UNIVERSITY OF CALIFORNIA, SAN DIEGO

Statistical Representations Of Microbial Systems

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science

by

Yoshiki Vázquez Baeza

Committee in charge:

Professor Rob Knight, Chair
Professor Pieter Dorrestein
Professor Rachel Dutton
Professor Jim Hollan
Professor Larry Smarr

2017

ProQuest Number: 10684635

ProQuest.

ProQuest 10684635

The dissertation of Yoshiki Vázquez Baeza is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2017

To my parents, my family and my friends.

EPIGRAPH

The purpose of computing is insight, not numbers

—Richard Hamming

TABLE OF CONTENTS

ANOSIM	Analysis of Similarities
AUC	Area Under the Curve
BCS	Body Condition Scores
BIOM	Biological Observation Matrix
BMI	Body Mass Index
CA	Correspondence Analysis
CC	Collagenous Colitis
CCD	Colonic Crohn's Disease
CD	Crohn's Disease
CDAI	Clinical Disease Activity Index
CDI	*Clostridium difficile* infection
CGR	Cardiac Glycoside Reductase
CIBDAI	Clinical Canine IBD Activity Index
DEXA	dual-energy X-ray Absorptiometry
ECG	Electrocardiograph
EMP	Earth Microbiome Project
ENA	European Nucleotide Archive
FMT	Fecal Microbiota Transplant

GI	Gastro Intestinal
GLM	Generalized Linear Mixed effects model
GLS	Genetic Load Scores
GNPS	Global Natural Products Social Molecular Networking
HC	Healthy Controls
HGT	Horizontal Gene Transfer
HMP	Human Microbiome Project
HP	Healthy Plane
HTML5	HyperText Markup Language, version 5
IBD	Inflammatory Bowel Disease
ICD	Ileal CD
ICoMM	International Census of Marine Microbes
IQR	InterQuantile Range
IRB	Institutional Review Board
KEGG	Kyoto Encyclopedia of Genes and Genomes
KiNG	Kinemage, Next Generation
LC	Lymphocytic Colitis
MD	Microbial Dysbiosis
MIxS	Minimum information about any (x) sequence
MP	Moving Pictures

MS	Mass Spectrometry
NMDS	Non-metric Multidimensional Scaling
NSAID	Non-steroidal anti-inflammatory drug
OTU	Operational Taxonomic Unit
PCA	Principal Component Analysis
PCoA	Principal Coordinates Analysis
PCR	Polymerase Chain Reaction
PERMANOVA	Permutational Multivariate Analysis of Variance
QIIME	Quantitative Insights into Microbial Ecology
R-CDI	recurrent *Clostridium difficile* infection (CDI)
ROC	Receiver Operating Characteristic
SHIME	Simulator of the Human Intestinal Microbial Ecosystem
SMP	Student Microbiome Project
SNP	Single Nucleotide Polymorphism
SRA	Sequence Read Archive
T2D	Type 2 Diabetes
UC	Ulcerative Colitis
UCL	Standard Upper Control Limit
WebGL	Web Graphics Library

WSAVA   World Small Animal Veterinary Association

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The work in this thesis rests upon the shoulders of giants. Over the past four years these magnificent giants facilitated, enabled, motivated, and supported me. In what follows, I seek to acknowledge the most personally transcendent and quintessential people to the realization of this thesis.

First of all thanks to my committee, for their comments, for their useful feedback and for their willingness to participate in the creation of this somewhat unorthodox thesis (Pieter Dorrestein, Rachel Dutton, Jim Hollan and Larry Smarr).

During these years, I grew as a person and scientist surrounded by dozens of stellar role models. The *Knight Laboratory* was the epicenter of my development. At its core, there's a group that keeps the lab moving financially and logistically. (Ulla Westerman, Jerry Kennedy, Gail Ackermann, Jeff DeReus, Michiko Souza, and Sarah Adams). The next layer is composed by a jovial, diverse, and clever swarm of graduate students. They were my peers and are the fuel steering scientific developments around the world. (Luke Ursell, Will VanTreuren, John Chase, Adam Robbins-Pianka, Justine Debelius, Joshua Shorenstein, Meg Pirrung, Sam Way, Jose Antonio Navas-Molina, Sophie Weiss, Lisa Marotz, Alex Washburne, Tony Walters, Anupriya Tripathi, Alison Vrbanac, Serene Lingjing Jiang, Jamie Morton, Bryn Taylor and Jake Minich). Postdoctoral researchers in the *Knight Laboratory* are role models and drivers of change. Without them mistakes are

made, wheels are re-invented, and progress is slowed down. Their mentoring and advice was essential for my scientific realization. (Juan Manuel Peralta, Sejin Song, Jon Sanders, Daniel McDonald, Greg Caporaso, Tomasz Kosciolek, Jose Carlos Clemente Litran, Jesse Stombaugh, Catherine Lozupone, Jessica Metcalf, Laura Parfrey, Embriette Hyde, Zhenjiang Xu, Luke Thompson, Chris Callewaert, Stefan Janssen, Christian Lauber, Jenya Kopylov, Amnon Amir and Qiyun Zhu). They say it is very easy to notice inadequate work, and hard to notice flawlessly executed work. By that definition, one would never be able to perceive the army of efficient scientists working in the wet lab. Their hard work is what makes computational analysis easy, not computers. (Greg Humphrey, Donna Berg-Lyons, Grant Gogul, James Gaffney, Elaine Wolfe, Caitriona Brennan, Rodolfo Antonio Salido Benítez, Karenina Sanders and Lindsay DeRight Goldasich).

It goes without saying that Rob Knight, my advisor, has had a major role in making this thesis a reality. The magnitude of his support is a perfect match to his contagious enthusiasm and scientific curiosity. I am most grateful for his constant efforts to give me a unique graduate experience.

Moving to a different country can be an intimidating experience. Without careful moral and professional guidance, you can feel lost and constantly confused. Antonio, Lucia, and Edgar knew this and without hesitation helped me as much as they could. It truly is impossible to think where I would be without their support.

Thanks to my friends Santiago and David for always being beacons of inspiration and clarity. Special thanks to my girlfriend Sarah (and the scruffiest most

lovable dog, Joni) for her help improving the quality of this document, and for her personal and professional support.

The force behind my parents' (*Jefesaurios*) will to drive my self-actualization is never-ending and full of love. As such, I have become a person fond of personal development and progress. The impact and importance of this work should fundamentally be credited to them.

Chapter 1, in full, is a reprint of the material as it appears in "Impacts of the human gut microbiome in therapeutics". Y. Vázquez-Baeza, C. Callewaert, J. Debelius, E. Hyde, C. Marotz, J. T. Morton, A. Swafford, A. Vrbanac, P. C. Dorrestein and R. Knight. *Annual Review of Pharmacology and Toxicology, 2017.* 58, 2017. The dissertation author was the primary investigator and author of this paper.

Chapter 2.1, in full, is a reprint of the material as it appears in "Emperor: a tool for visualizing high throughput microbial community data". Y. Vázquez-Baeza, A. Gonzalez, M. Pirrung, R. Knight. *GigaScience*, 2, 2013 The dissertation author was the primary investigator and author of this paper. In addition, we thank Jackson Chen, Jai Ram Rideout, Daniel McDonald, William Van Treuren, Jose Antonio Navas-Molina, Nickolas A. Bokulich, Adam Robbins-Pianka and Greg Caporaso for feedback and useful discussion regarding the design and implementation of the software package. This work was supported in part by the National Institutes of Health, the Crohns and Colitis Foundation of America, the Alfred P. Sloan Foundation, and the Howard Hughes Medical Institute. The dissertation

author was the primary investigator and author of this paper.

Chapter 2.2, in full, is a reprint of the material as it appears in "Bringing the Dynamic Microbiome to Life with Animations". Y. Vázquez-Baeza, A. Gonzalez, L. Smarr, D. McDonald, J. T. Morton, J. A. Navas-Molina, R. Knight. *Cell Host and Microbe*, 11, 2017. The dissertation author was the primary investigator and author of this paper.

Chapter 3.1, in full, is a reprint of the material as it appears in "Dog and human inflammatory bowel disease rely on overlapping yet distinct dysbiosis networks". Y. Vázquez-Baeza, E. R. Hyde, J. S. Suchodolski, R. Knight. *Nature Microbiology*, 1, 2016. The dissertation author was the primary investigator and author of this paper. We wish to acknowledge the support provided by the Crohn's and Colitis Foundation of American; the Templeton Foundation and the Keck Foundation (via the Earth Microbiome Project), the National institutes of Health. We wish to thank Zhenjiang Xu, Jon Sanders, Amnon Amir, Gail Ackermann, Jamie Morton, Luke Ursell, Jessica Metcalf, Antonio Gonzalez and Emma Schwager for their useful comments and feedback in the writing of this manuscript.

