure 16.**





**Figure 17.**

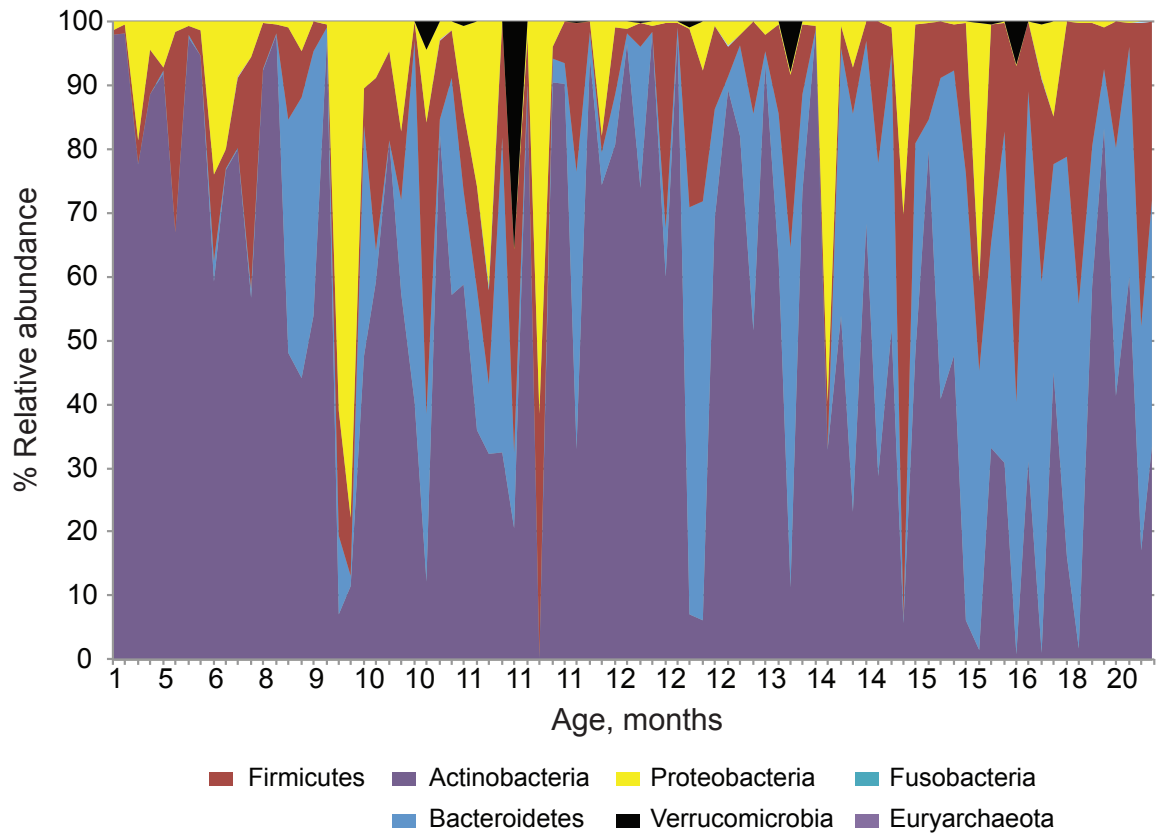
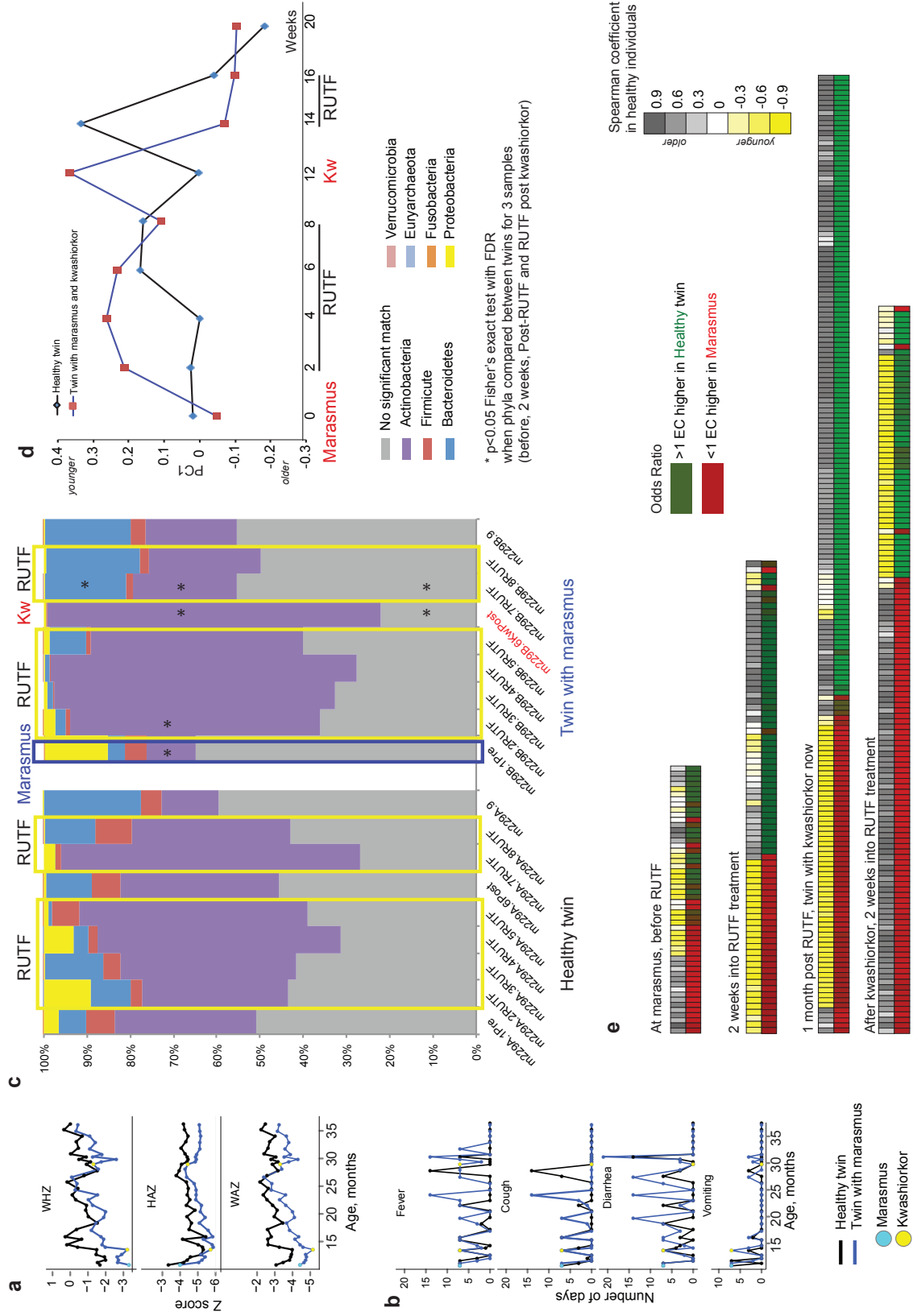


Figure 18.



## **Table Legends**

**Table 1.** (a) Number of twin pairs enrolled in study, their zygosity, gender and nutritional status. (b) An excerpt from part (a) showing number of MZ and DZ twins concordant or discordant for malnutrition used to estimate the significant of the relationship between the zygosity and concordance for malnutrition.

**Table 2.** Gender distribution in undernourished twin pairs.

**Table 3.** Characteristics of twins with malnutrition

**Table 4.** Characteristics of families whose fecal microbiomes were subjected to shotgun sequencing.

**Table 5.** Information on whole community DNA sequence datasets.

**Table 6.** List of the 127 reference sequenced human gut microbial genomes.

**Table 7.** ECs identified by Spearman correlation analysis that exhibit age-associated changes in their proportional representation in healthy fecal microbiomes.

**Table 8.** ECs whose representation is significantly different in the fecal microbiomes of healthy versus malnourished co-twins before and at the time of presentation with kwashiorkor, 2 weeks into RUTF treatment and 1 month after cessation of RUTF.

**Table 9.** ECs whose representation is significantly different in the gut microbiomes of a healthy co-twin versus his/her co-twin with kwashiorkor (pair k138). Twins were sampled before and at the time of presentation with kwashiorkor, 2 weeks into RUTF treatment and 1 month after cessation of RUTF.

**Table 10.** ECs whose representation in the fecal microbiome of a healthy co-twin was significantly different than in his/her co-twin with marasmus when sampled at the time of presentation with marasmus, 2 weeks into the RUTF treatment period, and 1 month after cessation of RUTF.

**Table 11.** ECs whose representation was significantly different in the fecal microbiome of a healthy twin versus his co-twin who presented with marasmus (pair m229). Twins were sampled at the time of presentation with marasmus, 2 weeks into the period of RUTF treatment, 1 month after cessation of RUTF, when the twin who had marasmus subsequently developed kwashiorkor, and 2 weeks after cessation of a second round of RUTF treatment.

**Tables**

**Table 1.**

**Table 1 (a) Number of twin pairs enrolled in study, their zygosity, gender and nutritional status. Each cell shows number of twin pairs in a category.**

Category	Both Healthy	Discordant for malnutrition		Concordant for malnutrition		Total number of twin pairs
		Kwashiorkor	Marasmus	Kwashiorkor	Marasmus	
<b>Total</b>	<b>159</b>	<b>135</b>		<b>23</b>		<b>317</b>
MZ	23	21		3		47
DZ	136	114		20		270
Same gender	100	95		15		210
Different gender	59	40		8		107
		Severe				
		Moderate				
<b>Total with malnutrition</b>		<b>81</b>	<b>21</b>	<b>4</b>	<b>5</b>	<b>144</b>
MZ		13	1	0	1	22
DZ		68	20	4	4	122
Same gender		57	19	4	4	110
Different gender		124	2	0	1	141

One twin died in a twin pair	19	11	3	9	1	0	43
One twin died in a mixed gender twin pair	9	2	0	1	0	0	12

**(b) An excerpt from part (a) showing number of MZ and DZ twins concordant or discordant for malnutrition used to estimate the significance of the relationship between the zygosity and concordance for malnutrition. Each cell shows number of twin pairs in a category.**

		Discordant for moderate or severe malnutrition		Discordant for severe malnutrition	
		MZ	DZ	MZ	DZ
Concordant	3	21	114	8	46
Concordant	1	3	20	1	8

**Table 2.**

**Table 2. Gender distribution in undernourished twin pairs.**

	# children	#females		#males		% females		% males		Number of children who were in the same gender twin pair	% total children
<b>Children with one type of malnutrition</b>											
Kwashiorkor	46	27	19	59	41	45	7.2				
Marasmus	16	7	9	44	56	16	2.5				
Moderate	89	54	35	61	39	85	13.8				
<b>Children with multiple types of malnutrition</b>											
Kwashiorkor and Marasmus	2	0	2	0	100	2	0.3				
Marasmus and Moderate	35	20	15	57	43	34	5.4				
Kwashiorkor and Moderate	20	12	8	60	40	16	3.1				
Kwashiorkor and Marasmus and Moderate	10	5	5	50	50	10	1.6				
Total number of children with malnutrition											218
Total number of children in a study											643
											10.4

**Table 3.**

**Table 3. Characteristics of twins with malnutrition**

Average	Children with 1 event of malnutrition						Average number of days with a symptom prior to a visit to the clinic						Number of children living in					
	Age	WHZ	HAZ	WAZ	MUAC*		Fever	Cough	Diarrhea	Vomiting	Chamba	Makwhira	Mayaka	Mbiza	Mitondo			
Kwashiorkor	16.3	-1.1	-3.6	-2.8	12.7		3.1	3.6	3.1	1.4	10	4	12	15	3			
Marasmus	10.6	-3.6	-3.7	-4.6	10.6		5.0	3.8	5.8	2.7	1	6	4	4	1			
Moderate	14.2	-2.2	-3.4	-3.5	11.9		3.3	2.9	2.9	1.1	16	18	20	20	24			
st dev											27	28	36	39	28			
Kwashiorkor	6.7	1.1	1.3	1.3	1.7		3.7	3.9	4.0	2.9								
Marasmus	4.1	0.6	1.3	1.0	0.7		3.6	2.9	4.2	3.7								
Moderate	6.6	0.5	1.2	0.9	1.0		3.4	3.6	4.0	2.2								
Dunn's multiple comparison test (after Kruskal-Wallis test)	*	***		**/*					*									

Average	Children with more than 1 event of malnutrition §						Average number of days with a symptom prior to a visit to the clinic						Number of children living in					
	Age	WHZ	HAZ	WAZ	MUAC §§		Fever	Cough	Diarrhea	Vomiting	Chamba	Makwhira	Mayaka	Mbiza	Mitondo			
12.7	-2.5	-4.0	-4.1	-4.1	11.3		3.2	3.0	3.5	1.7	17	42	37	31	25			
6.0	0.9	1.2	1.0	1.0	1.1		3.1	3.4	4.0	3.0								
*significant compared to children with 1 event of moderate malnutrition																		

§ For example when a child presented with kwashiorkor first, and later developed marasmus  
 §§ MUAC is mid-upper arm circumference

Table 4.

Table 4. Characteristics of families whose fecal microbiomes were subjected to shotgun sequencing

Twins	#twin pairs discordant throughout the study§	#twin pairs with samples before malnutrition	Total #samples sequenced	Gender of twin pairs			# twin pairs in village			Mothers characteristics (Averages)			Average ± SD at 1st presentation with severe malnutrition			Average ± SD throughout sampling			Average #days at malnutrition with											
				#boys	#girls	Mitondo	Makwhira	Mbiza	Chamba	Mayaka	Age	BMI	Height	MUAC	Subcategory	Age	WHZ	MUAC	Min	Max	WHZ	Min	Max	MUAC	During	Year	Months			
Both healthy	9	0	97	4	6	0	0	6	2	2	8	29	24.0	154.0	27.04	NA	NA	NA	11.6 ± 6.7	0.6	24.4	0.2 ± 1.2	-2.4	2.8	13.5 ± 1.5	1	0	2		
Discordant Kwashiorkor	13	9	207	8	5	1	2	3	3	4	11	28	22.7	151.9	26.13	Healthy twins	14.9 ± 5.5	-0.60	0.8	13.2 ± 1.0	13.6 ± 0.4	0.7	29.8	-0.4 ± 1.0	-2.8	2.2	13.4 ± 1.4	0	0	2
Discordant or marasmus	10	7	169	6	4	0	5	2	0	3	9	27	21.9	155.8	24.94	Kwashiorkor	14.9 ± 5.5	-0.71 ± 1.0	13.4 ± 1.7	14.9 ± 6.3	0.7	29.8	0.4 ± 1.2	-2.5	2.7	13.5 ± 1.7	2	1	3	
																Healthy twins	9.9 ± 3.3	-1.22 ± 0.7	12.3 ± 1.1	12.1 ± 4.3	0.9	22.7	-1.1 ± 0.9	-3.0	0.9	12.4 ± 1.3	4	1	3	
																Marasmus	9.9 ± 3.3	-3.2 ± 0.3	10.6 ± 0.7	12.0 ± 4.3	0.9	22.7	2.3 ± 1.1	-4.8	0.4	11.4 ± 1.0	7	3	6	

§ Mothers' characteristics were not significantly different

§ Healthy twin never developed malnutrition throughout the study



Table 5.

Site	Twin pairs	Subject health category	SampleID	Number of high quality read sequences	Family ID	Person ID	Gender	Sample number	Age	Z scores				Sample collection time						Number of days with a symptom prior to a visit				
										WHZ	WAZ	HAZ	MIUAC	Disease	RUTF	Month	Year	Fever	Cough		Diarrhea	Vomiting		
Mbiza	Both healthy	Healthy	h10A.2	193,914	h10	010A	Female	2	18.953	-0.38	-1.72	-1.1	14.6	NA	NA	Nov 2008	0	2	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h10A.3	86,497	h10	010A	Female	4	24.378	-0.9	-2.7	-2.13	14.2	NA	NA	Feb 2009	3	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h10A.4	97,578	h10	010A	Female	5	27.1376	-0.76	-3.13	-2.35	12.8	NA	NA	Apr 2009	2	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h10B.2	210,101	h10	010B	Female	2	18.953	-0.59	-1.72	-1.26	14.7	NA	NA	Nov 2008	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h10B.3	60,038	h10	010B	Female	3	21.6181	-0.28	-2.18	-1.32	14.6	NA	NA	Feb 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h10B.4	197,927	h10	010B	Female	4	24.378	-0.48	-2.86	-2.04	14.4	NA	NA	Apr 2009	6	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h10B.5	67,635	h10	010B	Female	5	27.1376	0.26	-2.98	-1.44	14.2	NA	NA	Jul 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h10M	150,839	h10	010M	Female	1	318	NA	NA	NA	NA	NA	NA	Jul 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Mother	h121A.1	99,688	h121	121A	Male	1	5.19097	2.01	-1.54	0.51	15.8	NA	NA	Oct 2008	0	3	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h121A.2	74,315	h121	121A	Male	2	7.95072	1.36	-2.74	-0.65	15	NA	NA	Jan 2009	7	5	4	2	0	0	0	
Mbiza	Both healthy	Healthy	h121A.3	113,817	h121	121A	Male	3	10.7105	2.12	-3.1	-0.08	16.6	NA	NA	Mar 2009	0	3	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h121A.4	66,748	h121	121A	Male	4	13.7002	2.61	-3.94	0.19	16.2	NA	NA	Jun 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h121A.5	152,887	h121	121A	Male	5	16.768	1.82	-2.62	0.09	16	NA	NA	Dec 2008	7	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h121A.6	45,460	h121	121A	Male	6	21.5195	1.82	-2.62	0.09	16	NA	NA	Dec 2008	7	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h121B.1	152,887	h121	121B	Male	1	5.19097	2.58	-2.03	1.46	15.3	NA	NA	Oct 2008	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h121B.2	46,377	h121	121B	Male	2	7.95072	2.77	-2.06	0.93	16	NA	NA	Jan 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h121B.3	107,319	h121	121B	Male	3	10.7105	2.38	-2.24	0.66	15.8	NA	NA	Mar 2009	3	3	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h121B.4	60,741	h121	121B	Male	4	13.4702	2.42	-2.23	0.82	16.4	NA	NA	Jun 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h121B.5	103,286	h121	121B	Male	5	18.7998	2.38	-2.37	0.73	17	NA	NA	Dec 2008	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h121B.6	142,538	h121	121B	Male	6	18.7998	2.38	-2.37	0.73	17	NA	NA	Dec 2008	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h209A.1	131,639	h209	209A	Female	2	6.18297	-0.46	-2.88	-1.89	14.3	NA	NA	Feb 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h209A.2	69,479	h209	209A	Female	3	8.6078	0.38	-2.65	-1.27	13.6	NA	NA	Feb 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h209A.3	59,957	h209	209A	Female	4	10.9076	0.06	-2.64	-1.4	13.6	NA	NA	Apr 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h209A.4	94,041	h209	209A	Female	5	12.7474	-0.15	-2.64	-1.46	13.6	NA	NA	Jun 2009	0	4	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h209A.5	42,466	h209	209A	Female	6	16.4271	0.63	-3.22	-1.12	13.8	NA	NA	Aug 2009	4	3	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h209A.6	98,369	h209	209A	Female	7	6.76797	-0.83	-3.22	-1.12	13.8	NA	NA	Dec 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h209B.2	17,965	h209	209B	Female	3	6.0078	0.24	-2.59	-1.35	12.8	NA	NA	Apr 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h209B.3	61,988	h209	209B	Female	4	8.6078	0.24	-2.59	-1.35	12.8	NA	NA	Apr 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h209B.4	47,188	h209	209B	Female	5	12.7474	-0.64	-3.39	-1.72	13.4	NA	NA	Aug 2009	4	5	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h209B.5	147,891	h209	209B	Female	6	16.4271	0.59	-3.39	-1.26	14	NA	NA	Aug 2009	4	5	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h209M.2	28,072	h209	209M	Female	2	4.89403	-0.62	-2.48	-2.7	11	NA	NA	NA	NA	NA	0	1	0	0	0	
Mbiza	Both healthy	Healthy	h257A.1	155,162	h257	257A	Male	2	2.33285	-0.62	-2.48	-2.7	11	NA	NA	Feb 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257A.2	85,786	h257	257A	Male	2	4.17248	0.96	-3.72	-2.36	11.8	NA	NA	Apr 2009	3	3	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257A.3	86,988	h257	257A	Male	3	6.01232	-0.12	-3.57	-2.51	12.4	NA	NA	Jun 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257A.4	87,865	h257	257A	Male	4	8.6078	0.24	-2.59	-1.35	12.8	NA	NA	Apr 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257A.5	82,789	h257	257A	Male	5	11.7618	-1.21	-3.67	-3.02	12.4	NA	NA	Aug 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257A.6	80,985	h257	257A	Male	6	11.7618	-1.21	-3.67	-3.02	12.4	NA	NA	Aug 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257A.7	52,169	h257	257A	Male	7	14.2916	-1.52	-4.56	-3.51	12.2	NA	NA	Dec 2010	3	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257A.8	82,789	h257	257A	Male	8	14.2916	-1.52	-4.56	-3.51	12.2	NA	NA	Dec 2010	3	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257B.1	111,466	h257	257B	Male	1	2.33285	0.47	-2.48	-1.97	11.6	NA	NA	Feb 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257B.2	59,679	h257	257B	Male	2	4.17248	1.02	-3.72	-2.31	12.4	NA	NA	Apr 2009	3	3	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257B.3	137,534	h257	257B	Male	3	6.01232	-0.1	-3.34	-2.39	12.6	NA	NA	Jun 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257B.4	87,865	h257	257B	Male	4	7.95216	-0.27	-2.9	-2.01	13.6	NA	NA	Aug 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257B.5	87,865	h257	257B	Male	5	11.7618	-1.21	-3.67	-3.02	12.4	NA	NA	Apr 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257B.6	65,076	h257	257B	Male	6	14.2916	-1.46	-4.86	-3.72	14.2	NA	NA	Aug 2009	3	2	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h257M.2	180,417	h257	257M	Female	2	291,222	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Mbiza	Both healthy	Healthy	h259A.1	144,014	h259	259A	Female	1	1.37988	0.1	-1.04	-0.87	12	NA	NA	Feb 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h259A.2	88,301	h259	259A	Female	2	3.21971	1.86	-2.05	-0.33	13.4	NA	NA	Apr 2009	1	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h259A.3	87,825	h259	259A	Female	3	5.05955	0.92	-1.86	-0.64	13.8	NA	NA	Jun 2009	0	3	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h259A.4	86,926	h259	259A	Female	4	10.5791	1.16	-2.29	-0.35	14.8	NA	NA	Jun 2009	0	3	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h259A.5	146,927	h259	259A	Female	5	13.7988	-0.83	-3.28	-3.02	14	NA	NA	Dec 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h259B.1	76,770	h259	259B	Female	2	3.21971	1.53	-3.23	-1.75	12	NA	NA	Apr 2009	1	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h259B.2	90,530	h259	259B	Female	3	5.05955	0.71	-2.99	-1.78	12.8	NA	NA	Apr 2009	0	1	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h259B.3	24,237	h259	259B	Female	4	6.89938	-0.15	-2.22	-1.58	13	NA	NA	Jun 2009	0	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h259B.4	72,110	h259	259B	Female	5	10.5791	-0.19	-2.7	-1.65	13.6	NA	NA	Aug 2009	3	0	0	0	0	0	0	
Mbiza	Both healthy	Healthy	h259B.5	81,818	h259	259B	Female	6	12.4189	0.17	-2.9	-2.01	13.6	NA	NA	Dec 2009	3	0						











Mbiza	Discordant for marasmus	m55A	Female	2	10,8747	-0.69	-4.42	-3.46	11	RUTF	Sep	2008	2	3	2	0
Mbiza	Healthy cotwin of marasmus	m55A_3RUTF	Female	3	11,3347	-0.41	-4.41	-3.04	11.5	RUTF	Sep	2008	2	1	2	0
Mbiza	Healthy cotwin of marasmus	m55A_4RUTF	Female	4	11,7947	-1.01	-4	-3.12	12	RUTF	Oct	2008	3	2	0	3
Mbiza	Healthy cotwin of marasmus	m55A_5RUTF	Female	5	12,2546	-0.26	-4.37	-2.8	12.4	RUTF	Nov	2008	2	1	2	0
Mbiza	Healthy cotwin of marasmus	m55A_6Post	Female	6	13,1745	-0.97	-4.52	-3.37	12	Marasmus	Nov	2008	3	2	7	2
Mbiza	Discordant for marasmus	m55B_1Mts	Female	1	10,4148	-3.44	-3.62	-4.5	9.8	Marasmus	Sep	2008	7	2	2	0
Mbiza	Discordant for marasmus	m55B_2RUTF	Female	2	10,8747	-2.58	-3.62	-4.08	10.3	RUTF	Sep	2008	2	3	2	0
Mbiza	Discordant for marasmus	m55B_3RUTF	Female	3	11,3347	-2.38	-4.01	-4.05	10.1	RUTF	Sep	2008	2	1	2	0
Mbiza	Discordant for marasmus	m55B_4RUTF	Female	4	11,7947	-2.01	-4.2	-3.92	10.5	RUTF	Oct	2008	3	2	0	3
Mbiza	Discordant for marasmus	m55B_5RUTF	Female	5	12,2546	-2.14	-3.8	-3.71	11.3	RUTF	Oct	2008	2	0	4	2
Mbiza	Discordant for marasmus	m55B_6MtsPost	Female	6	13,1745	-2.97	-3.76	-3.14	10.8	NA	Nov	2008	3	2	0	2
Mbiza	Discordant for marasmus	m93A_1	Male	1	15,4086	-3.18	-4.93	-4.63	11.4	Marasmus	NA	NA	NA	NA	7	NA
Mbiza	Discordant for marasmus	m93A_2	Male	2	16,8886	-2.56	-4.81	-4.11	12.1	Marasmus	Sep	2008	3	4	3	3
Mbiza	Discordant for marasmus	m93A_3RUTF	Male	3	16,8886	-2.23	-4.82	-4.03	13.2	RUTF	Oct	2008	2	2	3	0
Mbiza	Discordant for marasmus	m93A_4PostRUTF	Male	4	17,2485	-2.68	-3.96	-3.79	13	RUTF	Oct	2008	2	3	0	0
Mbiza	Discordant for marasmus	m93A_5PostRUTF	Male	5	18,1684	-2.08	-4.22	-3.57	13	RUTF	Nov	2008	3	3	5	1
Mbiza	Healthy cotwin of marasmus	m93B_1	Male	1	15,4086	0.83	-4.34	-1.65	14.3	RUTF	Dec	2008	0	0	1	0
Mbiza	Healthy cotwin of marasmus	m93B_2RUTF	Male	2	15,8686	0.88	-4.3	-1.57	14.5	RUTF	Dec	2008	0	2	0	0
Mbiza	Healthy cotwin of marasmus	m93B_3RUTF	Male	3	16,3285	0.74	-4.06	-1.51	15.1	RUTF	Oct	2008	0	3	2	0
Mbiza	Healthy cotwin of marasmus	m93B_4PostRUTF	Male	4	17,2485	0.66	-4.14	-1.61	15	RUTF	Oct	2008	0	0	0	0
Mbiza	Healthy cotwin of marasmus	m93B_5PostRUTF	Male	5	18,1684	0.85	-4.41	-1.62	13.3	NA	Nov	2008	2	3	1	0
Mbiza	Mother	m93M_1	Female	1	306	NA	NA	NA	NA	NA	Dec	2008	0	NA	NA	NA

**Table 6.****Table 6. List of the 127 reference human gut microbial genomes**

Genome name	Genbank ID	Genome size
<i>Actinomyces odontolyticus</i> ATCC 17982	NZ_AAYI00000000	2,393,758
<i>Akkermansia muciniphila</i> ATCC BAA-835	NC_010655	2,664,102
<i>Alistipes putredinis</i> DSM 17216	NZ_ABFK00000000	2,549,878
<i>Anaerococcus hydrogenalis</i> DSM 7454	NZ_ABXA00000000	1,889,366
<i>Anaerofustis stercorihominis</i> DSM 17244	NZ_ABIL00000000	2,284,603
<i>Anaerostipes caccae</i> DSM 14662	NZ_ABAX00000000	3,605,636
<i>Anaerotruncus colihominis</i> DSM 17241	NZ_ABGD00000000	3,718,888
<i>Bacteroides caccae</i> ATCC 43185	NZ_AAVM00000000	4,564,814
<i>Bacteroides capillosus</i> ATCC 29799	NZ_AAXG00000000	4,241,076
<i>Bacteroides cellulosilyticus</i> DSM 14838	NZ_ACCH00000000	6,726,268
<i>Bacteroides coprocola</i> DSM 17136	NZ_ABIY00000000	4,295,617
<i>Bacteroides coprophilus</i> DSM 18228	NZ_ACBW00000000	3,855,443
<i>Bacteroides dorei</i> DSM 17855	NZ_ABWZ00000000	5,487,768
<i>Bacteroides eggerthii</i> DSM 20697	NZ_ABVO00000000	4,157,980
<i>Bacteroides finegoldii</i> DSM 17565	NZ_ABXI00000000	4,881,901
<i>Bacteroides fragilis</i> 3_1_12	NZ_ABZX00000000	5,486,240
<i>Bacteroides fragilis</i> NCTC 9343	NC_003228	5,205,140
<i>Bacteroides fragilis</i> YCH46	NC_006347	5,277,274
<i>Bacteroides intestinalis</i> DSM 17393	NZ_ABJL00000000	6,052,596
<i>Bacteroides ovatus</i> ATCC 8483	NZ_AAXF00000000	6,463,169
<i>Bacteroides plebeius</i> DSM 17135	NZ_ABQC00000000	4,421,324
<i>Bacteroides</i> sp. 1_1_6	NZ_ACIC00000000	6,855,195
<i>Bacteroides</i> sp. D1	NZ_ACAB00000000	5,986,762
<i>Bacteroides</i> sp. D2	NZ_ACGA00000000	6,901,960
<i>Bacteroides stercoris</i> ATCC 43183	NZ_ABFZ00000000	4,009,229
<i>Bacteroides thetaiotaomicron</i> 3731	NC_Bthetaiotaomicron3731	7,098,445
<i>Bacteroides thetaiotaomicron</i> 7330	NC_Bthetaiotaomicron7330	6,894,436
<i>Bacteroides thetaiotaomicron</i> VPI-5482	NC_004663	6,260,361
<i>Bacteroides uniformis</i> ATCC 8492	NZ_AAYH00000000	4,717,497
<i>Bacteroides vulgatus</i> ATCC 8482	NC_009614	5,163,189
<i>Bacteroides</i> WH2	NC_BWH2	7,129,681
<i>Bacteroides xylanisolvens</i> XB1A	NC_BxylanisolvensXB1A	5,861,392
<i>Bifidobacterium adolescentis</i> ATCC 15703	NC_008618	2,089,645
<i>Bifidobacterium adolescentis</i> L2-32	NZ_AAXD00000000	2,385,710
<i>Bifidobacterium angulatum</i> DSM 20098	NZ_ABYS00000000	2,007,108
<i>Bifidobacterium animalis</i> subsp. lactis AD011	NC_011835	1,933,695
<i>Bifidobacterium animalis</i> subsp. lactis HN019	NZ_ABOT00000000	1,915,892
<i>Bifidobacterium breve</i> DSM 20213	NZ_ACCG00000000	2,297,799
<i>Bifidobacterium catenulatum</i> DSM 16992	NZ_ABXY00000000	2,058,429
<i>Bifidobacterium dentium</i>	NC_Bdentium	2,642,189
<i>Bifidobacterium gallicum</i> DSM 20093	NZ_ABXB00000000	2,019,802
<i>Bifidobacterium longum</i> DJO10A	NC_010816	2,375,792
<i>Bifidobacterium longum</i> NCC2705	NC_004307	2,256,640
<i>Bifidobacterium pseudocatenulatum</i> DSM 20438	NZ_ABXX00000000	2,304,808
<i>Blautia hansenii</i> DSM 20583	NZ_ABYU00000000	3,053,221
<i>Blautia hydrogenotrophica</i> DSM 10507	NZ_ACBZ00000000	3,565,428
<i>Bryantella formatexigens</i> DSM 14469	NZ_ACCL00000000	4,548,960
<i>Butyrivibrio crossotus</i> DSM 2876	NZ_ABWN00000000	2,482,791
<i>Catenibacterium mitsuokai</i> DSM 15897	NZ_ACCK00000000	2,671,313
<i>Citrobacter youngae</i> ATCC 29220	NZ_ABWL00000000	5,143,204
<i>Clostridium asparagiforme</i> DSM 15981	NZ_ACCJ00000000	6,224,391



<i>Clostridium bartlettii</i> DSM 16795	NZ_ABEZ00000000	2,971,856
<i>Clostridium bolteae</i> ATCC BAA-613	NZ_ABCC00000000	6,556,988
<i>Clostridium hiranonis</i> DSM 13275	NZ_ABWP00000000	2,423,348
<i>Clostridium hylemonae</i> DSM 15053	NZ_ABYI00000000	3,885,459
<i>Clostridium leptum</i> DSM 753	NZ_ABCB00000000	3,270,109
<i>Clostridium methylpentosum</i> DSM 5476	NZ_ACEC00000000	3,406,326
<i>Clostridium nexile</i> DSM 1787	NZ_ABWO00000000	3,861,016
<i>Clostridium ramosum</i> DSM 1402	NZ_ABFX00000000	3,234,795
<i>Clostridium scindens</i> ATCC 35704	NZ_ABFY00000000	3,619,905
<i>Clostridium</i> sp. L2-50	NZ_AAYW00000000	2,954,116
<i>Clostridium</i> sp. M62/1	NZ_ACFX00000000	3,836,694
<i>Clostridium</i> sp. SS2/1	NZ_ABGC00000000	3,141,381
<i>Clostridium spiroforme</i> DSM 1552	NZ_ABIK00000000	2,507,485
<i>Clostridium sporogenes</i> ATCC 15579	NZ_ABKW00000000	4,102,125
<i>Clostridium symbiosum</i>	NC_Csymbiosum	4,954,054
<i>Collinsella aerofaciens</i> ATCC 25986	NZ_AAVN00000000	2,439,869
<i>Collinsella intestinalis</i> DSM 13280	NZ_ABXH00000000	1,804,297
<i>Collinsella stercoris</i> DSM 13279	NZ_ABXJ00000000	2,399,821
<i>Coprococcus comes</i> ATCC 27758	NZ_ABVR00000000	3,238,915
<i>Coprococcus eutactus</i> ATCC 27759	NZ_ABEY00000000	3,102,087
<i>Desulfovibrio piger</i> ATCC 29098	NZ_ABXU00000000	2,826,240
<i>Desulfovibrio piger</i> GOR1	AF192152	2,597,386
<i>Dorea formicigenerans</i> ATCC 27755	NZ_AAXA00000000	3,186,031
<i>Dorea longicatena</i> DSM 13814	NZ_AAXB00000000	2,913,833
<i>Enterobacter cancerogenus</i>	NC_Ecancerogenus	4,605,129
<i>Escherichia coli</i> str. K-12 substr. MG1655	NC_000913	4,639,675
<i>Escherichia fergusonii</i> ATCC 35469	NC_011740	4,588,711
<i>Eubacterium bifforme</i> DSM 3989	NZ_ABYT00000000	2,415,920
<i>Eubacterium dolichum</i> DSM 3991	NZ_ABAW00000000	2,190,453
<i>Eubacterium eligens</i> ATCC 27750	NC_012778	2,144,190
<i>Eubacterium hallii</i> DSM 3353	NZ_ACEP00000000	3,290,996
<i>Eubacterium rectale</i> ATCC 33656	NC_012781	3,449,685
<i>Eubacterium rectale</i> DSM 17629	NC_Erectale_DSM17629	3,255,606
<i>Eubacterium ventriosum</i> ATCC 27560	NZ_AAVL00000000	2,869,695
<i>Faecalibacterium prausnitzii</i> A2-165	NZ_ACOP00000000	3,080,849
<i>Faecalibacterium prausnitzii</i> M21/2	NZ_ABED00000000	3,126,983
<i>Fusobacterium</i> sp. 4_1_13	NZ_ACDE00000000	2,268,505
<i>Fusobacterium varium</i> ATCC 27725	NZ_ACIE00000000	3,321,664
<i>Helicobacter pylori</i> HPAG1	NC_008086	1,596,366
<i>Holdemania filiformis</i> DSM 12042	NZ_ACCF00000000	3,803,745
<i>Lactobacillus casei</i> ATCC 334	NC_008526	2,895,264
<i>Lactobacillus delbrueckii</i> subsp. <i>bulgaricus</i> ATCC 1	NC_008054	1,864,998
<i>Lactobacillus reuteri</i> DSM 20016	NC_009513	1,999,618
<i>Lactococcus lactis</i> subsp. <i>cremoris</i> MG1363	NC_009004	2,529,478
<i>Lactococcus lactis</i> subsp. <i>cremoris</i> SK11	NC_008527	2,438,589
<i>Lactococcus lactis</i> subsp. <i>lactis</i> II1403	NC_002662	2,365,589
M23A	NC_M23A	4,338,875
<i>Methanobrevibacter smithii</i> ATCC 35061	CP000678.1	1,853,160
<i>Methanobrevibacter smithii</i> DSM 2374	NZ_ABYV00000000	1,727,775
<i>Methanobrevibacter smithii</i> DSM 2375	NZ_ABYW00000000	1,704,865
<i>Methanosphaera stadtmanae</i> DSM 3091	NC_007681	1,767,403
<i>Mitsuokella multacida</i> DSM 20544	NZ_ABWK00000000	2,574,556
<i>Parabacteroides distasonis</i> ATCC 8503	NC_009615	4,811,379
<i>Parabacteroides johnsonii</i> DSM 18315	NZ_ABYH00000000	4,612,238

Parabacteroides merdae ATCC 43184	NZ_AAXE00000000	4,431,877
Parvimonas micra ATCC 33270	NZ_ABEE00000000	1,703,772
Prevotella copri DSM 18205	NZ_ACBX00000000	3,507,873
Proteus penneri ATCC 35198	NZ_ABVP00000000	3,747,729
Providencia alcalifaciens DSM 30120	NZ_ABXW00000000	4,029,346
Providencia rettgeri DSM 1131	NZ_ACCI00000000	4,749,568
Providencia rustigianii DSM 4541	NZ_ABXV00000000	3,965,844
Providencia stuartii ATCC 25827	NZ_ABJD00000000	4,603,561
Roseburia intestinalis L1-82	NZ_ABYJ00000000	4,380,675
Ruminococcus bromii L263	NC_RbromiiL263	2,240,019
Ruminococcus gnavus ATCC 29149	NZ_AAYG00000000	3,501,911
Ruminococcus lactaris ATCC 29176	NZ_ABOU00000000	2,729,735
Ruminococcus obeum ATCC 29174	NZ_AAVO00000000	3,624,708
Ruminococcus torques ATCC 27756	NZ_AAVP00000000	2,739,406
Shigella sp. D9	NZ_ACDL00000000	4,717,340
Streptococcus infantarius subsp. infantarius ATCC	NZ_ABJK00000000	1,925,087
Streptococcus thermophilus CNRZ1066	NC_006449	1,796,226
Streptococcus thermophilus LMD-9	NC_008532	1,856,368
Streptococcus thermophilus LMG 18311	NC_006448	1,796,846
Subdoligranulum variabile DSM 15176	NZ_ACBY00000000	3,237,471
Victivallis vadensis ATCC BAA-548	NZ_ABDE00000000	5,294,868
Vibrio cholerae O1 biovar eltor str. N16961 (chromc	NC_002505, NC_002506	

**Tables 7 – 8.**

**Please reference provided CD for these tables.**

**Table 9.**

**Table 9. Significantly different ECs between healthy and malnourished twins 1;138 before and at the presentation with kwashiorkor, 2 weeks into RUTF treatment and 1 month after cessation of RUTF.**

Fig. 9's exact p-values and odds ratios are indicated in the healthy twin microbioms (f odds ratio >1). Fig. 9's exact p-values and odds ratios are indicated for each comparison. ECs are ordered in the healthy twin microbioms (f odds ratio >1). Spearman's correlation coefficients from the analysis of changes of ECs with age in healthy individuals are indicated: a positive correlation coefficient indicates an increase in proportional representation of the EC with increasing age (colored in grey), while a negative value indicates a decrease with increasing age (colored in yellow).