Chapter 4.1, in full, is a reprint of the material as it appears in "Dynamics of the human gut microbiome in inflammatory bowel disease". J. Halfvarson, C. J. Brislawn, R. Lamendella, Y. Vázquez-Baeza, W. A. Walters, L. M. Bramer, M. D'Amato, F. Bonfiglio, D. McDonald, A. Gonzalez, E. E. McClure, M. F. Dunklebarger, R. Knight, J. K. Jansson. *Nature Microbiology*, 2, 2017. The dissertation author was the primary analyst and a co-author of this paper.

Chapter 4.2, in full, is a reprint of the material as it appears in "Guiding longitudinal sampling in inflammatory bowel diseases cohorts". Y. Vázquez-Baeza, A. Gonzalez, Z. Zech Xu, A. Washburne, H. Herfarth, R. B. Sartor, R. Knight. *Gut.* 2017. The dissertation author was the primary investigator and author of this paper. We acknowledge grant support by NIH (P01-DK094779), the Crohn's and Colitis Foundation, and the research coordinators of the UNC Multidisciplinary IBD Center at UNC for patient recruitment. Additionally, we thank Gail Ackermann, Jamie Morton and Justin Silverman for their invaluable feedback and discussion provided in the preparation of this manuscript.

Chapter 5.1, in full, is a reprint of the material as it appears in "Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent Clostridium difficile infection". A. Weingarden, A. Gonzalez, Y. Vázquez-Baeza, S. Weiss, G. Humphry, D. Berg-Lyons, D. Knights, T. Unno, A. Bobr, J. Kang, A. Khoruts, R. Knight, M. J. Sadowsky. *Microbiome.* 3, 2015. The dissertation author was a co-primary investigator and author of this paper. This research was supported, in part, by NIH Grant R21AI091907 (to AK and MJS). ARW was supported by a Doctoral Dissertation Fellowship from the University of Minnesota Graduate School and by the Dennis W. Watson Fellowship in Microbiology. Work done in the Knight lab was funded by the NIH, Crohn's & Colitis Foundation of America, and the HHMI.

Chapter 5.2, in full, is a reprint of the material as it appears in "Changes in microbial ecology after fecal microbiota transplantation for recurrent C. diffi-

cile infection affected by underlying inflammatory bowel disease". S. Khanna, Y. Vázquez-Baeza, A. Gonzalez, S. Weiss, B. Schmidt, D. A. Muiz-Pedrogo, J. F. Rainey 3rd, P. Kammer, H. Nelson, M. Sadowsky, A. Khoruts, S. L. Farrugia, R. Knight, D. S. Pardi, P. C. Kashyap. *Microbiome.* 5, 2017. The dissertation author was the co-primary investigator and author of this paper.

2012	B. S. in Biomedical Engineering, Universidad Iberoamericana, Ciudad de México
2016	M. Sc. in Computer Science, University of California, San Diego
2017	Ph. D. in Computer Science, University of California, San Diego

PUBLICATIONS

Author names marked with † indicate shared first co-authorship.

Y. Vázquez-Baeza, A. Gonzalez, M. Pirrung, R. Knight. "Emperor: a tool for visualizing high throughput microbial community data", *GigaScience*, 2, 2013.

Y. Vázquez-Baeza, A. Gonzalez, L. Smarr, D. McDonald, J. T. Morton, J. A. Navas-Molina, R. Knight. "Bringing the Dynamic Microbiome to Life with Animations", *Cell Host and Microbe*, 11, 2017.

Y. Vázquez-Baeza, E. R. Hyde, J. S. Suchodolski, R. Knight. "Dog and human inflammatory bowel disease rely on overlapping yet distinct dysbiosis networks", *Nature Microbiology*, 1, 2016.

†S. Khanna, †**Y. Vázquez-Baeza**, A. Gonzalez, S. Weiss, B. Schmidt, D. A. Muiz-Pedrogo, J. F. Rainey 3rd, P. Kammer, H. Nelson, M. Sadowsky, A. Khoruts, S. L. Farrugia, R. Knight, D. S. Pardi, P. C. Kashyap. "Changes in microbial ecology after fecal microbiota transplantation for recurrent C. difficile infection affected by underlying inflammatory bowel disease", *Microbiome*. 5, 2017.

J. Halfvarson, C. J. Brislawn, R. Lamendella, **Y. Vázquez-Baeza**, W. A. Walters, L. M. Bramer, M. D'Amato, F. Bonfiglio, D. McDonald, A. Gonzalez, E. E. McClure, M. F. Dunklebarger, R. Knight, J. K. Jansson. "Dynamics of the human gut microbiome in inflammatory bowel disease", *Nature Microbiology*, 2, 2017.

†A. Weingarden, A. Gonzalez, †**Y. Vázquez-Baeza**, S. Weiss, G. Humphry, D. Berg-Lyons, D. Knights, T. Unno, A. Bobr, J. Kang, A. Khoruts, R. Knight, M. J. Sadowsky. "Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent Clostridium difficile infection". *Microbiome*. 3, 2015.

Y. Vázquez-Baeza, C. Callewaert, J. Debelius, E. Hyde, C. Marotz, J. T. Morton, A. Swafford, A. Vrbanac, P. C. Dorrestein and R. Knight. "Impacts of the human gut microbiome in therapeutics". *Annual Reviews.* 58, 2017.

Y. Vázquez-Baeza, A. Gonzalez, Z. Zech Xu, A. Washburne, H. Herfarth, R. B. Sartor, R. Knight. "Guiding longitudinal sampling in inflammatory bowel diseases cohorts". *Gut.* 2017.

The following publications were not included as part of this dissertation, but were also significant byproducts of my doctoral training.

C. A. Lozupone, J. Stombaugh, A. Gonzalez, G. Ackermann, D. Wendel, **Y. Vázquez-Baeza**, J. K. Jansson, J. I. Gordon, R. Knight. "Meta-analyses of studies of the human microbiota". *Genome Research.* 23, 2013.

J. A. Navas-Molina, J. M. Peralta-Sánchez, A. Gonzalez, P. J. McMurdie, **Y. Vázquez-Baeza**, Z. Xu, L. K Ursell, C. Lauber, H. Zhou, S. Jin Song, J. Huntley, G. L Ackermann, D. Berg-Lyons, S. Holmes,. Gregory Caporaso, R. Knight. "Advancing our understanding of the human microbiome using QIIME". *Methods in enzymology.* 531, 2013.

G. E Flores, G. Caporaso, J. B Henley, J. R. Rideout, D. Domogala, J. Chase, J. W Leff, **Y. Vázquez-Baeza**, A. Gonzalez, R. Knight, R. R. Dunn, N. Fierer. "Temporal variability is a personalized feature of the human microbiome". *Genome biology.* 15, 2014.

S. Lax, D. P Smith, J. Hampton-Marcell, S. M Owens, K. M Handley, N. M Scott, S. M Gibbons, P. Larsen, B. D Shogan, S. Weiss, J. L Metcalf, L. K Ursell, **Y. Vázquez-Baeza**, W. Van Treuren, N. A Hasan, M. K Gibson, R. Colwell, G. Dantas, R. Knight, J. A Gilbert. "Longitudinal analysis of microbial interaction between humans and the indoor environment". *Science.* 345, 2014.

D. Gevers, S. Kugathasan, L. A Denson, **Y. Vázquez-Baeza**, W. Van Treuren, B. Ren, E. Schwager, D. Knights, S. Jin Song, M. Yassour, X. C Morgan, A. D Kostic, C. Luo, A. Gonzalez, D. McDonald, Y. Haberman, T. Walters, S. Baker, J. Rosh, M. Stephens, M. Heyman, J. Markowitz, R. Baldassano, A. Griffiths, F. Sylvester, D. Mack, S. Kim, W. Crandall, J. Hyams, C. Huttenhower, R. Knight, R. J Xavier. "The treatment-naive microbiome in new-onset Crohn's disease". *Cell Host & Microbe.* 3, 2014.

†A. Gonzalez, †**Y. Vázquez-Baeza**, R. Knight. "SnapShot: the human microbiome". *Cell.* 158, 2014.

†A. Gonzalez, †**Y. Vázquez-Baeza**, J. B. Pettengill, A. Ottesen, D. McDonald, R. Knight. "Avoiding Pandemic Fears in the Subway and Conquering the Platypus". *mSystems.* 1, 2016.

R. A Quinn, J. A Navas-Molina, E. R Hyde, S. Jin Song, **Y. Vázquez-Baeza**, G. Humphrey, J. Gaffney, J. J Minich, A. V Melnik, J. Herschend, J. DeReus, A. Durant, R. J Dutton, M. Khosroheidari, C. Green, R. da Silva, P. C Dorrestein, R. Knight "From sample to Multi-Omics conclusions in under 48 Hours". *mSystems.* 1, 2016.

J. W Debelius, **Y. Vázquez-Baeza**, D. McDonald, Z. Xu, E. Wolfe, R. Knight. "Turning participatory microbiome research into usable data: lessons from the American Gut Project". *Journal of microbiology & biology education.* 1, 2016.

J. Debelius, S. Jin Song, **Y. Vázquez-Baeza**, Z. Zech Xu, A. Gonzalez, R. Knight. "Tiny microbes, enormous impacts: what matters in gut microbiome studies?". *Genome Biology.* 17, 2017.

J. T Morton, J. Sanders, R. A Quinn, D. McDonald, A. Gonzalez, **Y. Vázquez-Baeza**, J. A Navas-Molina, S. Jin Song, J. L Metcalf, E. R Hyde, M. Lladser, P. C Dorrestein, R. Knight. "Balance trees reveal microbial niche differentiation". *mSystems.* 2, 2017.

S. Weiss, Z. Zech Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, **Y. Vázquez-Baeza**, A. Birmingham, E. R Hyde, R. Knight "Normalization and microbial differential abundance strategies depend upon data characteristics". *Microbiome.* 5, 2017.

L. R. Thompson, J. G. Sanders, D. McDonald, A. Amir, J. Ladau, K. J. Locey, R. J. Prill, A. Tripathi, S. M. Gibbons, G. Ackermann, J. A. Navas-Molina, S. Janssen, E. Kopylova, **Y. Vázquez-Baeza**, A. Gonzalez, J. T. Morton, S. Mirarab, Z. Z. Xu, L. Jiang, M. F. Haroon, J. Kanbar, Q. Zhu, S. Song, T. Kosciolek, N. A. Bokulich J. Lefler, C. J. Brislawn, G. Humphrey, S. M. Owens, J. Hampton-Marcell, D. Berg-Lyons, V. McKenzie, N. Fierer, J. A. Fuhrman, A. Clauset, R. L. Stevens, A. Shade, K. S. Pollard, K. D. Goodwin, J. K. Jansson, J. A. Gilbert, R. Knight and The Earth Microbiome Project Consortium. "A communal catalogue reveals Earths multiscale microbial diversity". *Nature.* 2017.

ABSTRACT OF THE DISSERTATION

Statistical Representations Of Microbial Systems

by

Yoshiki Vázquez Baeza

Doctor of Philosophy in Computer Science

University of California, San Diego, 2017

Professor Rob Knight, Chair

Technological developments in the past thirty years have transformed sequencing-based microbiology into a data-intensive field. Here, computing and efficient representations are catalyzers of insight into omnipresent and complex microbial interactions. Notably, classical ecologists have set the foundations for the way we analyze these systems, with some techniques dating back to the beginning of the twentieth century. In this thesis, we expand and where possible reuse these techniques to unravel the hidden patterns comprising the human gut microbiome.

To set an appropriate motivation and context for the rest of this work, Chapter 1 reviews recent discoveries on the human microbiome and how the communities within can influence the effectiveness of therapeutic agents. Next, in Chapter 2, we introduce EMPeror, an interactive analysis and visualization tool that is crucial to the findings presented in later chapters.

The following three chapters study concrete examples where the microbiome has been implicated as a driver or marker for dysbiosis. Chapter 3 describes how the microbial signature associated with Crohn's Disease (CD) in humans, described in our previous work [61], is overlapping but distinct to that of dogs affected with Inflammatory Bowel Disease (IBD). Surprisingly, unlike with humans, dog fecal samples alone are strong indicators of the disease. In Chapter 4, we study IBD from a longitudinal perspective, revealing increased volatility in the gut microbiomes of subjects with IBD, a property that does not appear to be present in unaffected controls. Furthermore, we use this as a predicting feature of the disease, and improve on the classification accuracy possible through a single fecal sample. In Chapter 5, we study the effect of Fecal Microbiota Transplants (FMTs) to treat CDIs and, using the techniques described in Chapter 2, we show the first animated visualization of this process, a dramatic microbial transformation as the subjects recover from all CDI symptoms. In addition, for CDI patients who also suffer from a subtype of IBD, a treatment with a FMT results in an increased number of relapses and decreased microbial diversity.

The closing chapter discusses these results and their possible applications,

as well as future directions for computationally-centric microbiome research.

Chapter 1

The hidden world within

ourselves

Recent years have seen an increased interest in the study of microbial communities, partly driven by the affordability of the enabling technologies. This new focus provides a previously ignored axis of understanding to a vast number of research fields. In addition, this has uncovered a number of associations and demonstrated mechanisms, where microbial communities are key explanatory variables.

In fields where microbial communities were already taken into consideration, these advancements allowed experiments to be conducted at an unprecedented scale [83, 258, 63]. As a consequence, the underlying software and methods to analyze these studies needed to accommodate volumes of information previously unseen. Notably, most of the software used in this context has been developed to interpret

and analyze interdisciplinary experiments that span a broad range of seemingly disjoint fields (for example biofuel development [248], built environment [28, 44], forensics [163, 54], etc). As such, we motivate the rest of the work in this thesis by presenting a comprehensive survey focused on the human microbiome. The following section appeared in the journal *Annual Review of Pharmacology and Toxicology, 2017.*

1.1 Impacts of the human gut microbiome on therapeutics

The human microbiome contains a vast source of genetic and biochemical variation, and its impacts on therapeutic responses are just beginning to be understood. This expanded understanding is especially important because the human microbiome differs far more among different people than does the human genome, and it is also dramatically easier to change. Here we describe some of the major factors driving differences in the human microbiome among individuals and populations. We then describe some of the many ways in which gut microbes modify the action of specific chemotherapeutic agents, ranging from Non-steroidal anti-inflammatory drugs (NSAIDs) to cardiac glycosides, and outline the potential of Fecal Microbiota Transplant (FMT) as a therapeutic. Intriguingly, microbes also alter how hosts respond to therapeutic agents through various pathways acting at distal sites. Finally, we discuss some of the computational and practical issues surrounding use of the microbiome to stratify individuals for drug response, and envision a future where the microbiome will be modified to increase everyone's potential to benefit from therapy.

1.1.1 Introduction

Although extensive work has linked the human genome to drug response, the idea that the microbiome, the collection of genes in our microbial symbionts,

could affect drug response is much newer, with the majority of work conducted in the past ten years. This lack of attention is surprising. As far back as the 1970s, it was estimated that the number of microbial cells, mostly bacteria, associated with the human body was greater than the number of human cells carrying the human genome [198]. This observation has held to the present, albeit with narrowed error bars, leading to an estimate of 53% microbial cells to 47% human cells [205]. An important and more recent observation is that the unique microbial genes greatly outnumber the unique human genes, with estimates of the microbial genome catalog typically numbering into the millions [83, 185, 227] and dwarfing the 20,000 or so human genes [127]. What role all these microbial genes play in drug metabolism, and what effect drugs in turn have on the microbiome, are just starting to be explored.

A key barrier was the inability to identify the specific types of organisms associated with each individual. Traditionally, the limiting factor was the use of culture-dependent approaches which enabled only a small percentage of microbial strains from the gut to be cultured. Newer techniques, such as the use of initially germ-free mice as a culture medium [228] and advances in microdrop culturing [22], have greatly expanded this repertoire, but each system is still limited by the requirement for active growth. The ability to sequence DNA directly from the environment, pioneered by Pace and colleagues, revolutionized our view of the microbial world by eliminating the constraint of culture, with many studies focusing on the 16S rRNA gene as a universal taxonomic marker [177]. The first study

applying culture-independent analysis on a large scale to the human gastrointestinal tract [45] revealed dramatic differences among individuals, and sites in the distal large intestine. As in prior studies, stool served as a useful proxy for the intestinal contents (where most metabolism takes place) but failed to capture the diversity of any site perfectly [45]. Highly multiplexed assays with next-generation sequencing [74] allowed hundreds of microbiomes to be read out at once, dramatically decreasing the cost of assessing the microbiome. Further improvements in sequencing methodology [83, 185] and software techniques [23, 108, 172] dramatically expanded the accessibility of microbiome studies to a wide range of investigators, both at the marker gene level and at the shotgun metagenomic level [66], where total DNA is extracted and all the genes analyzed.

Results from these studies re-conceptualized our view of the microbiome. For example, rather than a common core of shared organisms, different people have very different collections of microbes, even if they share a home and/or genetic material. Changes in the microbiome throughout the lifespan are dramatic, especially in early life, and are greater than microbiome differences among adult mammalian species. Such differences in different parts of the body are comparable to the differences among microbial communities in different physical environments, such as soil and water. These large differences in species and gene composition and their implications for drug metabolism are illuminated by studies that combine DNA sequencing with other methods, including metabolomics analyses.

Now that identifying the taxa in a given microbiome sample is routine, and

identifying the gene repertoire increasingly accessible, one can link the individual diversity in gene content to drug metabolism and other gut-related associations (Figure 1.1.1). However, a complicating factor, especially for case/control study designs, is the issue of causality: many drugs, ranging from the obvious (antibiotics) to the surprising (metformin), dramatically affect the microbiome, so separating cause and effect can be difficult. It is perhaps most useful to think of the microbiome as a dynamic system in which the microbiome affects how drugs are metabolized and these products of metabolism in turn change the microbiome, which then changes future responses even to the same pharmaceutical. Such dynamic systems thinking is common in ecology, and although many current microbiome research techniques stem from this discipline, its application remains relatively rare in medicine.