EC	1.6 months before kwashiorkor		2 weeks into RUTF		1 month after cessation of RUTF		Spearman in healthy subjects	EC annotation	Pathway	Category	
	p value (adjusted with FDR)	Odds ratio	p value (adjusted with FDR)	Odds ratio	p value (adjusted with FDR)	Odds ratio					
EC2.3.1.-	1.0873E-06	0.4	1.26627E-21	0.22	0.02	1.39	-0.8	transferase activity, transferring acyl groups other than amino-acyl		AMINO ACID METABOLISM BIOSYNTHESIS OF POLYKETIDES AND NONRIBOSOMAL PEPTIDES BIOSYNTHESIS OF SECONDARY METABOLITES CARBOHYDRATE METABOLISM GENERAL FUNCTION BIOSYNTHESIS AND METABOLISM GLUCAN METABOLISM LIPID METABOLISM MEMBRANE AND INTRACELLULAR STRUCTURAL MOLECULES MEMBRANE TRANSPORT OTHER S PROTEIN FOLDING AND ASSOCIATED PROCESSING REPLICATION AND REPAIR TRANSLATION PROTEINS XENOBIOTICS BIODEGRADATION AND METABOLISM	
EC3.2.1.24	3.1734E-05	0.3	4.05065E-06	0.24			-0.8	alpha-mannosidase	Lysosome Other glycan degradation	GLUCAN BIOSYNTHESIS AND METABOLISM TRANSPORT AND CELLULAR FUNCTIONS	
EC3.2.1.21	0.00016128	0.6	0.031447482	0.77	0.03	0.82	0.5	beta-glucosidase	Cyanoamino acid metabolism Phenylpropanoid biosynthesis Starch and sucrose metabolism	BIOSYNTHESIS OF SECONDARY METABOLITES CARBOHYDRATE METABOLISM METABOLISM OF OTHER AMINO ACIDS	
EC2.3.2.-	0.00164878	2.4	0.001242056	0.24	9.9534E-06	2.59	-0.8	transferase activity, transferring amino-acyl groups	Butirosin and neomycin biosynthesis Peptidoglycan biosynthesis	BIOSYNTHESIS OF SECONDARY METABOLITES GLUCAN BIOSYNTHESIS AND METABOLISM	
EC2.7.1.69	0.00516443	1.4	1.16244E-12	1.77	1.1882E-19	2.18	0.005	0.5	protein-N(P)-phosphohistidine-sugar phosphotransferase	Amino sugar and nucleotide sugar metabolism Ascorbate and aldarate metabolism Deleted Fructose and mannose metabolism Galactose metabolism Glycolysis / Gluconeogenesis Phosphotransferase system (PTS) Starch and sucrose metabolism Purine metabolism	CARBOHYDRATE METABOLISM MEMBRANE TRANSPORT UNPROCESSED
EC3.2.2.1	0.02181532	0.4	0.003521534	0.30			-0.8	purine nucleosidase	DNA replication proteins Unclassified	METABOLISM OF COFACTORS AND VITAMINS NUCLEOTIDE METABOLISM	
EC2.7.7.49	1.38798E-20	5.10					0.6	RNA-directed DNA polymerase	DNA replication proteins Unclassified	REPLICATION AND REPAIR PROTEINS	
EC3.1.26.12	6.73006E-10	0.07	0.010	2.18			-0.8	Ribonuclease E	RNA degradation	FOLDING, SORTING AND DEGRADATION	
EC3.1.3.1	3.4626E-09	0.05					-0.7	alkaline phosphatase	Folate biosynthesis Two-component system gamma-Hexachlorocyclohexane degradation	METABOLISM OF COFACTORS AND VITAMINS SIGNAL TRANSDUCTION AND METABOLISM	
EC3.2.1.58	3.10845E-08	0.03					-0.8	glucan 1,3 beta-glucosidase	Starch and sucrose metabolism	BIODEGRADATION AND METABOLISM	
EC3.2.7.52	3.8496E-09	0.08					-0.9	glucan 1,3 beta-glucosyltransferase	Starch and sucrose metabolism	BIODEGRADATION AND METABOLISM	
EC3.2.1.86	2.06154E-06	2.93	1.5026E-06	3.33			-0.6	glucosyl-6-phospho-beta-D-glucosyltransferase	Glucosyl- / Glucosaminosyltransferase	GLUCAN TRANSPORT AND METABOLISM	
EC3.5.3.16	4.75254E-06	Inf					0.5	arginine deiminase	Arginine and proline metabolism	CARBOHYDRATE METABOLISM	
EC3.2.1.122	7.65035E-06	32.45					0.4	maltese-G-phosphate glucosidase	Starch and sucrose metabolism	AMINO ACID METABOLISM	
EC2.7.7.42	9.86661E-06	0.12	0.02	2.08			-0.8	[glutamate-ammonia-ligase] adenyllyltransferase	Unclassified	PROTEIN FOLDING AND ASSOCIATED PROCESSING	
EC4.1.1.31	1.46966E-05	0.26	0.009	1.89			-0.8	phosphoenolpyruvate carboxylase	Carbon fixation in photosynthetic organisms Pyruvate metabolism Reductive carboxylation cycle (CO2 fixation)	CARBOHYDRATE METABOLISM ENERGY METABOLISM	
EC1.1.1.14	2.64162E-05	9.82					0.4	L-iditol 2-dehydrogenase	Fructose and mannose metabolism	CARBOHYDRATE METABOLISM	
EC3.1.11.5	2.91452E-05	2.85					0.7	exodeoxyribonuclease V	DNA repair and recombination proteins Homologous recombination Unclassified	REPLICATION AND REPAIR PROTEINS	
EC3.4.22.40	4.94209E-05	0.37					-0.8	Bleomycin hydrolase	Peptidases	ENZYMATIC FAMILIES	

EC3.5.99.2									thiaminase	-0.6					Thiamine metabolism Transcription factors	METABOLISM OF COFACTORS AND VITAMINS TRANSCRIPTION
EC2.8.3.1									propionate CoA-transferase	0.2					Propionate metabolism Pyruvate metabolism Styrene degradation	CARBOHYDRATE METABOLISM XENOBIOTICS METABOLISM BIODEGRADATION AND METABOLISM
EC1.1.1.86									ketol-acid reductoisomerase	-0.6					Pantothenate and CoA biosynthesis Valine, leucine and isoleucine biosynthesis	AMINO ACID METABOLISM METABOLISM OF COFACTORS AND VITAMINS
EC2.7.9.1							0.45		pyruvate, phosphate dikinase	0.2					Carbon fixation in photosynthetic organisms Pyruvate metabolism	CARBOHYDRATE METABOLISM ENERGY METABOLISM
EC6.2.1.5									succinate-CoA ligase (ADP-forming)	-0.8					C5-Branched dibasic acid metabolism Citrate cycle (TCA cycle) Propionate metabolism Reductive carboxylate cycle (CO2 fixation)	CARBOHYDRATE METABOLISM ENERGY METABOLISM
EC1.5.3.1									sarcosine oxidase	0.8					Glycine, serine and threonine metabolism Lysine degradation Peroxisome	AMINO ACID METABOLISM TRANSPORT AND CATABOLISM
EC5.4.2.2									phosphoglucomutase	-0.8					Amino sugar and nucleotide sugar metabolism Galactose metabolism Glycolysis / gluconeogenesis Pentose phosphate pathway Starch and sucrose metabolism Streptomycin biosynthesis	BIOSYNTHESIS OF SECONDARY METABOLITES CARBOHYDRATE METABOLISM
EC2.2.1.1									transketolase	0.1					Biosynthesis of ansamycins Carbon fixation in photosynthetic organisms Pentose phosphate pathway	BIOSYNTHESIS OF POLYKETIDES AND NONRIBOSOMAL PEPTIDES CARBOHYDRATE METABOLISM ENERGY METABOLISM
EC6.2.1.3									long-chain fatty acid-CoA ligase	-0.7					Adipocytokine signaling pathway Fatty acid metabolism Lipid biosynthesis proteins PPAR signaling pathway Peroxisome	ENDOCRINE SYSTEM LIPID METABOLISM TRANSPORT AND CATABOLISM
EC2.7.13.3									protein histidine kinase	0.0					Bacterial chemotaxis Bacterial motility proteins MAPK signaling pathway - yeast Protein kinases Two-component system Unclassified Vibrio cholerae pathogenic cycle	CELL MOTILITY ENZYME FAMILIES INFECTIOUS DISEASES OTHERS SIGNAL TRANSDUCTION
EC2.7.11.1									Non-specific serine/threonine protein kinase	-0.8					Acute myeloid leukemia Prometastasis junction Adipocytokine signaling pathway Aldosterone regulated sodium reabsorption Apoptosis Axon guidance B cell receptor signaling pathway Bacterial secretion system Basal cell carcinoma Bladder cancer Cell cycle Cell cycle - yeast Chemokine signaling pathway Chromosome Chronic myeloid leukemia Circadian rhythm - fly Circadian rhythm - mammal Circadian rhythm - Part Colorectal cancer Cytoskeleton DNA damage sensing pathway DNA repair and recombination proteins DNA replication proteins Dilated cardiomyopathy (DCM) Endometrial cancer Epithelial cell signaling in Helicobacter pylori infection ERBB signaling pathway Fc epsilon RI signaling pathway Fc gamma R mediated phagocytosis Focal adhesion Gap junction Glioma GnRH signaling pathway Hedgehog signaling pathway Hypertrophic cardiomyopathy (HCM) Insulin signaling pathway Jak-STAT signaling pathway Leishmania infection Leukocyte transendothelial migration Long-term depression Long-term potentiation MAPK signaling pathway MAPK signaling pathway - yeast Meiosis - yeast Melanogenesis Melanoma NOD-like receptor signaling pathway Natural killer cell mediated cytotoxicity Neurotrophin signaling pathway Non-homologous end-joining Non-small cell lung cancer Oocyte meiosis PPAR signaling pathway Pancreatic cancer Parkinson's disease Pituitary tumor Proliferation Pathways of metabolism	CANCERS CELL COMMUNICATION CELL GROWTH AND DEATH CELL MOTILITY CIRCULATORY SYSTEMS CNS DNA REPLICATION ENDOCRINE SYSTEM ENVIRONMENTAL ADAPTATION ENZYME FAMILIES EXCRETORY SYSTEMS IMMUNE DISEASES MEMBRANE TRANSPORT NERVOUS SYSTEM EURODEGENERATIVE DISEASES REPLICATION AND REPAIR SIGNAL TRANSDUCTION SIGNAL TRANSDUCTION MECHANISMS SIGNALING MOLECULES AND INTERACTION SPORULATION TRANSCRIPTION TRANSPORT AND CATABOLISM
EC5.4.2.8									phosphomannomutase	0.8					Fructose 1,6-bisphosphate Fructose 6-phosphate Fructose and mannose metabolism	CARBOHYDRATE METABOLISM
EC2.7.1.11									6-phosphofructokinase	0.8					Fructose and mannose metabolism Galactose metabolism Glycolysis / Gluconeogenesis Pentose phosphate pathway	CARBOHYDRATE METABOLISM
EC1.1.1.202									1,3-propanediol dehydrogenase	0.6					Glycerolipid metabolism	LIPID METABOLISM
EC3.4.21.83									Oligopeptidase B	-0.8					Peptidases	ENZYME FAMILIES
EC5.3.3.10									5-carboxymethyl-2-hydroxymuconate delta-isomerase	-0.7					Benzoate degradation via hydroxylation Tyrosine metabolism	AMINO ACID METABOLISM XENOBIOTICS METABOLISM BIODEGRADATION AND METABOLISM
EC5.1.3.2									UDP-glucose 4-epimerase	-0.7					Amino sugar and nucleotide sugar metabolism Galactose metabolism	CARBOHYDRATE METABOLISM
EC1.5.1.20									methylene tetrahydrofolate reductase (NADPH)	0.3					Methane metabolism One carbon pool by folate	ENERGY METABOLISM METABOLISM OF COFACTORS AND VITAMINS

EC6.3.2.2	0.000496034	0.19							glutamate-cysteine ligase	Glutathione metabolism	METABOLISM OF OTHER AMINO ACIDS
EC2.7.7.19	0.000564954	0.20							polynucleotide adenylyltransferase	Chromosome RNA degradation	FOLDING, SORTING AND DEGRADATION REPLICATION AND REPAIR
EC1.1.1.261	0.000653007	22.20							glycerol-1-phosphate dehydrogenase [NAD(P)+]	Unclassified	ENERGY METABOLISM
EC3.5.4.13	0.000799743	0.06							dCTP deaminase	Pyrimidine metabolism	NUCLEOTIDE METABOLISM
EC1.6.5.3	0.000799743	1.87	0.012	0.34					NADH dehydrogenase (ubiquinone)	Alzheimer's disease Huntington's disease Oxidative phosphorylation Parkinson's disease	ENERGY METABOLISM NEURODEGENERATIVE DISEASES
EC1.6.1.2	0.001146334	0.14							NAD(P)+ transhydrogenase (AB-specific)	Nicotinate and nicotinamide metabolism	BIOSYNTHESIS OF COFACTORS AND VITAMINS
EC3.1.1.5	0.001394684	0.17							lysophospholipase	Biosynthesis of unsaturated fatty acids Glycan binding proteins Glycerophospholipid metabolism Lipid biosynthesis proteins Lysosome Unclassified	LIPID METABOLISM SIGNALING INTERACTION TRANSPORT AND CATABOLISM
EC6.4.1.2	0.001562344	2.62	9.266E-05	3.47					acetyl-CoA carboxylase	Fatty acid biosynthesis Insulin signaling pathway Propanoate metabolism Pyruvate metabolism Tetracycline biosynthesis	BIOSYNTHESIS OF SECONDARY METABOLITES CARBOHYDRATE METABOLISM ENDOCRINE SYSTEM LIPID METABOLISM
EC3.3.1.1	0.001971972	0.11							adenosylhomocysteinase	Cysteine and methionine metabolism Selenoamino acid metabolism	AMINO ACID METABOLISM METABOLISM OF OTHER AMINO ACIDS
EC3.5.4.1	0.002021115	0.18							cytosine deaminase	Arginine and proline metabolism Pyrimidine metabolism	AMINO ACID METABOLISM NUCLEOTIDE METABOLISM
EC1.8.1.9	0.002124993	0.49	0.04	1.50					thioredoxin-disulfide reductase	Pyrimidine metabolism	NUCLEOTIDE METABOLISM
EC3.2.1.35	0.002472401	2.15	5.3507E-09	0.30					alpha-L-arabinofuranosidase	Amino sugar and nucleotide sugar metabolism	CARBOHYDRATE METABOLISM
EC2.8.1.-	0.002864139	3.11							sulfurtransferase	Ubiquitin system Unclassified	FOLDING, SORTING AND DEGRADATION TRANSPORT AND METABOLISM OTHERS PROTEIN FOLDING AND ASSOCIATED PROCESSING
EC6.3.2.-	0.00299124	0.11							acid-amino acid ligase	Biosynthesis of siderophore group nonribosomal peptides Plant-pathogen interaction Transcription factors Unclassified	BIOSYNTHESIS OF POLYKETIDES AND NONRIBOSOMAL PEPTIDES ENVIRONMENTAL ADAPTATION MEMBRANE AND INTRACELLULAR STRUCTURAL MOLECULES METABOLISM OF COFACTORS AND VITAMINS OTHERS TRANSCRIPTION
EC3.6.3.17	0.004011301	1.99							monosaccharide-transporting ATPase	ABC transporters Transporters Unclassified	ENERGY METABOLISM MEMBRANE TRANSPORT
EC6.6.1.2	0.005662659	2.94	0.001	0.37					cobaltochelatase	Porphyrin and chlorophyll metabolism Unclassified	METABOLISM OF COFACTORS AND VITAMINS OTHERS
EC2.4.1.1.129	0.006105074	0.52							peptidoglycan glycosyltransferase	Chromosome Peptidoglycan biosynthesis	GLYCAN BIOSYNTHESIS AND METABOLISM REPLICATION AND REPAIR
EC3.4.16.4	0.006105074	2.02							serine-type D-Ala-D-Ala carboxypeptidase	Peptidases Peptidoglycan biosynthesis Unclassified	ENZYMES FAMILIES GLYCAN METABOLISM MEMBRANE AND INTRACELLULAR STRUCTURAL MOLECULES
EC3.5.1.28	0.006535529	2.03							N-acetylmuramoyl-L-alanine amidase	Chromosome Unclassified	MEMBRANE AND INTRACELLULAR STRUCTURAL MOLECULES REPLICATION AND REPAIR
EC3.5.3.12	0.006535529	5.46							arginine deiminase	Arginine and proline metabolism	AMINO ACID METABOLISM
EC4.3.1.12	0.006535529	17.07							ornithine cyclodeaminase	Arginine and proline metabolism	AMINO ACID METABOLISM
EC1.1.1.17	0.006535529	Inf							mammil-1-phosphate 5-dehydrogenase	Fructose and mannose metabolism	CARBOHYDRATE METABOLISM
EC3.4.11.2	0.00751702	0.43							Membrane alanyl aminopeptidase	Cellular antigens Glutathione metabolism Hematopoietic cell lineage Peptidases Renin-angiotensin system	ENDOCRINE SYSTEM ENZYMES FAMILIES IMMUNE SYSTEM METABOLISM OF OTHER AMINO ACIDS SIGNALING
EC2.3.1.157	0.007555351	0.35	0.008	2.44					glucosamine-1-phosphate N-acetyltransferase	Amino sugar and nucleotide sugar metabolism	CARBOHYDRATE METABOLISM
EC3.4.21.-	0.007555351	0.35	0.033	2.4					serine-type endopeptidase	CAM ligands Chaperones and folding catalysts Complement and coagulation cascade DNA repair and recombination proteins ECM-receptor interaction Focal adhesion Peptidases Secretion system Two-component system Unclassified	AMINO ACID METABOLISM CELL COMMUNICATION ENZYMES FAMILIES FOLDING, SORTING AND DEGRADATION IMMUNE SYSTEM MEMBRANE TRANSPORT REPLICATION AND REPAIR SIGNAL TRANSDUCTION SIGNALING MOLECULES AND INTERACTION

EC1.17.4.1						0.007555351	0.50					ribonucleoside-diphosphate reductase	Cytosolic DNA-sensing pathway DNA repair and recombination proteins Glutathione metabolism Purine metabolism Pyrimidine metabolism p53 signaling pathway	CELL GROWTH AND DEATH IMMUNE SYSTEM METABOLISM OF OTHER AMINO ACIDS NUCLEOTIDE METABOLISM REPLICATION AND REPAIR
EC3.6.4.-												Acting on acid anhydrides; involved in cellular and subcellular movement	Unknown	Unknown
EC3.4.21.102						0.007606934	0.71					C-terminal processing peptidase	Peptidases	ENZYMES FAMILIES
EC2.7.7.56						0.007606934	2.06	0.00012	0.41			tRNA nucleotidyltransferase	Unclassified	ENERGY METABOLISM METABOLISM OF OTHER AMINO ACIDS
EC1.8.99.2						0.00899828	4.39					adenylyl-sulfate reductase	Selenoamino acid metabolism Sulfur metabolism	Unknown
EC3.6.4.13						0.011447809	0.55					RNA helicase	Unknown	Unknown
EC3.6.3.5.4						0.014860603	1.86	0.025	2.9			asparagine synthase (glutamine-hydrolyzing)	Alanine, aspartate and glutamate metabolism Nitrogen metabolism Peptidases	AMINO ACID METABOLISM ENERGY METABOLISM ENZYMES FAMILIES
EC3.1.31.1						0.015565093	0.00					Micrococcal nuclease	Unclassified	REPLICATION, RECOMBINATION AND REPAIR PROTEINS
EC4.4.1.8						0.016868066	0.55					cystathionine beta-lyase	Cysteine and methionine metabolism Nitrogen metabolism Selenoamino acid metabolism Sulfur metabolism	AMINO ACID METABOLISM ENERGY METABOLISM METABOLISM OF OTHER AMINO ACIDS
EC4.2.1.11						0.016868066	1.83					phosphopyruvate hydratase	Glycolysis / Gluconeogenesis RNA degradation	CARBONHYDRATE METABOLISM
EC2.4.1.21						0.017993223	2.45					starch synthase	Starch and sucrose metabolism	CARBONHYDRATE METABOLISM
EC2.7.1.30						0.019114051	2.30					glycerol kinase	Glycerolipid metabolism PPAR signaling pathway Plant pathogen interaction	ADAPTATION LIPID METABOLISM
EC3.2.1.4						0.020401206	4.78					cellulase	Starch and sucrose metabolism	CARBONHYDRATE METABOLISM
EC3.5.4.-						0.020839153	5.55					hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	Atrazine degradation Cytosolic DNA-sensing pathway Spliceosome Unclassified	IMMUNE SYSTEM NUCLEOTIDE METABOLISM TRANSCRIPTION XENOBIOTICS BIODEGRADATION AND METABOLISM
EC1.1.1.44						0.022715761	0.40	0.04	1.85			phosphogluconate dehydrogenase (decarboxylating)	Glutathione metabolism Pentose phosphate pathway	CARBONHYDRATE METABOLISM OF OTHER AMINO ACIDS
EC5.3.1.5						0.026032239	0.19					xylase isomerase	Fructose and mannose metabolism Pentose and glucuronate interconversions	CARBONHYDRATE METABOLISM
EC1.1.1.6						0.026296031	2.67	0.010	5.31			glycerol dehydrogenase	Glycerolipid metabolism	LIPID METABOLISM
EC2.7.1.-						0.0297304	2.28					phosphotransferase activity, alcohol group as acceptor	Glycerolipid metabolism Lipopolysaccharide biosynthesis Lipopolysaccharide biosynthesis proteins Nicotinic and nicotinamide metabolism Pentose phosphate pathway Unclassified	CARBONHYDRATE METABOLISM FUNCTION UNKNOWN GLYCAN BIOSYNTHESIS AND METABOLISM LIPID METABOLISM METABOLISM OF COFACTORS AND VITAMINS SIGNAL TRANSDUCTION MECHANISMS TRANSLATION PROTEINS
EC3.1.6.1						0.0313774	3.10					arylsulfatase	Spingolipid metabolism Steroid hormone biosynthesis	LIPID METABOLISM
EC3.2.1.193						0.031447482	2.89					neoptulianase	Unclassified	CARBONHYDRATE METABOLISM
EC3.4.19.3						0.032020064	0.16	0.02	4.44			Pyroglutaryl-peptidase I	Peptidases	ENZYMES FAMILIES
EC1.2.7.4						0.032020064	Inf					carbon-monoxide dehydrogenase (ferredoxin)	Methane metabolism	ENERGY METABOLISM
EC2.3.1.169						0.032020064	Inf					CO-methylating acetyl-CoA synthase	Methane metabolism	ENERGY METABOLISM
EC3.2.1.22						0.032518282	1.47					alpha-galactosidase	Galactose metabolism Glycerolipid metabolism Glycosphingolipid biosynthesis - globoseries Lysosome Sphingolipid metabolism	CARBONHYDRATE METABOLISM GLYCAN BIOSYNTHESIS AND METABOLISM LIPID METABOLISM TRANSPORT AND CATABOLISM
EC2.1.1.-						0.033119744	1.32					methyltransferase	Bacterial motility proteins Bacterial secretion system Benzoxazinoid biosynthesis Carotenoid biosynthesis Chromosome DNA repair and recombination proteins Histidine metabolism Insect hormone biosynthesis Isoquinoline alkaloid biosynthesis Naphthalene and anthracene degradation Peptidases Phenylpropanoid biosynthesis Secretion system Selenoamino acid metabolism Transcription factors Tyrosine metabolism Ubiquinone and other terpenoid-quinone biosynthesis Unclassified	AMINO ACID METABOLISM BIOSYNTHESIS OF SECONDARY METABOLITES CELL MOTILITY ENZYMES FAMILIES MEMBRANE AND INTRACELLULAR STRUCTURAL MOLECULES MEMBRANE TRANSPORT METABOLISM OF COFACTORS AND VITAMINS METABOLISM OF OTHER AMINO ACIDS REPLICATION AND REPAIR TRANSCRIPTION TRANSLATION PROTEINS XENOBIOTICS BIODEGRADATION AND METABOLISM
EC3.4.13.3						0.033119744	2.30	0.03	0.47			Xaa-His dipeptidase	Arginine and proline metabolism Glutathione metabolism Histidine metabolism Peptidases beta-Alanine metabolism	AMINO ACID METABOLISM ENZYMES FAMILIES METABOLISM OF OTHER AMINO ACIDS

EC3.5.4.12	0.033530541	4.44					dCMP deaminase	0.7	Pyrimidine metabolism Secretion system	MEMBRANE TRANSPORT NUCLEOTIDE METABOLISM
EC3.1.1.11	0.034051218	5.12	0.03	0.18			pectinesterase	0.6	Pentose and glucuronate interconversions Starch and sucrose metabolism	CARBOHYDRATE METABOLISM
EC5.1.3.15	0.035526158	0.09					glucose-6-phosphate 1-epimerase	-0.8	Glycolysis / Gluconeogenesis	CARBOHYDRATE METABOLISM
EC2.5.1.48	0.035526158	0.28					cystathionine gamma-synthase	-0.8	Cysteine and methionine metabolism Selenoamino acid metabolism Sulfur metabolism	AMINO ACID METABOLISM ENERGY METABOLISM METABOLISM OF OTHER AMINO ACIDS
EC3.2.1.25	0.035526158	0.38					beta-mannosidase	0.1	Lysosome Other glycan degradation	GLYCAN BIOSYNTHESIS AND METABOLISM TRANSPORT AND CATABOLISM
EC1.4.1.1	0.035526158	7.68					alanine dehydrogenase	0.5	Alanine, aspartate and glutamate metabolism Reductive carboxylate cycle (CO2 fixation) Taurine and hypotaurine metabolism	AMINO ACID METABOLISM ENERGY METABOLISM METABOLISM OF OTHER AMINO ACIDS
EC4.1.1.11	0.035526158	7.68					aspartate 1-decarboxylase	0.8	Pantothenate and CoA biosynthesis beta-Alanine metabolism	METABOLISM OF COFACTORS AND VITAMINS METABOLISM OF OTHER AMINO ACIDS
EC2.7.1.71	0.038795465	0.37					shikimate kinase	-0.8	Phenylalanine, tyrosine and tryptophan biosynthesis metabolism CS-branched dibasic acid biosynthesis Panicoic acid and CoA biosynthesis Valine, leucine and isoleucine biosynthesis	AMINO ACID METABOLISM
EC2.2.1.6	0.038988108	0.63					acetolactate synthase	0.0	Shikimate cycle (TCA cycle) Glycine, serine and threonine metabolism Glyoxylate / Gluconate metabolism Pyruvate, metabolite Valine, leucine and isoleucine degradation	AMINO ACID METABOLISM CARBOHYDRATE METABOLISM METABOLISM OF COFACTORS AND VITAMINS
EC1.8.1.4	0.039034463	0.36					dihydropyridyl dehydrogenase	-0.7	Starch and sucrose metabolism	AMINO ACID METABOLISM CARBOHYDRATE METABOLISM
EC2.4.1.25	0.039441524	0.66					4-alpha-glucanotransferase	-0.7	Starch and sucrose metabolism	CARBOHYDRATE METABOLISM
EC6.1.1.16	0.041708804	0.53	0.2				cysteine-tRNA ligase	0.2	Aminoacyl-tRNA biosynthesis	TRANSLATION
EC2.8.1.6	0.041708804	2.39					biotin synthase	0.8	Biotin metabolism	METABOLISM OF COFACTORS AND VITAMINS
EC3.2.1.85	0.042084693	3.98					6-phospho-beta-galactosidase	-0.1	Galactose metabolism	CARBOHYDRATE METABOLISM
EC2.4.1.-	0.043984691	1.52					transferase activity, transferring hexosyl groups	0.8	Anthocyanin biosynthesis Butirosin and neomycin biosynthesis Carotenoid biosynthesis Cellular antigens Flavonoid biosynthesis Fructose and mannose metabolism Glycerolipid metabolism Glycosaminoglycan biosynthesis - keratan sulfate Glycosphingolipid biosynthesis - globoseries Glycosphingolipid biosynthesis - lacto and neolacto series Glycosylphosphatidylinositol (GPI)-anchor biosynthesis Glycosyltransferases High-mannose type N-glycan biosynthesis Lipopolysaccharide biosynthesis Lipopolymer biosynthesis Membrane and intracellular structural molecules N-Glycan biosynthesis O-Glycan biosynthesis O-Mannosyl glycan biosynthesis Peptidoglycan biosynthesis Sphingolipid metabolism Stibenoil, diarylheptanoid and gingerol biosynthesis Unclassified Zeaxin biosynthesis	BIO SYNTHESIS OF SECONDARY METABOLITES CARBOHYDRATE METABOLISM CELLULAR PROCESSES AND SIGNALING GLYCAN BIOSYNTHESIS AND METABOLISM LIPID METABOLISM MEMBRANE AND INTRACELLULAR STRUCTURAL MOLECULES SIGNALLING MOLECULES AND INTERACTION
EC5.3.1.8	0.047536501	2.56					mannose-6-phosphate isomerase	0.8	Amino sugar and nucleotide sugar metabolism Fructose and mannose metabolism	CARBOHYDRATE METABOLISM
EC2.4.1.5	1.5413E-12	4.7	1.1882E-19	28.38			dextranucrase	-0.5	Starch and sucrose metabolism Two-component system	CARBOHYDRATE METABOLISM SIGNAL TRANSDUCTION
EC2.4.1.12	0.00048933	12.1					cellulose synthase (UDP-forming)	0.0	Glycosyltransferases Starch and sucrose metabolism	CARBOHYDRATE METABOLISM GLYCAN BIOSYNTHESIS AND METABOLISM
EC3.2.2.1.23	0.00312345	0.7	0.03	0.81			beta-galactosidase	-0.1	Galactose metabolism Glycosaminoglycan degradation Glycosphingolipid biosynthesis - ganglioseries Lysosome Other glycan degradation Sphingolipid metabolism	CARBOHYDRATE METABOLISM GLYCAN BIOSYNTHESIS AND METABOLISM LIPID METABOLISM TRANSPORT AND CATABOLISM
EC4.1.2.22	0.00312345	0.3	0.00045	3.18			fructose-6-phosphate phosphoketolase	-0.7	Carbon fixation in photosynthetic organisms	ENERGY METABOLISM
EC5.4.99.9	0.00794126	0.5	0.00013	2.16			UDP-galactopyranose mutase	-0.6	Unclassified	MEMBRANE AND INTRACELLULAR STRUCTURAL MOLECULES
EC2.7.2.72	0.01653364	6.5					CCA tRNA nucleotidyltransferase	0.5	Unknown	Unknown
EC3.1.4.-	0.01653364	6.5					phosphoric diester hydrolase	0.5	DNA repair and recombination proteins Ether lipid metabolism Spleucosomes Unclassified	LIPID METABOLISM OTHERS REPLICATION AND REPAIR TRANSCRIPTION
EC3.2.1.11	0.02160746	14.1					dextranase	-0.3	Unclassified	OTHERS
EC3.1.4.16	0.02181532	3.0					2',3'-cyclic-nucleotide 2'-phosphodiesterase	-0.6	Purine metabolism Pyrimidine metabolism	NUCLEOTIDE METABOLISM
EC3.1.1.31	0.03438155	0.2					6-phosphogluconolactonase	-0.8	Pentose phosphate pathway	CARBOHYDRATE METABOLISM



EC5.1.1.3-																						BIOSYNTHESIS OF POLYKETIDES AND NONRIBOSOMAL PEPTIDES CARBOHYDRATE METABOLISM FUNCTION UNKNOWN/OTHERS
EC6.3.4.14																						amino sugar and nucleotide sugar metabolism corbate and aldinate metabolism acyl-sugar unit biosynthesis Unclassified
EC2.1.1.37																						Fatty acid biosynthesis
EC2.4.1.1																						Chromosome Cysteine and methionine metabolism DNA replication proteins
EC3.1.21.3																						Insulin signaling pathway Starch and sucrose metabolism
EC1.4.1.13																						RESTRICTION ENZYME
EC1.4.1.14																						Alanine, aspartate and glutamate metabolism Nitrogen metabolism
EC1.6.5.-																						Alanine, aspartate and glutamate metabolism Nitrogen metabolism
EC3.4.14.4																						Unclassified
EC3.2.1.31																						Enzyme families
EC6.1.1.23																						Drug metabolism - other enzymes Flavone and flavanol biosynthesis Glycosaminoglycan degradation Lysosome Pentose and glucuronate interconversions Porphyrin and chlorophyll metabolism Starch and sucrose metabolism
EC1.4.4.2																						beta-glucuronidase
EC2.7.2.7																						aspartate-tRNA(Asn) ligase
EC6.2.1.26																						Glycine, serine and threonine metabolism
EC4.1.3.6																						Butyrate metabolism
EC2.1.1.131																						Ubiquinone and other terpenoid-quinone biosynthesis
EC6.3.1.-																						Citrate cycle (TCA cycle) Two-component system
EC6.3.5.9																						Porphyrin and chlorophyll metabolism
EC1.8.1.8																						Porphyrin and chlorophyll metabolism
EC2.4.1.230																						Chaperones and folding catalysts Unclassified
EC2.8.3.10																						Unclassified
EC6.3.1.10																						Porphyrin and chlorophyll metabolism
EC2.3.1.47																						Biotin metabolism
EC5.4.99.2																						Propionate metabolism Valine, leucine and isoleucine degradation
EC5.3.1.27																						Methane metabolism Pentose and glucuronate interconversions
EC2.7.7.43																						Carbohydrate metabolism
EC1.11.1.9																						Carbohydrate metabolism
EC2.2.1.9																						Arachidonic acid metabolism Glutathione metabolism
EC3.2.1.65																						Ubiquinone and other terpenoid-quinone biosynthesis
EC4.2.1.47																						Starch and sucrose metabolism
EC5.3.1.12																						Amino sugar and nucleotide sugar metabolism Glucose and fructose interconversions
EC5.3.1.25																						Fructose and mannose metabolism
EC3.5.1.16																						Arginine and proline metabolism
EC2.7.4.16																						Thiamine metabolism
EC6.1.1.18																						Aminooacyl-tRNA biosynthesis
EC2.1.3.11																						Unclassified
EC1.1.5.8																						Pentose and glucuronate interconversions
EC4.1.5.3																						Glycerophospholipid metabolism
EC4.1.1.19																						Arginine and proline metabolism
EC5.1.7																						Lysine biosynthesis Lysine degradation
EC4.1.2.4																						Pentose phosphate pathway