In this review, we focus on the human gut microbiome, although, where appropriate, we refer to studies using animal models, especially gnotobiotic mice (germ-free mice colonized with a defined set of microbes, either pure strains or fecal material from a rodent or human donor). As noted above, different parts of the human body (e.g., mouth, skin, airways, and urogenital tract) have distinct microbiomes from the gut microbiome, but the gut has been most studied in terms of drug response, in part because it contains by far the greatest microbial biomass in the human body, and therefore, the greatest capacity for drug metabolism. We first provide an overview of the diversity of the human gut microbiome, identifying key factors that structure the microbiome in healthy and diseased populations. We

Continuum of Expected/Unexpected alterations of the Microbiome and their Associations

Circadian Clock Alterations (Leone et al. 2015)

Gut-Brain axis (Conn et al. 1977)

PTSD in Mice (Reber et al. 2016)

Gastric Cancer (Nomura et al. 1991)

Parkinson's Disease (Scheperjans et al 2015, Sampson et al. 2016)

Delivery Mode (Dominguez-Bello et al. 2010)

Metformin-induced alterations (Wu et al. 2017)

Lack of a Core Gut Microbiome (HMP Consortium, 2012)

Inflammatory Bowel Disease (Weinstein et al. 1961, Gevers et al. 2014)

Obesity (Let et al. 2005)

Celiac Disease (Wacklin et al. 2013)

Diet Changes (David et al. 2014)

Fecal Microbiota Transplantations (Van Nood et al. 2013)

Unexpected

Expected

Figure 1.1: A list of associations between the gut microbiome and various phenotypes, ranked by how expected or unexpected they were relative to previous knowledge [rank reflects consensus opinions of the co-authors of this Review, motivated by Figure 1 in [32]]

then describe connections between gut microbes and the action of several therapeutic agents, ranging from antibiotics to analgesics. Third, we describe how the microbiome can alter host response to drugs, though not necessarily metabolizing the drugs themselves. Fourth, we discuss prospects for identifying which patients will respond to a given drug based on the whole microbiome and prospects for exploiting the microbiome as a companion diagnostic to increase efficacy or reduce adverse events. Finally, we provide an outlook for the field in terms of improved patient stratification, and raise the intriguing prospect that patients might be converted from non-responders to responders via the use of companion therapies that alter the microbiome to increase efficacy of a specific drug.

1.1.2 Diversity of the human gut microbiome across populations and life stages

There are very few germ-free humans; therefore, understanding differences among individuals colonized with different microbes is key for predicting those who will respond similarly to drugs. However, many of the factors affecting the microbiome are highly counterintuitive. Here we focus primarily on a recent review [39], five large-scale studies [9, 51, 83, 258, 266], and two large-scale meta-analyses [141, 234] that summarize these findings.

As noted above, different sites in the human body host highly distinct microbial communities: on a Principal Coordinates Analysis 'map' of the micro-

biome, they emerge as different 'continents' [34, 83]. However, the differences in microbiome composition between, for example, the mouth and the gut of the same person are much greater than the differences between soil samples collected from different continents, and instead are more comparable to the difference between soil and seawater [133]. Although different body sites are highly distinct at the level of microbial taxa, they are relatively similar in terms of major microbial metabolic functions, at least as measured by Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways or other broad levels of gene function [83, 228].

This pattern of microbial biogeography extends to the human gut [12, 34, 45, 265]. The microbiome changes dramatically along the gastrointestinal tract, with distinct populations in the oral cavity, esophagus, stomach, and small and large intestine. The most abundant genera of the oral cavity include *Actinomyces*, *Streptococcus*, *Neisseria*, and *Veillonella*, while the throat and stomach are dominated by *Streptococcus* and *Prevotella* species. In the stomach, roughly half of individuals are colonized by *Helicobacteria pylori*, which dominates the gastric microbiome in these individuals [11]. The small intestine harbors fast-growing, G+C rich organisms specialized in digesting simple carbohydrates [267] and is often dominated by genera such as *Peptostreptococcus*, which are rare in the large intestine. The bulk of the microbiome and microbial metabolism is in the large intestine. Typical bacterial loads are 105 CFU/ml in the jejunum, 106 CFU/ml in the ileum, but 108 CFU/ml in the cecum [149], and 1011-1012 CFU/g in the feces [212]. This, together with stool being a relatively good proxy for the large

intestine luminal contents but much more convenient to sample, means that most attention has focused on stool. However, for some applications such as detecting inflammatory bowel disease, microbiome analyses from biopsies dramatically outperform analyses of stool in their ability to distinguish healthy from diseased subjects [61]. The range of study designs where biopsies are essential is still hotly debated; nonetheless, stool biomarkers are highly informative for many applications, as noted below.

The process of development from infancy to early childhood has the greatest known effect on the stool microbiome. This effect is so profound that it traverses the vast gulf in microbiome configuration among body sites (Figure 1.1.2). As they pass through the birth canal, newborns are typically colonized by bacteria resembling the constituents of the mother's vaginal microbiome. By contrast, newborns delivered by C-section without labor lack these co-evolved vaginal microbes. Instead their microbiome resembles the skin, either from direct contact with caregivers and medical staff or from dust, which largely consists of human skin flakes and their associated microbes in indoor environments [14, 31, 43]. The first few weeks of life are chaos in the infant gut microbiome [83, 178] with very large differences at successive time points and among children. The subsequent overall process of development to the adult state takes about 2-3 years [14, 83, 258], although smaller changes continue throughout childhood [258]. Intriguingly, although the process of development to the adult state is consistent cross-culturally, the state that is reached varies across populations [258]. These major differences

10

in the microbiome may explain some of the differences in drug response between children and adults that are not explained by other factors.

Drugs that work in one human population may fail for unknown reasons in other populations, though the reasons for this are often obscure. Large efforts (and amounts of money) have sought the answer by studying human genomes. However, human genomes are 99.9% the same but gut microbiomes can be 90% different [73]. Given the rate of microbial metabolism and diversity of microbially-encoded enzymes, it makes sense to inquire whether differences in drug responses are linked to the gut microbiome. Indeed, different human populations have markedly different microbiomes, which are largely thought to be linked to diet, with an axis that trades *Prevotella* in diets rich in sugar and grains for *Bacteroides* in diets rich in animal protein. These trends have been observed in the US population [255] and among US, South American, and African populations [258]. However, diet cannot be the sole driver, or between-population microbiome variation would not exceed the within-population variation, yet this is what one observes.

Even with hundreds of individuals, it is not yet possible to ascertain the factors at the population level that are most important in structuring the adult gut microbiome, because diet, environmental (including drug) exposures, host genetics, and other factors vary among populations in complex ways. Studies encompassing dozens of populations, with detailed tracking of individual- and population-level information, are likely required to resolve these questions, just as for such questions in human population genetics [183]. Hunter-gatherer populations, such as the
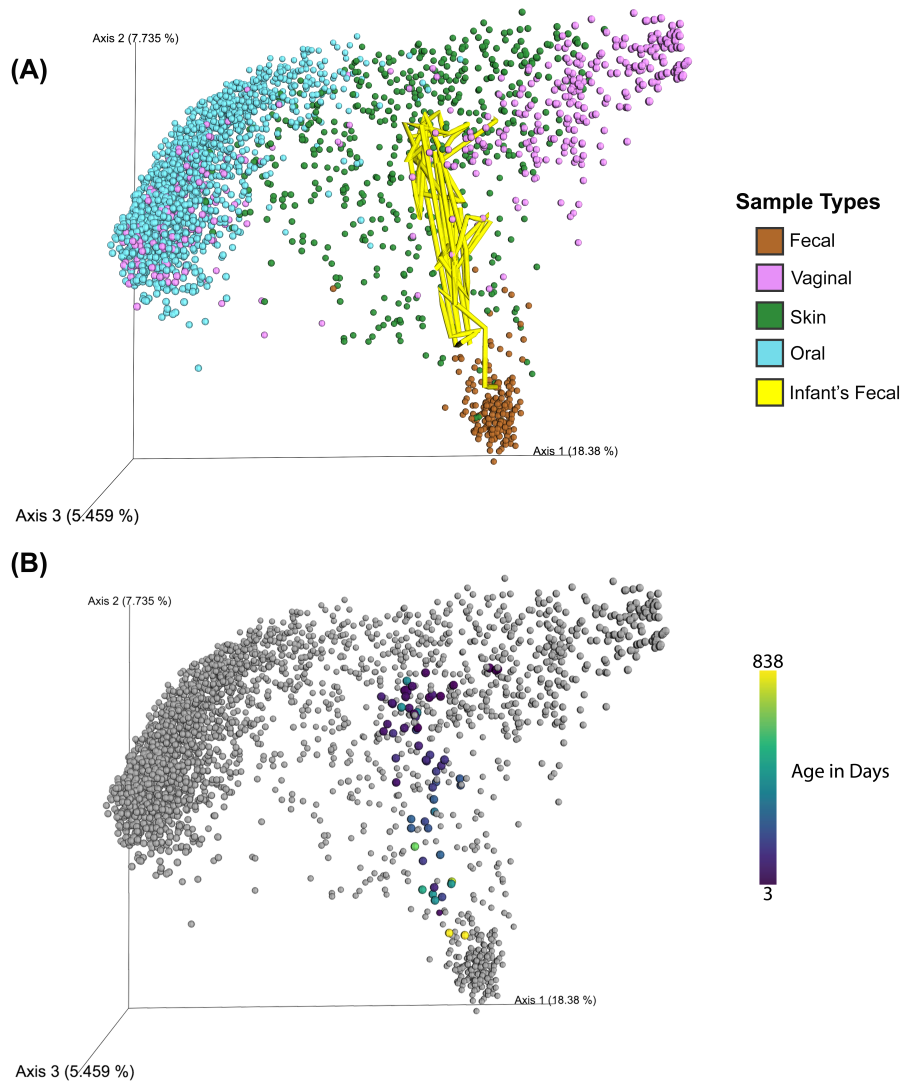
Figure 1.2: Principal coordinates analysis on unweighted UniFrac distances of the process of microbiome assembly in the gut of one infant, contextualized with the Human Microbiome Project. Panel (A) colored by the sample type, and (B) colored by infant's age. As time goes by, the microbial communities in the infant's stool shift from resembling healthy adult vaginal and skin communities to resemble stool composition. This demonstrates how the process of gut microbiome development spans some of the most profound changes in the microbiome, i.e. the distance across different body sites.

Yanomami in the Amazon basin and the Hadza in sub-Saharan Africa, have distinct microbiomes harboring entire phyla, such as Spirochetes, that are even absent from rural farming populations [33, 201, 229]. This suggests that the loss of diversity from the microbiome of modern populations may have begun with the advent of agriculture. However, other factors such as C-section, antibiotics, and hygiene practices may be leading to continual depletion of our natural microbial symbionts [19].

Several population-based studies that examine many variables have begun to provide insights into factors that shape microbiomes within populations. Drugs, especially antibiotics, but also metformin and proton pump inhibitors, have a surprisingly large effect on the microbiome and must be controlled for in studies that seek to link microbiomes to disease [65]. Long-term diet has a large effect, especially the balance of carbohydrates and protein. Age and season have large effects, although it is not known whether these are mediated in part by diet. Having a dog has an intermediate effect, smaller in the gut than in the skin and in the indoor environment [44, 214]. Many other factors, including sex, Body Mass Index (BMI), and even amounts of sleep and exercise can be identified as differences among groups that contain hundreds of subjects per group but cannot be used to classify individuals. Host genetics has a surprisingly small impact in humans compared to other factors: although monozygotic twins are slightly more similar to one another than are dizygotic twins, hundreds of twin pairs are needed to see this effect. However, some specific components of the microbiome are highly

heritable and correlate with phenotype. For example, Christensenella, identified by microbiome-wide heritability analysis, correlated with low BMI, including in slimmed down germ-free mice that received an otherwise obesogenic microbiome from an obese human donor. Replicating this feat to manipulate weight or other phenotypes in humans remains a goal for future studies.

1.1.3 Connections between specific gut microbes and action of therapeutic agents

The literature demonstrating how gut microbes affect the efficacy or toxicity of specific drugs is expanding rapidly. Here we provide a few illustrative examples see also Table 1.1 for a brief summary. For additional examples, readers should consult recent reviews [217, 47].

Analgesics

NSAIDs are used for pain relief in many conditions. Multiple studies have demonstrated that NSAID use is associated with damage to the small intestine; the severity of this damage depends on the microbial community present at the time of NSAID administration [83, 176]. Increased Gram-negative taxa have been associated with more severe damage [176, 259]. However, the precise link between NSAID-induced damage and the microbiome remains poorly understood. Antibiotic treatment can alter the existing microbial community enough to reduce NSAID-associated small intestinal damage in humans [200], and can alter micro-

bial metabolism of aspirin to potentiate its anti-thrombotic effects in mice [109]. However, other known negative effects of antibiotic treatment suggest risks in using antibiotics to reduce adverse events from specific combinations of microbiome state and NSAID use. For example, antibiotic treatment of mice transplanted with a microbiome considered to be 'protective' against NSAID damage produced higher mortality with NSAID administration [257], underscoring that the association between NSAID usage, microbiome community structure, and bowel damage reduction is complex and must be approached carefully. Probiotic strains have been deployed to prevent NSAID-induced bowel damage. For example, protective effects of *Lactobacillus gasseri* against aspirin-induced small intestinal damage have been assessed in human subjects [223], and although results were promising, studies in larger populations are needed to evaluate the efficacy of this approach. In addition, the impact of the gut microbiome on the bioavailability, efficacy and toxicity of orally administered drugs, xenobiotics and dietary substances was studied for >30 pro-drugs [216]. Many prodrugs that treat gut inflammation are converted to active forms by the gut microbiome. These include sulfasalazine, olsalazine, ipsalazide and balsalazide, for which the aminosalicylic acid needs to be released in order to become active [181]. Gut bacteria of the genera Clostridia and Eubacteria exhibit this ability [187].

Antibiotics

Although antibiotics influence the composition of the microbiome, the effects of antibiotic treatment depend on its starting composition. The extent of cellular damage induced by ex vivo antibiotic exposure of fecal samples depends on the individual's original microbiota [156]. This is consistent with previous data demonstrating species-specific susceptibility to antibiotics [171]. Healthy adults with similar microbial communities prior to treatment with cefprozil had similar alterations in the abundance of *Enterobacter cloaceae* following treatment [189]. The gut microbiome may also regulate the effect of antibiotics at distal sites of infection. Resistance genes that lead to chemical modifications, such as hydrolysis or chemical modification, may decrease the bioavailability of the antibiotic [18], potentially leading to a lower effective concentration at the primary site of infection. Future work is needed to investigate whether the overall gut microbiome composition influences these resistance genes.

Cardiac glycosides

Cardiac glycosides are a family of organic compounds that increase the contractility of the heart and are prescribed mainly for congestive heart failure and cardiac arrhythmia. The bioavailability of one of the most common cardiac glycosides, digoxin, varies greatly among patients. Early experiments showed that some patients' gut microbiota can metabolize digoxin into inactive metabolites and that pretreatment with antibiotics in individuals harboring these microbiota increased

the plasma level of digoxin [135]. Detailed metagenomic analysis identified a Cardiac Glycoside Reductase (CGR) operon in some, but not other, strains of the commensal gut bacteria *Eggerthella lenta*, which is responsible for the degradation of this drug [70]. Studies in animal models revealed that supplementation with the amino acid arginine can suppress the cgr operon and increase digoxin bioavailability. These experiments underline the importance of the microbiome in drug metabolism and provide an example where targeted therapeutic alternatives, in this case arginine supplementation in place of antibiotic treatment, can improve clinical outcomes.

Metformin

Metformin is used in the treatment of Type 2 Diabetes (T2D). Since its introduction in the 1950s [153], many mechanisms of action have been proposed to explain its effects on the liver and insulin-sensitive tissues [110]. However, a recent report of trans-species transmission of the therapeutic effect of improved glucose tolerance from the stool of metformin-treated humans into germ-free mice [256] suggests that the insulin-sensitizing effects of the drug may result from a metabolic shift in the gut microbiome rather than direct effects on the liver or other tissues. Further evidence for this mechanism is suggested by the fact that most common side-effects of metformin are GI-centric (nausea, vomiting and diarrhea), and delivery of metformin into the portal vein did not change hepatic glucose production [158].

Fecal microbiota transplants (FMT)

Microbiome transplants could potentially be used for many indications in the gut, but also in skin, oral, eye and vaginal health. The most prominent example is the treatment of recurrent *Clostridium difficile* infection (CDI). The standard of care for recurrent CDI is antibiotics, which inadvertently deplete other microbes necessary for healthy gut function. FMT have outstanding efficacy in treating CDIs [233]. The procedure relies on finding a healthy donor (often a close-relative or friend, although the evidentiary basis for this choice is weak), who donates fecal microbes for transplant into the diseased subject. Changes occur in the microbial communities of the recipient in less than a day, and disease-associated symptoms quickly disappear [244]. A meta-analysis of longitudinal datasets revealed that post-FMT subjects also recover dynamic features common to healthy humans [16]. FMTs are not 100% effective and there is even anecdotal evidence of unintended changes, such as sudden excessive weight gain post-FMT [10]. This has led researchers to seek optimal conditions for effective FMT, such as antibiotic pre-treatment [90, 99]. From an ecological standpoint, the logic behind this is straightforward, because newly introduced communities need not compete with an established community of bacteria, and can develop interactions likely already present in the donor's gut. Outside the gut, adding microbes from healthy individuals to skin afflicted with atopic dermatitis reduces symptoms [168] and vaginal microbe transplants from healthy individuals may overcome bacterial vaginosis

[clinical trial[1]]. The microbiome field is still young so other beneficial transplant therapies will likely be developed in the future. However, public enthusiasm for such treatments and for probiotics still greatly outstrips the available scientific evidence.