EC2.7.7.22									amino sugar and nucleotide sugar metabolism Fructose and mannose metabolism	CARBOHYDRATE METABOLISM
EC3.2.1.99	mannose-1-phosphate guanylyltransferase (GDP)	0.7		2.3547E-05	0.29				Unclassified	CARBOHYDRATE METABOLISM
	arabinan endo-1,5-alpha-L-arabinosidase	0.5		0.03	0.31				1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) degradation 1- and 2-Methylnaphthalene degradation 3-Chloroacrylic acid degradation Benzoate degradation via hydroxylation Biphenyl degradation Butirosin and neomycin biosynthesis Fluorene degradation Phenylalanine metabolism Phenylpropanoid biosynthesis Purine metabolism Pyruvate metabolism Tyrosine metabolism Ubiquinone and other terpenoid-quinone biosynthesis	AMINO ACID METABOLISM BIOSYNTHESIS OF SECONDARY METABOLITES ENERGY METABOLISM BIOSYNTHESIS OF COFACTORS AND VITAMINS
EC4.1.1.1-	carboxyl-lyase	0.7		0.01	0.32					AMINO ACID METABOLISM BIOSYNTHESIS OF SECONDARY METABOLITES ENERGY METABOLISM BIOSYNTHESIS OF COFACTORS AND VITAMINS
										AMINO ACID METABOLISM BIOSYNTHESIS OF SECONDARY METABOLITES ENERGY METABOLISM BIOSYNTHESIS OF COFACTORS AND VITAMINS
EC2.5.1.1-	transferase activity, transferring alkyl or aryl (other than methyl) groups	0.6		0.04	0.32				Carotenoid biosynthesis Cysteine and methionine metabolism Oxidative phosphorylation Porphyrin and chlorophyll metabolism Riboflavin metabolism Terpenoid backbone biosynthesis Ubiquinone and other terpenoid-quinone biosynthesis Unclassified	AMINO ACID METABOLISM BIOSYNTHESIS OF SECONDARY METABOLITES ENERGY METABOLISM BIOSYNTHESIS OF COFACTORS AND VITAMINS
EC2.7.1.45	2-dehydro-3-deoxyglucosylkinase	0.8		0.00039	0.33				Pentose and glucuronate interconversions Pentose phosphate pathway	CARBOHYDRATE METABOLISM
EC6.3.1.1	aspartate ammonia lyase	0.6		0.03	0.35				Alaspartate and glutamate metabolism Cyanosulfonamide acid metabolism Nitrogen metabolism	AMINO ACID METABOLISM ENERGY METABOLISM METABOLISM OF OTHER AMINO ACIDS
EC1.1.1.22	UDP-glucose 6-dehydrogenase	0.7		0.00080	0.36				Amino sugar and nucleotide sugar metabolism Ascorbate and aldarate metabolism Pentose and glucuronate interconversions Starch and sucrose metabolism	CARBOHYDRATE METABOLISM
EC4.2.1.75	uroporphyrinogen-III synthase	0.7		0.04	0.36				Porphyrin and chlorophyll metabolism	METABOLISM OF COFACTORS AND VITAMINS
EC3.5.4.25	GTP cyclohydrolase II	-0.1		0.03	0.38				Riboflavin metabolism	METABOLISM OF COFACTORS AND VITAMINS
EC4.1.99.12	3,4-dihydroxy-2-butanone-4-phosphate synthase	0.0		0.03	0.38				Riboflavin metabolism	METABOLISM OF COFACTORS AND VITAMINS
EC1.2.7.3	2-oxoglutarate synthase	0.8		0.007	0.39				Citrate cycle (TCA cycle) Reductive carboxylate cycle (CO2 fixation)	CARBOHYDRATE METABOLISM ENERGY METABOLISM
EC1.10.3.-	oxidoreductase activity, acting on diphenols and related substances	0.4		0.04	0.40				Oxidative phosphorylation	ENERGY METABOLISM
EC1.4.7.1	glutamate synthase (ferredoxin)	0.7		0.008	0.40				Nitrogen metabolism	ENERGY METABOLISM
EC3.4.15.5	peptidyl-dipeptidase Bcp	0.7		0.02	0.41				Peptidases	ENZYMIC FAMILIES
EC3.4.11.9	Xaa-Pro aminopeptidase	0.5		0.00019	0.41				Peptidases	ENZYMIC FAMILIES
EC2.1.1.13	methionine synthase	0.8		0.004	0.42				Cysteine and methionine metabolism One carbon pool by folate	AMINO ACID METABOLISM METABOLISM OF COFACTORS AND VITAMINS
EC3.4.21.53	Endopeptidase La	0.6		0.005	0.42				Peptidases	ENZYMIC FAMILIES
EC2.6.1.83	L,L-diaminopimelate aminotransferase	0.8		0.02	0.42				Lysine biosynthesis	AMINO ACID METABOLISM
EC4.2.1.2	fumarate hydratase	0.7		0.02	0.43				Citrate cycle (TCA cycle) Pathways in cancer Reductive carboxylate cycle (CO2 fixation) Renal cell carcinoma	CANCERS CARBOHYDRATE METABOLISM ENERGY METABOLISM
EC3.1.5.1	dGTPase	-0.1		0.04	0.43				Purine metabolism	NUCLEOTIDE METABOLISM
EC1.7.2.-	oxidoreductase activity, acting on other nitrogenous compounds, as	0.8		0.04	0.43				Unclassified	ENERGY METABOLISM OTHERS
EC2.1.1.72	site-specific DNA-methyltransferase (adenine-specific)	0.9		1.5026E-06	0.48				DNA repair and recombination proteins DNA replication proteins Mismatch repair Unclassified	REPLICATION AND REPAIR PROLIFERATION AND CELL DIVISION ENZYMIC TRANSFERASES PROTEIN METABOLISM ENERGY METABOLISM
EC4.3.1.1	aspartate ammonia-lyase	-0.8		0.04	0.48				Alanine aspartate and glutamate metabolism Nitrogen metabolism	AMINO ACID METABOLISM ENERGY METABOLISM
EC1.1.1.267	1-deoxy-D-xylulose-5-phosphate reductoisomerase	-0.1		0.03	0.51				Terpenoid backbone biosynthesis	BIOSYNTHESIS OF SECONDARY METABOLITES
EC3.5.99.6	glucosamine-6-phosphate deaminase	0.3		0.005	0.52				Amino sugar and nucleotide sugar metabolism Unclassified	CARBOHYDRATE METABOLISM
EC2.7.7.24	glucose-1-phosphate thymidyltransferase	0.8		0.03	0.52				Polyketide sugar unit biosynthesis Streptomycin biosynthesis	BIOSYNTHESIS OF POLYKETIDES AND NONRIBOSOMAL PEPTIDES BIOSYNTHESIS OF SECONDARY METABOLITES
EC3.1.3.11	fructose 1,6-bisphosphate 1-phosphatase	0.7		0.05	0.53				Carbon fixation in photosynthetic organisms Fructose and mannose metabolism Glycolysis / Gluconeogenesis Insulin signaling pathway Pentose phosphate pathway	CARBOHYDRATE METABOLISM ENDOCRINE SYSTEM ENERGY METABOLISM
EC5.1.3.13	dTDP-4-dehydrothiamose 3,5-epimerase	0.7		0.04	0.56				Polyketide sugar unit biosynthesis Streptomycin biosynthesis	BIOSYNTHESIS OF POLYKETIDES AND NONRIBOSOMAL PEPTIDES BIOSYNTHESIS OF SECONDARY METABOLITES

EC3.2.1.20	0.001	0.57			0.6	alpha-glucosidase	Galactase metabolism Lysosome Starch and sucrose metabolism	CARBOHYDRATE METABOLISM TRANSPORT AND CATABOLISM
EC5.99.1.-	0.05	0.64			0.7	NA	Chromosome DNA replication proteins	REPLICATION AND REPAIR
EC3.2.1.51	0.03	0.64			-0.7	alpha-L-fucosidase	Other glycan degradation	GLYCAN BIOSYNTHESIS AND METABOLISM
EC5.2.1.8	0.02	0.66			0.6	peptidyl-prolyl ds-trans isomerase	Calcium signaling pathway Chaperones and folding catalysis Parkinson's disease Parkinson's disease RC3-like receptor signaling pathway Proteasome Ubiquitin mediated proteolysis Ubiquitin system Unclassified	FOLDING, SORTING AND DEGRADATION IMMUNE SYSTEM NEURODEGENERATIVE DISEASES PROTEIN FOLDING AND ASSOCIATED PROCESSING SIGNAL TRANSDUCTION TRANSCRIPTION
EC5.99.1.2	0.05	0.76			0.8	DNA topoisomerase type I	DNA repair and recombination proteins DNA replication proteins Homologous recombination	REPLICATION AND REPAIR
EC2.3.1.54	0.05	1.45			0.2	formate C-acetyltransferase	Butanoate metabolism Propanoate metabolism Pyruvate metabolism	CARBOHYDRATE METABOLISM
EC3.6.3.2	0.02	1.65			-0.1	magnesium-importing ATPase	Unclassified	ENERGY METABOLISM
EC1.1.1.1	0.002	1.72			-0.7	alcohol dehydrogenase (NAD)	1- and 2-Methylxanthine degradation 3-Chloroacrylic acid degradation Drug metabolism - Cytochrome P450 Fatty acid metabolism Glycine, serine and threonine metabolism Glycolysis / Gluconeogenesis Metabolism of xenobiotics by Cytochrome P450 Methane metabolism Retinol metabolism Tyrosine metabolism	METABOLISM CARBOHYDRATE METABOLISM ENERGY METABOLISM LIPID METABOLISM METABOLISM IMMUNE COFACTORS AND VITAMINS XENOBIOTICS BIODEGRADATION AND METABOLISM
EC3.2.1.10	0.003	1.81			-0.5	oligo-1,6-glucosidase	Starch and sucrose metabolism	CARBOHYDRATE METABOLISM
EC1.2.1.10	0.001	1.92			-0.7	acetaldehyde dehydrogenase (acetylating)	Butanoate metabolism Pyruvate metabolism	CARBOHYDRATE METABOLISM
EC5.4.2.10	0.03	1.93			-0.6	phosphoglucoamine mutase	Amino sugar and nucleotide sugar metabolism	CARBOHYDRATE METABOLISM
EC2.7.7.23	0.04	1.97			-0.8	UDP-N-acetylglucosamine 6-phosphorylase	Amino sugar and nucleotide sugar metabolism	CARBOHYDRATE METABOLISM
EC2.1.3.3	0.04	2.12			-0.3	ornithine carbamoyltransferase	Arginine and proline metabolism	AMINO ACID METABOLISM
EC6.3.4.18	0.03	2.66			-0.6	5-(Carboxyamino)imidazole ribonucleotide synthase	Purine metabolism	NUCLEOTIDE METABOLISM
EC2.7.1.35	0.001	3.34			-0.5	pyridoxal kinase	Vitamin B6 metabolism	METABOLISM OF COFACTORS AND VITAMINS
EC3.4.14.11	0.02	3.86			-0.4	Xaa-Pro dipeptidyl-peptidase	Peptidases	ENZYMES FAMILIES
EC6.1.1.24	0.10	4.18			0.0	glutamate-RNA (Gln) ligase	Aminoacyl-tRNA biosynthesis	NUCLEOTIDE METABOLISM
EC3.6.1.11	0.04	4.34			0.4	ecsoptophosphatase	Purine metabolism	NUCLEOTIDE METABOLISM
EC4.1.4.40	0.04	5.15			0.0	tartrate-bisphosphate aldolase	Galactose metabolism	CARBOHYDRATE METABOLISM
EC2.4.1.52	0.04	6.76			-0.4	poly(glycerol-phosphate) alpha-glucosyltransferase	Glycosyltransferases	GLYCAN BIOSYNTHESIS AND METABOLISM
EC3.5.1.5	5.1201E-09	7.09			-0.8	urease	Arginine and proline metabolism Atrazine degradation Epithelial cell signaling in Helicobacter pylori infection Purine metabolism	AMINO ACID METABOLISM METABOLISM INFECTION DISEASES NUCLEOTIDE METABOLISM XENOBIOTICS BIODEGRADATION AND METABOLISM
EC5.3.3.2	0.02	7.72			-0.2	sopenteryl-alpha-phosphate delta-isomerase	Terpenoid backbone biosynthesis	BIOSYNTHESIS OF SECONDARY METABOLITES
EC3.4.13.9	0.00027	9.01			0.2	Xaa-Pro dipeptidase	Peptidases	ENZYMES FAMILIES
EC3.5.2.10	0.04	11.59			-0.1	creatinease	Arginine and proline metabolism	AMINO ACID METABOLISM
EC1.6.4.-	0.02	13.52			-0.5	NA	Unclassified	PROTEIN FOLDING AND ASSOCIATED PROCESSING
EC1.1.1.88	0.006	15.45			-0.2	hydroxymethylglutaryl-CoA reductase	Terpenoid backbone biosynthesis	BIOSYNTHESIS OF SECONDARY METABOLITES
EC6.1.1.13	0.001	18.35			-0.2	D-alanine-poly(phosphoribitol) ligase	D-Alanine metabolism	METABOLISM OF OTHER AMINO ACIDS
EC2.3.2.3	0.00051	20.28			-0.3	lysyltransferase	Unknown	Unknown
EC3.4.21.96	0.00013	Inf			0.1	Lactocypin	Chaperones and folding catalysts Peptidases	ENZYMES FAMILIES FOLDING, SORTING AND DEGRADATION
EC2.4.1.8	0.00022	Inf			0.4	maltose phosphorylase	Starch and sucrose metabolism	CARBOHYDRATE METABOLISM
EC2.3.3.10	0.001	Inf			-0.4	hydroxymethylglutaryl-CoA synthase	Butanoate metabolism Synthesis and degradation of terpenoid backbone biosynthesis Valine, leucine and isoleucine degradation	AMINO ACID METABOLISM BIOSYNTHESIS OF SECONDARY METABOLITES CARBOHYDRATE METABOLISM LIPID METABOLISM
EC4.1.2.9	0.003	Inf			0.5	phosphoketolase	Carbon fixation in photosynthetic organisms Methane metabolism Pentose phosphate pathway	CARBOHYDRATE METABOLISM ENERGY METABOLISM

**Table 9. Significantly different ECs between healthy and malnourished twins k138 before and at the presentation with kwashiorkor, 2 weeks into RUTF treatment and 1 month after cessation of RUTF.**

Fisher's exact test was used to identify the significance of ECs. P values adjusted with false discovery rate (FDR), less than 0.05, and Odds ratios are indicated for each comparison. EC was overrepresented in the healthy twin microbiome if odds ratio > 1. Spearman coefficients from the analysis of changes of ECs with age in healthy individuals are indicated: a positive correlation coefficient indicates an increase in proportional representation of the EC with increasing age (colored in grey), while a negative value indicates a decrease with increasing age (colored in yellow).

EC	1.6 months before kwashiorkor		At the presentation with kwashiorkor		2 weeks into RUTF		1 month after cessation of RUTF		EC annotation	Pathway	Category
	p value (adjusted with FDR)	Odds ratio	p value (adjusted with FDR)	Odds ratio	p value (adjusted with FDR)	Odds ratio	p value (adjusted with FDR)	Odds ratio			
EC2.3.1.-	1.0873E-06	0.4	1.26627E-21	0.22	0.02	1.39			transferase activity, transferring acyl groups other than amino-acyl metabolism Glycerophospholipid metabolism Glycosphingolipid biosynthesis - ganglio series Limonene and pinene degradation Lipid biosynthesis proteins Lipopysaccharide biosynthesis Lipopolysaccharide biosynthesis Protein tyrosine phosphorylation Protein tyrosine phosphorylation Unclassified	AMINO ACID METABOLISM BIOSYNTHESIS OF POLYKETIDES AND NONRIBOSOMAL PEPTIDES BIOSYNTHESIS OF SECONDARY METABOLITES CARBOHYDRATE METABOLISM GENERAL FUNCTION PREDICTION ONLY GLYCAN BIOSYNTHESIS AND METABOLISM LIPID METABOLISM MEMBRANE AND INTRACELLULAR STRUCTURAL MOLECULES MEMBRANE TRANSPORT OTHERS PROTEIN OCCURRENCE ASSOCIATED PROCESSES REGULATION AND REPAIR TRANSLATION PROTEIN YENOROTICS	
EC3.2.1.24	3.1734E-05	0.3	4.05065E-06	0.24				-0.8	lysosome Other glycan degradation	GLYCAN BIOSYNTHESIS AND METABOLISM TRANSPORT AND CATABOLISM	
EC3.1.1.21	0.00016128	0.6	0.031447482	0.77	0.03	0.82		0.5	Cyanoamino acid metabolism Phenylpropanoid biosynthesis Starch and sucrose metabolism	BIOSYNTHESIS OF SECONDARY METABOLITES CARBOHYDRATE METABOLISM METABOLISM OF OTHER AMINO ACIDS	
EC2.3.2.-	0.00164878	2.4	0.001242056	0.24	9.9534E-06	2.59		-0.8	Butirosin and neomycin biosynthesis Peptidoglycan biosynthesis	BIOSYNTHESIS OF SECONDARY METABOLITES GLYCAN BIOSYNTHESIS AND METABOLISM	
EC2.7.1.69	0.00516443	1.4	1.16244E-12	1.77	1.1882E-19	2.18	0.005	0.5	Amino sugar and nucleotide sugar metabolism Ascorbate and alderate metabolism Deleted Fructose and mannose metabolism Galactose metabolism Glycolysis / Gluconeogenesis Phosphotransferase system (PTS) Starch and sucrose metabolism Purine metabolism	CARBOHYDRATE METABOLISM MEMBRANE TRANSPORT UNPROCESSED	
EC3.2.2.1	0.02181532	0.4	0.003521534	0.30				-0.8	purine nucleosidase	METABOLISM OF COFACTORS AND VITAMINS NUCLEOTIDE METABOLISM	
EC2.7.7.49			1.38798E-20	5.10				0.6	DNA replication proteins Unclassified	REPLICATION AND REPAIR PROTEINS	
EC3.1.26.12			6.73006E-10	0.07	0.010	2.18		-0.8	RNA degradation	FOLDING, SORTING AND DEGRADATION	
EC3.1.3.1			3.4626E-09	0.05				-0.7	Folate biosynthesis Two-component system gamma-Peactinon Cyclodextrane degradation	METABOLISM OF COFACTORS AND VITAMINS SIGNALLING PHYSIOLOGICS	
EC3.2.1.58			3.10845E-08	0.00				-0.8	Starch and sucrose metabolism	BIOSYNTHESIS OF SECONDARY METABOLITES MEMBRANE TRANSPORT UNPROCESSED	
EC2.7.1.59			1.84964E-07	0.08				-0.7	Two-component system	BIOSYNTHESIS OF SECONDARY METABOLITES MEMBRANE TRANSPORT UNPROCESSED	
EC3.2.1.86			2.06156E-06	2.93	1.5026E-06	3.33		-0.6	Glycolysis / Gluconeogenesis	CARBOHYDRATE METABOLISM	
EC3.5.3.16			4.7524E-06	Inf				0.5	Arginine and proline metabolism	CARBOHYDRATE METABOLISM	
EC3.2.1.122			7.65035E-06	32.45				0.4	Starch and sucrose metabolism	AMINO ACID METABOLISM	
EC2.7.7.42			9.86661E-06	0.12	0.02	2.08		-0.8	Unclassified	PROTEIN FOLDING AND ASSOCIATED PROCESSING	
EC4.1.1.31			1.46966E-05	0.26	0.009	1.89		-0.8	Carbon fixation in photosynthetic organisms Pyruvate metabolism Reductive carboxylate cycle (CO2 fixation)	CARBOHYDRATE METABOLISM	
EC1.1.1.14			2.64162E-05	9.82				0.4	L-iditol 2-dehydrogenase	METABOLISM ENERGY METABOLISM	
EC3.1.11.5			2.91452E-05	2.85				0.7	Fructose and mannose metabolism	CARBOHYDRATE METABOLISM	
EC3.4.22.40			4.94209E-05	0.37				-0.8	DNA repair and recombination proteins Homologous recombination Unclassified	REPLICATION AND REPAIR PROTEINS	
									Peptidases	ENZYMATIC FAMILIES	

**Tables 10 11.**

**Please reference provided CD for these tables.**

## **Chapter 5**

### **Conclusions and Future Directions**

Together, the studies I have carried out during my thesis work demonstrate that while the organismal and functional composition of the gut microbiota is highly variable across humans, common patterns of assembly of this microbial ‘organ’ starting at birth can be deciphered. My work emphasizes the importance of further understanding the complex and dynamic interrelationship between diet and nutritional status and microbiome function, both in infants and children when microbial community assembly is occurring and in adults. My studies illustrate the importance of understanding the influence of varying cultural traditions and geography on our human microbiomes. The findings reported in my thesis provide a starting point for follow-up testing of a variety of the hypotheses presented in Chapters 2-4.

#### **Enhancing the nutritional value of food via intra-familial probiotics**

In the Chapter 4 of my thesis, I described the composition and gene content of the microbiota in twins who are discordant for severe forms of malnutrition. Much work is still needed to elucidate the mechanisms by which the gut microbiota contributes to severe childhood malnutrition. Studies in gnotobiotic animals, where potentially confounding variables that are difficult to control in human studies could be controlled, offer an opportunity to move beyond *in silico* predictions of the functional potential of a person’s microbiome to direct tests of its functional activities, and of its contributions to host phenotypes. A number of experiments involving fecal samples obtained from twin pairs who were part of the cohort described in the Chapter 4 have already begun. The idea of these experiments is based on work initiated in our lab several years ago. Peter Turnbaugh, studying twins who were concordant for obesity or leanness, showed that it was possible to take a frozen fecal microbiota and transplant the community from a single donor to multiple recipient mice (Turnbaugh et al. 2009). Remarkably, the donor community was replicated to a large degree in the recipients; importantly, the communities established in each mouse were highly similar to one another. In essence, this procedure allowed a single human microbial com-

munity, obtained at a given moment in a person's life, to be reproduced multiple times. The resulting 'humanized' gnotobiotic mice could then be followed over time, fed various human diets, and be subjected to a variety of dietary or other perturbations under highly controlled conditions. The transplant recipient could be a wild-type mouse or a mouse with an engineered mutation thought to play a role in modulating or effecting host-microbiome interactions. A key benefit of this approach was that it provided an opportunity to directly test the degree to which a donor's physiologic or pathologic phenotype can be transferred to and thus assigned to his or her microbiota.

With these thoughts in mind, Michelle Smith, a post-doctoral fellow in the lab, has transplanted fecal microbiota from four twin pairs discordant for kwashiorkor into groups of adult C57Bl/6J germ-free mice. Transplant recipients have been given a corn-based macro- and micronutrient deficient diet typically consumed in Malawi ('Malawi diet'), followed by RUTF, followed by a return to the Malawi diet. She has found that mice receiving microbiota from co-twins with kwashiorkor experienced a significantly greater degree of weight loss than those harboring a microbiota from the healthy co-twin. RUTF only partially rescues the weight loss phenotype. The combination of a kwashiorkor microbiota and Malawi diet also disrupts mucosal barrier function leading to immune activation. As in humans, the response of a transplanted kwashiorkor microbiota to RUTF is more dramatic than that of the transplanted healthy co-twins microbiota and withdrawal of the therapeutic food is accompanied by a regression in the gut microbial community's organismal and functional gene configuration towards a pre-treatment state.

A procedure to generate taxonomically defined, clonally arrayed 'personalized culture collections' composed of bacterial isolates from a person's stool has been recently developed in the lab (Goodman et al. 2011). The method captures 99% of the order-level and over 50% of species level bacterial taxa found in the original fecal sample. Culture collections can be generated from a healthy and kwashiorkor co-twins. Transplantation of these culture collections into gnotobiotic mice fed the sequence of diets described above would



allow us to test whether the culturable component of a kwashiorkor microbiota can transfer the phenotypes transmitted by the intact uncultured community. If this is observed, then a sequence of experiments could be performed. Cultured communities from kwashiorkor and healthy co-twins could be mixed together prior to transplantation into germ-free recipients to ascertain whether a healthy community could ameliorate phenotypes transmitted by the kwashiorkor community. Alternatively, mice harboring each of the two communities could be co-housed together with a germ-free mouse to study phenotype transfer/amelioration. In yet another derivative, the ability of a mother harboring a transplanted culture collection to transfer that collection and a microbiota-associated phenotype to her offspring could be determined; if transfer does occur, then cross fostering experiments could be performed involving litters where mothers either harbor a kwashiorkor or a healthy co-twin's culture collection.

Another set of studies could focus on systematic tests of which components of a kwashiorkor or healthy co-twin's complete community or culture collection are responsive to existing or experimental therapeutic foods. Responsive taxa would be identified by feeding these foods to mice harboring a complete uncultured human microbiota from kwashiorkor or healthy co-twin donors or the corresponding culture collections and monitoring community responses by 16S rRNA profiling. The effects of removing these taxa from the culture collection before transplant on host phenotypic responses to the therapeutic foods could then be ascertained. These leave 'one or more taxa out' experiments would be followed by 'add one or more back' experiments. This type of approach would represent a preclinical pipeline for identifying new pre- and probiotics or mixtures of the two ('synbiotics'). The results could ultimately give rise to new therapeutic strategies for treating the donors.

An additional attraction of using gnotobiotic mice harboring complete communities or culture collections is that it allows virtual clinical trials to be performed using microbiota from children with or without malnutrition living in other areas of the world: mice

harboring different donor microbiota would be fed the diets of the donors or diets from other populations, and the effects of a given therapeutic food, fed for various periods of time, tested in various microbiota/native diet contexts.

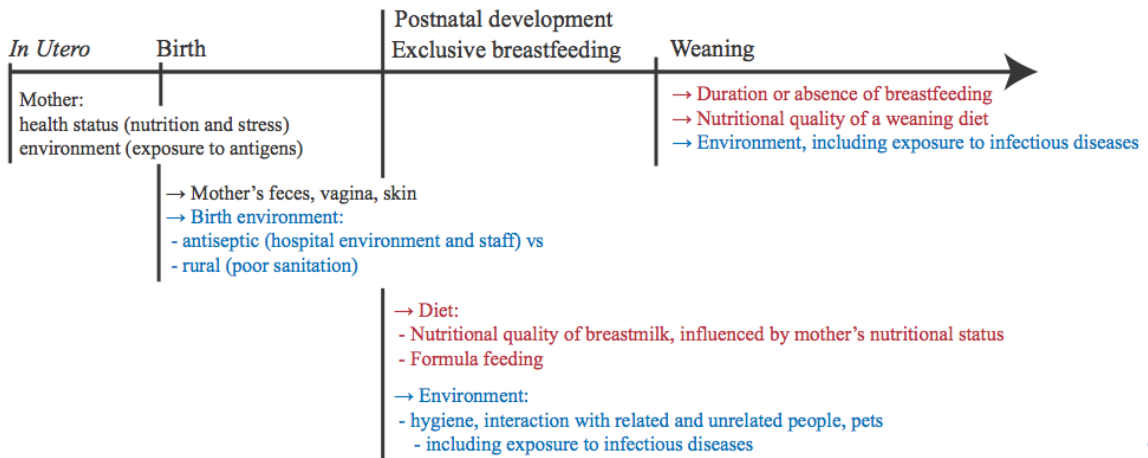
Another advantage of using sequenced personal culture collections is that it allows a custom probiotic consortium to be selected and subsequently manufactured as a therapeutic agent. As the efficiency of generating these collections increases, one source may be a healthy family member (sibling or mother who has already adapted to the local diet).

Finally, by 2100 our planet is expected to be home to 10 billion human beings (United Nations 2010). This raises great challenges for human nutrition, especially given concerns about sustainable agriculture in the face of climate change, diminishing land and water resources. One hope is that nutritional recommendations and even decisions about what crops to grow can be predicated on deeper knowledge of the consumer's microbiome and/or by increasing the nutritional value of existing foods through manipulation of a human gut microbiota. The experiments described above could be extended to other human populations, especially where the diet is monotonous and poor in quality (which in fact may simplify enrichment of the most responsive taxa). Despite the large interpersonal variation in gut microbiome configurations described in my thesis, the findings presented in Chapter 3 indicate that features of gut microbiomes distinctive to given human populations can be identified, and thus population-specific 'probiotics' designed.

### **Filling the gaps in our understanding of the assembly of the gut microbiota**

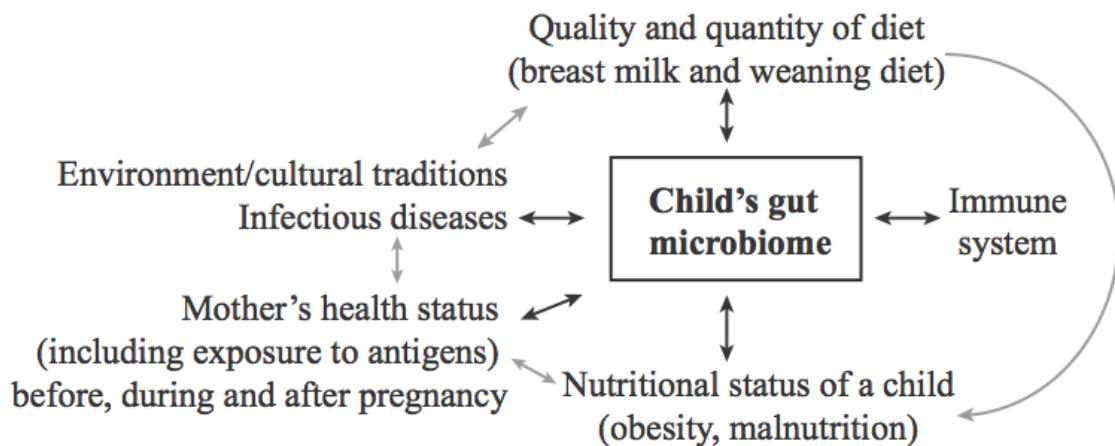
While my studies provide insights into the patterns of organismal and functional maturation of gut communities in infants across multiple human populations, our knowledge of factors that influence assembly is still incomplete. Understanding the variables that shape an infant's microbiome could lead to development of tools for manipulation of the microbiome early in life while the community is more 'flexible'. **Figure 1** represents my view of the factors that influence 'maturation' of the microbiome in children.

**Figure 1.** Influence of mother’s health status, diet and environment on the microbial colonization of an infant gut.



Remarkably, not only the nutritional status of a mother is important, but her exposure to various microbial antigens during pregnancy influences the development of her child’s immune system (Ege et al. 2006; Schaub et al. 2009). The innate and adaptive arms of the immune system influence the composition and function of the gut microbiome (**Fig. 2**). To my knowledge, there have not been extensive studies addressing the question of the role of nutritional status of a mother, including the composition of breast milk over the course of lactation, on the microbiota of her offspring.

**Figure 2.** Interaction of factors influencing ‘maturation’ of a child’s gut microbiome.



## **Initial inoculum**

During birth, microbes are thought to be transmitted primarily from the mother. Role of other variables, such as the environment in which birth and subsequent perinatal exposures have occurred (e.g., who handles the child in different cultural traditions, the presence or absence of household pets) has been difficult to study systematically. In addition, the importance of initial versus later microbial exposures in shaping the organismal and functional makeup of the microbiota is unclear. Although a few reports exist about the transmission of bacterial strains from a mother to a child (Mändar and Mikelsaar 1996; Dominguez-Bello et al. 2010), more work is clearly needed. Now that DNA sequencing is becoming more affordable, and computational tools for handling massive metagenomic datasets are in hand, it will be interesting to establish microbial observatory projects in which mothers are enrolled in the third trimester of pregnancy. The microbial communities of mothers would be thoroughly sampled in multiple body habitats (skin, mouth, vaginal) prior, immediately after, and at multiple intervals following delivery. Breast milk would also be sampled, not only for definition of immune and nutrient content but also for microbes. The child would be similarly sampled at daily intervals beginning at birth for the first month, as would all of his/her human contacts and immediate environment. Deep sequencing of 16S rRNA genes would be required to estimate the fraction of shared phylotypes between a mother and her child over the course of initial colonization. Currently, sequencing of 16S rRNA amplicons on the Illumina HiSeq instrument allows the most cost-effective way for monitoring diversity in high-resolution time series studies (Caporaso et al. 2011).

## **The co-development of the gut microbiota and breast milk**

The composition of breast milk varies over the course of lactation between women (Neville et al. 1984; Thurl et al. 2010). How does this variation contribute to intra- and interpersonal variations in the gut microbiome in babies is unknown. The nutritional status of a mother affects the nutritional content of her breast milk (Brenna et al. 2007; Qian et al. 2010). A

large diversity of oligosaccharides present in the human milk escape digestion in the infant small intestine, but represent an energy source for distal gut microbes. The preferences of different components of the infant gut microbiota towards these human milk oligosaccharides (HMOs) remains ill-defined. Methods for HMO purification have been and are being developed (German et al. 2008). This should allow a series of experiments to be performed using sequenced personal culture collections generated from infant gut microbiota. For example, purified oligosaccharides could supplement the diets of gnotobiotic mice colonized with culture collections (Goodman et al. 2011) created from a fecal microbiota of a breast-fed baby. Responsive taxa could be recovered from the arrayed culture collections, their genomes sequenced and their transcriptional and metabolic responses to the purified HMOs defined *in vitro*. These responsive taxa represent potential new generation probiotics. To illustrate this point further, the diversity of oligosaccharides in the breast milk of mothers living in Western versus non-Western societies has not been well characterized. If milk samples could be obtained from mothers living in Malawi and USA, comparative analyses could be performed to identify glycan species whose representation is affected by maternal diet and nutritional status. The response of members of culture collections generated from the infants of representative sampled mothers to purified differentially represented breast milk glycans could be ascertained *in vitro*, and *in vivo* using gnotobiotic mouse models harboring culture collections from these infants (and in follow up experiments from infants living in other areas of the world). Responsive taxa could then be omitted from or added to culture collections prior to their transplantation to mice and the effects of these glycan-responsive taxa on host nutritional status, microbiome and host metabolism, immune function including gut integrity, and other parameters could be defined. This workflow could represent a pipeline for generating new generation pre-, pro- and synbiotics for babies at risk for undernutrition because their mother's nutritional status is compromised.

## References

- Brenna J.T., B. Varamini, R. G. Jensen, D. A. Diersen-Schade, J.A. Boettcher, and L. M. Arterburn. Docosahexaenoic and arachidonic acid concentrations in human breast milk worldwide. *The American Journal of Clinical Nutrition* **85**, 1457-1464 (2007).
- Caporaso J.G., C.L. Lauber, E.K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, J.I. Gordon, and R. Knight. Moving pictures of the human microbiome. *Genome Biol* **12**, R50 (2011).
- Dominguez-Bello M.G., E.K. Costello, M. Contreras, M. Magris, G. Hidalgo, N. Fierer, and R. Knight. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 11971-11975 (2010).
- Ege M.J., C. Bieli, R. Frei, R. T. van Strien, J. Riedler, E. Üblagger, D. Schram-Bijkerk, B. Brunekreef, M. van Hage, A. Scheynius, G. Pershagen, M.R. Benz, R. Lauener, E. von Mutius, C. Braun-Fahrlander, and the PARSIFAL Study team. Prenatal farm exposure is related to the expression of receptors of the innate immunity and to atopic sensitization in school-age children. *Journal of Allergy and Clinical Immunology* **117**, 817-823 (2006)
- German J.B., S.L. Freeman, C.B. Lebrilla, and D.A. Mills. Human Milk Oligosaccharides: Evolution, Structures and Bioselectivity as Substrates for Intestinal Bacteria. *Nestle Nutr Workshop Ser Pediatr Program* **62**, 205-222 (2008).
- Goodman A.L., G. Kallstrom, J.J. Faith, A. Reyes, A. Moore, G. Dantas, and J.I. Gordon. Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 6252-6257 (2011).
- Mändar R., and M. Mikelsaar. Transmission of mother's microflora to the newborn at birth. *Biol. Neonate* **69**, 30-35 (1996).

- Neville M.C., R.P. Keller, J. Seacat, C.E. Casey, J.C. Allen, and P. Archer. Studies on human lactation. I. Within-feed and between-breast variation in selected components of human milk. *Am. J. Clin. Nutr.* **40**, 635-646 (1984).
- Qian J., T. Chen, W. Lu, S. Wu, and J. Zhu. Breast milk macro- and micronutrient composition in lactating mothers from suburban and urban Shanghai. *Journal of Paediatrics and Child Health* **46**, 115-120 (2010).
- Schaub B., J. Liu, S. Höppler, I. Schleich, J. Huehn, S. Olek, G. Wiczorek, S. Illi, and E. von Mutius. Maternal farm exposure modulates neonatal immune mechanisms through regulatory T cells. *Journal of Allergy and Clinical Immunology* **123**, 774-782.e5 (2009).
- Thurl S., M. Munzert, J. Henker, G. Boehm, B. Müller-Werner, J. Jelinek, and B. Stahl. Variation of human milk oligosaccharides in relation to milk groups and lactational periods. *Br. J. Nutr.* **104**, 1261-1271 (2010)
- Turnbaugh P.J., V.K. Ridaura, J. J. Faith, F.E. Rey, R. Knight, and J.I. Gordon. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci Transl Med* **1**, 6ra14 (2009)
- United Nations. World Population Prospects, the 2010 Revision.

## **Appendices**



## Appendix A

Turnbaugh PJ, Hamady M, **Yatsunenko T**, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI.

“A core gut microbiome in obese and lean twins.”

*Nature*. **2009** Jan 22; 457 (7228): 480-4.

## LETTERS

# A core gut microbiome in obese and lean twins

Peter J. Turnbaugh<sup>1</sup>, Micah Hamady<sup>3</sup>, Tanya Yatsunenkov<sup>1</sup>, Brandi L. Cantarel<sup>5</sup>, Alexis Duncan<sup>2</sup>, Ruth E. Ley<sup>1</sup>, Mitchell L. Sogin<sup>6</sup>, William J. Jones<sup>7</sup>, Bruce A. Roe<sup>8</sup>, Jason P. Affourtit<sup>9</sup>, Michael Egholm<sup>9</sup>, Bernard Henrissat<sup>5</sup>, Andrew C. Heath<sup>2</sup>, Rob Knight<sup>4</sup> & Jeffrey I. Gordon<sup>1</sup>

The human distal gut harbours a vast ensemble of microbes (the microbiota) that provide important metabolic capabilities, including the ability to extract energy from otherwise indigestible dietary polysaccharides<sup>1–6</sup>. Studies of a few unrelated, healthy adults have revealed substantial diversity in their gut communities, as measured by sequencing 16S rRNA genes<sup>6–8</sup>, yet how this diversity relates to function and to the rest of the genes in the collective genomes of the microbiota (the gut microbiome) remains obscure. Studies of lean and obese mice suggest that the gut microbiota affects energy balance by influencing the efficiency of calorie harvest from the diet, and how this harvested energy is used and stored<sup>3–5</sup>. Here we characterize the faecal microbial communities of adult female monozygotic and dizygotic twin pairs concordant for leanness or obesity, and their mothers, to address how host genotype, environmental exposure and host adiposity influence the gut microbiome. Analysis of 154 individuals yielded 9,920 near full-length and 1,937,461 partial bacterial 16S rRNA sequences, plus 2.14 gigabases from their microbiomes. The results reveal that the human gut microbiome is shared among family members, but that each person's gut microbial community varies in the specific bacterial lineages present, with a comparable degree of co-variation between adult monozygotic and dizygotic twin pairs. However, there was a wide array of shared microbial genes among sampled individuals, comprising an extensive, identifiable 'core microbiome' at the gene, rather than at the organismal lineage, level. Obesity is associated with phylum-level changes in the microbiota, reduced bacterial diversity and altered representation of bacterial genes and metabolic pathways. These results demonstrate that a diversity of organismal assemblages can nonetheless yield a core microbiome at a functional level, and that deviations from this core are associated with different physiological states (obese compared with lean).

We characterized gut microbial communities in 31 monozygotic twin pairs, 23 dizygotic twin pairs and, where available, their mothers ( $n = 46$ ) (Supplementary Tables 1–5). Monozygotic and dizygotic co-twins and parent–offspring pairs provided an attractive model for assessing the impact of genotype and shared early environmental exposures on the gut microbiome. Moreover, genetically 'identical'<sup>9</sup> monozygotic twin pairs gain weight in response to overfeeding in a more reproducible way than unrelated individuals<sup>10</sup> and are more concordant for body mass index (BMI) than dizygotic twin pairs<sup>11</sup>.