1.1.4   Microbiome effects on host responses to drugs

An important emerging area of investigation in drug response is the effects of the microbiome on host responses to chemotherapeutics. Although sequencing tumors and pharmacogenomics plays a key role in precision medicine, our microbial genomes can also affect host gene expression throughout the body, altering drug metabolism and efficacy. Because the gut microbiome varies markedly among individuals, host-drug interactions mediated by microbial metabolism warrant close consideration in tailoring patient treatment.

Studying drug 'co-metabolism' by the gut microbiota and the human host is complicated by interindividual variability in human genetics, diet, and other clinical factors. Additionally, the human intestine is relatively inaccessible for direct examination. Animal and in vitro models can provide useful tools to simplify the study of these interactions. Even non-mammalian models can be important for exploring drug metabolism. For example, fluropyrimidines are among the most widely used chemotherapeutics for treating colon cancer. Two detailed mechanistic studies exploiting the power of large population sizes in *C. elegans* combined

---

[1]https://clinicaltrials.gov/ct2/show/NCT02236429

Table 1.1: Summary of selected studies where the microbiome is known to be implicated in the outcome of a treatment. In these cases, the presence or absence of a group of bacteria may hinder or promote the efficacy of a procedure

Treatment	Interaction	Reported
Non-steroidal anti-inflammatory drugs (NSAIDs)	Increased gram negative bacteria are associated with increased damage to the small intestine. These lesions can be ameliorated if pre-treated with antibiotics	[259, 200]
Antibiotics	Treatment effect depends on the host's starting composition and antibiotic class. Subjects with similar profiles had correlated abundance alterations of Enterobacter *cloaceae*.	[156, 189]
Cardiac Glycosides	The bioavailability of digoxin is affected by the presence of *Eggerthella lenta*, this can be regulated by suppressing the reductase responsible for degrading this drug.	[135, 70]
Metformin	The mechanism driving the effect of this drug is a result of metabolic shifts of the gut microbiome.	[256]
Fecal Microbiome Transplant	The efficacy of this treatment is known to benefit from pre-treatment with antibiotics.	[90, 99]
Chemotherapeutics	Negative drug-drug interactions by bacterial metabolism, protection from chemotherapy-related bloodstream infection.	[166, 170]

with the ability to associate specific components of the microbiome and mutant bacterial strains allowed elucidation of the precise pathways by which components of the microbiome such as *E. coli* and *Comamonas upp* enable the action of this chemotherapeutic agent via bacterial pyrimidine metabolism, using proliferation of gonad cells (and hence reproduction) as a bioassay for side-effects [60, 203]. This is a new innovation as mammalian (mouse, rat, guinea pig, hamster, rabbit) models are more widely used to study the interaction of microbial communities and common pharmaceuticals.

Although human studies are the gold standard, they have many limitations. They require Institutional Review Board (IRB) approval, are invasive, costly, and there is high and uncontrolled inter-individual variability. Close medical supervision is required and toxicity must be carefully monitored. Animal studies are less expensive, have logistical advantages with control of age and dietary variability, and all tissues can be directly accessed. However, the results do not always translate to humans (in part because the gastrointestinal physiology, diet and microbiome all differ markedly) and coprophagy can be an issue. Even so, animal studies have considerable cost and ethical issues, especially when large sample sizes are needed to set dosage ranges and identify sources of inter-individual variability. In vitro systems are therefore attractive models and preferred in early preclinical phases of drug discovery and screening. They are cheaper and provide easy access to materials, logistical and scale advantages, and incur no ethical concerns [216]. However, in vitro systems are often oversimplified and lack intestinal secre-

tion, measures of absorption, and host-microbe interactions. Nonetheless, several in vitro models have been developed and have had considerable utility. One is a Simulator of the Human Intestinal Microbial Ecosystem (SHIME) consisting of a six-stage reactor, simulating stomach, duodenum/jejunum, ileum, ascending, transverse and descending colon [231] and provides a simpler model for growing human fecal bacteria at scale [152]. Each experimental method has its own advantages and disadvantages; in vivo studies provide insight in the real life colonic metabolism, while in vitro models are more cost-effective and do not have ethical drawbacks. No perfect model exists and therefore a combination of different methods is preferred to clarify the role of the gut microbiome in drug metabolism.

Diet can modify the gut microbiome and gene expression of the host and thus influence epigenetic changes associated with diseases; such changes affect host gene expression through histone modifications, DNA methylation and non-coding RNAs [85]. The epigenetic changes are heritable and persist from one cell generation to the next. Nutrition contributes to a large extent to the gut microbiome and the gut microbiome has an important role in human metabolism. The gut microbiome is thus potentially the largest environmental factor affecting the human epigenome. Indeed, comparisons of germ-free mice with conventionally raised mice show differences in gene expression patterns throughout the body, including in the liver and brain. Initially germ-free mice colonized early with microbes resembled conventionally raised mice in physiology, gene expression and behavior, whereas those colonized later resembled germ-free mice, suggesting microbiome-mediated

critical periods for development [42]. However, the mechanistic basis for these effects is unknown. One specific interaction that has been well characterized is that bacteria are responsible for short chain fatty acid formation, especially butyrate, in the gut. Butyrate promotes cell turnover of colon epithelium cells. In colonic cancer cells, glucose rather than butyrate is used as growth substrate. Butyrate accumulates in the nucleus, inhibits cell proliferation and induces cell apoptosis [144]. In addition to their impact on the colonic epithelium, many microbial products are absorbed into the blood and lymph and can alter gene expression systemically.

In the human body small-molecule drugs are largely metabolized by human cytochrome P450 enzymes in the liver (CYP450). These enzymes can also convert prodrugs into their active molecules. P450 enzymes are highly polymorphic in humans and greatly impact drug response [262]. The variability in CYP450 is not limited to humans: bacterial CYP450 enzymes vary in copy number, function, and substrate [98]. Germ-free mice have different CYP450 expression profiles than conventionally raised mice and alterations in expression of other host genes linked to xenobiotic metabolism [17]. Although differences in the microbiome have not been linked to CYP450 activity in humans, accurate prediction of drug metabolism may ultimately require knowledge of both human CYP and the gut microbiota genetics, since one of the functional roles of the gut is to partially digest and absorb nutrients and therefore is a major component of metabolismi. Microbially-encoded cytochromes can also be directly important for drug metabolism. For example, as noted above, a classic example of microbiota-drug interactions is inactivation of

the cardiac glycoside digoxin by *Eggerthella lenta* [135] with metabolism by the cytochrome-encoding cgr operon in strains of *E. lenta* [71].

Bile acids are produced by hepatocytes and released by the gall bladder into the duodenum to absorb fatty acids, cholesterol, and fat-soluble compounds [154]. They are highly efficient detergents that digest, solubilize, and absorb dietary lipids and vitamins, and are widely used in drug delivery systems with increasing interest in using them directly as therapeutic agents [53]. Bile acids are metabolized by the gut microbiota into secondary bile acids [191], which are reabsorbed via the enterohepatic circulation and again back into the liver, biliary tract and gut. The gut microbiome not only influences secondary bile acid formation in gut but also modulates the enterohepatic system by regulating the bile acid synthesis in the liver [199]. High levels of secondary bile acids are linked to gastrointestinal diseases, colon cancer and gallstones [161]. For example, a meat-based diet is associated with higher levels of taurine conjugation to bile acids and gut production of hydrogen sulfide [145], which can lead to colon cancer. A high intake of saturated fats leads to a higher abundance of Clostridium clusters XI/XIVa and sulfate- or sulfite-reducing bacteria, which are linked to an increase in secondary bile acids produced by the microbiota [208]. Similarly the gut microbiome can contribute to obesity and type 2 diabetes since lipid and glucose metabolism are regulated by secondary bile acids in the gut [242]. Considerable potential exists for using probiotics to reduce the levels of specific secondary bile acids in the colon, although clinical trials and validated outcomes are still lacking. In vitro models have shown the

24

potential of lactobacilli and bifidobacteria to assimilate cholic acid in their cells [125]. Thus modifying secondary bile acid production by modifying the microbiome has promise, although validation in humans is yet to be established.

1.1.5 Prospects for stratifying patient response based on the whole microbiome

The considerable interindividual variation of the gut microbiome, together with the impact of the microbiome on drug metabolism opens up possibilities for stratifying patients on the basis of predicted response to a given intervention.

Microbiome studies, including stratification, have been dramatically transformed by computational tool development [20, 46, 122, 139, 190]. Most developments address only a specific problem within a long sequence of steps. Tools such as Quantitative Insights into Microbial Ecology (QIIME) (Quantitative Insights Into Microbial Ecology) [23] provide an integrated platform that allows scientists to go directly from DNA sequences to actionable and interpretable results. More recently these platforms are available in the form of web applications (MG-RAST [67], MicrobiomeAnalyst [41], VAMPS [87], Qiita[2]. Such integrated platforms allow analysis and storage of microbiome data and reduce infrastructure problems such as storage, data deposition and fault tolerance. These challenges must be overcome even for small microbiome studies.

These microbiome analytic platforms have been used to look for differences

---

[2]https://qiita.ucsd.edu/

in the microbiome and to stratify patients in terms of their therapeutic response in a number of diseases. For example, Inflammatory Bowel Disease (IBD), a gastrointestinal autoimmune condition, has a history [195] and evidence supporting a microbial-driven origin or regulation [61, 91, 251]. In the studies cited, microbial and metabolic features were inferred from microbiome and metabolomic samples to distinguish between healthy and IBD-affected subjects. This approach is of value because current diagnostic methods rely on expensive tests, such as the quantification of calprotectin present in fecal samples [117]. Further development of microbiome biomarkers could provide non-invasive, affordable methods to detect a group of bacteria or molecules. However, the classification accuracy of this method can be too low with fecal samples and even with rectal or ileal biopsies (which show the best performance) [61]. A promising approach uses a dog model of IBD [235], in which methods established and tested in humans were applied to a cohort of IBD and non-IBD affected dogs. In contrast to the sensitivity and specificity of the approach in humans, the microbial signal present in a dog's fecal sample was sufficient to classify IBD in dogs, with better performance than that obtained from biopsy samples in humans. These types of predictive models, using the microbiome data and certain data classifiers, have been useful for predicting which patients will best respond to specific drugs for Parkinson's Disease [82], and for predicting which dietary items have the greatest patient-specific impact on blood glucose response[264].

Advances in untargeted metabolomics may be key for understanding mi-

crobial metabolism and its impact on drug responses. For Mass Spectrometry (MS), the advent of large and community-curated reference knowledgebases like Global Natural Products Social Molecular Networking (GNPS) [239] provides a crowd-sourced data repository and peer-reviewed family of annotations for natural products, xenobiotics and metabolites. This is an important step towards the democratization and adoption of MS in a broad range of fields. GNPS further provides an analytic infrastructure to help generate molecular networks [243] that can be used to study drug metabolism [186]. Integrating MS and microbiome data, whether 16S rRNA profiling or shotgun metagenomics, is not a simple task but will ultimately provide key insights into complex problems for which the microbial characteristics alone are not useful, and where, for example, the presence/absence of a molecule could break ties or strengthen inferences in an algorithmic context. Co-occurrence and co-exclusion detection models could potentially benefit, because the current state of the art presents formidable challenges [126, 246]. However, statistical obstacles must be addressed so as to improve the ability to correlate microbiome and metabolite datasets. These dataset may have far more features than samples, and applying pairwise correlations leads to many spurious correlations. Finally, interpreting the results of these methods remains a major challenge, because these methods do not output correlations for individual microbiomes and molecules, but rather correlations between abstract axes that are composed of complex and often uninterpretable combinations of microbes and molecules. Enabling straightforward interpretation will be key for streamlining these methods so they

can be applied clinically. Such interpretation will also become easier as MS reference data and data sets of microbes and microbial molecules become available, a gap GNPS is aiming to fulfill.

1.1.6 Outlook

Recent advances have enabled the characterization of the human body from an integrative perspective across many data axes ('omes': genome, gene expression, microbiome, metabolome, virome, proteome, exposome, etc). When these heterogeneous data layers are combined, their full potential can address systems-level problems even when mechanisms are not fully understood. An outstanding example of this type of deep characterization is glycemic responses [264]: by combining 16S rRNA microbial profiling, food frequency questionnaires, anthropometric data, and glucose monitoring, one can create a model that successfully predicts each individual's blood glucose response to a specific meal and can create personalized diets that to increase or decrease the glycemic response of a set of new subjects not in the original study. These results are of utmost importance in this context, as the glycemic responses are known to be a personalized feature [237, 238]. In a separate study by the same group, researchers demonstrated that transitioning an individual's diet between industrial white bread, and whole grain sourdough bread, resulted in small changes in the microbiome and glycemic responses that differed on an individual-by-individual basis [119]. These experiments, which could be applied equally well to drug responses, are leading examples of the possibilities in the

as-yet nascent ability to personalize treatments, not only based on the traditional one-dimensional approach (anthropometric data), but by modeling the body with a holistic integrative approach.

The microbes and molecules present on and in our bodies can produce a dysbiotic or healthy environment. Therefore, we must characterize the archetypical organizations of our microbiome in large populations. We caution against attractive yet incorrectly oversimplified stratification approaches [13] that are very sensitive to arbitrary parameter selections [114, 121]. Instead, one must embrace the complexity of the microbiome, including its dynamic features (for example the rate of change of groups of bacteria). It was recently shown that patients with active IBD have microbiomes that experience more variability over time than healthy controls [72] while microbiome 'motion' of healthy subjects was much more restricted. Properly establishing a mechanism that describes these dynamics and creating appropriate models that take them into account remains a challenge, given the number of patients and samples obtained longitudinally that will be required for validation.

Better models of drug action may require a microbiome component. In a large-scale study analyzing over 1,000 Dutch participants, a series of lifestyle descriptors (diet, smoking habits, medications, etc.) were used to model their bacterial composition, and these descriptors could recapture almost 20% of the total microbial variation when using metagenomic sequences [266]. However, 80% of variation remains unexplained. The microbiome may serve as a proxy to covari-

ates that may not be readily available in all scenarios. For example, that study also found that decreased microbial and functional diversity was linked to a higher hemoglobin levels. Such studies need to unravel underlying mechanisms driving these associations and must show consistency across cohorts.

Is it possible to change a patient's microbiome to improve drug response? An intriguing example of this is that bioavailability of the antiarrhythmic drug amiodarone can be increased by concomitant administration E. coli Nissle 1917 in rats [47, 155]. Similarly, in germ-free mice colonized with the *Eggerthella lenta* DSM2243, access to a high protein diet rich in arginine inhibits digoxin inactivation [70, 142], suggesting that dietary changes could also affect the metabolism of the native gut microbiota to modify efficacy. Cheaper microbiome sampling and analysis, combined with algorithms and visual displays to better interpret results, could therefore open up a whole new area of modifying or drugging the microbiome to convert drug non-responders into responders.

1.1.7 Acknowledgments

Chapter 1.1, in full, is a reprint of the material as it appears in "Impacts of the human gut microbiome in therapeutics". Y. Vázquez-Baeza, C. Callewaert, J. Debelius, E. Hyde, C. Marotz, J. T. Morton, A. Swafford, A. Vrbanac, P. C. Dorrestein and R. Knight. *Annual Reviews*. 58, 2017. The dissertation author was the primary investigator and author of this paper.

Chapter 2

Exploratory microbiome data analysis

As we learnt in Chapter 1, there's a wealth of information that can be acquired through the study of microbial communities. The acquisition of this knowledge, however, is often only possible through the usage of specialized software for analysis and visualization. QIIME, introduced previously, is a successful[1] microbiome analysis pipeline that integrates a collection of bioinformatics tools and makes these programs available through a unified user-interface. Until its fifth release, QIIME relied on Kinemage, Next Generation (KiNG) to represent three-dimensional β-diversity plots. KiNG was originally developed as a molecular viewer and was *hacked* to act as a scatter-plot viewer. As datasets got larger and

---

[1]At the moment of this writing the original paper has been cited over 8,000 times according to Google Scholar, and the Python package has been downloaded over 34,000 times, data from https://pypi.org and https://anaconda.org.

more common, it became clear that we needed to develop our own solution, thus we created Emperor.

Emperor is an interactive β-diversity viewer, tailored to fit the QIIME ecosystem. β-diversity plots, commonly referred to as ordinations or dimensionality reductions, are often the starting point when analyzing a microbiome dataset. Their ability to overview the relative differences of an unlimited number of samples with an unlimited number of covariates makes this representation invaluable to diagnose and troubleshoot any problems that may have occurred during sample collection, preparation, or processing. For example, batch effects are often easy to notice in this representation [62] but more challenging to see in feature-level analyses, especially if those specific features are not affected by the batch effects.

As use-cases arose, these were integrated as part of the software, something previously not possible with KiNG. A notable feature allowed users to animate longitudinal sampling schemes, of key importance for the work in Chapter 4 and Chapter 5. The following two sections introduce Emperor, further discussing some of its statistical capabilities, and the use of animated ordinations as a way to interact with microbiome data.

Chapter 2.1 was published in the journal *GigaScience, 2013* and Chapter 2.2 was published in the journal *Cell Host & Microbe, 2017*. As the lead contributor of these two sections, I co-wrote the text, generated the main figures, wrote the software for the Python package, wrote documentation, and still maintain the project in the context of QIIME and provide support through the QIIME Forum.