Twin pairs who had been enrolled in the Missouri Adolescent Female Twin Study (MOAFTS<sup>12</sup>) were recruited for this study (mean period of enrolment in MOAFTS,  $11.7 \pm 1.2$  years; range, 4.4–13.0 years). Twins were 21–32 years old, of European or African ancestry, and were generally concordant for obesity (BMI  $\geq 30$  kg m<sup>-2</sup>) or

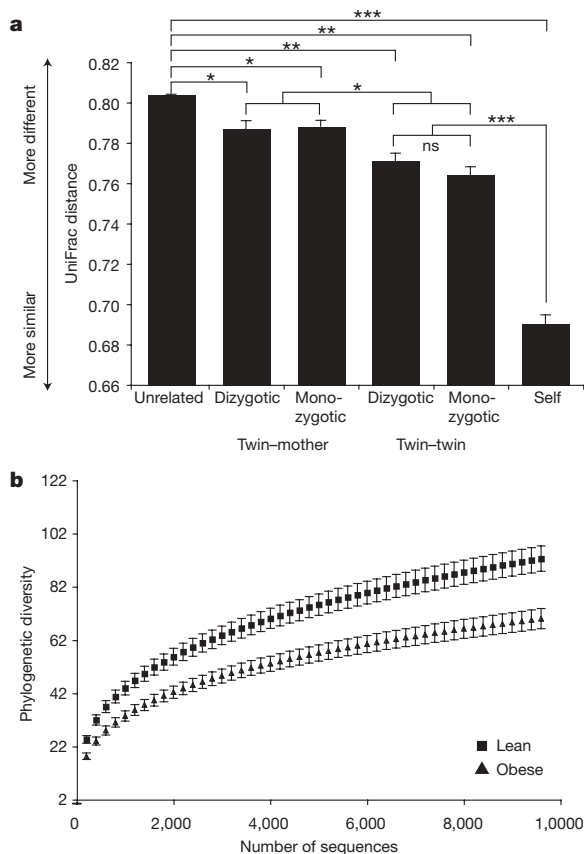
leanness (BMI = 18.5–24.9 kg m<sup>-2</sup>) (one twin pair was lean/overweight (overweight defined as BMI  $\geq 25$  and  $< 30$ ) and six pairs were overweight/obese). They had not taken antibiotics for at least  $5.49 \pm 0.09$  months. Each participant completed a detailed medical, lifestyle and dietary questionnaire: study enrollees were broadly representative of the overall Missouri population for BMI, parity, education and marital status (see Supplementary Results). Although all were born in Missouri, they currently live throughout the USA: 29% live in the same house, but some live more than 800 km apart. Because faecal samples are readily attainable and representative of interpersonal differences in gut microbial ecology<sup>7</sup>, they were collected from each individual and frozen immediately. The collection procedure was repeated again with an average interval between sampling of  $57 \pm 4$  days.

To characterize the bacterial lineages present in the faecal microbiotas of these 154 individuals, we performed 16S rRNA sequencing, targeting the full-length gene with an ABI 3730xl capillary sequencer. Additionally, we performed multiplex pyrosequencing with a 454 FLX instrument to survey the gene's V2 variable region<sup>13</sup> and its V6 hypervariable region<sup>14</sup> (Supplementary Tables 1–3).

Complementary phylogenetic and taxon-based methods were used to compare 16S rRNA sequences among faecal communities (see Methods). No matter which region of the gene was examined, individuals from the same family (a twin and her co-twin, or twins and their mother) had a more similar bacterial community structure than unrelated individuals (Fig. 1a and Supplementary Fig. 1a, b), and shared significantly more species-level phylotypes (16S rRNA sequences with  $\geq 97\%$  identity comprise each phylotype) ( $G = 55.2$ ,  $P < 10^{-12}$  (V2);  $G = 12.3$ ,  $P < 0.001$  (V6);  $G = 11.3$ ,  $P < 0.001$  (full-length)). No significant correlation was seen between the degree of physical separation of family members' current homes and the degree of similarity between their microbial communities (defined by UniFrac<sup>15</sup>). The observed familial similarity was not due to an indirect effect of the physiological states of obesity versus leanness; similar results were observed after stratifying twin pairs and their mothers by BMI category (concordant lean or concordant obese individuals; Supplementary Fig. 2). Surprisingly, there was no significant difference in the degree of similarity in the gut microbiotas of adult monozygotic compared with dizygotic twin pairs (Fig. 1a). However, we could not assess whether monozygotic and dizygotic twin pairs had different degrees of similarities at earlier stages of their lives.

Multiplex pyrosequencing of V2 and V6 amplicons allowed higher levels of coverage compared with what was feasible using Sanger sequencing, reaching on average  $3,984 \pm 232$  (V2) and  $24,786 \pm 1,403$  (V6) sequences per sample. To control for differences

<sup>1</sup>Center for Genome Sciences. <sup>2</sup>Department of Psychiatry, Washington University School of Medicine, St Louis, Missouri 63108, USA. <sup>3</sup>Department of Computer Science. <sup>4</sup>Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado 80309, USA. <sup>5</sup>CNRS, UMR6098, Marseille, France. <sup>6</sup>Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA. <sup>7</sup>Environmental Genomics Core Facility, University of South Carolina, Columbia, South Carolina 29208, USA. <sup>8</sup>Department of Chemistry and Biochemistry and the Advanced Center for Genome Technology, University of Oklahoma, Norman, Oklahoma 73019, USA. <sup>9</sup>454 Life Sciences, Branford, Connecticut 06405, USA.



**Figure 1 | 16S rRNA gene surveys reveal familial similarity and reduced diversity of the gut microbiota in obese individuals.** **a**, Average unweighted UniFrac distance (a measure of differences in bacterial community structure) between individuals over time (self), twin pairs, twins and their mother, and unrelated individuals (1,000 sequences per V2 data set; Student's *t*-test with Monte Carlo; \* $P < 10^{-5}$ ; \*\* $P < 10^{-14}$ ; \*\*\* $P < 10^{-41}$ ; mean  $\pm$  s.e.m.). **b**, Phylogenetic diversity curves for the microbiota of lean and obese individuals (based on 1–10,000 sequences per V6 data set; mean  $\pm$  95% confidence intervals shown).

in coverage, all analyses were performed on an equal number of randomly selected sequences (200 full-length, 1,000 V2 and 10,000 V6). At this level of coverage, there was little overlap between the sampled faecal communities. Moreover, the number of 16S rRNA gene sequences belonging to each phylotype varied greatly between faecal microbiotas (Supplementary Tables 6–8).

Because this apparent lack of overlap could reflect the level of coverage (Supplementary Tables 1–3), we subsequently searched all hosts for bacterial phylotypes present at high abundance using a sampling model based on a combination of standard Poisson and binomial sampling statistics. The analysis allowed us to conclude that no phylotype was present at more than about 0.5% abundance in all of the samples in this study (see Supplementary Results). Finally, we sub-sampled our data set by randomly selecting 50–3,000 sequences per sample; again, no phylotypes were detectable in all individuals sampled within this range of coverage (Supplementary Fig. 3).

Samples taken from the same individual at the initial collection point and  $57 \pm 4$  days later were consistent with respect to the specific phylotypes found (Supplementary Figs 4 and 5), but showed variations in relative abundance of the major gut bacterial phyla (Supplementary Fig. 6). There was no significant association between UniFrac distance and the time between sample collections. Overall, faecal samples from the same individual were much more similar to one another than samples from family members or unrelated individuals (Fig. 1a), demonstrating that short-term temporal changes in community structure within an individual are minor compared with inter-personal differences.

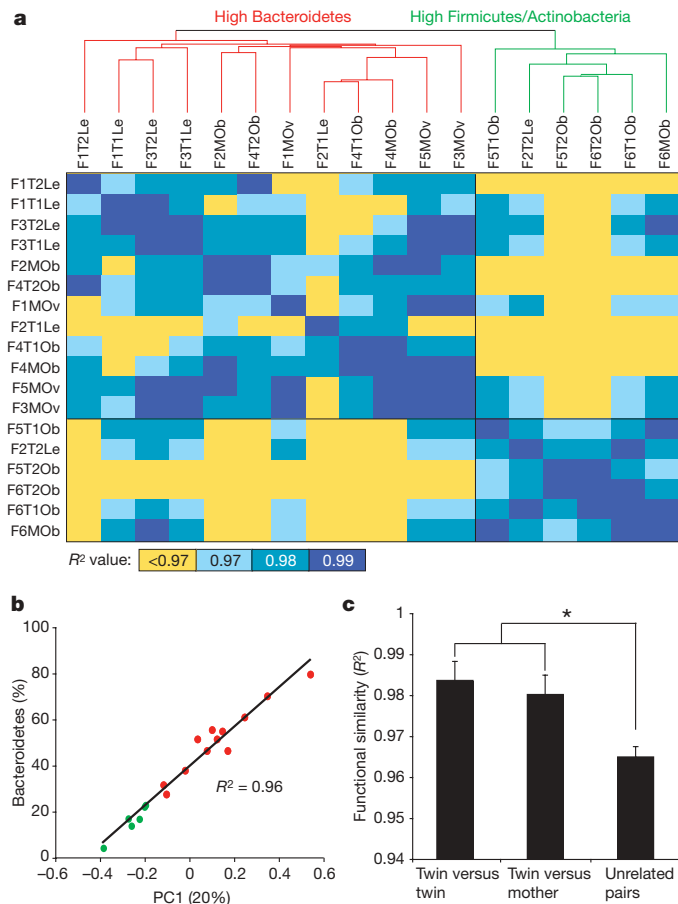
Analysis of 16S rRNA data sets produced by the three PCR-based methods, plus shotgun sequencing of community DNA (see below), revealed a lower proportion of Bacteroidetes and a higher proportion of Actinobacteria in obese compared with lean individuals of both ancestries (Supplementary Table 9). Combining the individual *P* values across these independent analyses using Fisher's method disclosed significantly fewer Bacteroidetes ( $P = 0.003$ ), more Actinobacteria ( $P = 0.002$ ) but no significant difference in Firmicutes ( $P = 0.09$ ). These findings agree with previous work showing comparable differences in both taxa in mice<sup>2</sup> and a progressive increase in the representation of Bacteroidetes when 12 unrelated, obese humans lost weight after being placed on one of two reduced-calorie diets<sup>6</sup>.

Across all methods, obesity was associated with a significant decrease in the level of diversity (Fig. 1b and Supplementary Fig. 1c–f). This reduced diversity suggests an analogy: the obese gut microbiota is not like a rainforest or reef, which are adapted to high energy flux and are highly diverse; rather, it may be more like a fertilizer runoff where a reduced-diversity microbial community blooms with abnormal energy input<sup>16</sup>.

We subsequently characterized the microbial lineage and gene content of the faecal microbiomes of 18 individuals representing six of the families (three lean and three obese European ancestry monozygotic twin pairs and their mothers) through shotgun pyrosequencing (Supplementary Tables 4 and 5) and BLASTX comparisons against several databases (KEGG<sup>17</sup> (version 44) and STRING<sup>18</sup>) plus a custom database of 44 reference human gut microbial genomes (Supplementary Figs 7–10 and Supplementary Results). Our analysis parameters were validated using control data sets comprising randomly fragmented microbial genes with annotations in the KEGG database<sup>17</sup> (Supplementary Fig. 11 and Supplementary Methods). We also tested how technical advances that produce longer reads might improve these assignments by sequencing faecal community samples from one twin pair using Titanium pyrosequencing methods (average read length of  $341 \pm 134$  nucleotides (s.d.) versus  $208 \pm 68$  nucleotides for the standard FLX method). Supplementary Fig. 12 shows that the frequency and quality of sequence assignments is improved as read length increases from 200 to 350 nucleotides.

The 18 microbiomes were searched to identify sequences matching domains from experimentally validated carbohydrate-active enzymes (CAZymes). Sequences matching 156 total CAZy families were found within at least one human gut microbiome, including 77 glycoside hydrolase, 21 carbohydrate-binding module, 35 glycosyl-transferase, 12 polysaccharide lyase and 11 carbohydrate-esterase families (Supplementary Table 10). On average,  $2.62 \pm 0.13\%$  of the sequences in the gut microbiome could be assigned to CAZymes (a total of 217,615 sequences), a percentage that is greater than the most abundant KEGG pathway ('Transporters';  $1.20 \pm 0.06\%$  of the filtered sequences generated from each sample) and indicative of the abundant and diverse set of microbial genes directed towards accessing a wide range of polysaccharides.

Category-based clustering of the functions from each microbiome was performed using principal components analysis (PCA) and hierarchical clustering<sup>19</sup>. Two distinct clusters of gut microbiomes were identified based on metabolic profile, corresponding to samples with an increased abundance of Firmicutes and Actinobacteria, and samples with a high abundance of Bacteroidetes (Fig. 2a). A linear regression of the first principal component (PC1, explaining 20% of the functional variance) and the relative abundance of the Bacteroidetes showed a highly significant correlation ( $R^2 = 0.96$ ,  $P < 10^{-12}$ ; Fig. 2b). Functional profiles stabilized within each individual's microbiome after 20,000 sequences had been accumulated (Supplementary Fig. 13). Family members had more similar profiles than unrelated individuals (Fig. 2c), suggesting that shared bacterial community structure ('who's there' based on 16S rRNA analyses) also translates into shared community-wide relative abundance of metabolic pathways. Accordingly, a direct comparison of functional



**Figure 2 | Metabolic-pathway-based clustering and analysis of the human gut microbiome of monozygotic twins.** **a**, Clustering of functional profiles based on the relative abundance of KEGG metabolic pathways. All pairwise comparisons were made of the profiles by calculating each  $R^2$  value. Sample identifier nomenclature: family number, twin number or mother, and BMI category (Le, lean; Ov, overweight; Ob, obese; for example, F1T1Le stands for family 1, twin 1, lean). **b**, The relative abundance of Bacteroidetes as a function of the first principal component derived from an analysis of KEGG metabolic profiles. **c**, Comparisons of functional similarity between twin pairs, between twins and their mother, and between unrelated individuals. Asterisk indicates significant differences (Student's *t*-test with Monte Carlo;  $P < 0.01$ ; mean  $\pm$  s.e.m.).

and taxonomic similarity (see Supplementary Methods) disclosed a significant association: individuals with similar taxonomic profiles also share similar metabolic profiles ( $P < 0.001$ ; Mantel test).

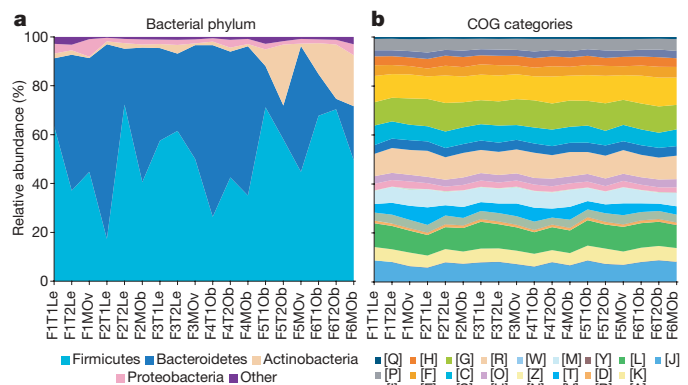
Functional clustering of phylum-wide sequence bins representing microbiome reads assigned to 23 human gut Firmicutes and 14 Bacteroidetes reference genomes showed discrete clustering by phylum (Supplementary Figs 14a and 15). Bootstrap analyses of the relative abundance of metabolic pathways in the microbiome-derived Firmicutes and Bacteroidetes sequence bins disclosed 26 pathways with a significantly different relative abundance (Supplementary Fig. 14a). The Bacteroidetes bins were enriched for several carbohydrate metabolism pathways, whereas the Firmicutes bins were enriched for transport systems. This finding is consistent with our CAZyme analysis, which revealed a significantly higher relative abundance of glycoside hydrolases, carbohydrate-binding modules, glycosyltransferases, polysaccharide lyases and carbohydrate esterases in the Bacteroidetes sequence bins (Supplementary Fig. 14b).

One of the major goals of the International Human Microbiome Project(s) is to determine whether there is an identifiable 'core microbiome' of shared organisms, genes or functional capabilities found in a given body habitat of all or the vast majority of humans<sup>1</sup>. Although all of the 18 gut microbiomes surveyed showed a high level

of  $\beta$ -diversity with respect to the relative abundance of bacterial phyla (Fig. 3a), analysis of the relative abundance of broad functional categories of genes and metabolic pathways (KEGG) revealed a generally consistent pattern regardless of the sample surveyed (Fig. 3b and Supplementary Table 11): the pattern is also consistent with results we obtained from a meta-analysis of previously published gut microbiome data sets from nine adults<sup>20,21</sup> (Supplementary Fig. 16). This consistency is not simply due to the broad level of these annotations, as a similar analysis of Bacteroidetes and Firmicutes reference genomes revealed substantial variation in the relative abundance of each category (see Supplementary Fig. 17). Furthermore, pairwise comparisons of metabolic profiles obtained from the 18 microbiomes in this study revealed an average value of  $R^2$  of  $0.97 \pm 0.002$  (Fig. 2a), indicating a high level of functional similarity.

Overall functional diversity was compared using the Shannon index<sup>22</sup>, a measurement that combines diversity (the number of different metabolic pathways) and evenness (the relative abundance of each pathway). The human gut microbiomes surveyed had a stable and high Shannon index value ( $4.63 \pm 0.01$ ), close to the maximum possible level of functional diversity (5.54; see Supplementary Methods). Despite the presence of a small number of abundant metabolic pathways (listed in Supplementary Table 11), the overall functional profile of each gut microbiome is quite even (Shannon evenness of  $0.84 \pm 0.001$  on a scale of 0–1), demonstrating that most metabolic pathways are found at a similar level of abundance. Interestingly, the level of functional diversity in each microbiome was significantly linked to the relative abundance of the Bacteroidetes ( $R^2 = 0.81$ ,  $P < 10^{-6}$ ); microbiomes enriched for Firmicutes/Actinobacteria had a lower level of functional diversity. This observation is consistent with an analysis of simulated metagenomic reads generated from each of 36 Bacteroidetes and Firmicutes genomes (Supplementary Fig. 18): on average, the Bacteroidetes genomes have a significantly higher level of both functional diversity and evenness (Mann–Whitney *U*-test,  $P < 0.01$ ).

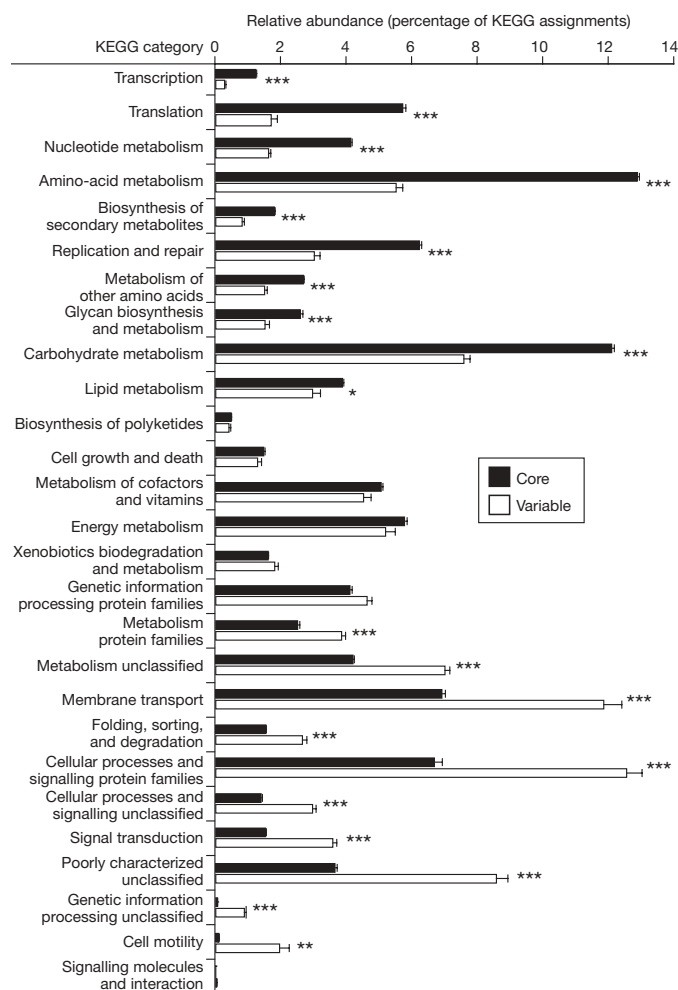
At a finer level, 26–53% of 'enzyme'-level functional groups (KEGG/CAZy/STRING) were shared across all 18 microbiomes, whereas 8–22% of the groups were unique to a single microbiome (Supplementary Fig. 19a–c). The 'core' functional groups present in all microbiomes were also highly abundant, representing 93–98% of the total sequences. Given the higher relative abundance of these 'core' groups, more than 95% were found after 26.11  $\pm$  2.02 megabases of sequence were collected from a given microbiome, whereas the 'variable' groups continued to increase substantially with each additional megabase of sequence. Of course, any estimate of the total size of the core microbiome will depend on sequencing effort, especially for



**Figure 3 | Comparison of taxonomic and functional variations in the human gut microbiome.** **a**, Relative abundance of major phyla across 18 faecal microbiomes from monozygotic twins and their mothers, based on BLASTX comparisons of microbiomes and the National Center for Biotechnology Information non-redundant database. **b**, Relative abundance of categories of genes across each sampled gut microbiome (letters correspond to categories in the COG database).

functional groups found at a low abundance. On average, our survey achieved more than 450,000 sequences per faecal sample, which, assuming an even distribution, would allow us to sample groups found at a relative abundance of  $10^{-4}$ . To estimate the total size of the core microbiome based on the 18 individuals, we randomly sub-sampled each microbiome in 1,000 sequence intervals (Supplementary Fig. 19d). Based on this analysis, the core microbiome is approaching a total of 2,142 total orthologous groups (one site binding (hyperbola) curve fit,  $R^2 = 0.9966$ ), indicating that we identified 93% of functional groups (defined by STRING) found within the core microbiome of the 18 individuals surveyed. Of these core groups, 71% (CAZy), 64% (KEGG) and 56% (STRING) were also found in the nine previously published, but much lower coverage, data sets generated by capillary sequencing of adult faecal DNA<sup>20,21</sup> (average of  $78,413 \pm 2,044$  bidirectional reads per sample; see Supplementary Methods).

Metabolic reconstructions of the 'core' microbiome revealed significant enrichment for several expected functional categories, including those involved in transcription and translation (Fig. 4). Metabolic profile-based clustering indicated that the representation of 'core' functional groups was highly consistent across samples (Supplementary Fig. 20), and included several pathways that are



**Figure 4 | KEGG categories enriched or depleted in the core versus variable components of the gut microbiome.** Sequences from each of the 18 faecal microbiomes were binned into the 'core' or 'variable' microbiome based on the co-occurrence of KEGG orthologous groups (core groups were found in all 18 microbiomes whereas variable groups were present in fewer (<18) microbiomes; see Supplementary Fig. 19a). Asterisks indicate significant differences (Student's *t*-test, \* $P < 0.05$ , \*\* $P < 0.001$ , \*\*\* $P < 10^{-3}$ ; mean  $\pm$  s.e.m.).

likely important for life in the gut, such as those for carbohydrate and amino-acid metabolism (for example, fructose/mannose metabolism, amino-sugar metabolism and N-glycan degradation). Variably represented pathways and categories include cell motility (only a subset of Firmicutes produce flagella), secretion systems and membrane transport (for example, phosphotransferase systems involved in the import of nutrients, including sugars; Fig. 4 and Supplementary Fig. 20).

The distribution of CAZy glycoside hydrolase and glycosyltransferase families was compared between each pair of microbiomes (see Supplementary Table 10 for CAZy families with a relative abundance greater than 1%). This analysis revealed that all individuals had a similar profile of glycosyltransferases ( $R^2 = 0.96 \pm 0.003$ ), whereas the profiles of glycoside hydrolases were significantly more variable, even between family members ( $R^2 = 0.80 \pm 0.01$ ;  $P < 10^{-30}$ , paired Student's *t*-test). This suggests that the number and spectrum of glycoside hydrolases is affected by 'external' factors such as diet more than the glycosyltransferases.

To identify metabolic pathways associated with obesity, only non-core associated (variable) functional groups were included in a comparison of the gut microbiomes of lean versus obese twin pairs. A bootstrap analysis<sup>23</sup> was used to identify metabolic pathways that were enriched or depleted in the variable obese gut microbiome. For example, similar to a mouse model of diet-induced obesity<sup>4</sup>, the obese human gut microbiome was enriched for phosphotransferase systems involved in microbial processing of carbohydrates (Supplementary Table 12). All gut microbiome sequences were compared with the custom database of 44 human gut genomes: an odds ratio analysis revealed 383 genes that were significantly different between the obese and lean gut microbiome ( $q$  value  $< 0.05$ ; 273 enriched and 110 depleted in the obese microbiome; Supplementary Tables 13 and 14). By contrast, only 49 genes were consistently enriched or depleted between all twin pairs (see Supplementary Methods).

These obesity-associated genes were representative of the taxonomic differences described above: 75% of the obesity-enriched genes were from Actinobacteria (compared with 0% of lean-enriched genes; the other 25% are from Firmicutes) whereas 42% of the lean-enriched genes were from Bacteroidetes (compared with 0% of the obesity-enriched genes). Their functional annotation indicated that many are involved in carbohydrate, lipid and amino-acid metabolism (Supplementary Tables 13 and 14). Together, they comprise an initial set of microbial biomarkers of the obese gut microbiome.

Our finding that the gut microbial community structures of adult monozygotic twin pairs had a degree of similarity that was comparable to that of dizygotic twin pairs, and only slightly more similar than that of their mothers, is consistent with an earlier fingerprinting study of adult twins<sup>24</sup>, and with a recent microarray-based analysis, which revealed that gut community assembly during the first year of life followed a more similar pattern in a pair of dizygotic twins than 12 unrelated infants<sup>25</sup>. Intriguingly, another fingerprinting study of monozygotic and dizygotic twins in childhood showed a slightly reduced similarity profile in dizygotic twins<sup>26</sup>. Thus, comprehensive time-course studies, comparing monozygotic and dizygotic twin pairs from birth through adulthood, as well as intergenerational analyses of their families' microbiotas, will be key to determining the relative contributions of host genotype and environmental exposures to (gut) microbial ecology.

The hypothesis that there is a core human gut microbiome, definable by a set of abundant microbial organismal lineages that we all share, may be incorrect: by adulthood, no single bacterial phylotype was detectable at an abundant frequency in the guts of all 154 sampled humans. Instead, it appears that a core gut microbiome exists at the level of shared genes, including an important component involved in various metabolic functions. This conservation suggests a high degree of redundancy in the gut microbiome and supports an ecological view of each individual as an 'island' inhabited by unique

collections of microbial phylotypes: as in actual islands, different species assemblages converge on shared core functions provided by distinctive components. Our findings raise the question of how core functionality is assembled in this body habitat. Understanding the underlying principles should provide insights about microbial adaptation to, and mutualistic community assembly within, a wide range of environments.

## METHODS SUMMARY

Faecal samples were collected from each individual. Community DNA was prepared and used for pyrosequencing (454 Life Sciences), as well as for PCR and sequencing of bacterial 16S rRNA genes. Shotgun reads were mapped to reference genomes using the National Center for Biotechnology Information 'non-redundant' database, KEGG<sup>17</sup>, STRING<sup>18</sup>, CAZy (<http://www.cazy.org/>) and a 44-member human-gut microbial genome database. Metabolic reconstructions were performed based on CAZy, KEGG and STRING annotations. The relative abundance of KEGG metabolic pathways is referred to as a 'metabolic profile'.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

Received 29 June; accepted 14 October 2008.

Published online 30 November 2008.

- Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449**, 804–810 (2007).
- Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl Acad. Sci. USA* **102**, 11070–11075 (2005).
- Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027–1031 (2006).
- Turnbaugh, P. J., Bäckhed, F., Fulton, L. & Gordon, J. I. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe* **3**, 213–223 (2008).
- Bäckhed, F. *et al.* The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl Acad. Sci. USA* **101**, 15718–15723 (2004).
- Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Human gut microbes associated with obesity. *Nature* **444**, 1022–1023 (2006).
- Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308**, 1635–1638 (2005).
- Frank, D. N. *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl Acad. Sci. USA* **104**, 13780–13785 (2007).
- Bruder, C. E. *et al.* Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am. J. Hum. Genet.* **82**, 763–771 (2008).
- Bouchard, C. *et al.* The response to long-term overfeeding in identical twins. *N. Engl. J. Med.* **322**, 1477–1482 (1990).
- Maes, H. H., Neale, M. C. & Eaves, L. J. Genetic and environmental factors in relative body weight and human adiposity. *Behav. Genet.* **27**, 325–351 (1997).
- Heath, A. C. *et al.* Ascertainment of a mid-western US female adolescent twin cohort for alcohol studies: assessment of sample representativeness using birth record data. *Twin Res.* **5**, 107–112 (2002).
- Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J. & Knight, R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods* **5**, 235–237 (2008).
- Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl Acad. Sci. USA* **103**, 12115–12120 (2006).
- Lozupone, C., Hamady, M. & Knight, R. UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7**, 371 (2006).
- Watson, S. B., McCauley, E. & Downing, J. A. Patterns in phytoplankton taxonomic composition across temperate lakes of differing nutrient status. *Limnol. Oceanogr.* **42**, 487–495 (1997).
- Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
- von Mering, C. *et al.* STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35**, D358–D362 (2007).
- de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453–1454 (2004).
- Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355–1359 (2006).
- Kurokawa, K. *et al.* Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* **14**, 169–181 (2007).
- Dinsdale, E. A. *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452**, 629–632 (2008).
- Rodríguez-Brito, B., Rohwer, F. & Edwards, R. A. An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7**, 162 (2006).
- Zoetand, E. G., Akkermans, A. D. L., Akkermans-van Vliet, W. M., de Visser, J. A. & de Vos, W. M. The host genotype affects the bacterial community in the human gastrointestinal tract. *Microb. Ecol. Health Dis.* **13**, 129–134 (2001).
- Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5**, e177 (2007).
- Stewart, J. A., Chadwick, V. S. & Murray, A. Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. *J. Med. Microbiol.* **54**, 1239–1242 (2005).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank: S. Wagoner and J. Manchester for technical support; S. Marion and D. Hopper for recruitment of participants and sample collection; A. Goodman, B. Muegge, and M. Mahowald for suggestions; S. Huse (Marine Biological Laboratory), F. Niazi and S. Attiya (454 Life Sciences), C. Markovic, L. Fulton, B. Fulton, E. Mardis and R. Wilson (Washington University Genome Sequencing Center) and S. Macmil, G. Wiley, C. Qu, and P. Wang (University of Oklahoma) for their assistance with sequencing; and P. M. Coutinho (Université de Provence, France) for help with the CAZy analysis. Deep draft assemblies of reference gut genomes were generated as part of a National Human Genome Research Institute (NHGRI)-sponsored human gut microbiome initiative ([http://genome.wustl.edu/pub/organism/Microbes/Human\\_Gut\\_Microbiome/](http://genome.wustl.edu/pub/organism/Microbes/Human_Gut_Microbiome/)). This work was supported in part by the National Institutes of Health (DK78669/ES012742/AA09022/HD049024), the National Science Foundation (OCE0430724), the W.M. Keck Foundation, and the Crohn's and Colitis Foundation of America.

**Author Contributions** P.J.T., A.C.H., R.K. and J.I.G. designed the experiments. P.J.T., T.Y., A.D., R.E.L., M.L.S., W.J.J., B.A.R., J.P.A. and M.E. generated the data. P.J.T., M.H., M.L.S., B.L.C., A.D., B.H., A.C.H., R.K. and J.I.G. analysed the data. P.J.T., A.C.H., R.K. and J.I.G. wrote the manuscript with input from the other members of the team.

**Author Information** This Whole Genome Shotgun project is deposited in DDBJ/EMBL/GenBank under accession number 32089. 454 pyrosequencing reads are deposited in the NCBI Short Read Archive. Nearly full-length 16S rRNA gene sequences are deposited in GenBank under accession numbers FJ362604–FJ372382. Annotated sequences are also available in MG-RAST (<http://metagenomics.nmpdr.org/>). 454-generated 16S rRNA sequences with sample identifiers are also available at <http://gordonlab.wustl.edu/SuppData.html>. Correspondence and requests for materials should be addressed to J.I.G. ([jgordon@wustl.edu](mailto:jgordon@wustl.edu)).

## METHODS

**Community DNA preparation.** Faecal samples were frozen immediately after they were produced. De-identified samples were stored at  $-80^{\circ}\text{C}$  before processing. Ten to twenty grams of each sample was pulverized in liquid nitrogen with a mortar and pestle. An aliquot (approximately 500 mg) of each sample was then suspended, while frozen, in a solution containing 500  $\mu\text{l}$  of extraction buffer (200 mM Tris (pH 8.0), 200 mM NaCl, 20 mM EDTA), 210  $\mu\text{l}$  of 20% SDS, 500  $\mu\text{l}$  of a mixture of phenol:chloroform:isoamyl alcohol (25:24:1, pH 7.9), and 500  $\mu\text{l}$  of a slurry of 0.1-mm diameter zirconia/silica beads (BioSpec Products). Microbial cells were subsequently lysed by mechanical disruption with a bead beater (BioSpec Products) set on high for 2 min at room temperature, followed by extraction with phenol:chloroform:isoamyl alcohol, and precipitation with isopropanol. DNA obtained from three separate aliquots of each faecal sample were pooled ( $\geq 200$   $\mu\text{g}$  DNA) and used for pyrosequencing (see below).

**16S rRNA gene-sequence-based surveys.** Complementary phylogenetic- and taxon-based methods were used to compare 16S rRNA sequences among faecal communities. Phylogenetic clustering with UniFrac<sup>15</sup> is based on the principle that communities can be compared in terms of their shared evolutionary history, as measured by the degree to which they share branch length on a phylogenetic tree. We complemented this approach with taxon-based methods<sup>27</sup>, which disregard some of the information contained in the phylogenetic tree of the taxa in question, but have the advantage that specific taxa unique to, or shared among, groups of samples can be identified (for example, those from lean or obese individuals). Before both types of analysis, we grouped 16S rRNA gene sequences into operational taxonomic units (OTUs/phylotypes) using both cd-hit<sup>28</sup> and the furthest-neighbour-like algorithm, with a sequence identity threshold of 97%, which is commonly used to define 'species'-level phylotypes. Taxonomy was assigned using the best-BLAST-hit against Greengenes<sup>29</sup> ( $E$  value cutoff of  $10^{-10}$ , minimum 88% coverage, 88% identity) and the Hugenholtz taxonomy (downloaded from [http://greengenes.lbl.gov/Download/Sequence\\_Data/Greengenes\\_format/](http://greengenes.lbl.gov/Download/Sequence_Data/Greengenes_format/) on 12 May 2008, excluding sequences annotated as chimaeric).

**Selection of operational taxonomic units.** 16S rRNA gene-derived pyrosequencing data were pre-processed to remove sequences with low-quality scores, sequences with ambiguous characters or sequences outside the length bounds ( $V6 < 50$  nucleotides,  $V2 < 200$  nucleotides), and binned according to sample-specific barcode (see, for example, ref. 13). Similar sequences were identified using Megablast<sup>30</sup> and cd-hit, with the following parameters:  $E$  value  $10^{-10}$  (Megablast only); minimum coverage 99%; minimum pairwise identity 97%. Candidate OTUs were identified as sets of sequences connected to each other at this level using a maximum of 4,000 hits per sequence. Each candidate OTU was considered valid if the average density of connection was above threshold; otherwise, it was broken up into smaller connected components<sup>27</sup>.

**Tree building and UniFrac clustering for PCA analysis.** A relaxed neighbour-joining tree was built from one representative sequence per OTU using Clearcut<sup>31</sup>, employing the Kimura correction (the PH Lane mask was applied to V2 and full-length data), but otherwise with default comparisons. Unweighted UniFrac<sup>15</sup> was run using the resulting tree. PCA was performed on the resulting matrix of distances between each pair of samples. To determine if the UniFrac distances were on average significantly different for pairs of samples (that is, between twin pairs, between twins and their mother, or between unrelated individuals), we performed a  $t$ -test on the UniFrac distance matrix, and generated a  $P$  value for the  $t$ -statistic by permutation of the rows and columns as in the Mantel test, regenerating the  $t$ -statistic for 1,000 random samples, and using the distribution to obtain an empirical  $P$  value.

**Rarefaction and phylogenetic diversity measurements.** To determine which individuals had the most diverse communities of gut bacteria, rarefaction plots and phylogenetic diversity measurements, as described by Faith<sup>32</sup>, were made for each sample. Phylogenetic diversity is the total amount of branch length in a phylogenetic tree constructed from the combined 16S rRNA data sets, leading to the sequences in a given sample. To account for differences in sampling effort between individuals, and to estimate how far we were from sampling the diversity of each individual completely, we plotted the accumulation of phylogenetic diversity (branch length) with sampling effort, in a manner analogous to rarefaction curves. We generated the phylogenetic diversity rarefaction curve

for each individual by applying custom python code (<http://bmf2.colorado.edu/unifrac/about.psp>) to the Arb parsimony insertion tree<sup>27</sup>.

**Pyrosequencing of total community DNA.** Shotgun sequencing runs were performed on the 454 FLX pyrosequencer from total faecal community DNA. Two samples were also analysed in a single run using Titanium extra-long-read pyrosequencing technology (see Supplementary Tables 4 and 5). Sequencing reads with degenerate bases ('Ns') were removed along with all duplicate sequences, as sequences of identical length and content are a common artefact of the pyrosequencing methodology. Finally, human sequences were removed by identifying sequences homologous to the *Homo sapiens* reference genome (BLASTN  $E < 10^{-5}$ , %identity  $> 75$ , score  $> 50$ ).

**CAZyme analysis.** Metagenomic sequence reads were searched against a library of modules derived from all entries in the carbohydrate-active enzymes (CAZy) database ([www.cazy.org](http://www.cazy.org) using FASTY<sup>33</sup>,  $E < 10^{-6}$ ). This library consists of approximately 180,000 previously annotated modules (catalytic modules, carbohydrate-binding modules and other non-catalytic modules or domains of unknown function) derived from about 80,000 protein sequences. The number of sequencing reads matching each CAZy family was divided by the number of total sequences assigned to CAZymes and multiplied by 100 to calculate a relative abundance. An  $R^2$  value was calculated for each pair of CAZy profiles. We then compared the distribution of glycoside hydrolase similarity scores with the distribution of glycosyltransferase similarity scores.

**Statistical analyses.** Xipe<sup>23</sup> (version 2.4) was used for bootstrap analyses of pathway enrichment and depletion, using the parameters sample size = 10,000 and confidence level = 0.95. Linear regressions were performed in Excel (version 11.0, Microsoft). Mann-Whitney and Student's  $t$ -tests were used to identify statistically significant differences between two groups (Prism version 4.0, GraphPad; Excel version 11.0, Microsoft). The Bonferroni correction was used to correct for multiple hypotheses. The Mantel test was used to compare distance matrices: the matrix of each pairwise comparison of the abundance of each reference genome, and the abundance of each metabolic pathway, were compared (Mantel program in Python using PyCogent<sup>34</sup>; 10,000 replicates). Data are represented as mean  $\pm$  s.e.m. unless otherwise indicated.

Microbiome sequences were compared against the custom database of 44 gut genomes (BLASTX  $E < 10^{-5}$ , bitscore  $> 50$ , and %identity  $> 50$ ). A gene-by-sample matrix was then screened to identify genes 'commonly-enriched' in either the obese or lean gut microbiome (defined by an odds ratio greater than 2 or less than 0.5 when comparing the pooled obese twin microbiomes with the pooled lean twin microbiomes, and when comparing each individual obese twin microbiome with the aggregate lean twin microbiome, or vice versa). The statistical significance of enriched or depleted genes was then calculated using a modified  $t$ -test ( $q$  value  $< 0.05$ ; calculated with code supplied by M. Pop and J.R. White, University of Maryland). We also searched for genes that were consistently enriched or depleted in all six monozygotic twin pairs. A gene-by-sample matrix was generated based on BLASTX comparisons of each microbiome with our custom 44-genome database, to calculate an odds ratio based on the frequency of each gene in each twin versus the respective co-twin. The analysis revealed only 49 genes (odds ratio  $> 2$  or  $< 0.5$ ): they represent a variety of taxonomic groups, including Firmicutes, Bacteroidetes and Actinobacteria, and did not show any clear functional trends.

27. Ley, R. E. *et al.* Evolution of mammals and their gut microbes. *Science* **320**, 1647–1651 (2008).
28. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
29. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
30. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214 (2000).
31. Sheneman, L., Evans, J. & Foster, J. A. Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics* **22**, 2823–2824 (2006).
32. Faith, D. P. Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* **61**, 1–10 (1992).
33. Pearson, W. R., Wood, T., Zhang, Z. & Miller, W. Comparison of DNA sequences with protein sequences. *Genomics* **46**, 24–36 (1997).
34. Knight, R. *et al.* PyCogent: a toolkit for making sense from sequence. *Genome Biol.* **8**, R171 (2007).

## Appendix B

Turnbaugh PJ, Quince C, Faith JJ, McHardy AC, **Yatsunenko T**, Niazi F, Affourtit J, Egholm M, Henrissat B, Knight R, Gordon JI.

“Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins.”

*Proceedings of the National Academy of Sciences USA*. **2010**. 107(16): 7503-8.



# Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins

Peter J. Turnbaugh<sup>a,1</sup>, Christopher Quince<sup>b</sup>, Jeremiah J. Faith<sup>a</sup>, Alice C. McHardy<sup>c</sup>, Tanya Yatsunenka<sup>a</sup>, Faheem Niazi<sup>d</sup>, Jason Affourtit<sup>d</sup>, Michael Egholm<sup>d</sup>, Bernard Henrissat<sup>e</sup>, Rob Knight<sup>f</sup>, and Jeffrey I. Gordon<sup>a,2</sup>

<sup>a</sup>Center for Genome Sciences, Washington University School of Medicine, St. Louis, MO 63108; <sup>b</sup>Department of Civil Engineering, University of Glasgow, Glasgow, United Kingdom; <sup>c</sup>Max-Planck Institute for Informatics, 66123 Saarbrücken, Germany; <sup>d</sup>454 Life Sciences, Branford, CT 06405; <sup>e</sup>Architecture et Fonction des Macromolécules Biologiques, Centre National de la Recherche Scientifique, Marseille, France; and <sup>f</sup>Howard Hughes Medical Institute and Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309

Contributed by Jeffrey I. Gordon, February 25, 2010 (sent for review February 1, 2010)

We deeply sampled the organismal, genetic, and transcriptional diversity in fecal samples collected from a monozygotic (MZ) twin pair and compared the results to 1,095 communities from the gut and other body habitats of related and unrelated individuals. Using a new scheme for noise reduction in pyrosequencing data, we estimated the total diversity of species-level bacterial phylotypes in the 1.2–1.5 million bacterial 16S rRNA reads obtained from each deeply sampled cotwin to be ~800 (35.9%, 49.1% detected in both). A combined 1.1 million read 16S rRNA dataset representing 281 shallowly sequenced fecal samples from 54 twin pairs and their mothers contained an estimated 4,018 species-level phylotypes, with each sample having a unique species assemblage ( $53.4 \pm 0.6\%$  and  $50.3 \pm 0.5\%$  overlap with the deeply sampled cotwins). Of the 134 phylotypes with a relative abundance of  $>0.1\%$  in the combined dataset, only 37 appeared in  $>50\%$  of the samples, with one phylotype in the Lachnospiraceae family present in 99%. Non-gut communities had significantly reduced overlap with the deeply sequenced twins' fecal microbiota ( $18.3 \pm 0.3\%$ ,  $15.3 \pm 0.3\%$ ). The MZ cotwins' fecal DNA was deeply sequenced (3.8–6.3 Gbp/sample) and assembled reads were assigned to 25 genus-level phylogenetic bins. Only 17% of the genes in these bins were shared between the cotwins. Bins exhibited differences in their degree of sequence variation, gene content including the repertoire of carbohydrate active enzymes present within and between twins (e.g., predicted cellulases, dockerins), and transcriptional activities. These results provide an expanded perspective about features that make each of us unique life forms and directions for future characterization of our gut ecosystems.

microbial phylogenetic analyses | microbiota | transcriptomics | carbohydrate active enzymes

Human microbiome projects are being initiated throughout the world, with the goal of correlating human physiological phenotypes with the structures and functions of their indigenous microbial communities. Substantial insight into the patterns of variation in the microbiota between body habitats and individuals has been gained using shallow sequencing of 16S rRNA gene amplicons and community DNA. Because of limitations imposed by sequencing costs and throughput, these studies have examined the more abundant species or genes. A timely question is this: What additional insights about the microbial diversity present within a body habit are obtained with deeper sequencing? Moreover, how much of the observed organismal diversity is an artifact of noise introduced during PCR and sequencing of 16S rRNA genes (1–3)? Therefore, in the current study we use a variety of experimental and computational approaches to explore the level of diversity and interpersonal variation in bacterial phylotypes, microbial genes, and their expressed mRNA transcripts within the human gut, home to our largest community of microorganisms.

## Results and Discussion

**Study Design and Data Collection.** Total community DNA and RNA was initially isolated from two fecal samples, each obtained from 26-year-old, obese, MZ female cotwins (body mass index, 39 and 45 kg/m<sup>2</sup>). Both cotwins (designated TS28 and TS29) had been vaginally delivered; neither cotwin had any history of intestinal disease, and neither had used antibiotics at least 6 months before providing fecal samples, at which time the cotwins lived 5 km apart (4). A 454 pyrosequencing method was used to obtain 1.2–1.5 million sequencing reads from PCR-amplified V2 regions of bacterial 16S rRNA genes present in each fecal sample (average read length ~232 nt), and 3.8–6.3 Gbp of single- and paired-end shotgun reads from total fecal community DNA (Table S1). Using a method for rRNA depletion based on a combination of size selection (to remove 5S rRNA and tRNA), and streptavidin bead-based pull-down of biotinylated oligonucleotides hybridized to domains conserved among gut bacterial rRNA genes (5), we enriched for fecal mRNA and then generated 12–16 million sequencing reads representing expressed genes in their microbiomes (Table S2).

**Analysis of Bacterial Diversity Present in the Gut Microbiota. Algorithms for denoising pyrosequencing data: tests using mixtures of bacterial strains.** We analyzed test datasets composed of an unequal mixture of DNA from 90 cloned bacterial 16S rRNA gene sequences (2) or DNA purified from 67 bacterial strains cultured from the human gut and pooled together over a range of relative concentrations (Table S3). These test datasets were used to establish a set of procedures for removing noise from 16S rRNA datasets that arise from PCR and pyrosequencing (SI Text).

**Comparison of the fecal microbiota of the deeply sampled MZ co-twins.** Using these procedures, we determined that most species-level phylotypes were present at low abundance [species defined as organisms sharing  $\geq 97\%$  sequence identity (%ID) in their 16S rRNA genes; Fig. S1]; ~100,000 16S rRNA sequences were required to observe 60% of the total phylotypes (Fig. 1A). At the

Author contributions: P.J.T., C.Q., R.K., and J.I.G. designed research; P.J.T., J.J.F., and T.Y. performed research; P.J.T., F.N., J.A., and M.E. contributed new reagents/analytic tools; P.J.T., C.Q., J.J.F., A.C.M., B.H., R.K., and J.I.G. analyzed data; and P.J.T., C.Q., R.K., and J.I.G. wrote the paper.

The authors declare no conflict of interest.

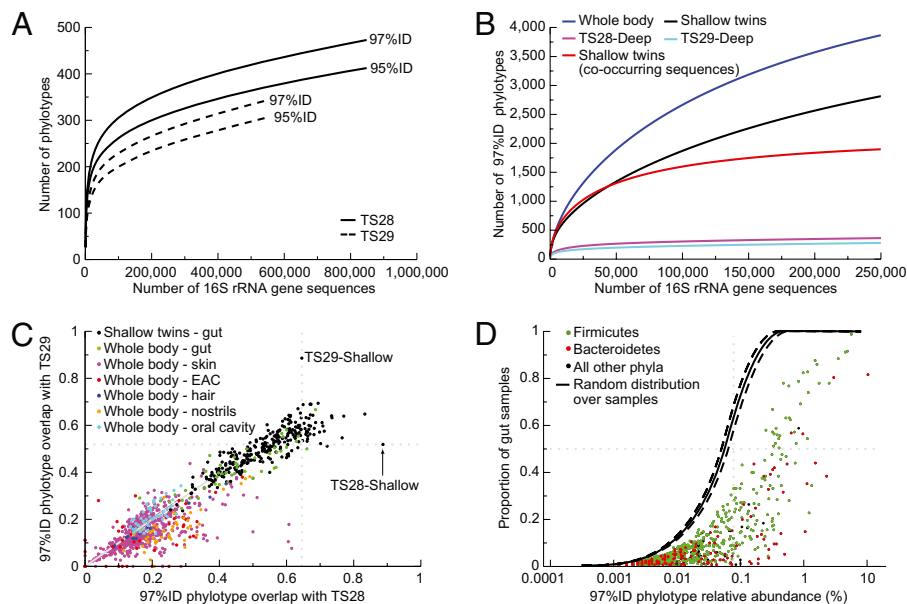
Data deposition: Datasets from shotgun sequencing projects have been deposited at DDBJ/EMBL/GenBank under accession number 43253. 454 and Illumina sequencing reads have been deposited in the NCBI Short Read Archive.

Freely available online through the PNAS open access option.

<sup>1</sup>Current address: FAS Center for Systems Biology, Harvard University, Cambridge, MA 02138.

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

This article contains supporting information online at [www.pnas.org/cgi/content/full/1002355107/DCSupplemental](http://www.pnas.org/cgi/content/full/1002355107/DCSupplemental).



**Fig. 1.** Measurements of bacterial diversity in the human fecal microbiota. (A) Rarefaction curves at 97%ID and 95%ID phylotype cutoffs are shown for the deeply sequenced TS28 and TS29 MZ cotwin (“Deep Twins”) datasets. Sequences were classified as chimeric at the 50% probability cutoff. (B) Comparison of diversity within and between gut microbial communities. Curves at 97%ID phylotype cutoff are shown for 250 fecal samples taken from 146 individuals (“Shallow twins”; 1,000 16S rRNA gene sequences were randomly selected from each sample), 250 samples taken from multiple body habitats (“Whole body”; 1,000 randomly selected sequences per sample), and the two deeply sequenced fecal samples (“TS28-Deep” and “TS29-Deep”). Phylotypes found in multiple fecal samples are labeled “co-occurring.” (C) Plot of proportion of 97%ID phylotypes found in TS28 and TS29 across 277 fecal samples (black circles) and 814 samples taken from multiple body habitats in nine individuals [habitat groups are colored green (fecal), purple (skin), red (external auditory canal; EAC), blue (hair), orange (nostrils), and light blue (oral cavity)]. Four EAC and one skin sample did not contain any shared phylotypes with TS28 and TS29. (D) The proportion of the 250 fecal samples containing each 97%ID phylotype plotted as a function of the relative abundance (%) of each phylotype in the combined dataset. Phylotypes are colored according to phylum: Bacteroidetes (red), Firmicutes (green), and other (black). The expected proportion of samples containing each phylotype, assuming a random distribution across samples, is shown (median  $\pm$  95% confidence interval).

95%ID and 97%ID phylotype cutoffs, rarefaction curves did not completely saturate even when  $>10^6$  sequences were collected (Fig. 1A), indicating that additional phylotypes remain uncharacterized even at this high level of coverage.

The total estimated diversity of species-level bacterial phylotypes (97%ID) in the TS28 and TS29 datasets was lower than expected (878 and 768, respectively; Table 1 and Table S4), based on previous studies that did not account for noise. There was notable variation even between these genetically identical cotwins: 35.9% and 49.1% of the species-level phylotypes found in the fecal communities of TS28 and TS29, respectively, were shared between the two samples (39.0% and 52.8% were shared at the 95%ID level).

However, these values do not account for phylotypes that may be abundant in one sample and rare in another. Overall, shared phylotypes showed a small but positive correlation in relative abundance between samples, and rarely varied by more than two orders

of magnitude ( $R^2 = 0.18$  for 97%ID and  $R^2 = 0.27$  for 95%ID). This observation allowed us to define a normalized overlap between the samples by considering only phylotypes found at a sufficient relative abundance in each sample that they are unlikely to have been missed because of variations in their relative abundance (“Normalized overlap” in SI Text). With this normalization, 68% and 79% of 97% ID phylotypes in TS28 and TS29 were designated as being shared in the other cotwin’s microbiota (76.7% and 86.0% at 95%ID).

**Comparisons to more shallowly sampled fecal samples obtained from other twin pairs.** To test whether the deep sampling of these cotwins allowed us to capture the bacterial diversity present in fecal samples obtained from other families containing twins, we extended our survey to include 1.1 million bacterial V2 16S rRNA sequencing reads from 281 fecal samples procured from 31 MZ and 23 dizygotic (DZ) twin pairs and their mothers [ $3,984 \pm 232$  (mean  $\pm$  SEM) reads/sample] (4). Like the deeply sampled cot-

**Table 1.** Number of species-level (97%ID) and 95%ID bacterial phylotypes in the deep and shallow sequenced fecal microbiota of twins, and in the whole body sampling datasets

Dataset	16S rRNA seqs	Observed phylotypes (97%ID)	Estimated phylotypes (97%ID Chao) <sup>a</sup>	Observed phylotypes (95%ID)	Estimated phylotypes (95%ID Chao)
TS28-Deep	848,512	473	627	413	538
TS29-Deep	553,416	344	558	307	514
TS28-Shallow	3,288	135	375	121	329
TS29-Shallow	1,178	81	127	70	130
TSAll-Shallow	250,000	2,815	4,018	1,974	2,498
TSAll-Co-occur	250,000	1,898	2,043	1,221	1,283
WholeBody	250,000	3,869	4,949	2,957	3,646

<sup>a</sup>Chao’s nonparametric total diversity estimates are given. Phylotypes are grouped based on the degree of sequence identity in the V2 regions of their 16S rRNA genes.

TS28- and TS29-Deep, deeply sequenced cotwin fecal samples; TS28- and TS29-Shallow, shallow sequenced cotwin fecal samples; TSAll-Shallow, 1,000 randomly selected sequences from 250 fecal samples; TSAll-Co-occur, restricted to co-occurring sequences from 250 fecal samples; WholeBody, 250 randomly selected samples from a total of 814 samples obtained from 27 body sites from 9 individuals, 1,000 sequences/sample).

wins, these other twin pairs were born in Missouri, ranged in age from 25 to 32 years, did not have a history of GI pathology, and had not consumed antibiotics before sampling. All 16S rRNA pyrosequencing reads were preprocessed as done above to remove noise and chimeras.

A comparison of the total bacterial diversity found across these fecal samples and the two deeply sequenced samples underscored the much higher level of inter- compared with intrapersonal variation when considering a single body habitat (Fig. 1B). The combined “Shallow” fecal datasets had an estimated 4,018 97%ID phylotypes and 2,498 95%ID phylotypes (2,815 and 1,974 observed, respectively). These values are ~5-fold higher than in each deeply sequenced fecal microbiota (Table 1). In addition, each sample had a unique collection of 97%ID phylotypes (mean  $\pm$  SEM,  $53.4 \pm 0.6\%$  and  $50.3 \pm 0.5\%$  overlap with TS28 and TS29), whereas the fraction of phylotypes from each sample that were shared with TS28 correlated with the fraction shared with TS29 ( $R^2 = 0.73$ ; Fig. 1C).

Our initial analysis of these fecal samples had indicated that there was no core set of abundant species-level phylotypes found in all individuals (4); this was confirmed after removing PCR and sequencing noise. The proportion of samples containing each phylotype was lower than expected by chance at all levels of relative abundance (Fig. 1D), but within each level of abundance there was a large spread. Only a few phylotypes appeared in the majority of samples: of the 134 species-level phylotypes that had a relative abundance in the combined dataset  $>0.1\%$ , only 37 appeared in  $>50\%$  of the samples (28% of the phylotypes, compared with 100% expected by chance). Phylotypes assigned to the Firmicutes phylum were more evenly spread than the Bacteroidetes: 33% with  $>0.1\%$  relative abundance appeared in 50% of samples, compared with only 12% of the Bacteroidetes phylotypes (Fig. 1D). In addition, one nearly ubiquitous phylotype belonging to the family Lachnospiraceae (phylum Firmicutes) was found in 99% of the samples, representing 5.7% of the sequences in the combined dataset.

#### Comparisons to bacterial phylotypes present in other human body habitats.

To determine whether phylotypes present in the gut microbiota were detectable in other body habitats, we surveyed V2 16S rRNA sequencing reads obtained from nine unrelated healthy individuals (male and female) who had been sampled at 27 sites, including feces, twice over a 24-h period on two occasions, each occasion separated by 3 months (age range, 30–35 years with the exception of one individual 60 years of age; no recent history of antibiotic use; mean  $\pm$  SD  $1,315 \pm 420$  reads per sample) (6). All data were subjected to the same denoising procedures described above.

A comparison of the total diversity found across the 27 body habitats to the shallowly sequenced fecal samples and the two deeply sequenced fecal samples demonstrated higher levels of diversity when comparing across multiple body habitats vs. comparisons of the same habitat across multiple individuals (Fig. 1B). The combined 27-body habitats dataset contained an estimated 4,949 species-level phylotypes (97%ID) and 3,646 95%ID phylotypes (3,869 and 2,957 observed, respectively) (Table 1). Although the range of overlapping species-level phylotypes for the fecal samples from the 27-body habitat survey was comparable to the twin fecal cohort (mean  $\pm$  SEM  $45.1 \pm 1.9\%$  and  $41.7 \pm 1.4\%$ ), the other nongut body habitats showed a significantly reduced overlap (mean  $\pm$  SEM  $18.3 \pm 0.3\%$  and  $15.3 \pm 0.3\%$  with TS28 and TS29;  $P < 10^{-17}$ , Student's *t* test; Fig. 1C). As with the fecal samples from the shallowly sampled twins, the fraction of phylotypes from each sample that were shared with TS28 correlates with the fraction shared with TS29 ( $R^2 = 0.42$ ).

**Conclusions.** Together, these results emphasize the following: (i) despite large interpersonal variations in the composition of the gut microbiota and the absence of a core set of abundantly represented universally shared phylotypes, common phylotypes can be identified through deep sequencing of a small number of individuals; (ii)

a surprising amount of phylotypes are shared between distinct body habitats across unrelated individuals (i.e., only five samples did not contain any phylotypes from the deeply sequenced TS28 and TS29 gut microbial communities); and (iii) it seems feasible that future studies that broadly sample humans living in distinct cultural settings will be able to define population-wide gut phylotypes and, as a result, provide a rationale for selecting cultured representatives of these phylotypes for genome sequencing (e.g., start with phylotypes in the top right portion of Fig. 1D).

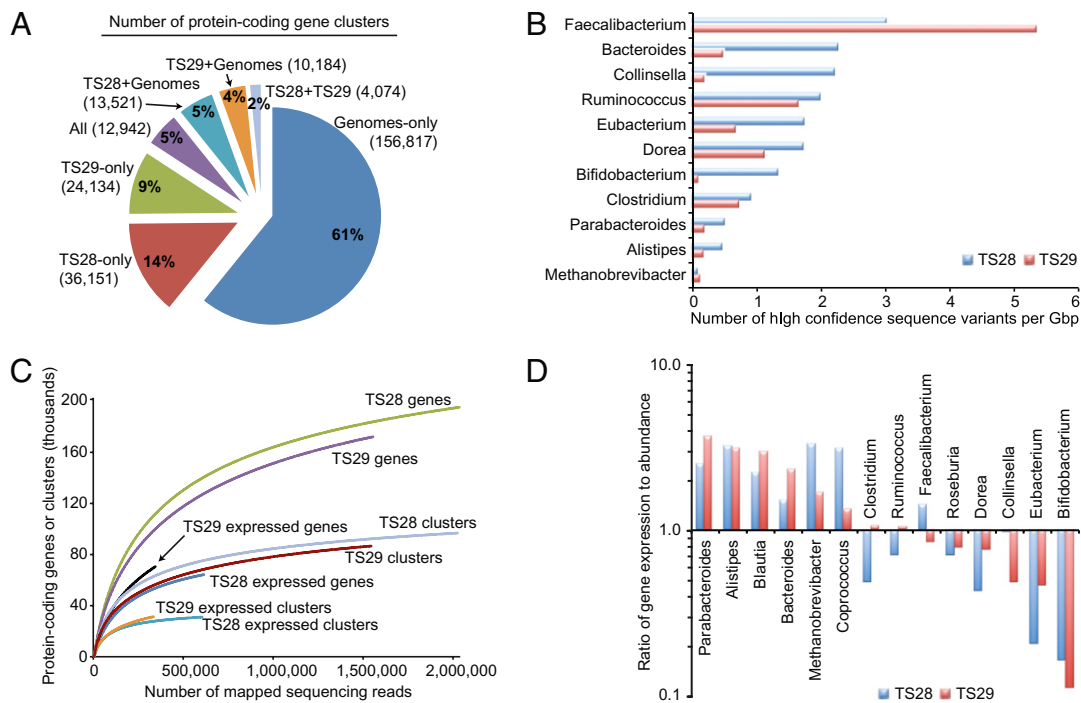
#### Deep Shotgun Sequencing of the Fecal Microbiome of the MZ Cotwins:

**Analyses of Genus-level Phylogenetic Bins.** We turned next to the following questions: Does deep sequencing enable the assembly and binning of “population genomes” from complex microbial communities? How diverse is the gut microbiome in terms of gene content, and how unique are these genes relative to those contained in 122 genomes from cultured human gut isolates? What can we infer about the similarities and differences between MZ cotwins when interrogating their deeply sequenced microbiomes?

Deep shotgun sequencing of total fecal community DNA allowed us to assemble and bin large scaffolds from the TS28 and TS29 microbiomes (Tables S5 and S6 and *Phylogenetic binning of microbiome scaffolds* in *SI Text*). A combined assembly of single- and paired-end pyrosequencing reads from TS28 and TS29 yielded 92,104 and 61,460 contigs  $>500$  bp per sample, with 11,780 and 6,392 scaffolds, respectively (scaffolds represent one or more contigs ordered and oriented using paired-end reads). PhyloPythia, a phylogenetic classifier that uses a multiclass Support Vector Machine (SVM) for composition-based characterization of sequence fragments at different taxonomic ranks (7), was trained on 1,775 finished or draft microbial genomes, in addition to 5,548 and 3,391 contigs from TS28 and TS29, respectively, that mapped with high confidence to gut microbial genomes (Table S7). After training, PhyloPythia was used to accurately bin all scaffolds  $>2$  Kbp at the genus- and family-level, resulting in 24–25 bins of scaffolds per fecal sample; these bins contained from 2.0 Kbp to 22.4 Mbp of total sequence (Figs. S2B and S3 and Table S6).

The total number of genes across all microbiome bins from the TS28 and TS29 fecal samples was 88,316 and 64,453, respectively. Clustering of protein sequences from these bins and the 122 gut microbial genomes, revealed 180,550, 257,823, and 334,211 total protein-coding gene clusters at 40%, 60%, and 80% identity cut-offs, respectively (Fig. 2A and Fig. S24). The largest group of gene clusters at all cutoffs was unique to the reference genomes, whereas 25% of the clusters were found only in the TS28 or TS29 microbiome bins. Overall, 36% of the gut microbiome gene clusters had a representative (60%ID) in the 122 gut microbial genome database, indicating that although sequencing reference genomes from culturable members of the microbiota has already uncovered a substantial proportion of the gene content present in the fecal communities of these cotwins, more reference genome and microbiome sequencing is clearly needed.

A total of 25 genus- and family-level bins were identified in the TS28 fecal microbiome dataset, and 24 in the TS29 dataset; 22 of these bins were found in both samples (bins unique to one sample only contained nine of the 16,554 total scaffolds). There were strong correlations between the two fecal microbiomes with respect to the number of scaffolds, their aggregate length, and the number of genes found in each bin ( $R^2 = 0.94, 0.74, \text{ and } 0.69$ , respectively; Table S6). As expected from our bacterial 16S rRNA analyses, the genus-level bins with the largest number of scaffolds were the *Ruminococcus*, *Bacteroides*, *Clostridium*, and *Eubacterium* (members of the Bacteroidetes and Firmicutes phyla). However, substantial assemblies were also obtained from *Methanobrevibacter* [*M. smithii* is reported to be the dominant archaeon in the human gut microbiome; (8)] and from *Bifidobacterium* (the former is missed with primers for amplification of bacterial 16S rRNA genes, whereas the current version of V2-directed bacterial primers miss members of the latter



**Fig. 2.** Diversity of the human fecal microbiome and its metatranscriptome. (A) Distribution of gene clusters across gut microbial genomes and microbiome bins. All protein sequences from 122 gut genomes and the microbiome bins were clustered using cd-hit at 60%ID. (B) Number of sequence variants in each microbiome bin (values normalized by Gbp in bin; all genus-level bins with >100 scaffolds are shown). (C) Rarefaction analysis of the number of genes, gene clusters, expressed genes, and expressed gene clusters in the fecal microbial communities of TS28 and TS29 as a function of sequencing depth. The total number of protein-coding genes in the set of 122 gut genomes and the microbiome bins is 525,329, representing 257,823 gene clusters. (D) Ratio of gene expression to gene abundance (relative abundance of cDNA sequences divided by relative abundance of DNA sequences) mapped to a subset of the bacterial taxa in the fecal microbiome. Taxa with >1,000 mapped cDNA and DNA sequencing reads in both samples are shown.

taxa; Fig. S4). When sequencing reads from each sample were mapped to the microbiome bins from that sample to identify high-confidence sequence variants in each bin, we found that the *Faecalibacterium* had the highest relative level of variation, whereas the *Methanobrevibacter* had the lowest (Fig. 2B).

Taken together, these results suggest that “population genomes” can be constructed and reliably binned even from diverse microbial communities given enough sequencing depth, although rare members of the community will be missed (e.g., the TM7 phylum). The bins provided the basis for a more in-depth analysis, annotation, and transcriptional profiling than a standard gene-centric (i.e., sequencing read-based) approach, revealing 36,151 and 24,134 gene clusters unique to TS28 and TS29, respectively, and not represented in any of the 122 reference gut genomes (Fig. 2A). Comparisons of the abundance of shared clusters between TS28 and TS29 revealed a stronger average correlation than the shared species-level phylotypes (mean  $R^2 = 0.37$  vs.  $R^2 = 0.18$ ). Rarefaction analysis disclosed that the number of genes and gene clusters in the gut microbiomes continues to increase even after 2 million mapped reads (Fig. 2C), with an estimated plateau of 242,023 and 234,661 genes, corresponding to 115,216 and 112,522 gene clusters in the TS28 and TS29 fecal microbiomes, respectively (Table S8).

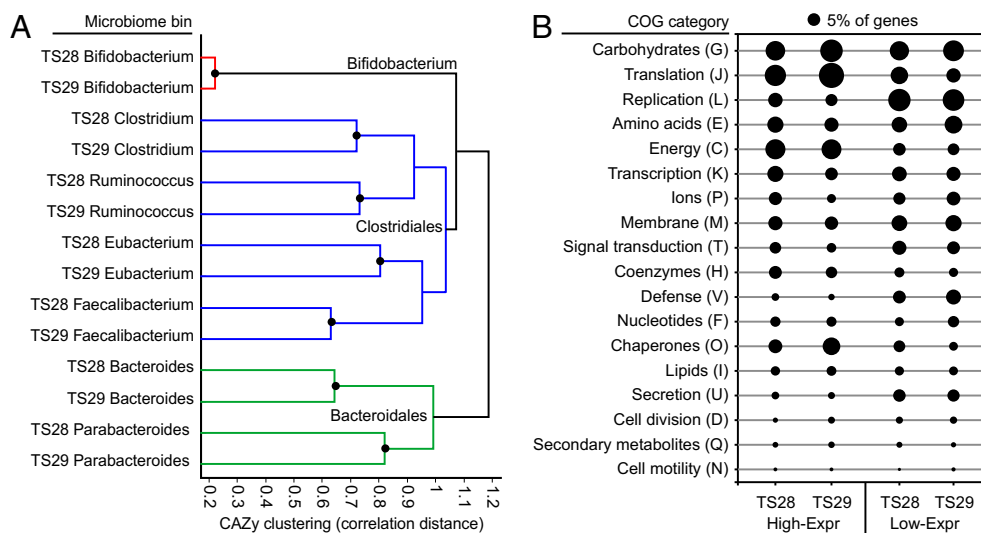
#### The Diversity of Carbohydrate Active Enzymes in the Human Gut Microbiome and Evidence of Genes with Predicted Cellulolytic Activity.

The human genome lacks the large repertoire of glycoside hydrolases and polysaccharide lyases required to cleave the many glycosidic linkages present in complex dietary polysaccharides (9). Because processing of these polysaccharides is a major function of the distal gut microbiota (10), we annotated the predicted proteins from each genus- and family-level microbiome bin using procedures described in the Carbohydrate-Active EnZyme database

[CAZy (9)] (Table S9 and S10 and Fig. S5). In total, we observed 143 CAZy families representing 5,145 genes in the gut microbiomes of these cotwins.

In general, the relative abundance of genes assigned to each CAZy family was consistent across genus-level bins from both individuals (Fig. 3A and Table S10). However, one notable exception was found: the *Faecalibacterium* bin from TS28 contained 42 genes predicted to encode dockerins, which are small proteins involved in the assembly of extracellular cellulosomes (11). None of these genes were identified in the *Faecalibacterium* bin from her cotwin’s fecal microbiome, nor in the genome of *F.prausnitzii* isolate M21/2. However, 30 dockerins were identified across the *Ruminococcus* and *Eubacterium* bins of the two samples (Table S9). In agreement with the predicted formation of cellulosomes, the *Faecalibacterium* dockerins from TS28 were found with a number of genes predicted to encode cellulases (GH5, GH9, GH44, GH48), beta-mannanases (GH26), xyloglucanases (GH74), and polysaccharide lyases (PL), none of which were observed in the *Faecalibacterium* bin from TS29 ( $\chi^2$  test,  $P < 10^{-4}$ ). Finally, a cohesin-encoding gene (the cognate molecule for dockerins) was identified in the *Faecalibacterium* bin from TS28, further supporting the existence of human gut cellulosomes.

To assess the distribution of genes predicted to encode dockerins across microbiomes from other twins, we compared 18 fecal microbiome datasets (mean  $\pm$  SEM 535,232  $\pm$  23,294 sequencing reads per sample; 118.7  $\pm$  8.7 Mb/sample) obtained from six MZ twin-pairs and their mothers (4) to the protein-coding gene sequences from the microbiome bins obtained from the deeply sampled MZ twins. This analysis revealed that the identified dockerin-encoding genes are widely distributed across gut microbiomes but vary in abundance: all 18 microbiomes contained reads with significant sequence similarity to these genes (mean number of genes 12.4, range 1–55 genes; and mean number of sequencing reads



**Fig. 3.** Clustering of fecal microbiome bins and the annotation of differentially expressed genes. (A) UPGMA clustering was performed on the relative abundance of CAZY families across each microbiome bin. Number of genes assigned to each CAZY family was normalized to the total number of genes in each sequence bin (all bins with >30 CAZY family assignments in both samples are shown). Black circles represent clustered nodes after z-score normalization across all bins (inconsistency threshold = 0.75, “cluster” function in Matlab v7.7.0). (B) Percentage of genes with high relative expression (High-Expr) or low relative expression (Low-Expr) assigned to each COG category. Percentages are represented by the area of each circle (black circle labeled 5% provides reference).

54.3, range 2–222 reads). However, only sequences from TS28 contained reads matching the identified cohesin-encoding gene.

Together, these results expand the known diversity of CAZymes in the human gut microbiome and reveal a suite of genes with predicted cellulolytic activity. The fact that the latter genes were highly enriched in the Faecalibacterium bins found in the microbiome of TS28 and not in her genetically identical cotwin highlights another level of genetic variation between humans. Future research will be necessary to characterize the enzymatic activity of these systems, the breadth of their organismal distribution, the host and environmental parameters (including diet) that determine their abundance in a given human gut microbiome, and their contributions to host nutrient/energy harvest.

**The Metatranscriptome Viewed from the Perspective of Phylogenetic Bins.** To characterize gene expression in the gut microbiome, we analyzed cDNA and DNA datasets obtained from sequencing total community cDNA and DNA prepared from the two fecal samples of TS28 and TS29. All sequencing reads were mapped against the database of 122 gut microbial genomes and the microbiome bins (*Metatranscriptome analysis in SI Text*). The results revealed marked differences in gene abundance and expression (Figs. S6 and S7). In all cases, technical replicates of each microbiome and metatranscriptome ( $n = 3-4$ ) clustered together; this clustering was robust to subsampling by COG functional categories (Fig. S6). Microbiome profiles showed the highest average correlation between individuals ( $R^2 = 0.37$ ), relative to metatranscriptomes ( $R^2 = 0.12$ ) and the relative abundance of species-level phylotypes ( $R^2 = 0.18$ ). As with the microbiome, rarefaction analysis of the metatranscriptome revealed that the number of expressed genes and gene clusters continues to increase even after 500,000 mapped reads (Fig. 2C), with an estimated plateau of 85,099 and 173,309 genes, corresponding to 35,781 and 58,339 gene clusters in TS28 and TS29, respectively (Table S8).

We subsequently calculated the ratio of the relative abundance of cDNA sequences in each microbiome bin to the relative abundance of DNA sequences in that bin, for each fecal community (12). Even at the genus-level, there were detectable differences in relative gene expression: six bins showed higher relative expression than gene abundance, whereas the Bifidobacterium had the lowest level of relative expression in both microbiomes (Fig. 2D).

We then compared cDNA and DNA profiles at the level of individual genes to determine the relative expression of each gene compared with its abundance (12). Genes were defined as “High Relative Expression” (High-Expr) or Low-Expr based on the ratio of cDNA to DNA relative abundance. A 10-fold difference was

chosen as the threshold cutoff based on all pairwise comparisons of technical replicate datasets obtained from cDNA or DNA sequencing of each sample (Fig. S8A,  $n = 3-4$  replicates per sample per method).

These comparisons revealed 6,961 genes with high or low relative expression in the fecal microbiome of TS28 (4,816 High-Expr and 2,145 Low-Expr) and 7,893 genes in TS29 (5,476 High-Expr and 2,417 Low-Expr; Tables S11 and S12). As expected, many of these genes came from bins with an overall higher relative expression (Fig. 2D), including Parabacteroides, Alistipes, Methanobrevibacter, and Bacteroides, or bins with a lower relative expression (the Bifidobacterium bin contained 962 Low-Expr genes in sample TS29 and 112 in TS28). However, some notable exceptions were found; the Bacteroides had 1,416 High-Expr genes in the TS28 microbiome, despite having overall similar levels of cDNA and DNA assignments across the entire bin (ratio 1.5).

The distribution of genes assigned to COG functional categories was then calculated using each set of High- or Low-Expr genes (Fig. 3B), as well as the set of genes that were observed only with cDNA or DNA sequencing (Fig. S9A). A disproportionate number of High-Expr genes encoded hypothetical proteins without predicted functions [33.9% (TS28) and 31.2% (TS29) of the High-Expr genes, comprising 77.9% (TS28) and 75.1% (TS29) of the total hypothetical genes with either a high or low relative expression]. High-Expr genes from both microbiomes were more frequently assigned to COG categories for translation (J), energy metabolism (C), and chaperones (O) (Fig. 3B and Tables S11 and S12), whereas Low-Expr genes were more frequently assigned to COG categories for secretory systems (U), replication, recombination, and repair (L), and membrane proteins (M) (Fig. 3B). In addition, many of these High-Expr genes have predicted functions related to fermentation and carbohydrate metabolism: e.g., ABC-type transport systems for carbohydrate import and metabolism plus genes involved in methanogenesis and acetogenesis (key pathways in the clearance of the hydrogen end-product of fermentation, and thus important determinants of fermentation efficiency).

To better characterize specific pathways represented by genes with high or low relative expression, we annotated each gene in the 122 gut microbial genomes and the microbiome bins using the KEGG annotation scheme (v52) (13). The relative abundance of KEGG pathways was tallied across genes defined as High- or Low-Expr in TS28 and TS29 or found to be unique to the cDNA or DNA datasets, and used for UPGMA clustering. Both microbiomes showed consistent trends, including high relative expression of genes assigned to pathways for essential cell processes, e.g., “RNA polymerase,” “Ribosome,” “Pyruvate metabolism,” and

“Glycolysis” (Fig. S8B). We extended these analyses to five additional samples from two sets of MZ cotwins and one unrelated individual (Samples labeled “TSDA” in Fig. S9 and *Additional microbiomes and meta-transcriptomes* in *SI Text*) and found similar results, including the higher relative expression of genes assigned to COG categories for transcription, energy metabolism, defense mechanisms, and chaperones (Fig. S9A), in addition to KEGG pathways involved in carbohydrate metabolism (e.g., fructose/ mannose metabolism), nucleotide metabolism, and vitamin metabolism/biosynthesis (e.g., folate biosynthesis) (Fig. S9B).

**Prospectus.** Our results indicate that a majority of species-level phylotypes are shared between these deeply sampled MZ cotwins, despite large variations in the abundance of each phylotype. The genetic and transcriptional diversity of the human gut microbiome is remarkable. Much of this diversity has not been previously identified through sequencing cultured human gut isolates; 64% of the gene clusters present in our microbiome bins had no representative in a set of 122 human gut microbial genomes, and only 17% were shared between the two cotwins. This diversity, even between genetically identical individuals, provides an expanded view of our multicellularity and interpersonal genetic variation. Features of the genus-level bins within the gut microbiome were distinctive in many ways, ranging from differences in gene content and transcriptional activity, to the extent of sequence variation within each population. Identifying the factors that determine such between-taxon differences will provide an important step toward understanding the functions (niches) of these organisms in the human gut microbial community, with the ultimate goal of linking the presence of specific organisms to gene content and activity. Our results and the accompanying datasets also provide a framework for future studies of human and environmental microbiomes. As noted above, 16S rRNA gene sequence datasets can be used to prioritize genomes for isolation and sequencing, starting with the most abundant phylotypes found across the most individuals, and working toward the rare members of the gut microbiota. The reduced level of organismal diversity in a single individual implies that it may be soon be possible to identify all strains present in a single gut (fecal) microbiota. The fraction of shared phylotypes between MZ cotwins, between unrelated individuals, and between body habitats provides an important context for designing studies of the assembly, dynamic operations, and host effects of “model” human gut microbiota/microbiomes, composed of sequenced cultured gut isolates, in gnotobiotic mice. Finally, the application of transcriptional profiling to the study of human body habitat-associated microbial communities will enable correlations to be made between genes expressed by our microbiomes and our physiologic and metabolic phenotypes.

## Materials and Methods

**Sequencing of 16S rRNA Gene Amplicons.** Fecal samples were stored at  $-80^{\circ}\text{C}$  before processing. DNA was extracted by bead beating followed by phenol-chloroform extraction as described previously (4). The V2 region was targeted for amplification by PCR (with primers 8F-338R) and multiplex GS FLX pyrosequencing (4). In addition, six control pools were constructed with equimolar or variable concentrations of purified genomic DNA from 67 cultured reference human gut-derived strains; the V2 regions of 16S rRNA genes present in these pools were then amplified and sequenced.

**Assembly of the Human Gut Microbiome.** Shotgun sequencing runs were performed on libraries prepared from total fecal community DNA using the 454 GS FLX Titanium single- and paired-end protocols. For all analyses involving unassembled reads, sequencing reads with degenerate bases (“Ns”) were removed along with all replicate sequences using the following parameters: 0.9 (90%ID), length difference requirement = 0, and 3 beginning bases checked (14). Each deeply sequenced dataset (TS28 and TS29) was assembled separately using the 454 GS de novo assembler software (Newbler v2.0.00.22), and all scaffolds were used for subsequent analysis. High-confidence sequence variants were identified using the 454 GS Reference Mapper software (v2.0.00.20).

**Metatranscriptome Analysis.** Microbial RNA sequencing (RNA-Seq) was performed as described previously (5). Briefly, total RNA was extracted from each fecal sample. The sample was subjected to rigorous DNase digestion to remove residual gDNA, depleted for rRNA and tRNA, converted to cDNA, and sequenced using the Illumina GALL platform. A total of 36 nucleotide reads produced from the each run were trimmed at their beginning and ends to remove bases with a quality score  $<20$ . Adapter sequences and sequencing reads with a length  $<20$  nucleotides were subsequently eliminated from further analysis. All trimmed reads were mapped with SSAHA2 (15) to phylogenetic bins constructed from microbiome scaffolds and to 122 sequenced human gut-associated microbial genomes (SSAHA2 parameters: -best 1 -score 20 -solexa). Gene clusters were defined by grouping all protein sequences from the database using the program cd-hit [parameter -c 0.6 -n 4 (16)]. Gene and gene cluster counts were normalized based on the total number of mapped sequencing reads. Genes from the database with significant homology (BLASTN e-value  $<10^{-30}$ ) to non-coding transcripts from the 122 gut microbial genomes were excluded from subsequent analysis. Ties representing sequences matching multiple reference genes with the same score were split evenly, whereas ties matching multiple gene clusters were weighted according to the frequency of unique (nontie) matches to each cluster.

Details concerning (i) phylogenetic binning of microbiome scaffolds, (ii) analysis of gene, bin, and transcript abundance, (iii) development and validation of methods for 16S rRNA gene sequence analysis, and (iv) additional cohorts of humans analyzed are given in *SI Text*.

**ACKNOWLEDGMENTS.** We thank S. Wagoner, J. Manchester, J. Hoisington-López, A. Heath, S. Marion, and D. Riches for their technical and other assistance during various phases of this work, as well as A. Goodman and A. Reyes for many helpful suggestions. This work was supported in part by National Institutes of Health Grants DK78669 and DK70977 (to J.I.G.), the Crohn's and Colitis Foundation of America, and Groupe Danone.

- Lahr DJ, Katz LA (2009) Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* 47:857–866.
- Quince C, et al. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 6:639–641.
- Kunin V, Engelbrekton A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. *Environ Microbiol* 12:118–123.
- Turnbaugh PJ, et al. (2009) A core gut microbiome in obese and lean twins. *Nature* 457:480–484.
- Turnbaugh PJ, et al. (2009) The effect of diet on the human gut microbiome: A metagenomic analysis in humanized gnotobiotic mice. *Sci Transl Med*, 1: 6ra14.
- Costello EK, et al. (2009) Bacterial community variation in human body habitats across space and time. *Science* 326:1694–1697.
- McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* 4:63–72.
- Eckburg PB, et al. (2005) Diversity of the human intestinal microbial flora. *Science* 308:1635–1638.
- Cantarel BL, et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): An expert resource for Glycogenomics. *Nucleic Acids Res* 37:D233–D238.
- Flint HJ, Bayer EA, Rincon MT, Lamed R, White BA (2008) Polysaccharide utilization by gut bacteria: Potential for new insights from genomic analysis. *Nat Rev Microbiol* 6:121–131.
- Bayer EA, Lamed R, White BA, Flint HJ (2008) From cellulosomes to cellulosomes. *Chem Rec* 8:364–377.
- Frias-Lopez J, et al. (2008) Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci USA* 105:3805–3810.
- Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32:D277–D280.
- Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J* 3:1314–1317.
- Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: A fast search method for large DNA databases. *Genome Res* 11:1725–1729.
- Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.

## Appendix C

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

“QIIME allows analysis of high-throughput community sequencing data.”

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

## QIIME allows analysis of high-throughput community sequencing data

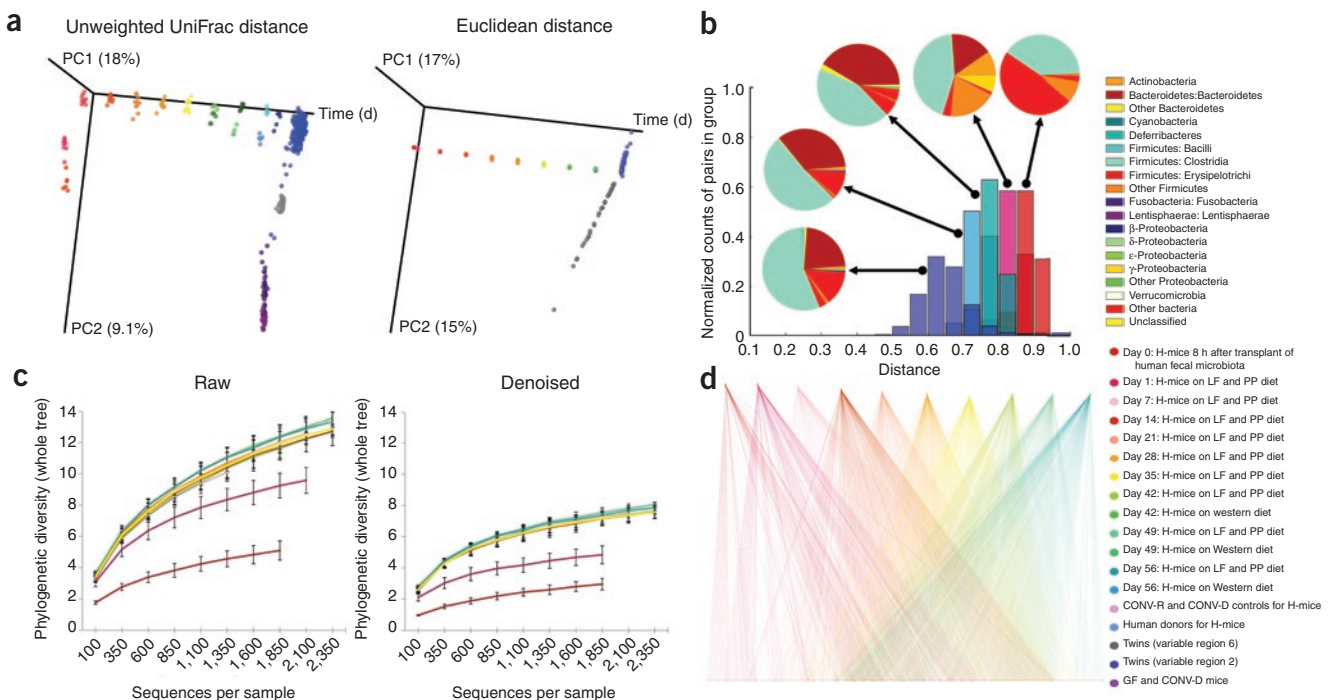
**To the Editor:** High-throughput sequencing is revolutionizing microbial ecology studies. Efforts like the Human Microbiome Projects<sup>1</sup> and the US National Ecological Observatory Network<sup>2</sup> are helping us to understand the role of microbial diversity in habitats within our own bodies and throughout the planet.

Pyrosequencing using error-correcting, sample-specific barcodes allows hundreds of communities to be analyzed simultaneously in multiplex<sup>3</sup>. Integrating information from thousands of samples, including those obtained from time series, can reveal large-scale patterns that were inaccessible with lower-throughput sequencing methods. However, a major barrier to achieving such insights has been the lack of software that can handle these increasingly massive datasets. Although tools exist to perform library demultiplexing and taxonomy assignment<sup>4,5</sup>, tools for downstream analyses are scarce.

Here we describe ‘quantitative insights into microbial ecology’ (QIIME; pronounced ‘chime’), an open-source software pipe-

line built using the PyCogent toolkit<sup>6</sup>, to address the problem of taking sequencing data from raw sequences to interpretation and database deposition. QIIME, available at <http://qiime.sourceforge.net/>, supports a wide range of microbial community analyses and visualizations that have been central to several recent high-profile studies, including network analysis, histograms of within- or between-sample diversity and analysis of whether ‘core’ sets of organisms are consistently represented in certain habitats. QIIME also provides graphical displays that allow users to interact with the data. Our implementation is highly modular and makes extensive use of unit testing to ensure the accuracy of results. This modularity allows alternative components for functionalities such as choosing operational taxonomic units (OTUs), sequence alignment, inferring phylogenetic trees and phylogenetic and taxon-based analysis of diversity within and between samples (including incorporation of third-party applications for many steps) to be easily integrated and benchmarked against one another (Supplementary Fig. 1).

We applied the QIIME workflow to a combined analysis of previously collected data (see Supplementary Discussion) for distal gut bacterial communities from conventionally raised mice, adult



**Figure 1** | QIIME analyses of the distal gut microbiotas of conventionally raised and conventionalized mice, gnotobiotic mice colonized with a human fecal gut microbiota (H-mice), and human adult mono- and dizygotic twins. **(a)** Principal coordinates analysis plots for mice, H-mice and twins. Colors correspond to separate samples by species and time point, and are consistent throughout the panels. **(b)** Unweighted UniFrac distance histograms between the data for fecal microbiota of human twins; human donors for the H-mice study; day 56 post-transplant H-mice on a low-fat (LF) and plant polysaccharide-rich (PP) diet; day 1 H-mice (LF and PP diet); and day 0 H-mice. Taxonomic classifications are presented at the class level. **(c)** Alpha diversity rarefaction plots of phylogenetic diversity for the H-mice samples. **(d)** OTU network connectivity of H-mice time series data. CONV-D, conventionalized mice; CONV-R, conventionally raised mice; and GF, germ-free mice.



human monozygotic and dizygotic twins and their mothers, and a time series study of adult germ-free mice after they received human fecal microbiota (Fig. 1, Supplementary Table 1 and Supplementary Discussion). This analysis combined ten full 454 FLX runs and one partial run, totalling 3.8 million bacterial 16S rRNA sequences from previously published studies, including reads from different regions of the 16S rRNA gene.

QIIME is thus a robust platform for combining heterogeneous experimental datasets and for rapidly obtaining new insights about various microbial communities. Because QIIME scales to millions of sequences and can be used on platforms from laptops to high-performance computing clusters, we expect it to keep pace with advances in sequencing technology and to facilitate characterization of microbial community patterns ranging from normal variations to pathological disturbances in many human, animal and other environmental ecosystems.

Note: Supplementary information is available on the Nature Methods website.

#### ACKNOWLEDGMENTS

We thank our collaborators for their helpful suggestions on features, documentation and the manuscript, and our funding agencies for their commitment to open-source software. This work was supported in part by Howard Hughes Medical Institute and grants from the Crohn's and Colitis Foundation of America, the German Academic Exchange Service, the Bill and Melinda Gates Foundation, the Colorado Center for Biofuels and Biorefining and the US National Institutes of Health (DK78669, GM65103, GM8759, HG4872 and its ARRA supplement, HG4866, DK83981 and LM9451).

#### COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at <http://www.nature.com/naturemethods/>.

J Gregory Caporaso<sup>1,12</sup>, Justin Kuczynski<sup>2,12</sup>, Jesse Stombaugh<sup>1,12</sup>, Kyle Bittinger<sup>3</sup>, Frederic D Bushman<sup>3</sup>, Elizabeth K Costello<sup>1</sup>, Noah Fierer<sup>4</sup>, Antonio Gonzalez Peña<sup>5</sup>, Julia K Goodrich<sup>5</sup>, Jeffrey I Gordon<sup>6</sup>, Gavin A Huttley<sup>7</sup>, Scott T Kelley<sup>8</sup>, Dan Knights<sup>5</sup>, Jeremy E Koenig<sup>9</sup>, Ruth E Ley<sup>9</sup>, Catherine A Lozupone<sup>1</sup>, Daniel McDonald<sup>1</sup>, Brian D Muegge<sup>6</sup>, Meg Pirrung<sup>1</sup>, Jens Reeder<sup>1</sup>, Joel R Sevinsky<sup>10</sup>, Peter J Turnbaugh<sup>6</sup>, William A Walters<sup>2</sup>, Jeremy Widmann<sup>1</sup>, Tanya Yatsunenkov<sup>6</sup>, Jesse Zaneveld<sup>2</sup> & Rob Knight<sup>1,11</sup>

<sup>1</sup>Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado, USA. <sup>2</sup>Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado, USA. <sup>3</sup>Department of Microbiology, University of Pennsylvania, Philadelphia, Pennsylvania, USA. <sup>4</sup>Cooperative Institute for Research in Environmental Sciences and Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado, USA. <sup>5</sup>Department of Computer Science, University of Colorado, Boulder, Colorado, USA. <sup>6</sup>Center for Genome Sciences, Washington University School of Medicine, St. Louis, Missouri, USA. <sup>7</sup>Computational Genomics Laboratory, John Curtin School of Medical Research, The Australian National University, Canberra, Australian Capital Territory, Australia. <sup>8</sup>Department of Biology, San Diego State University, San Diego, California, USA. <sup>9</sup>Department of Microbiology, Cornell University, Ithaca, New York, USA. <sup>10</sup>Luca Technologies, Golden, Colorado, USA. <sup>11</sup>Howard Hughes Medical Institute, Boulder, Colorado, USA. <sup>12</sup>These authors contributed equally to this work. e-mail: [rob.knight@colorado.edu](mailto:rob.knight@colorado.edu)

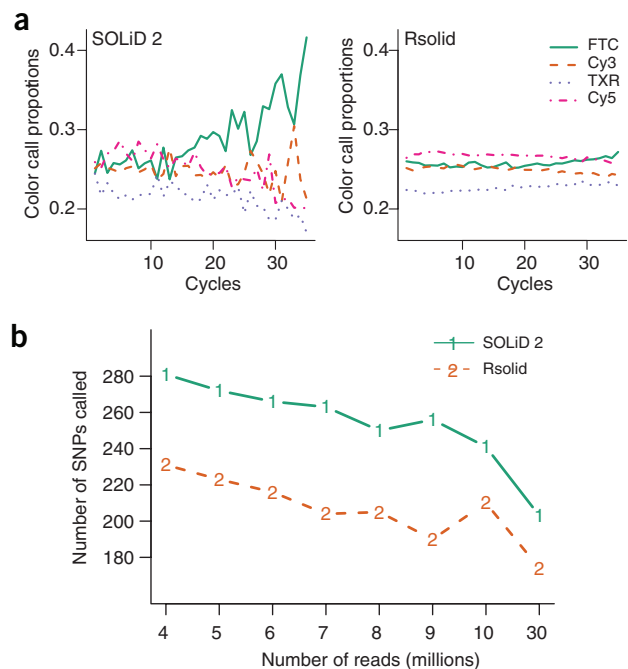
PUBLISHED ONLINE 11 APRIL 2010; DOI:10.1038/NMETH.F.303

- National Institutes of Health Human Microbiome Project Working Group *et al.* *Genome Res.* **19**, 2317–2323 (2009).
- Hopkin, M. *Nature* **444**, 420–421 (2006).
- Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J. & Knight, R. *Nat. Methods* **5**, 235–237 (2008).
- Cole, J.R. *et al.* *Nucleic Acids Res.* **37**, D141–D145 (2009).
- Schloss, P.D. *et al.* *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
- Knight, R. *et al.* *Genome Biol.* **8**, R171 (2007).

## Intensity normalization improves color calling in SOLiD sequencing

To the Editor: Applied Biosystems' SOLiD system<sup>1</sup> is a commonly used massively parallel DNA sequencing platform for applications from genotyping and structural variation analysis<sup>1</sup> to transcriptome quantification and reconstruction<sup>2</sup>. Like other sequencing technologies, it measures fluorescence intensities from dye-labeled molecules to determine the sequence of DNA fragments. Ultimately, sequences are determined by complicated statistical manipulations of noisy intensity measurements, and systematic biases may mislead downstream analysis<sup>3</sup>. Several proposed methods improve base calling and quality metrics for other sequencing technologies<sup>3–5</sup>, and we now present Rsolid, software implementing an intensity normalization strategy for the SOLiD platform that substantially improves yield and accuracy at small computational costs (6% increase in total matches, 13% increase in perfect matches, 5% reduced error rate and a substantial reduction in false positive single-nucleotide polymorphism (SNP) calls in an *Escherichia coli* genomic DNA sample).

In the SOLiD system, the proportions of color calls across sequencing cycles are extremely variable (Fig. 1a), even though they should be equal across sequencing cycles and proportional to the dinucleotide content of the library (Supplementary Methods). This bias can be traced to the fluorescence intensity measurements used to make the color calls (Supplementary Fig. 1). The distributions of intensities are similar across channels in early sequencing cycles, but a color bias starts to appear in later cycles. The Rsolid method uses a simple and computationally efficient procedure to normalize the color-channel



**Figure 1** | Effect of normalization on color proportions and SNP calling. (a) Color proportions in sample of *E. coli* genomic DNA on each sequencing cycle. Color calls as reported by the SOLiD 2 system (left) and after normalization by Rsolid (right). FTX, TXR, Cy3 and Cy5 are dyes used by SOLiD. (b) Number of false positive SNPs called in *E. coli* at various coverage. After normalization, fewer SNPs were called even at high coverage (30 M reads correspond to ~100-fold coverage).

## Appendix D

Garrett WS, Gallini CA, **Yatsunenko T**, Michaud M, DuBois A, Delaney ML, Punit S, Karlsson M, Bry L, Glickman JN, Gordon JI, Onderdonk AB, Glimcher LH.

“Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis.”

*Cell Host and Microbe*. **2010** Sep 16; 8(3): 292-300.

# Enterobacteriaceae Act in Concert with the Gut Microbiota to Induce Spontaneous and Maternally Transmitted Colitis

Wendy S. Garrett,<sup>1,2,\*</sup> Carey A. Gallini,<sup>1</sup> Tanya Yatsunenکو,<sup>3</sup> Monia Michaud,<sup>1</sup> Andrea DuBois,<sup>4</sup> Mary L. Delaney,<sup>4</sup> Shivesh Punit,<sup>1,6</sup> Maria Karlsson,<sup>3</sup> Lynn Bry,<sup>4</sup> Jonathan N. Glickman,<sup>4,7</sup> Jeffrey I. Gordon,<sup>3</sup> Andrew B. Onderdonk,<sup>4</sup> and Laurie H. Glimcher<sup>1,2,5,\*</sup>

<sup>1</sup>Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA 02115, USA

<sup>2</sup>Department of Medicine, Harvard Medical School, Boston, MA 02115, USA

<sup>3</sup>Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108, USA

<sup>4</sup>Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA

<sup>5</sup>Ragon Institute of MGH/Harvard/MIT, Boston, MA 02129, USA

<sup>6</sup>Present address: Saban Research Institute, Children's Hospital Los Angeles, Los Angeles, CA 90027, USA and Department of Genetics and Molecular Cellular Biology, University of Southern California, Los Angeles, CA 90089, USA

<sup>7</sup>Present address: GI Pathology, Boston Caris Diagnostics, 320 Needham Street, Suite 200, Newton, MA 02464, USA

\*Correspondence: wgarrett@hsph.harvard.edu (W.S.G.), lglimche@hsph.harvard.edu (L.H.G.)

DOI 10.1016/j.chom.2010.08.004

## SUMMARY

Disruption of homeostasis between the host immune system and the intestinal microbiota leads to inflammatory bowel disease (IBD). Whether IBD is instigated by individual species or disruptions of entire microbial communities remains controversial. We characterized the fecal microbial communities in the recently described *T-bet*<sup>-/-</sup> × *Rag2*<sup>-/-</sup> ulcerative colitis (TRUC) model driven by T-bet deficiency in the innate immune system. 16S rRNA-based analysis of TRUC and *Rag2*<sup>-/-</sup> mice revealed distinctive communities that correlate with host genotype. The presence of *Klebsiella pneumoniae* and *Proteus mirabilis* correlates with colitis in TRUC animals, and these TRUC-derived strains can elicit colitis in *Rag2*<sup>-/-</sup> and WT adults but require a maternally transmitted endogenous microbial community for maximal intestinal inflammation. Cross-fostering experiments indicated a role for these organisms in maternal transmission of disease. Our findings illustrate how gut microbial communities work in concert with specific culturable colitogenic agents to cause IBD.

## INTRODUCTION

The human intestine is populated with up to 10<sup>12</sup> microbes per gram of luminal contents. Coexistence with this microbial community (microbiota) demands a well-regulated homeostasis between the host immune system and the microbiota (Duerkop et al., 2009; Hill and Artis, 2009). Inflammatory bowel disease (IBD) can occur when this homeostasis is disrupted (Sartor, 2009). Whether individual pathogenic species or entire microbial communities instigate inflammation still remains controversial

(Frank and Pace, 2008; Hansen et al., 2010). Defining features of the microbiota and host that are associated with or initiate IBD is critical.

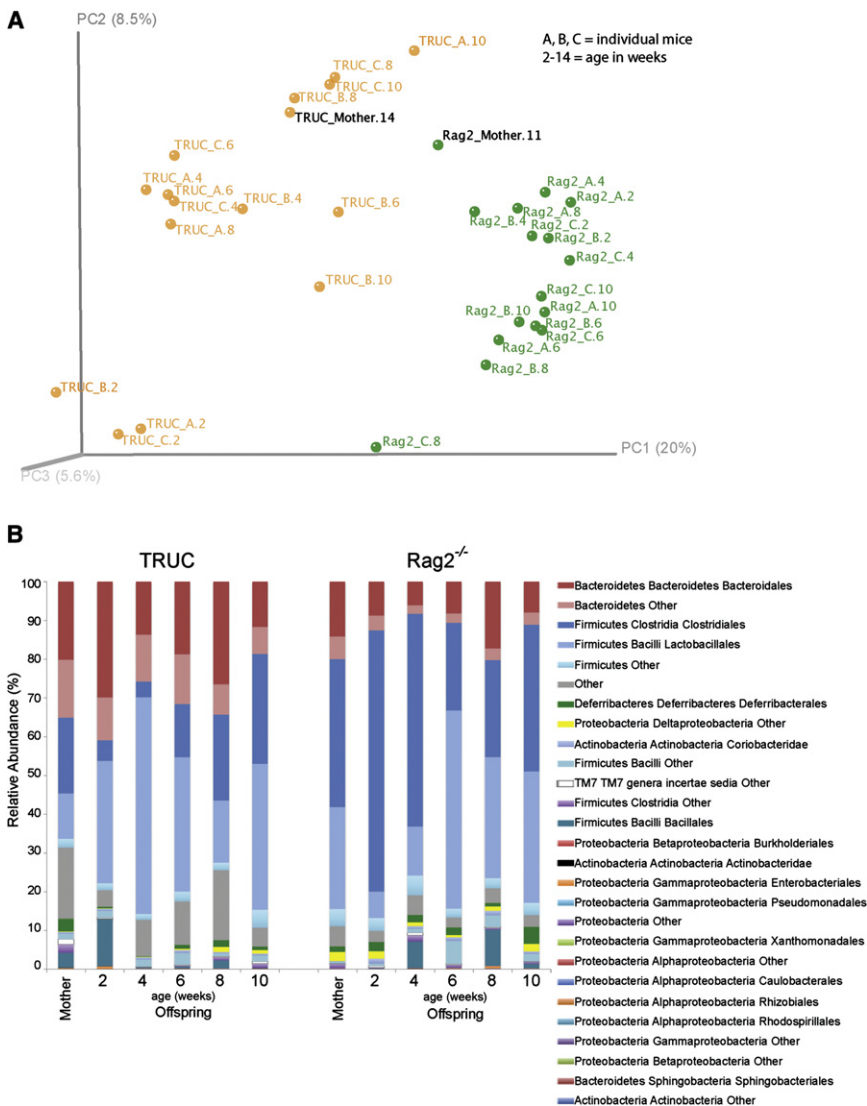
In the absence of adaptive immunity, loss of the transcription factor T-bet in conventionally raised *T-bet*<sup>-/-</sup> × *Rag2*<sup>-/-</sup> knockout mice results in a spontaneous and highly penetrant colitis that shares histologic features with ulcerative colitis in humans. *T-bet*<sup>-/-</sup> × *Rag2*<sup>-/-</sup> ulcerative colitis (TRUC) is associated with altered colonic barrier function, elevated TNF-α levels, and dysfunctional dendritic cells (DCs) (Garrett et al., 2007, 2009). It is transmissible to WT hosts when they are cross-fostered or cohoused with TRUC mice (Garrett et al., 2007). TRUC mice provide an opportunity to probe the host-microbe relationship in a model that displays both the immunodeficiency and hyperimmunity observed in humans with IBD.

Here, we show that the presence of *Proteus mirabilis* and *Klebsiella pneumoniae* correlates with colitis in TRUC mice and that TRUC-derived strains, in conjunction with an endogenous microbial community, incite colitis in WT mice. These studies revealed the utility of using both culture-independent and -dependent approaches to interrogate the contribution of community members to disease pathogenesis. This model also provides a foundation for defining how gut microbial communities work in concert with specific culturable colitogenic agents to cause IBD and creates an opportunity to evaluate preventative or therapeutic measures directed at components of the gut microbiota and/or host.

## RESULTS

### 16S rRNA-Based Time Series Analysis of TRUC versus *Rag2*<sup>-/-</sup> Fecal Microbiota

A pilot experiment—using offspring of conventionally raised, specified-pathogen-free (SPF) *T-bet*<sup>-/-</sup> × *Rag2*<sup>-/-</sup> and *Rag2*<sup>-/-</sup> mothers—analyzed fecal samples collected from mothers at a single time point and from their female pups (n = 3/genotype)



**Figure 1. 16S rRNA-Based Time Series Analysis of TRUC versus *Rag2*<sup>-/-</sup> Fecal Microbiota**

(A) Principal coordinates analysis (PCoA) of unweighted UniFrac distances from 2- to 10-week-old TRUC ( $n = 3$ ) and *Rag2*<sup>-/-</sup> ( $n = 3$ ) mice and their mothers. Host genotype influences microbial community structure. Abbreviations: A, B, C, individual pups colored by genotype, followed over time (A.2, A.4, A.6, A.8, and A.10 refer to mouse A sampled at 2, 4, 6, and 10 weeks of age).

(B) Distribution of order-level phylotypes in TRUC and *Rag2*<sup>-/-</sup> fecal microbial communities. Relative abundance (%) is plotted for each age group.

### *K. pneumoniae* and *P. mirabilis* Correlate with the Presence of Colitis in TRUC Mice

We also performed quantitative cultures to obtain independent verification of differences in bacterial burden for defined species and to have culturable isolates available to test the specific effects of individual strains on disease initiation and progression. A total of 57 bacterial species were recovered and identified from fecal pellets obtained from three TRUC and three *Rag2*<sup>-/-</sup> mice surveyed at 2, 4, 6, 8, and 10 weeks of age and from their mothers (Figure 2A and Table S2).

Experiments administering oral antibiotics (Abx) helped further refine potential classes of colitogenic commensal organisms. Gentamicin (gent) or metronidazole (metro) but not vancomycin (vanco) were highly effective in ameliorating TRUC colitis and resulted in clinically and statistically significant changes in colitis scores (mean colitis score  $0.5 \pm 0.52$ ,  $p < 0.0001$

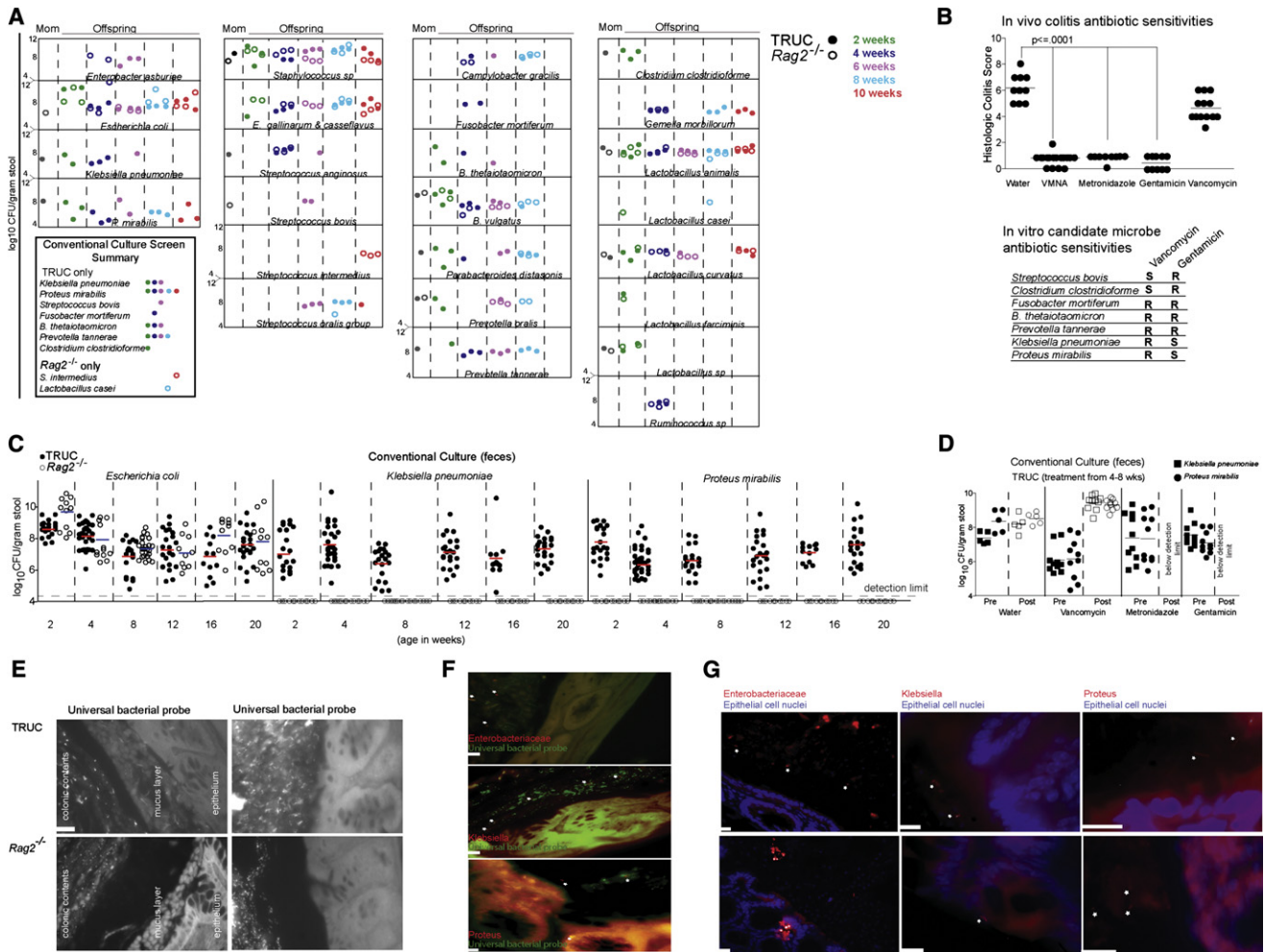
at 2 week time intervals from 2 to 10 weeks. A culture-independent survey of their fecal microbiota by multiplex pyrosequencing of V2 region amplicons of bacterial 16S rRNA genes ( $n = 32$  samples;  $2348 \pm 343$  reads/sample) analyzed by principal coordinates analysis (PCoA) plots based on unweighted UniFrac measurements disclosed a correlation between host genotype and community phylogeny at all ages surveyed (Figure 1A). A total of 69 species-level phylogenetic types, belonging to four major bacterial phyla, exhibited significant differences at various ages in the fecal communities of mice of the two genotypes (Table S1). Compared to *Rag2*<sup>-/-</sup> controls, TRUC mice had a significantly higher proportional representation of species-level operational taxonomic units (OTUs) belonging to the order *Bacteroidales* (phylum Bacteroidetes;  $p = 0.00643$ ) and significantly lower proportional representation of OTUs belonging to the orders *Clostridiales* (phylum Firmicutes;  $p = 0.0201$ ) and *Deltaproteobacteria* (phylum Proteobacteria;  $p = 0.0299$ ) ( $p$  values by Mann-Whitney test with Bonferroni correction) (Figure 1B).

compared to water control) (Figure 2B). This result pointed us to a role for Gram-negative facultative organisms in TRUC.

The in vitro Abx resistance profiles of the commensal strains selectively recovered from TRUC but not *Rag2*<sup>-/-</sup> fecal samples (Figures 2A and 2B, lower panel) corresponded to the in vivo Abx sensitivity of the colitis (Figure 2B, upper panel) since *K. pneumoniae* and *P. mirabilis*, both facultative enterics, were sensitive to gent but resistant to vanco (Figure 2B, lower panel).

A more extensive, culture-based survey of a larger number of TRUC and *Rag2*<sup>-/-</sup> mice to determine if these bacteria were present in afflicted but absent from healthy mice revealed that *K. pneumoniae* and *P. mirabilis* were culturable in all TRUC mice tested ( $n = 126$ ) at every time point (Figure 2C). In contrast, both species were below our limit of detection ( $<4.4 \log_{10}$  cfu/g fecal material) in *Rag2*<sup>-/-</sup> mice at each time point (Figure 2C).

We treated 4-week-old TRUC mice with Abx using the protocol shown previously to ameliorate colitis and cultured feces obtained 1 day before and 1 day after Abx administration. *K. pneumoniae* and *P. mirabilis* fecal levels fell below our



**Figure 2. The Presence of *K. pneumoniae* and *P. mirabilis* Correlates with the Presence of Colitis in TRUC Mice**

(A) Culture-based identification of bacteria present in fecal samples from Figure 1 mice and time points. Species observed in >1 mouse or in 1 mouse at >1 time point are shown. Summary of species-level differences in the fecal microbiota of TRUC versus *Rag2*<sup>-/-</sup> mice is observed in the inset.

(B) Upper panel: In vivo Abx sensitivities of TRUC colitis. Each dot represents one mouse treated for 4 weeks with the indicated Abx. VMNA: vanco, metro, neomycin, and ampicillin. Horizontal bars represent the mean. p value ≤ 0.0001 by Mann-Whitney test. Lower panel: Summary of in vitro Abx sensitivities for species selectively detected in TRUC fecal microbiota.

(C) Culture-based survey of Gram-negative aerobes in fecal samples from TRUC (shaded circles) and *Rag2*<sup>-/-</sup> (open circles) at 2–20 weeks of age.

(D) In vivo sensitivity of *K. pneumoniae* (squares) and *P. mirabilis* (circles), as defined by culture-based surveys of TRUC fecal samples collected 1 day before (shaded symbol) and 1 day after (open symbol) Abx treatment. Each dot represents data from a fecal sample obtained from one mouse. Horizontal bars represent the mean value.

(E) FISH using an Oregon-Green 488-conjugated “universal bacterial” 16S rRNA-directed oligonucleotide probe (EUB338) demonstrates the presence of bacteria in the mucus layer and directly adjacent to the epithelium in TRUC mice. Upper panels, TRUC; lower panels, *Rag2*<sup>-/-</sup>. A 10 μm scale bar for the panel is shown in the lower left of the first image.

(F) *Enterobacteriaceae* (red), *Klebsiella* (red), and *Proteus* (red) were visualized adjacent to the epithelium in TRUC mice using Fluor-conjugated 16S rRNA or 23S rRNA oligonucleotide probes (pB-00914 [*Enterobacteriaceae*], pB-00352 [*Klebsiella pneumoniae*], pB-02110 [*Proteus mirabilis*]). Sections were also hybridized with the EUB338 universal bacterial probe (green). Scale bars (10 μm) are shown for each image.

(G) *Enterobacteriaceae* (red), *Klebsiella* (red), and *Proteus* (red) probe signals are seen adjacent to or along the epithelium in TRUC mice. Epithelial cell nuclei were stained with DAPI. White star symbols mark bacteria in (F) and (G). Scale bars (10 μm) are shown for each image.

limit of detection when mice were treated with gent or metro, but treatment with vanco neither abolished colitis nor reduced levels of these bacteria (Figure 2D).

To test the hypothesis that *K. pneumoniae* and *P. mirabilis* play a role in TRUC pathogenesis, we determined the physical location of these species using 16S and 23S rDNA fluorescence in situ hybridization (FISH) oligonucleotide probes on whole

colonic sections. We focused on the degree of colonization in the lumen, the mucus layer over the epithelium, and the mucosa. As has been reported in IBD patients using a universal bacterial 16S rRNA FISH probe (Swidsinski et al., 2005), we observed that the colonic mucus of colitic TRUC mice harbored numerous bacteria and that there was a consistent loss of a “bacterial-free zone” adjacent to the colonic epithelium (Figure 2E). Healthy

(noncolitic) *Rag2*<sup>-/-</sup> mice did not exhibit any of these phenotypes (Figure 2E). Using a set of probeBase consortium 23S and 16S rDNA probes to detect *K. pneumoniae* and *P. mirabilis* (Loy et al., 2007), we visualized a small number of organisms adjacent to the epithelium (Figures 2F and 2G). Hence, *K. pneumoniae* and *P. mirabilis* may have invasive potential, or the proximity of their bacterial products to the apical epithelial surface may trigger inflammatory responses without frank invasion. Either could explain their role in TRUC colitis, as access to the mucosa would increase the opportunity for eliciting a host proinflammatory response.

### ***K. pneumoniae* and *P. mirabilis* Elicit Colitis but Require a Maternally Transmitted Endogenous Microbial Community for Maximal Intestinal Inflammation**

Following postnatal exposure to a TRUC dam, WT and *Rag2*<sup>-/-</sup> mice develop histopathologic features of colitis (penetration of phenotype: 94% at 8 weeks of age) (Garrett et al., 2007). We asked if this maternally transmitted disease had a pattern of Abx sensitivity similar to spontaneous TRUC colitis. We cross-fostered TRUC, *Rag2*<sup>-/-</sup>, and WT pups on TRUC mothers who received water, gent, metro, or vanco from preconception through weaning. Gent and metro markedly improved the colitis score for all mice in a statistically significant fashion, while vanco did not, similar to what we observed in spontaneous TRUC colitis (n = 2 foster mothers/genotype; 2–4 pups/litter surveyed) (Figure 3A).

We cultured fecal samples from WT and *Rag2*<sup>-/-</sup> mice that developed transmissible colitis from cross-fostering (Figure 3B). *K. pneumoniae* and *P. mirabilis* were detected in all fecal samples obtained from 8-week-old TRUC-fostered *Rag2*<sup>-/-</sup> and WT pups at levels comparable to age-matched TRUC-fostered TRUC mice. In contrast, neither organism was detected in any control *Rag2*<sup>-/-</sup>-fostered *Rag2*<sup>-/-</sup> or WT-fostered WT animals (n = 2 foster mothers/genotype; 2–3 pups/litter surveyed) (Figure 3B). Neither TRUC mice fostered on *Rag2*<sup>-/-</sup> nor WT mothers had histologic evidence of colitis or *K. pneumoniae* or *P. mirabilis* at 8 weeks of age (Figures S2 and 3B). The presence of *K. pneumoniae* and *P. mirabilis* in colitic TRUC mice and TRUC fostered *Rag2*<sup>-/-</sup> and WT mice and the lack of detectable bacteria in the fecal microbiota of healthy *Rag2*<sup>-/-</sup>, WT, and WT or *Rag2*<sup>-/-</sup>-fostered TRUC provided additional evidence for an association between the presence of these bacteria and colitis.

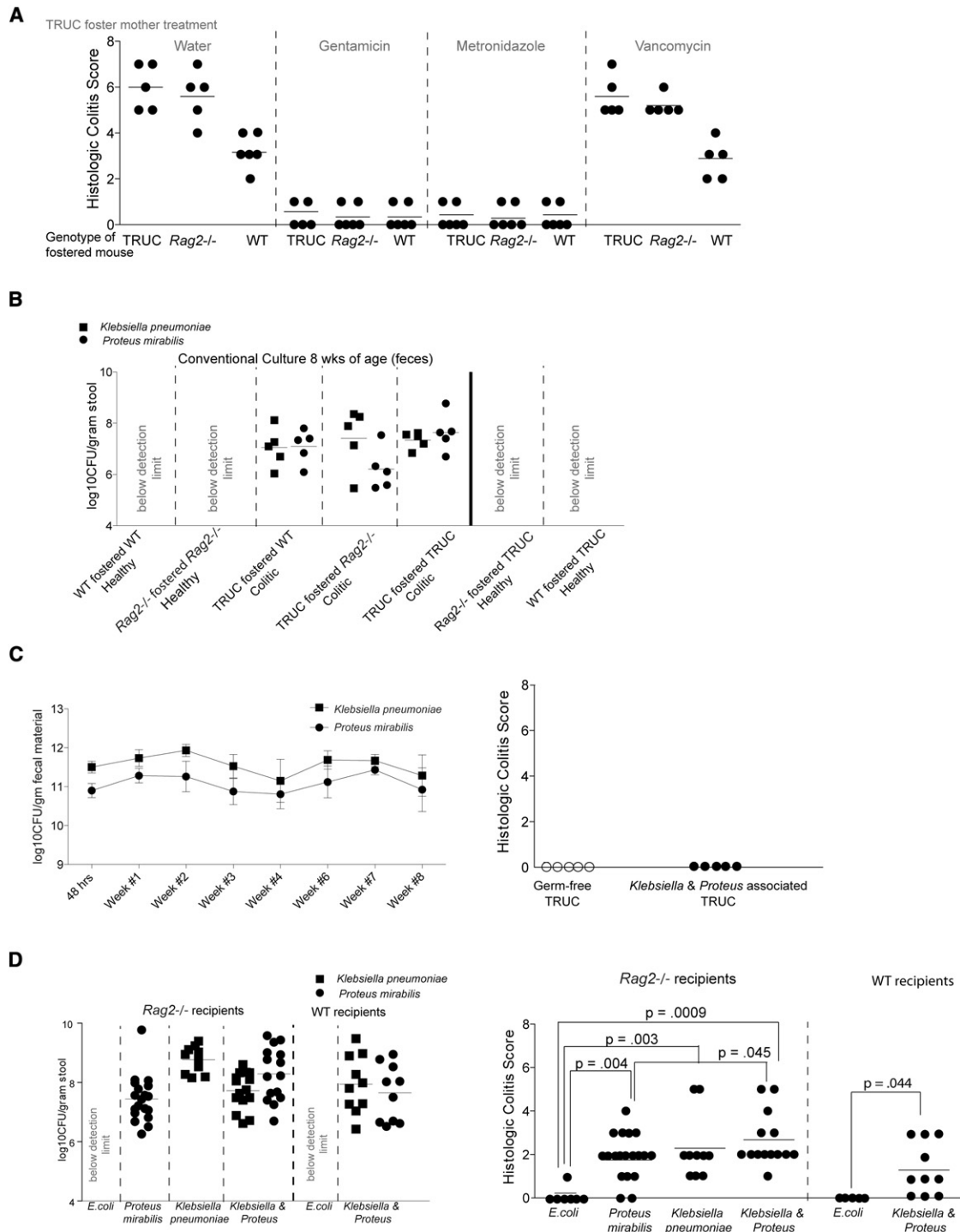
One possibility is that the presence of *K. pneumoniae* and *P. mirabilis* is a consequence rather than a cause of inflammation. Intestinal inflammation caused by *Citrobacter rodentium* may drive blooms of *Enterobacteriaceae*, although this result is controversial (Hoffmann et al., 2009; Lupp et al., 2007). To investigate the effects of inflammation on intestinal colonization by *K. pneumoniae* and *P. mirabilis*, we treated 8-week-old WT and *Rag2*<sup>-/-</sup> mice with the mucosal toxin dextran sulfate sodium to induce colitis (n = 8 mice/genotype). We did not detect culturable *K. pneumoniae* or *P. mirabilis* in the fecal microbiota of any of these mice during our period of surveillance (n = 8 mice/genotype; samples collected before and 1 day after completion of a 1 week treatment) (Figure S1), arguing against an inflammatory response causing expansion of *K. pneumoniae* and *P. mirabilis* in the TRUC gut microbiota.

To directly test the colitogenic potential of *K. pneumoniae* and *P. mirabilis*, we rederived conventionally raised TRUC mice as germ-free and cocolonized the mice with these two *Enterobacteriaceae* at 8 weeks of age for 8 weeks (n = 5 mice). Both organisms established themselves in the guts of all recipients (mean value 10<sup>11.29 ± 0.46</sup> cfu/microbial species/g dry weight of feces; assayed 48 hr and weekly after the initial gavage) (Figure S2). Colonic inflammation did not develop in these cocolonized gnotobiotic TRUC mice, suggesting that interactions among *K. pneumoniae*, *P. mirabilis*, and other members of a gut microbial community are required to ignite the immunoinflammatory cascade that leads to colitis. To evaluate this possibility, we colonized 2 week SPF WT and *Rag2*<sup>-/-</sup> mice with *K. pneumoniae*, *P. mirabilis*, or a combination of *K. pneumoniae* and *P. mirabilis* (recovered from feces from the female TRUC mother in Figure 2A and administered by direct oral instillation of 10<sup>7</sup> cfu and by addition of 10<sup>7</sup> cfu to the drinking water every other day for 8 weeks; n = 5–18 mice/treatment group). Control groups of mice received a TRUC-derived *E. coli* strain. Both *K. pneumoniae* and *P. mirabilis* established themselves in the gut microbiota of both *Rag2*<sup>-/-</sup> and WT (as defined by cfu assays of feces obtained 2 days after the completion of 8 weeks of treatment [Figure 3C]). Feces from WT and *Rag2*<sup>-/-</sup> hosts contain *E. coli*, but we did not have the tools to distinguish these indigenous strains from the exogenously administered TRUC-associated *E. coli* strain. While no colonic inflammation was observed after inoculation with *E. coli* (Figure 3D), treatment with *P. mirabilis*, *K. pneumoniae*, or a combination of the two organisms induced inflammation in both WT and *Rag2*<sup>-/-</sup> mice, with colitis severity being significantly greater in *Rag2*<sup>-/-</sup> mice exposed to both species compared to *P. mirabilis* alone (Figure 3D). We conclude that two *Enterobacteriaceae*, in concert with members of the microbiota, are able to elicit colitis, even in mice not genetically predisposed to developing immunopathologic responses.

The penetrance and severity of colitis observed in the cocolonization experiments with *K. pneumoniae* and *P. mirabilis* were decreased compared to that observed in the spontaneous TRUC model (e.g., Figure 2C) and in neonatal cross-fostering experiments (TRUC-fostered *Rag2*<sup>-/-</sup> mean colitis score 5.6 ± 1.14 and TRUC-fostered WT 3.17 ± 0.75) (Figure 3A). Instead, it resembled experiments where adult TRUC mice were co-housed with adult *Rag2*<sup>-/-</sup> or WT mice (Garrett et al., 2007), speaking to a possible role of maternal/foster bacterial and nonbacterial factors in structuring microbial communities in the neonate. Consistent with this, we found that TRUC milk has a proinflammatory cytokine profile (Figure S5) and that the microbiota of 2-week-old TRUC mice clusters in a distinct group, as judged by PCoA plots of UniFrac measurements of 16S rRNA-defined communities (Figure 1A).

### ***K. pneumoniae* and *P. mirabilis* Colonization Patterns Change in Response to Immunotherapy, and Both Strains Induce TNF- $\alpha$ Production in *T-bet*<sup>-/-</sup> *Rag2*<sup>-/-</sup> *MyD88*<sup>-/-</sup> Bone Marrow-Derived DCs**

We next asked whether *K. pneumoniae* and *P. mirabilis* colonization patterns might change in response to two immunotherapeutic interventions previously shown to cure TRUC colitis, i.e., TNF- $\alpha$  neutralization and T-regulatory cell (T-reg) transfer (Garrett et al., 2007). We used quantitative culture-based



**Figure 3. *K. pneumoniae* and *P. mirabilis* Elicit Colitis but Require a Maternally Transmitted Endogenous Microbial Community for Maximal Intestinal Inflammation**

(A) The Abx sensitivities of colitis transmitted via TRUC cross-fostering are the same as spontaneous TRUC colitis. Abx-treated pregnant TRUC females were used as foster mothers and treated with Abx in their water until weaning. Histologic colitis scores are shown for the fostered mice at 8 weeks of age.

(B) *K. pneumoniae* (squares) and *P. mirabilis* (circles) are detected in the fecal microbiota of TRUC cross-fostered *Rag2*<sup>-/-</sup> and WT mice at 8 weeks of age but not in 8-week-old TRUC mice fostered by *Rag2*<sup>-/-</sup> or WT mice. TRUC-fostered TRUC, *Rag2*<sup>-/-</sup>-fostered *Rag2*<sup>-/-</sup>, and WT-fostered WT are shown as controls. Limits of detection: 10<sup>4.4</sup> cfu/g dry weight of feces. Each symbol represents a fecal sample from a different mouse.

(C) Left panel: Fecal bacterial counts for cocolonized gnotobiotic TRUC mice. Mean values ± 1 SD are shown for *K. pneumoniae* (squares) and *P. mirabilis* (circles) (n = 5 mice). Right panel: Histologic colitis scores of germ-free TRUC and germ-free TRUC mice cocolonized with *K. pneumoniae* and *P. mirabilis* from the TRUC mother in Figure 1.

methods to assay *K. pneumoniae* and *P. mirabilis* levels in feces prior to treatment of 4-week-old TRUC mice with anti-TNF- $\alpha$ , during weekly treatment for 4 weeks, and for 6 weeks after the last dose (Figure 4A) ( $n = 10$  mice, anti-TNF- $\alpha$ ;  $n = 10$ , isotype control) (histologic colitis scores in Figure S3). Significant differences in fecal *K. pneumoniae* levels between the TNF- $\alpha$  neutralization and isotype control groups were observed after 7 weeks of treatment (age 11 weeks;  $p = 0.0172$ ; Mann-Whitney test) and for *P. mirabilis* after a shorter period of treatment ( $p = 0.008$ ,  $p = 0.0012$ ,  $p = 0.0004$ , and  $p = 0.0403$  at 7, 8, 9, and 10 weeks of age). Two-way ANOVA revealed that anti-TNF- $\alpha$  neutralization accounted for 10.7% of the total variance observed in fecal *P. mirabilis* levels (adjusting for matching:  $F = 22.83$ ,  $DFn = 1$ ,  $DFd = 18$ ,  $p = 0.0002$ ). TNF- $\alpha$  did not directly affect the growth kinetics of either *K. pneumoniae* or *P. mirabilis* under in vitro monoculture conditions (Figure S4).

We performed a similar analysis in TRUC mice that had received 75,000 purified WT T-reg cells at 4 weeks of age (histologic colitis scores at 12 weeks of age in Figure S3). Surprisingly, while T-reg infusion ameliorated this colitis (Garrett et al., 2007), it did not affect fecal levels of either of these two bacterial species (Figure 4B). These results demonstrate that *K. pneumoniae* and *P. mirabilis* levels are not simply associated with inflammation, as both these modalities reduced host inflammation but did not uniformly alter Enterobacteriaceal representation. Our results illustrate that certain host-directed treatments may exert their effects not only by altering host inflammatory pathways but also by directly impacting the microbiota.

To begin to identify cell-based mechanisms by which TRUC-derived Enterobacteriaceae elicit a host immune response, TNF- $\alpha$  production was measured in *T-bet*<sup>-/-</sup> *Rag2*<sup>-/-</sup> *MyD88*<sup>-/-</sup> bone marrow-derived DCs cocultured with the *K. pneumoniae* and *P. mirabilis* TRUC strains, as DCs and TNF- $\alpha$  production are key features of the immunopathogenesis in TRUC mice, and the TRUC inflammatory response is independent of MyD88 (Garrett et al., 2007, 2009). Both live and heat-killed bacteria stimulated TNF- $\alpha$  production from *T-bet*<sup>-</sup>, *Rag2*<sup>-</sup>, and *MyD88*-deficient DCs (Figure 4C). These latter findings set the stage for future bacterial cell fractionation experiments where the microbial molecular determinants of host responses can be characterized using in vitro systems composed of genetically manipulated immune cells.

## DISCUSSION

Defining microbial features that are associated with or initiate IBD is complicated by host genetics, inflammatory state, and diet (Peterson et al., 2008). Designing prospective studies in human IBD to identify microbial communities that instigate inflammation has not been feasible, even in genetically susceptible populations. Thus, we performed a time series screen in a mouse model of IBD that shares several pathophysiologic features of human IBD, including immunodeficiency, compromised host barrier function, and hyperimmunity, to characterize

a colitogenic microbiota. We established that host genotype influenced the global structure of the associated microbial community detected in feces and observed a number of significant order- and species-level differences. Combined with culture-dependent methods, we were able to identify bacterial species whose role we could test in the development of disease. Our experiments demonstrate that *K. pneumoniae* and *P. mirabilis*, together with other members of the endogenous microbiota, can elicit colitis even in WT mice. It will be important to determine if significant associations are noted between these Enterobacteriaceae species and ulcerative colitis or Crohn's disease in ongoing (e.g., Qin et al., 2010) and future metagenomic studies of gut microbial ecology in various populations of patients with IBD.

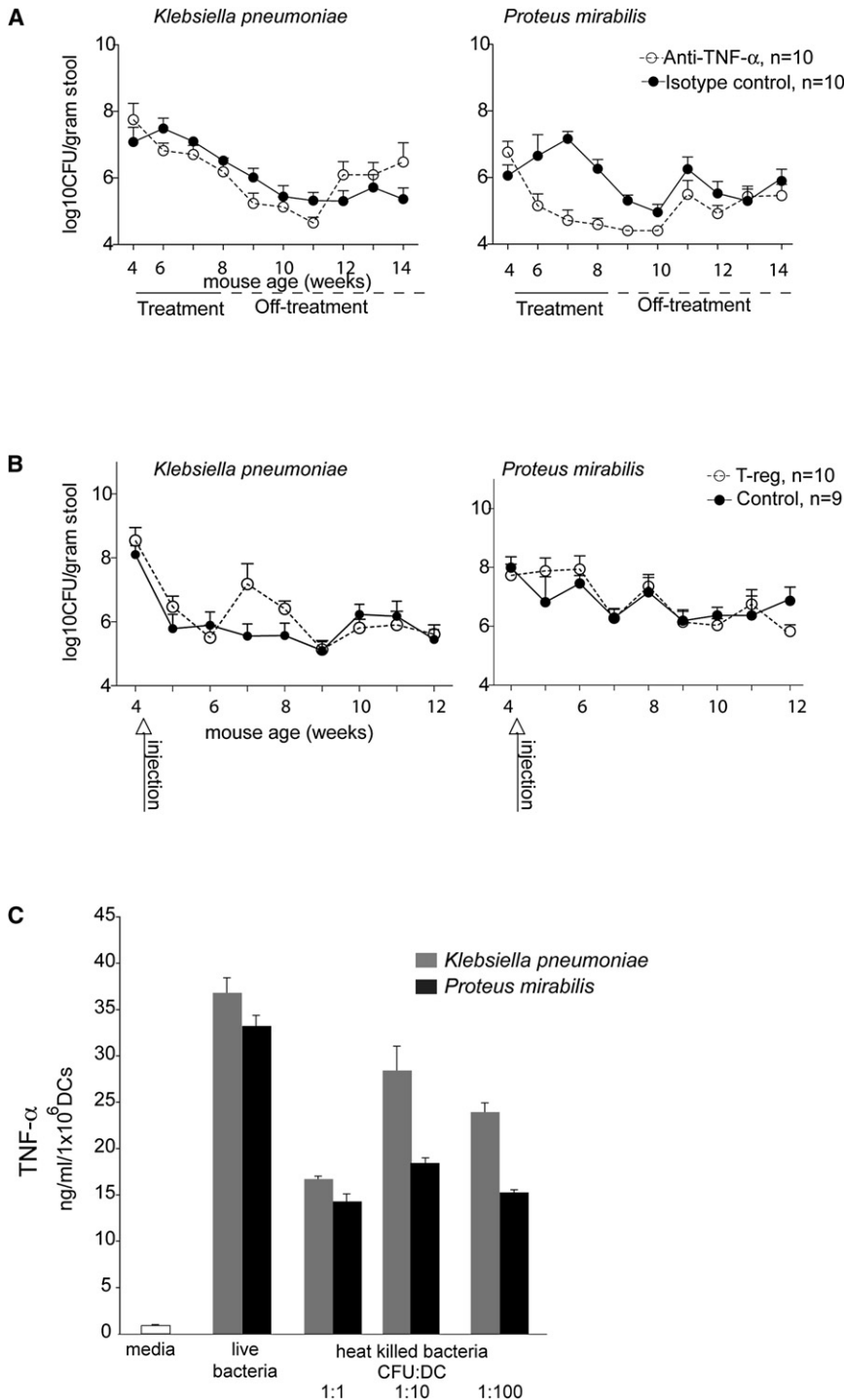
*K. pneumoniae* and *P. mirabilis* can colonize mouse and human intestine (Lau et al., 2008). Notably, we only recovered these microbes from TRUC mice in our colony, not in *Rag2*<sup>-/-</sup> or WT animals. While there was individual variation in bacterial counts, the colonization pattern of these species across the TRUC population over time was not significantly different and did not vary as colitis worsened with age. Inciting inflammation with dextran sulfate sodium in WT mice in the colony did not result in an emergence of these bacteria. In contrast, Abx treatment had a dramatic effect on the degree of host colonization with *K. pneumoniae* and *P. mirabilis*, as expected. The increased counts observed in response to vanco suggest that in the untreated host, members of the Gram-positive flora affect the degree of colonization by members of the Enterobacteriaceae.

A key feature of the colonic inflammation in TRUC mice is elevated TNF- $\alpha$ . While both neutralizing antibodies and T-reg infusion reduce mucosal TNF- $\alpha$  levels, these interventions had disparate effects on *K. pneumoniae* and *P. mirabilis* fecal counts. Cytokines may interact with bacteria, and TNF- $\alpha$  has been shown to affect *Salmonella typhimurium* replication in vivo (Romanova et al., 2002). While TNF- $\alpha$  did not appear to affect the proliferation of TRUC-derived *K. pneumoniae* and *P. mirabilis* in vitro, in vivo there were significant effects in response to TNF-neutralizing antibodies. Neutralizing antibodies and infusion of T-regs both lower TNF- $\alpha$  levels in TRUC mice, but through different mechanisms. T-regs also produce both IL-10 and TGF- $\beta$  (Izcue et al., 2009). T-regs and neutralizing antibodies may have direct but distinct effects on microbial populations or indirect effects through their differential effects on colonic DCs.

Opportunistic infection with *K. pneumoniae* and *P. mirabilis* is well recognized in the respiratory and urinary tracts. However, *Klebsiella oxytoca* but not *K. pneumoniae* has been tied to intestinal pathology (Högenauer et al., 2006). *Klebsiella* and *Proteus* species are observed more frequently in the stool of ulcerative colitis patients than healthy controls (Dorofeyev et al., 2009; Kanareykina et al., 1987), and elevated titers of Enterobacteriaceal antibodies have been reported in IBD patients (Cooper et al., 1988; Ibbotson et al., 1987; Tiwana et al., 1998). Genome sequencing of these isolates and comparisons to other sequenced isolates obtained from other body and environmental habitats could yield

(D) Left panel: *K. pneumoniae* and *P. mirabilis* fecal cfu in *Rag2*<sup>-/-</sup> and WT mice treated every other day from 2 to 10 weeks of age with  $10^7$  cfu of *E. coli*, *P. mirabilis*, *K. pneumoniae*, or a combination of both *P. mirabilis* and *K. pneumoniae* added to their drinking water (all strains isolated from the TRUC mother in Figure 1). Right panel: Histologic scores for colitis as assayed at sacrifice at 10 weeks of age. Each circle represents a separate animal in the treatment group.  $p$  values shown were calculated by the Mann-Whitney test.





**Figure 4. *K. pneumoniae* and *P. mirabilis* Colonization Patterns Change in Response to Immunotherapies, and Both Strains Induce TNF- $\alpha$  Production in *T-bet*<sup>-/-</sup> *Rag2*<sup>-/-</sup> *MyD88*<sup>-/-</sup> Bone Marrow-Derived DCs**

(A) Immunotherapy by TNF- $\alpha$  blockade alters levels of culturable fecal *Enterobacteriaceae*. TRUC mice were treated with anti-TNF- $\alpha$  (15 mg/kg every week) (open circles) or isotype control (shaded circles) for 4 weeks, and then therapy was stopped. *Enterobacteriaceae* levels were defined by culture of fecal samples obtained 1 day before, during, and after treatment (up to 14 weeks of age). Circles represent the mean value of anti-TNF- $\alpha$  mice (n = 10) and isotype controls (n = 10). Error bars represent 1 SD.

(B) Immunotherapy by T-reg infusion does not produce statistically significant differences in the levels of culturable *Enterobacteriaceae* species compared to vehicle-treated controls. TRUC mice were injected once with 75,000 T-regs (n = 10) or PBS (n = 9).

(C) TNF- $\alpha$  production from *T-bet*<sup>-/-</sup> *Rag2*<sup>-/-</sup> *MyD88*<sup>-/-</sup> bone marrow-derived DCs cocultured with heat-killed and live *K. pneumoniae* and *P. mirabilis* strains. Bars represent the mean value of triplicate determinations/sample. Error bars are 1 SD. Data are representative of three independent experiments.

Gut microbes help to structure the mucosal immune system, and the mucosal immune system shapes microbial community structure (Smith et al., 2007; Hooper and Macpherson, 2010). Microbial community members may be needed for the development of particular immune subsets or appropriate localization of immune cell subsets within the mucosa to generate proinflammatory responses to *K. pneumoniae* and *P. mirabilis*. For example, adherent cecal segmented filamentous bacteria have recently been shown to play a central role in the development of IL-17-producing CD4<sup>+</sup> T helper cells in mice (Ivanov et al., 2009; Gaboriau-Routhiau et al., 2009). CD11c<sup>+</sup> DCs are necessary for TRUC colitis (Garrett et al., 2007, 2009), and lamina propria CD11c<sup>+</sup> CX3CR1<sup>+</sup> DCs are markedly reduced in germ-free mice (Niess and Adler 2010).

testable hypotheses about genetic determinants that may underlie their ability to drive development of an IBD phenotype. Future studies can take advantage of the fact that both heat-killed and live *K. pneumoniae* and *P. mirabilis* induce TNF- $\alpha$  in *T-bet*<sup>-/-</sup> *Rag2*<sup>-/-</sup> *MyD88*<sup>-/-</sup> DCs to identify the responsible bacterial molecules and their host receptors. However, it is important to also emphasize that an endogenous microbial community is required for *K. pneumoniae* and *P. mirabilis* to exert their colitogenic effects.

In addition, interactions between *K. pneumoniae* and *P. mirabilis* and microbial community members may result in the acquisition of traits by these two *Enterobacteriaceae* (e.g., invasion) or by other community members that elicit intestinal inflammation. Convergence of host genetic susceptibility and microbial community features could also affect the behavior of these *Enterobacteriaceae* and the immune response to them, as we have observed in the TRUC model.

Elevated TNF- $\alpha$  and beneficial responses to TNF- $\alpha$ -neutralizing antibodies are common to both human IBD and several experimental colitis models. Host factors, like elevated TNF- $\alpha$ , may have virulence-promoting effects on these microbes. This notion is not without precedent, as the *Pseudomonas aeruginosa* protein OprF binds the proinflammatory cytokine IFN- $\gamma$ , resulting in expression of PA-I lectin, a quorum sensing-dependent virulence determinant (Wu et al., 2005).

In summary, future studies need to be directed at defining the genomic features of TRUC-associated *K. pneumoniae* and *P. mirabilis* strains, identifying co-occurring culturable members of the microbiota that contribute to disease pathogenesis in conventionally raised and gnotobiotic mouse models, characterizing host factors that drive these microbes to become colitogenic, and determining the microbial-associated molecular patterns and pattern recognition receptors involved in spontaneous and transmitted TRUC colitis. Together, these efforts may provide mechanistic insights about how gut microbial communities, working in concert with specific colitogenic agents, contribute to initiation and perpetuation of IBD in susceptible human hosts and provide the foundation for proof-of-concept tests of preventative or therapeutic measures. An additional benefit may be to help elucidate the association between IBD and increased risk for tumorigenesis, since the majority of TRUC mice spontaneously develop colonic dysplasia and rectal adenocarcinoma (Garrett et al., 2009).

## EXPERIMENTAL PROCEDURES

### Husbandry of Conventionally Raised Mice

*Rag2*<sup>-/-</sup>, *T-bet*<sup>-/-</sup>  $\times$  *Rag2*<sup>-/-</sup>, and *MyD88*<sup>-/-</sup>  $\times$  *T-bet*<sup>-/-</sup>  $\times$  *Rag2*<sup>-/-</sup> mice and their genotyping have been described (Garrett et al., 2009). Mice were housed in microisolator cages in a barrier facility at the Harvard School of Public Health under a 12 hr light cycle.

### 16S rRNA-Based Analyses of Fecal Microbial Communities

#### Community DNA Preparation

Fecal samples were flash frozen on collection and stored at -80°C before processing. DNA was extracted by bead-beating as described (Turnbaugh et al., 2009).

#### Sequencing and Analysis of 16S rRNA Gene Amplicons

The V2 region (primers 8F-338R) of bacterial 16S rRNA genes was targeted for amplification and multiplex pyrosequencing with error-correcting barcodes (Hamady et al., 2008). A total of 75,145 high-quality reads were generated from 32 samples (2348  $\pm$  343 reads per sample). See also Supplemental Experimental Procedures.

### Culture-Based Studies of Fecal Microbial Community Structure

#### Stool Collection

A minimum of three fecal pellets was collected from each mouse in a laminar flow hood. Each mouse (three females/genotype; TRUC and *Rag2*<sup>-/-</sup>) was sampled every 2 weeks at the same time of day from 2 to 10 weeks of age. Mothers were sampled once when their pups were 2 weeks old.

#### Culture

Fecal pellets were collected into tubes of PBS with 0.05% cysteine HCl. Serial 10-fold dilutions were made and plated on nonselective media and selective media. Anaerobes were incubated at 37°C in a Coy Anaerobic chamber for a minimum of 5 days. Aerobes were incubated for 24–48 hr at 37°C.

### Fecal Collection and Culture of Gram-Negative Aerobes

Mice were singly placed in autoclaved plastic cages. Four to six pellets were collected/ mouse/ time point. *Rag2*<sup>-/-</sup> mice in Figure 2F were sampled twice over a 3 day period for each weekly time point. Pellets were resuspended in

sterile PBS; 10-fold serial dilutions were generated, plated on MacConkey's medium, and incubated in ambient air at 37°C overnight. The lower limit of detection for these studies was 10<sup>4.4</sup> cfu/gram fecal dry weight.

### Histology

Colons were harvested and prepared for histology as described (Garrett et al., 2007). See Supplemental Experimental Procedures for more detail.

### Antibiotic Treatment

Mice were treated with the following Abx dissolved in their autoclaved drinking water as indicated: ampicillin (1 g/l; Roche), vancomycin (500 mg/l; Sigma), neomycin sulfate (1 g/l; Sigma), metronidazole (1 g/l; Sigma; solubilized with 15 ml of 0.1 N acetic acid/l), and gentamicin (2 g/l; Cell Gro). Fluid intake was monitored.

### Fluorescence In Situ Hybridization

Colons harvested from 16 *Rag2*<sup>-/-</sup> and 15 TRUC (3- to 8-week-old) mice were fixed in Carnoy's solution overnight and embedded in paraffin, and 5  $\mu$ m thick sections prepared (Swidsinski et al., 2005). The sequences of the following FISH probes were obtained from probeBase (<http://www.microbial-ecology.net/probebase/>) (Loy et al., 2007): the "universal" bacterial probe-EUB338 (pB-00159), *Enterobacteriaceae* targeted probe (pB-00914), *K. pneumoniae*-directed probe (pB-00352), and *P. mirabilis* probe (pB-02110).

### Cross-Fostering

On the day of birth, the mother was removed from the birthing cage and placed in a clean cage. A litter of pups with the designated genotype was then put into the cage. Pups were weaned on postnatal day 21 (Garrett et al., 2007).

### Gnotobiotic Mouse Experiments

All protocols related to the generation and husbandry of germ-free mice were approved by the Washington University (Wash U) Animal Studies Committee. Conventionally raised SPF *T-bet*<sup>-/-</sup>  $\times$  *Rag2*<sup>-/-</sup> mice were rederived as germ-free in the gnotobiotic facility at Wash U. Subsequent experiments were carried out at the Harvard Digestive Disease Center (HDDC) gnotobiotic facility. Five mice were maintained germ-free, and another five mice (3 female and 2 male) were cocolonized by introducing 4.8  $\times$  10<sup>8</sup> cfu of *K. pneumoniae* and 9.2  $\times$  10<sup>8</sup> cfu of *P. mirabilis* into their oral cavity and simultaneously spreading an equivalent amount of organisms on their fur and anus. See Supplemental Experimental Procedures for more detail.

### Invasion Experiments

*K. pneumoniae*, *P. mirabilis*, *E. coli*, or both *K. pneumoniae* and *P. mirabilis* (2  $\times$  10<sup>7</sup> cfu each; all isolated from the TRUC mother in Figure 1) were instilled into the oral cavity of each mouse using a sterile pipette tip, and 1  $\times$  10<sup>7</sup> cfu was placed into a new container of their drinking water every other day.

### Anti-TNF- $\alpha$ Treatment

Mice were injected with anti-TNF- $\alpha$  (clone TN3-19.12), a hamster anti-mouse TNF- $\alpha$ -neutralizing IgG1 antibody, and control Ab (hamster anti-GST IgG1) (Leinco Technologies, St. Louis) (15 mg/kg) once a week for 4 weeks (Garrett et al., 2007).

### Adoptive Transfer of T-Regulatory Cells

FACS-sorted lymph node CD4<sup>+</sup> CD62L<sup>hi</sup> CD25<sup>+</sup> cells (T-reg, 75,000 cells) or PBS were injected per mouse at 4 weeks of age (n = 10 for T-regs; n = 9 for PBS) (Garrett et al., 2007). This experiment was terminated by euthanasia at 12 weeks because two control group mice became moribund from colitis.

### Coculturing Bone Marrow-Derived Dendritic Cells and Bacterial Strains

Mouse bone marrow-derived DCs were generated as described (Garrett et al., 2007) and purified using anti-mouse CD11c-coupled magnetic beads. *K. pneumoniae* or *P. mirabilis* was cocultured with DCs at a ratio of 1 cfu/DC for 4 hr at 37°C in a cell culture incubator at 5% CO<sub>2</sub>. Gent (50  $\mu$ g/ml) was then added to the media for 1 hr, and cells were collected, washed, and incubated with medium containing gent (20  $\mu$ g/ml) for an additional 16 hr. Bacteria were also heat-killed (incubation at 100°C for 3 min followed by plating to

confirm killing) and added to cultures of DCs at ratios of 1:1, 10:1, and 100:1. Cells were cocultured for 20 hr. TNF- $\alpha$  levels in supernatants collected from centrifuged live and heat-killed cocultures were determined using the mouse OptEIA ELISA kit (BD Biosciences) and expressed as ng/ml/ $1 \times 10^6$  DCs.

### Statistical Analysis

The Prism graphing and analysis program was used for calculation of statistical measures including mean values, standard deviations, p values (Mann-Whitney test), and two-way ANOVA.

### SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, Supplemental References, two tables, and four figures and can be found with this article online at doi:10.1016/j.chom.2010.08.004.

### ACKNOWLEDGMENTS

We thank J. Braun (UCLA), D. Relman (Stanford U.), A. Swidsinski (Charité Humboldt U.), G. Hansson (U. of Gothenburg), and members of the Gordon and Glimcher labs for helpful discussions. We thank J. Ramirez for care of our SPF mice and S. Wagoner, V. Yesiliseyev, and C. Belzer for care of germ-free TRUC. This work was supported by the NIH (CA112663 to L.H.G.), Danone Research (L.H.G.), and the Crohn's and Colitis Foundation of America (J.I.G.), plus career development awards from the Burroughs Wellcome Fund and the NIH (K08AI078942) to W.S.G. The HDDC germ-free core is supported by P30-DK03485. 454 pyrosequencing reads have been deposited in the NCBI Short Read Archive. L.H.G. declares that she is a member of the Board of Directors of the Bristol-Myers Squibb Corporation (BMSC) and holds equity in BMSC.

Received: February 9, 2010

Revised: May 25, 2010

Accepted: July 22, 2010

Published: September 15, 2010

### REFERENCES

- Cooper, R., Fraser, S.M., Sturrock, R.D., and Gemmell, C.G. (1988). Raised titres of anti-klebsiella IgA in ankylosing spondylitis, rheumatoid arthritis, and inflammatory bowel disease. *Br. Med. J. (Clin. Res. Ed.)* **296**, 1432–1434.
- Dorofeyev, A.E., Vasilenko, I.V., and Rassokhina, O.A. (2009). Joint extraintestinal manifestations in ulcerative colitis. *Dig. Dis.* **27**, 502–510.
- Duerkop, B.A., Vaishnava, S., and Hooper, L.V. (2009). Immune responses to the microbiota at the intestinal mucosal surface. *Immunity* **31**, 368–376.
- Frank, D.N., and Pace, N.R. (2008). Gastrointestinal microbiology enters the metagenomics era. *Curr. Opin. Gastroenterol.* **24**, 4–10.
- Gaboriau-Routhiau, V., Rakotobe, S., Lécuyer, E., Mulder, I., Lan, A., Bridonneau, C., Rochet, V., Pisi, A., De Paepe, M., Brandi, G., et al. (2009). The key role of segmented filamentous bacteria in the coordinated maturation of gut helper T cell responses. *Immunity* **31**, 677–689.
- Garrett, W.S., Lord, G.M., Punit, S., Lugo-Villarino, G., Mazmanian, S.K., Ito, S., Glickman, J.N., and Glimcher, L.H. (2007). Communicable ulcerative colitis induced by T-bet deficiency in the innate immune system. *Cell* **131**, 33–45.
- Garrett, W.S., Punit, S., Gallini, C.A., Michaud, M., Zhang, D., Sigrist, K.S., Lord, G.M., Glickman, J.N., and Glimcher, L.H. (2009). Colitis-associated colorectal cancer driven by T-bet deficiency in dendritic cells. *Cancer Cell* **16**, 208–219.
- Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J., and Knight, R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods* **5**, 235–237.
- Hansen, R., Thomson, J.M., El-Omar, E.M., and Hold, G.L. (2010). The role of infection in the aetiology of inflammatory bowel disease. *J. Gastroenterol.* **45**, 266–276.
- Hill, D.A., and Artis, D. (2009). Maintaining diplomatic relations between mammals and beneficial microbial communities. *Sci. Signal.* **2**, pe77.
- Hoffmann, C., Hill, D.A., Minkah, N., Kim, T., Troy, A., Artis, D., and Bushman, F. (2009). Community-wide response of the gut microbiota to enteropathogenic *Citrobacter rodentium* infection revealed by deep sequencing. *Infect. Immun.* **77**, 4668–4678.
- Högenauer, C., Langner, C., Beubler, E., Lippe, I.T., Schicho, R., Gorkiewicz, G., Krause, R., Gerstgrasser, N., Krejs, G.J., and Hinterleitner, T.A. (2006). *Klebsiella oxytoca* as a causative organism of antibiotic-associated hemorrhagic colitis. *N. Engl. J. Med.* **355**, 2418–2426.
- Hooper, L.V., and Macpherson, A.J. (2010). Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nat. Rev. Immunol.* **10**, 159–169.
- Ibbotson, J.P., Pease, P.E., and Allan, R.N. (1987). Serological studies in Crohn's disease. *Eur. J. Clin. Microbiol.* **6**, 286–290.
- Ivanov, I.I., Atarashi, K., Manel, N., Brodie, E.L., Shima, T., Karaoz, U., Wei, D., Goldfarb, K.C., Santee, C.A., Lynch, S.V., et al. (2009). Induction of intestinal Th17 cells by segmented filamentous bacteria. *Cell* **139**, 485–498.
- Izcue, A., Coombes, J.L., and Powrie, F. (2009). Regulatory lymphocytes and intestinal inflammation. *Annu. Rev. Immunol.* **27**, 313–338.
- Kanareykina, S.K., Misautova, A.A., Zlatkina, A.R., and Levina, E.N. (1987). *Proteus* dysbioses in patients with ulcerative colitis. *Nahrung* **31**, 557–561.
- Lau, H.Y., Huffnagle, G.B., and Moore, T.A. (2008). Host and microbiota factors that control *Klebsiella pneumoniae* mucosal colonization in mice. *Microbes Infect.* **10**, 1283–1290.
- Loy, A., Maixner, F., Wagner, M., and Horn, M. (2007). probeBase—an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic Acids Res.* **35** (Database issue), D800–D804.
- Lupp, C., Robertson, M.L., Wickham, M.E., Sekirov, I., Champion, O.L., Gaynor, E.C., and Finlay, B.B. (2007). Host-mediated inflammation disrupts the intestinal microbiota and promotes the overgrowth of Enterobacteriaceae. *Cell Host Microbe* **2**, 204.
- Niess, J.H., and Adler, G. (2010). Enteric flora expands gut lamina propria CX3CR1+ dendritic cells supporting inflammatory immune responses under normal and inflammatory conditions. *J. Immunol.* **184**, 2026–2037.
- Peterson, D.A., Frank, D.N., Pace, N.R., and Gordon, J.I. (2008). Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases. *Cell Host Microbe* **3**, 417–427.
- Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al; MetaHIT Consortium. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65.
- Romanova, Y.M., Scheglovitova, O.N., Boshnakov, R.H., Alekseeva, N.V., Stepanova, T.V., Tomova, A.S., and Gintsburg, A.L. (2002). TNF-alpha and gamma-irradiation induced activation of the *Salmonella typhimurium* reproduction in the organs of infected animals. *Russ. J. Immunol.* **7**, 129–134.
- Sartor, R.B. (2009). Microbial-host interactions in inflammatory bowel diseases and experimental colitis. *Nestle Nutr. Workshop Ser. Pediatr. Program.* **64**, 121–132, discussion 132–137, 251–257.
- Smith, K., McCoy, K.D., and Macpherson, A.J. (2007). Use of axenic animals in studying the adaptation of mammals to their commensal intestinal microbiota. *Semin. Immunol.* **19**, 59–69.
- Swidsinski, A., Weber, J., Loening-Baucke, V., Hale, L.P., and Lochs, H. (2005). Spatial organization and composition of the mucosal flora in patients with inflammatory bowel disease. *J. Clin. Microbiol.* **43**, 3380–3389.
- Tiwana, H., Walmsley, R.S., Wilson, C., Yiannakou, J.Y., Ciclitira, P.J., Wakefield, A.J., and Ebringer, A. (1998). Characterization of the humoral immune response to *Klebsiella* species in inflammatory bowel disease and ankylosing spondylitis. *Br. J. Rheumatol.* **37**, 525–531.
- Turnbaugh, P.J., Hamady, M., Yatsunenkov, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484.
- Wu, L., Estrada, O., Zaborina, O., Bains, M., Shen, L., Kohler, J.E., Patel, N., Musch, M.W., Chang, E.B., Fu, Y.X., et al. (2005). Recognition of host immune activation by *Pseudomonas aeruginosa*. *Science* **309**, 774–777.

## Appendix E

Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PS, Charlson E, Costello EK, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman JA, Gallery RE, Gevers D, Gibbs RA, San Gil I, Gonzalez A, Gordon JI, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau AL, Kelley ST, Kennedy J, Knights D, Koren O, Kuczynski J, Kyrpides N, Larsen R, Lauber CL, Legg T, Ley RE, Lozupone CA, Ludwig W, Lyons D, Maguire E, Methé BA, Meyer F, Muegge B, Nakielny S, Nelson KE, Nemergut D, Neufeld JD, Newbold LK, Oliver AE, Pace NR, Palanisamy G, Peplies J, Petrosino J, Proctor L, Pruesse E, Quast C, Raes J, Ratnasingham S, Ravel J, Relman DA, Assunta-Sansone S, Schloss PD, Schriml L, Sinha R, Smith MI, Sodergren E, Spo A, Stombaugh J, Tiedje JM, Ward DV, Weinstock GM, Wendel D, White O, Whiteley A, Wilke A, Wortman JR, **Yatsunenko T**, Glöckner FO.

“Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications.”

*Nature Biotechnology*. **2011** May; 29(5): 415-20.

# Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications

Pelin Yilmaz<sup>1,2\*</sup>, Renzo Kottmann<sup>1</sup>, Dawn Field<sup>3</sup>, Rob Knight<sup>4,5</sup>, James R Cole<sup>6,7</sup>, Linda Amaral-Zettler<sup>8</sup>, Jack A Gilbert<sup>9–11</sup>, Ilene Karsch-Mizrachi<sup>12</sup>, Anjanette Johnston<sup>12</sup>, Guy Cochrane<sup>13</sup>, Robert Vaughan<sup>13</sup>, Christopher Hunter<sup>13</sup>, Joonhong Park<sup>14</sup>, Norman Morrison<sup>3,15</sup>, Philippe Rocca-Serra<sup>16</sup>, Peter Sterk<sup>3</sup>, Manimozhiyan Arumugam<sup>17</sup>, Mark Bailey<sup>3</sup>, Laura Baumgartner<sup>18</sup>, Bruce W Birren<sup>19</sup>, Martin J Blaser<sup>20</sup>, Vivien Bonazzi<sup>21</sup>, Tim Booth<sup>3</sup>, Peer Bork<sup>17</sup>, Frederic D Bushman<sup>22</sup>, Pier Luigi Buttigieg<sup>1,2</sup>, Patrick S G Chain<sup>7,23,24</sup>, Emily Charlson<sup>22</sup>, Elizabeth K Costello<sup>4</sup>, Heather Huot-Creasy<sup>25</sup>, Peter Dawyndt<sup>26</sup>, Todd DeSantis<sup>27</sup>, Noah Fierer<sup>28</sup>, Jed A Fuhrman<sup>29</sup>, Rachel E Gallery<sup>30</sup>, Dirk Gevers<sup>19</sup>, Richard A Gibbs<sup>31,32</sup>, Inigo San Gil<sup>33</sup>, Antonio Gonzalez<sup>34</sup>, Jeffrey I Gordon<sup>35</sup>, Robert Guralnick<sup>28,36</sup>, Wolfgang Hankeln<sup>1,2</sup>, Sarah Highlander<sup>31,37</sup>, Philip Hugenholtz<sup>38</sup>, Janet Jansson<sup>23,39</sup>, Andrew L Kau<sup>35</sup>, Scott T Kelley<sup>40</sup>, Jerry Kennedy<sup>4</sup>, Dan Knights<sup>34</sup>, Omry Koren<sup>41</sup>, Justin Kuczynski<sup>18</sup>, Nikos Kyrpides<sup>23</sup>, Robert Larsen<sup>4</sup>, Christian L Lauber<sup>42</sup>, Teresa Legg<sup>28</sup>, Ruth E Ley<sup>41</sup>, Catherine A Lozupone<sup>4</sup>, Wolfgang Ludwig<sup>43</sup>, Donna Lyons<sup>42</sup>, Eamonn Maguire<sup>16</sup>, Barbara A Methé<sup>44</sup>, Folker Meyer<sup>10</sup>, Brian Muegge<sup>35</sup>, Sara Nakielny<sup>4</sup>, Karen E Nelson<sup>44</sup>, Diana Nemergut<sup>45</sup>, Josh D Neufeld<sup>46</sup>, Lindsay K Newbold<sup>3</sup>, Anna E Oliver<sup>3</sup>, Norman R Pace<sup>18</sup>, Giriprakash Palanisamy<sup>47</sup>, Jörg Peplies<sup>48</sup>, Joseph Petrosino<sup>31,37</sup>, Lita Proctor<sup>21</sup>, Elmar Pruesse<sup>1,2</sup>, Christian Quast<sup>1</sup>, Jeroen Raes<sup>49</sup>, Sujeevan Ratnasingham<sup>50</sup>, Jacques Ravel<sup>25</sup>, David A Relman<sup>51,52</sup>, Susanna Assunta-Sansone<sup>16</sup>, Patrick D Schloss<sup>53</sup>, Lynn Schriml<sup>25</sup>, Rohini Sinha<sup>22</sup>, Michelle I Smith<sup>35</sup>, Erica Sodergren<sup>54</sup>, Aymé Spor<sup>41</sup>, Jesse Stombaugh<sup>4</sup>, James M Tiedje<sup>7</sup>, Doyle V Ward<sup>19</sup>, George M Weinstock<sup>54</sup>, Doug Wendel<sup>4</sup>, Owen White<sup>25</sup>, Andrew Whiteley<sup>3</sup>, Andreas Wilke<sup>10</sup>, Jennifer R Wortman<sup>25</sup>, Tanya Yatsunenko<sup>35</sup> & Frank Oliver Glöckner<sup>1,2</sup>

Here we present a standard developed by the Genomic Standards Consortium (GSC) for reporting marker gene sequences—the minimum information about a marker gene sequence (MIMARKS). We also introduce a system for describing the environment from which a biological sample originates. The ‘environmental packages’ apply to any genome sequence of known origin and can be used in combination with MIMARKS and other GSC checklists. Finally, to establish a unified standard for describing sequence data and to provide a single point of entry for the scientific community to access and learn about GSC checklists, we present the minimum information about any (x) sequence (MIxS). Adoption of MIxS will enhance our ability to analyze natural genetic diversity documented by massive DNA sequencing efforts from myriad ecosystems in our ever-changing biosphere.

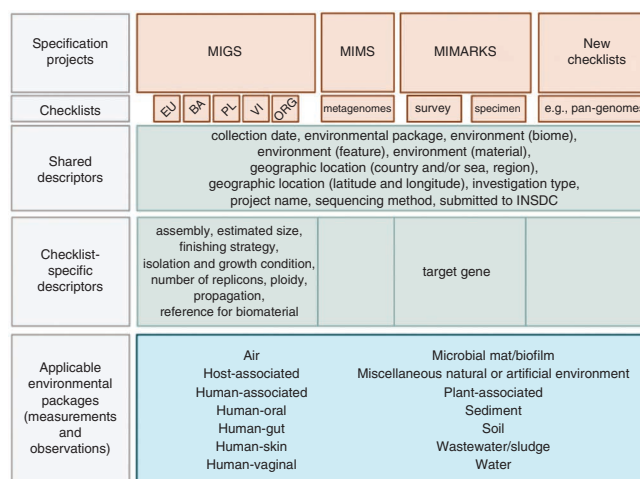
Without specific guidelines, most genomic, metagenomic and marker gene sequences in databases are sparsely annotated with the information required to guide data integration, comparative studies and

knowledge generation. Even with complex keyword searches, it is currently impossible to reliably retrieve sequences that have originated from certain environments or particular locations on Earth—for example, all sequences from ‘soil’ or ‘freshwater lakes’ in a certain region of the world. Because public databases of the International Nucleotide Sequence Database Collaboration (INSDC; comprising DNA Data Bank of Japan (DDBJ), the European Nucleotide Archive (EBI-ENA) and GenBank (<http://www.insdc.org/>)) depend on author-submitted information to enrich the value of sequence data sets, we argue that the only way to change the current practice is to establish a standard of reporting that requires contextual data to be deposited at the time of sequence submission. The adoption of such a standard would elevate the quality, accessibility and utility of information that can be collected from INSDC or any other data repository.

The GSC has previously proposed standards for describing genomic sequences—the “minimum information about a genome sequence” (MIGS)—and metagenomic sequences—the “minimum information about a metagenome sequence” (MIMS)<sup>1</sup>. Here we introduce an extension of these standards for capturing information about marker genes. Additionally, we introduce ‘environmental packages’ that standardize sets of measurements and observations describing particular habitats that are applicable across all GSC checklists and beyond<sup>2</sup>. We define ‘environment’ as any location in which a sample or organism

\*A list of affiliations appears at the end of the paper.

**Figure 1** Schematic overview about the GSC MxS standard (brown), including combination with specific environmental packages (blue). Shared descriptors apply to all MxS checklists; however, each checklist has its own specific descriptors as well. Environmental packages can be applied to any of the checklists. EU, eukarya; BA, bacteria/archaea; PL, plasmid; VI, virus; ORG, organelle.



is found, e.g., soil, air, water, human-associated, plant-associated or laboratory. The original MIGS/MIMS checklists included contextual data about the location from which a sample was isolated and how the sequence data were produced. However, standard descriptions for a more comprehensive range of environmental parameters, which would help to better contextualize a sample, were not included. The environmental packages presented here are relevant to any genome sequence of known origin and are designed to be used in combination with MIGS, MIMS and MIMARKS checklists.

To create a single entry point to all minimum information checklists from the GSC and to the environmental packages, we propose an overarching framework, the MxS standard ([http://gensc.org/gc\\_wiki/index.php/MxS](http://gensc.org/gc_wiki/index.php/MxS)). MxS includes the technology-specific checklists from the previous MIGS and MIMS standards, provides a way of introducing additional checklists such as MIMARKS, and also allows annotation of sample data using environmental packages. A schematic overview of MxS along with the MxS environmental packages is shown in **Figure 1**.

### Development of MIMARKS and the environmental packages

Over the past three decades, the 16S rRNA, 18S rRNA and internal transcribed spacer gene sequences (ITS) from *Bacteria*, *Archaea* and microbial *Eukaryotes* have provided deep insights into the topology of the tree of life<sup>3,4</sup> and the composition of communities of organisms that live in diverse environments, ranging from deep sea hydrothermal vents to ice sheets in the Arctic<sup>5–16</sup>. Numerous other phylogenetic marker genes have proven useful, including RNA polymerase subunits (*rpoB*), DNA gyrases (*gyrB*), DNA recombination and repair proteins (*recA*) and heat shock proteins (*HSP70*)<sup>3</sup>. Marker genes can also reveal key metabolic functions rather than phylogeny; examples include nitrogen cycling (*amoA*, *nifH*, *ntcA*)<sup>17,18</sup>, sulfate reduction (*dsrAB*)<sup>19</sup> or phosphorus metabolism (*phnA*, *phnI*, *phnJ*)<sup>20,21</sup>. In this paper we define all phylogenetic and functional genes (or gene fragments) used to profile natural genetic diversity as ‘marker genes’. MIMARKS (**Table 1**) complements the MIGS/MIMS checklists for genomes and metagenomes by adding two new checklists, a MIMARKS survey, for uncultured diversity marker gene surveys, and a MIMARKS specimen, for marker gene sequences obtained from any material identifiable by means of specimens. The MIMARKS extension adopts and incorporates the standards being developed by the Consortium for the Barcode of Life (CBOL)<sup>22</sup>. Therefore, the checklist can be universally applied to any marker gene, from small subunit rRNA to cytochrome oxidase I (COI), to all taxa, and to studies ranging from single individuals to complex communities.

Both MIMARKS and the environmental packages were developed by collating information from several sources and evaluating it in the framework of the existing MIGS/MIMS checklists. These include four independent community-led surveys, examination of the parameters reported in published studies and examination of compliance with optional features in INSDC documents. The overall goal of these activities was to design the backbone of the MIMARKS checklist, which describes the most important aspects of marker gene contextual data.

### Results of community-led surveys

Four online surveys about descriptors for marker genes have been conducted to determine researcher preferences for core descriptors.

The Department of Energy Joint Genome Institute and SILVA<sup>23</sup> surveys focused on general descriptor contextual data for a marker gene, whereas the Ribosomal Database Project (RDP)<sup>24</sup> focused on prevalent habitats for rRNA gene surveys, and the Terragenome Consortium<sup>25</sup> focused on soil metagenome project contextual data (**Supplementary Results 1**). The above recommendations were combined with an extensive set of contextual data items suggested by an International Census of Marine Microbes (ICoMM) working group that met in 2005. These collective resources provided valuable insights into community requests for contextual data items to be included in the MIMARKS checklist and the main habitats constituting the environmental packages.

### Survey of published parameters

We reviewed published rRNA gene studies, retrieved from SILVA and the ICoMM database MICROBIS (The Microbial Oceanic Biogeographic Information System, <http://icomm.mbl.edu/microbis/>) to further supplement contextual data items that are included in the respective environmental packages. In total, 39 publications from SILVA and >40 ICoMM projects were scanned for contextual data items to constitute the core of the environmental package subtables (**Supplementary Results 1**).

In a final analysis step, we surveyed usage statistics of INSDC source feature key qualifier values of rRNA gene sequences contained in SILVA (**Supplementary Results 1**). Notably, <10% of the 1.2 million 16S rRNA gene sequences (SILVA release 100) were associated with even basic information such as latitude and longitude, collection date or PCR primers.

### The MIMARKS checklist

The MIMARKS checklist provides users with an ‘electronic laboratory notebook’ containing core contextual data items required for consistent reporting of marker gene investigations. MIMARKS uses the MIGS/MIMS checklists with respect to the nucleic acid sequence source and sequencing contextual data, but extends them with further experimental contextual data such as PCR primers and conditions, or target gene name.

For clarity and ease of use, all items within the MIMARKS checklist are presented with a value syntax description, as well as a clear definition of the item. Whenever terms from a specific ontology are required as the value of an item, these terms can be readily found in the respective ontology browsers linked by URLs in the item definition. Although this version of the MIMARKS checklist does not

**Table 1 The core items of the MIMARKS checklists, along with the value types, descriptions and requirement status**

Item	Description	Report type	
		MIMARKS survey	MIMARKS specimen
<b>Investigation</b>			
Submitted to INSDC <sup>[boolean]</sup>	Depending on the study (large-scale, e.g., done with next-generation sequencing technology, or small-scale) sequences have to be submitted to SRA (Sequence Read Archives), DRA (DDBJ Sequence Read Archive) or through the classical Webin/Sequin systems to GenBank, ENA and DDBJ	M	M
Investigation type <sup>[mimarks-survey or mimarks-specimen]</sup>	Nucleic Acid Sequence Report is the root element of all MIMARKS compliant reports as standardized by Genomic Standards Consortium (GSC). This field is either MIMARKS survey or MIMARKS specimen	M	M
Project name	Name of the project within which the sequencing was organized	M	M
<b>Environment</b>			
Geographic location (latitude and longitude <sup>[float, point, transect and region]</sup> )	The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and in WGS84 system	M	M
Geographic location (depth <sup>[integer, point, interval, unit]</sup> )	Please refer to the definitions of depth in the environmental packages	E	E
Geographic location (elevation of site <sup>[integer, unit]</sup> ; altitude of sample <sup>[integer, unit]</sup> )	Please refer to the definitions of either altitude or elevation in the environmental packages	E	E
Geographic location (country and/or sea <sup>[INSDC or GAZ]</sup> ; region <sup>[GAZ]</sup> )	The geographical origin of the sample as defined by the country or sea name. Country, sea or region names should be chosen from the INSDC list ( <a href="http://insdc.org/country.html">http://insdc.org/country.html</a> ), or the GAZ (Gazetteer, v1.446) ontology ( <a href="http://bioportal.bioontology.org/visualize/40651">http://bioportal.bioontology.org/visualize/40651</a> )	M	M
Collection date <sup>[ISO8601]</sup>	The time of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated, that is, all of these are valid times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008; except for 2008-01 and 2008, all are ISO6801 compliant	M	M
Environment (biome <sup>[EnvO]</sup> )	In environmental biome level are the major classes of ecologically similar communities of plants, animals and other organisms. Biomes are defined based on factors such as plant structures, leaf types, plant spacing and other factors like climate. Examples include desert, taiga, deciduous woodland or coral reef. Environment Ontology (EnvO) (v1.53) terms listed under environmental biome can be found at <a href="http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00000428">http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00000428</a>	M	M
Environment (feature <sup>[EnvO]</sup> )	Environmental feature level includes geographic environmental features. Examples include harbor, cliff or lake. EnvO (v1.53) terms listed under environmental feature can be found at <a href="http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00002297">http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00002297</a>	M	M
Environment (material <sup>[EnvO]</sup> )	The environmental material level refers to the matter that was displaced by the sample, before the sampling event. Environmental matter terms are generally mass nouns. Examples include: air, soil or water. EnvO (v1.53) terms listed under environmental matter can be found at <a href="http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00010483">http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00010483</a>	M	M
<b>MIGS/MIMS/MIMARKS extension</b>			
Environmental package <sup>[air, host-associated, human-associated, human-skin, human-oral, human-gut, human-vaginal, microbial mat/biofilm, miscellaneous natural or artificial environment, plant-associated, sediment, soil, wastewater/sludge, water]</sup>	MIGS/MIMS/MIMARKS extension for reporting of measurements and observations obtained from one or more of the environments where the sample was obtained. All environmental packages listed here are further defined in separate subtables. By giving the name of the environmental package, a selection of fields can be made from the subtables and can be reported	M	M
<b>Nucleic acid sequence source</b>			
Isolation and growth conditions <sup>[PMID, DOI or URL]</sup>	Publication reference in the form of PubMed ID (PMID), digital object identifier (DOI) or URL for isolation and growth condition specifications of the organism/material	–	M
<b>Sequencing</b>			
Target gene or locus (e.g., 16S rRNA, 18S rRNA, nif, amoA, rpo)	Targeted gene or locus name for marker gene study	M	M
Sequencing method (e.g., dideoxysequencing, pyrosequencing, polony)	Sequencing method used, e.g., Sanger, pyrosequencing, ABI-solid	M	M

Items for the MIMARKS specification and their mandatory (M), status for both MIMARKS-survey and MIMARKS-specimen checklists. Furthermore, “–” denotes that an item is not applicable for a given checklist. E denotes that a field has environment-specific requirements. For example, whereas “depth” is mandatory for the environments water, sediment or soil, it is optional for human-associated environments. MIMARKS-survey is applicable to contextual data for marker gene sequences, obtained directly from the environment, without culturing or identification of the organisms. MIMARKS-specimen, on the other hand, applies to the contextual data for marker gene sequences from cultured or voucher-identifiable specimens. Both MIMARKS-survey and specimen checklists can be used for any type of marker gene sequence data, ranging from 16S, 18S, 23S, 28S rRNA to COI, hence the checklists are universal for all three domains of life. Item names are followed by a short description of the value of the item in parentheses and/or value type in brackets as a superscript. Whenever applicable, value types are chosen from a controlled vocabulary (CV) or an ontology from the Open Biological and Biomedical Ontologies (OBO) foundry (<http://www.obofoundry.org/>). This table only presents the very core of MIMARKS checklists, that is, only mandatory items for each checklist. **Supplementary Results 2** contains all MIMARKS items, the tables for environmental packages in the MIGS/MIMS/MIMARKS extension and GenBank structured comment name that should be used for submitting MIMARKS data to GenBank. In case of submitting to EBI-ENA, the full names can be used.



contain unit specifications, we recommend all units to be chosen from and follow the International System of Units (SI) recommendations. In addition, we strongly urge the community to provide feedback regarding the best unit recommendations for given parameters. Unit standardization across data sets will be vital to facilitate comparative studies in future. An Excel version of the MIMARKS checklist is provided on the GSC web site ([http://gensc.org/gc\\_wiki/index.php/MIMARKS](http://gensc.org/gc_wiki/index.php/MIMARKS)).

### The MIxS environmental packages

Fourteen environmental packages provide a wealth of environmental and epidemiological contextual data fields for a complete description of sampling environments. The environmental packages can be combined with any of the GSC checklists (**Fig. 1** and **Supplementary Results 2**). Researchers within The Human Microbiome Project<sup>26</sup> contributed the host-associated and all human packages. The Terragenome Consortium contributed sediment and soil packages. Finally, ICoMM, Microbial Inventory Research Across Diverse Aquatic Long Term Ecological Research Sites and the Max Planck Institute for Marine Microbiology contributed the water package. The MIMARKS working group developed the remaining packages (air, microbial mat/biofilm, miscellaneous natural or artificial environment, plant-associated and wastewater/sludge). The package names describe high-level habitat terms in order to be exhaustive. The miscellaneous natural or artificial environment package contains a generic set of parameters, and is included for any other habitat that does not fall into the other thirteen categories. Whenever needed, multiple packages may be used for the description of the environment.

### Examples of MIMARKS-compliant data sets

Several MIMARKS-compliant reports are included in **Supplementary Results 3**. These include a 16S rRNA gene survey from samples obtained in the North Atlantic, an 18S pyrosequencing tag study of anaerobic protists in a permanently anoxic basin of the North Sea, a *pmoA* survey from Negev Desert soils, a *dsrAB* survey of Gulf of Mexico sediments and a 16S pyrosequencing tag study of bacterial diversity in the western English Channel (SRA accession no. SRP001108).

### Adoption by major database and informatics resources

Support for adoption of MIMARKS and the MIxS standard has spread rapidly. Authors of this paper include representatives from genome sequencing centers, maintainers of major resources, principal investigators of large- and small-scale sequencing projects, and individual investigators who have provided compliant data sets, showing the breadth of support for the standard within the community.

In the past, the INSDC has issued a reserved 'barcode' keyword for the CBOL<sup>7</sup>. Following this model, the INSDC has recently recognized the GSC as an authority for the MIxS standard and issued the standard with official keywords within INSDC nucleotide sequence records<sup>27</sup>. This greatly facilitates automatic validation of the submitted contextual data and provides support for data sets compliant with previous versions by including the checklist version as a keyword.

GenBank accepts MIxS metadata in tabular format using the sequin and tbl2asn submission tools, validates MIxS compliance and reports the fields in the structured comment block. The EBI-ENA Webin submission system provides prepared web forms for the submission of MIxS compliant data; it presents all of the appropriate fields with descriptions, explanations and examples, and validates the data entered. One tool that can aid submitting contextual data is

MetaBar<sup>28</sup>, a spreadsheet and web-based software, designed to assist users in the consistent acquisition, electronic storage and submission of contextual data associated with their samples in compliance with the MIxS standard. The online tool CDinFusion (<http://www.megx.net/CDinFusion>) was created to facilitate the combination of contextual data with sequence data, and generation of submission-ready files.

The next-generation Sequence Read Archive (SRA) collects and displays MIxS-compliant metadata in sample and experiment objects. There are several tools that are already available or under development to assist users in SRA submissions. The myRDP SRA PrepKit allows users to prepare and edit their submissions of reads generated from ultra-high-throughput sequencing technologies. A set of suggested attributes in the data forms assist researchers in providing metadata conforming to checklists such as MIMARKS. The Quantitative Insights Into Microbial Ecology (QIIME) web application (<http://www.microbio.me/qiime>) allows users to generate and validate MIMARKS-compliant templates. These templates can be viewed and completed in the users' spreadsheet editor of choice (e.g., Microsoft Excel). The QIIME web-platform also offers an ontology lookup and geo-referencing tool to aid users when completing the MIMARKS templates. The Investigation/Study/Assay (ISA) is a software suite that assists in the curation, reporting and local management of experimental metadata from studies using one or a combination of technologies, including high-throughput sequencing<sup>29</sup>. Specific ISA configurations (<http://isa-tools.org/tools.html>) have been developed to ensure MIxS compliance by providing templates and validation capability. Another tool, ISAconverter, produces SRA.xml documents, facilitating submission to the SRA repository. MIxS checklists are also registered with the BioSharing catalog of standards (<http://biosharing.org/>), set to progressively link minimal information specifications to the respective exchange formats, ontologies and compliant tools.

Further detailed guidance for submission processes can be found under the respective wiki pages ([http://gensc.org/gc\\_wiki/index.php/MIxS](http://gensc.org/gc_wiki/index.php/MIxS)) of the standard.

### Maintenance of the MIxS standard

To allow further developments, extensions and enhancements of MIxS, we set up a public issue tracking system to track changes and accomplish feature requests (<http://mixs.gensc.org/>). New versions will be released annually. Technically, the MIxS standard, including MIMARKS and the environmental packages, is maintained in a relational database system at the Max Planck Institute for Marine Microbiology Bremen on behalf of the GSC. This provides a secure and stable mechanism for updating the checklist suite and versioning. In the future, we plan to develop programmatic access to this database to allow automatic retrieval of the latest version of each checklist for INSDC databases and for GSC community resources. Moreover, the Genomic Contextual Data Markup Language is a reference implementation of the GSC checklists by the GSC and now implements the full range of MIxS standards. It is based on XML Schema technology and thus serves as an interoperable data exchange format for infrastructures based on web services<sup>30</sup>.

### Conclusions and call for action

The GSC is an international body with a stated mission of working towards richer descriptions of the complete collection of genomes and metagenomes through the MIxS standard. The present report extends the scope of GSC guidelines to marker gene sequences and environmental packages and establishes a single portal where experimentalists



can gain access to and learn how to use GSC guidelines. The GSC is an open initiative that welcomes the participation of the wider community. This includes an open call to contribute to refinements of the MIxS standards and their implementations.

The adoption of the GSC standards by major data providers and organizations, as well as the INSDC, supports efforts to contextually enrich sequence data and complements recent efforts to enrich other (meta) 'omics data. The MIxS standard, including MIMARKS, has been developed to the point that it is ready for use in the publication of sequences. A defined procedure for requesting new features and stable release cycles will facilitate implementation of the standard across the community. Compliance among authors, adoption by journals and use by informatics resources will vastly improve our collective ability to mine and integrate invaluable sequence data collections for knowledge- and application-driven research. In particular, the ability to combine microbial community samples collected from any source, using the universal tree of life as a measure to compare even the most diverse communities, should provide new insights into the dynamic spatiotemporal distribution of microbial life on our planet and on the human body.

Note: Supplementary information is available on the Nature Biotechnology website.

#### ACKNOWLEDGMENTS

Funding sources are listed in the **Supplementary Note**.

#### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/nbt/index.html>.

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

- Field, D. *et al.* The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.* **26**, 541–547 (2008).
- Taylor, C.F. *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* **26**, 889–896 (2008).
- Ludwig, W. & Schleifer, K.H. in *Microbial Phylogeny and Evolution, Concepts and Controversies*. (ed. Sapp, J.) 70–98 (Oxford University Press, New York, 2005).
- Ludwig, W. *et al.* Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* **19**, 554–568 (1998).
- Giovannoni, S.J., Britschgi, T.B., Moyer, C.L. & Field, K.G. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60–63 (1990).
- Stahl, D.A. Analysis of hydrothermal vent associated symbionts by ribosomal RNA sequences. *Science* **224**, 409–411 (1984).
- Ward, D.M., Weller, R. & Bateson, M.M. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* **345**, 63–65 (1990).
- DeLong, E.F. Archaea in coastal marine environments. *Proc. Nat. Acad. Sci. USA* **89**, 5685–5689 (1992).
- Diez, B., Pedros-Alio, C. & Massana, R. Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl. Environ. Microbiol.* **67**, 2932–2941 (2001).
- Fuhrman, J.A., McCallum, K. & Davis, A.A. Novel major archaeobacterial group from marine plankton. *Nature* **356**, 148–149 (1992).
- Hewson, I. & Fuhrman, J.A. Richness and diversity of bacterioplankton species along an estuarine gradient in Moreton Bay, Australia. *Appl. Environ. Microbiol.* **70**, 3425–3433 (2004).
- Huber, J.A., Butterfield, D.A. & Baross, J.A. Temporal changes in archaeal diversity and chemistry in a mid-ocean ridge seafloor habitat. *Appl. Environ. Microbiol.* **68**, 1585–1594 (2002).
- Lopez-Garcia, P., Rodriguez-Valera, F., Pedros-Alio, C. & Moreira, D. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**, 603–607 (2001).
- Moon-van der Staay, S.Y., De Wachter, R. & Vaulot, D. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607–610 (2001).
- Pace, N.R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
- Rappe, M.S. & Giovannoni, S.J. The uncultured microbial majority. *Annu. Rev. Microbiol.* **57**, 369–394 (2003).
- Francis, C.A., Beman, J.M. & Kuypers, M.M.M. New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *ISME J.* **1**, 19–27 (2007).
- Zehr, J.P., Mellon, M.T. & Zani, S. New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of nitrogenase (*nifH*) genes. *Appl. Environ. Microbiol.* **64**, 3444–3450 (1998).
- Minz, D. *et al.* Diversity of sulfate-reducing bacteria in oxic and anoxic regions of a microbial mat characterized by comparative analysis of dissimilatory sulfite reductase genes. *Appl. Environ. Microbiol.* **65**, 4666–4671 (1999).
- Gilbert, J.A. *et al.* The seasonal structure of microbial communities in the Western English Channel. *Environ. Microbiol.* **11**, 3132–3139 (2009).
- Martinez, A.W., Tyson, G. & DeLong, E.F. Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ. Microbiol.* **12**, 222–238 (2009).
- Hanner, R. Data Standards for BARCODE Records in INSDC (BRIs) (Database Working Group, Consortium for the Barcode of Life, 2009). <[http://www.barcodeoflife.org/sites/default/files/legacy/pdf/DWG\\_data\\_standards-Final.pdf](http://www.barcodeoflife.org/sites/default/files/legacy/pdf/DWG_data_standards-Final.pdf)>.
- Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**, 7188–7196 (2007).
- Cole, J.R. *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37**, D141–D145 (2009).
- Vogel, T.M. *et al.* TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat. Rev. Microbiol.* **7**, 252 (2009).
- Turnbaugh, P.J. *et al.* The Human Microbiome Project. *Nature* **449**, 804–810 (2007).
- Benson, D.A. *et al.* GenBank. *Nucleic Acids Res.* **36**, D25–D30 (2008).
- Hankeln, W. *et al.* MetaBar—a tool for consistent contextual data acquisition and standards compliant submission. *BMC Bioinformatics* **11**, 358 (2010).
- Rocca-Serra, P. *et al.* ISA infrastructure: supporting standards-compliant experimental reporting and enabling curation at the community level. *Bioinformatics* **26**, 2354–2356 (2010).
- Kottmann, R. *et al.* A standard MIGS/MIMS compliant XML schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* **12**, 115–121 (2008).

<sup>1</sup>Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany. <sup>2</sup>Jacobs University Bremen gGmbH, Bremen, Germany. <sup>3</sup>Natural Environment Research Council Environmental Bioinformatics Centre, Wallington CEH, Oxford, UK. <sup>4</sup>Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado, USA. <sup>5</sup>Howard Hughes Medical Institute, San Francisco, California, USA. <sup>6</sup>Ribosomal Database Project, Michigan State University, East Lansing, Michigan, USA. <sup>7</sup>Center for Microbial Ecology, Michigan State University, East Lansing, Michigan, USA. <sup>8</sup>The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts, USA. <sup>9</sup>Plymouth Marine Laboratory, Plymouth, UK. <sup>10</sup>Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA. <sup>11</sup>Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA. <sup>12</sup>National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA. <sup>13</sup>European Molecular Biology Laboratory (EMBL) Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. <sup>14</sup>WCU Center for Green Metagenomics, School of Civil and Environmental Engineering, Yonsei University, Seoul, Republic of Korea. <sup>15</sup>School of Computer Science, University of Manchester, Manchester, UK. <sup>16</sup>Oxford e-Research Centre, University of Oxford, Oxford, UK. <sup>17</sup>Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. <sup>18</sup>Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado, USA. <sup>19</sup>Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA. <sup>20</sup>Department of Medicine and the Department of Microbiology, New York University Langone Medical Center, New York, New York, USA. <sup>21</sup>National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA. <sup>22</sup>Department of Microbiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA. <sup>23</sup>DOE Joint Genome Institute, Walnut Creek, California, USA. <sup>24</sup>Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico, USA. <sup>25</sup>Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, USA. <sup>26</sup>Department of Applied Mathematics and Computer Science, Ghent University, Ghent, Belgium. <sup>27</sup>Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, Berkeley, California, USA. <sup>28</sup>Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado, USA. <sup>29</sup>Department of Biological Sciences, University of Southern California, Los Angeles, California, USA. <sup>30</sup>National Ecological Observatory Network, Boulder, Colorado, USA. <sup>31</sup>Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. <sup>32</sup>Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA. <sup>33</sup>Department of Biology, University of New Mexico, LTER Network Office, Albuquerque, New Mexico, USA. <sup>34</sup>Department of Computer Science, University of Colorado, Boulder, Colorado, USA. <sup>35</sup>Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri, USA. <sup>36</sup>University of Colorado Museum of Natural History, University of Colorado,



Boulder, Colorado, USA. <sup>37</sup>Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, Texas, USA. <sup>38</sup>Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Australia. <sup>39</sup>Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. <sup>40</sup>Department of Biology, San Diego State University, San Diego, California, USA. <sup>41</sup>Department of Microbiology, Cornell University, Ithaca, New York, USA. <sup>42</sup>Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA. <sup>43</sup>Lehrstuhl für Mikrobiologie, Technische Universität München, Freising, Germany. <sup>44</sup>J. Craig Venter Institute, Rockville, Maryland, USA. <sup>45</sup>Department of Environmental Sciences, University of Colorado, Boulder, Colorado, USA. <sup>46</sup>Department of Biology, University of Waterloo, Ontario, Canada. <sup>47</sup>Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA. <sup>48</sup>Ribocon GmbH, Bremen, Germany. <sup>49</sup>VIB - Vrije Universiteit Brussel, Brussels, Belgium. <sup>50</sup>Canadian Centre for DNA Barcoding, Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada. <sup>51</sup>Departments of Microbiology and Immunology and Department of Medicine, Stanford University School of Medicine, Stanford, California, USA. <sup>52</sup>Veterans Affairs Palo Alto Health Care System, Palo Alto, California, USA. <sup>53</sup>Department of Microbiology and Immunology, Ann Arbor, Michigan, USA. <sup>54</sup>The Genome Center, Department of Genetics, Washington University in St. Louis School of Medicine, St. Louis, Missouri, USA. Correspondence should be addressed to F.O.G. (fog@mpi-bremen.de).