2.1   EMPeror: a tool for visualizing high-through-put microbial community data

Background

As microbial ecologists take advantage of high-throughput sequencing technologies to describe microbial communities across ever-increasing numbers of samples, new analysis tools are required to relate the distribution of microbes among larger numbers of communities, and to use increasingly rich and standards-compliant metadata to understand the biological factors driving these relationships. In particular, the Earth Microbiome Project drives these needs by profiling the genomic content of tens of thousands of samples across multiple environment types.

Findings

Features of EMPeror include: ability to visualize gradients and categorical data, visualize different principal coordinates axes, present the data in the form of parallel coordinates, show taxa as well as environmental samples, dynamically adjust the size and transparency of the spheres representing the communities on a per-category basis, dynamically scale the axes according to the fraction of variance each explains, show, hide or recolor points according to arbitrary metadata including that compliant with the Minimum information about any (x) sequence (MIxS) family of standards developed by the Genomic Standards Consortium, display jackknifed-resampled data to assess statistical confidence in clustering, perform

coordinate comparisons (useful for procrustes analysis plots), and greatly reduce loading times and overall memory footprint compared to existing approaches. Additionally, ease of sharing, given Emperor's small output file size, enables agile collaboration by allowing users to embed these visualizations via e-mails or web pages without the need for extra plugins.

Conclusions

Here we present EMPeror, an open source and web browser enabled tool with a versatile command line interface that allows researchers to perform rapid exploratory investigations of 3D visualizations of microbial community data such as the widely used principal coordinates plots. EMPeror includes a rich set of controllers to modify features as a function of the metadata. By being specifically tailored to the requirements of microbial ecologists, EMPeror thus increases the speed with which insight can be gained from large microbiome datasets.

2.1.1 Background

Rapid increases in sequencing capacity are greatly expanding our ability to understand the microbial world: scaling from a handful of samples to hundreds, or thousands, allows a rich picture of trends over temporal and spatial scales that were previously unattainable. Human microbiome studies are not the only beneficiaries of this ability to perform increased sampling: large-scale patterns are now being discovered in communities ranging from soils [129] to oceans [77] including

the efforts from the International Census of Marine Microbes (ICoMM). We can now process thousands of samples in a single sequencing run [24], and in turn computational tools must also scale to fulfill these needs [68].

Although data visualization is an empowering tool that allows an efficient understanding of information [173], it remains a major challenge in this area of study, specifically because with more samples comes richer information relating the samples to one another (this contextual information is often referred to as 'sequence metadata') and to the study design itself. When analyzing large numbers of samples, researchers need to know the patterns that link specific samples or microbes to overall patterns of diversity, and to different metadata variables: this is typically critical for usable visualizations. A well know ecological metric to quickly compare the microbial composition of the samples is beta diversity, which collates them by creating a distance matrix of these differences. Ordination methods such as Principal Coordinates Analysis (PCoA) [69] are useful for dimensionality reduction and widely used in different fields to conceptualize distance matrices, however determining how to visualize the samples to reveal clear patterns often remains a challenge. Figure 2.1 A shows the samples colored by the body site each belong to, a common approach that will make evident the main differences explained in the first two axes of variation, however when integrating metadata in the coloring patterns (Fig 2.1 B-1, B-2), the plot clearly shows the age differences between the samples of an infant, compared to the samples belonging to healthy human adults.

There are several existing methods for displaying PCoA results, but none to
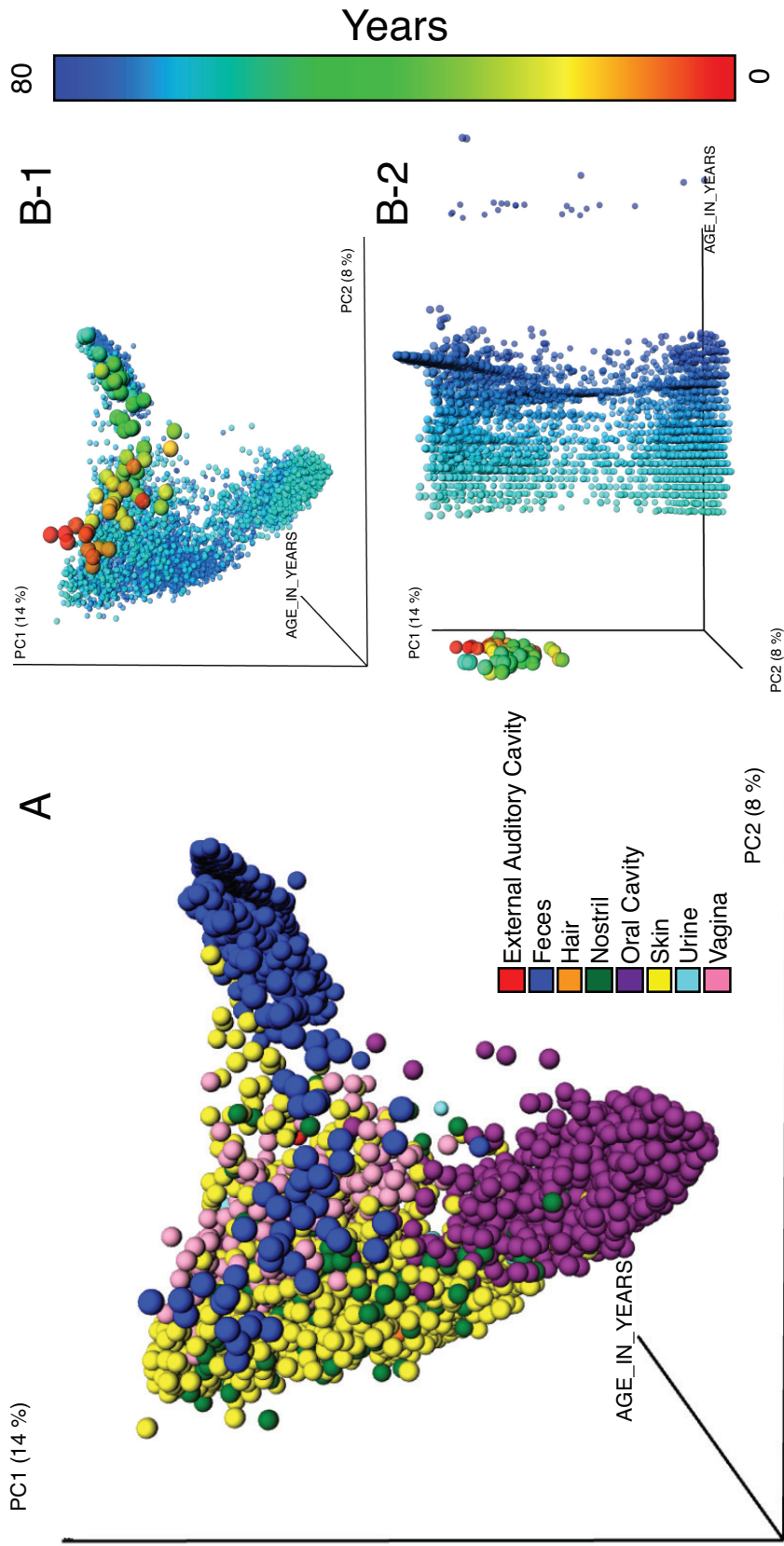
Figure 2.1: EMPeror display showing the combination of the datasets described in [24, 34, 83, 83], consisting of 5740 samples representing human auditory canal, skin, nostril, feces, vagina, urine, hair and oral body habitats. (A) Data colored by body habitat; (B-1) Principal coordinate 1 (PC1) vs. PC2 with the same the data colored according to the age of the subjects (a continuous variable). (B-2) PC1 vs. an explicit time axis. The results allow us to see by immediate visual inspection that the body habitats are remarkably different between each other and that this is consistent through time as a human reaches adulthood

date specifically designed to account for the common use cases in this research field; furthermore each of the most representative solutions allots different limitations. For example, QIIME [23], an open source framework for upstream and downstream analysis of microbial community samples generated via high-throughput sequencing instruments, typically generates 3D plots using KiNG [29] originally designed as a molecular graphics viewer, which requires static files containing each metadata field to be produced in advance, replicating the coordinates for each of these categories and resulting in long load times and large file sizes when the metadata are rich. SpotFire [7] is a very expensive commercial solution, beyond the budget of many research laboratories. Generic packages that provide 3D plotting functionalities such as MATLAB [5], Mathematica [1], R [6], Excel [3] or Matplotlib [86] can always be used, but custom code or manual approaches are typically required to relate each point to a specific visual feature intended to highlight a given variable. Consequently, this could become a time-consuming process, which as a side effect compromises its reliability, reusability and reproducibility. Moreover, none of the previously mentioned applications are specifically modeled to support the workflows of the modern microbial ecologist. Allowing the user to choose among metadata coloring dynamically, and separating coloring from visibility, has a surprisingly large effect in encouraging interactive exploration, understanding and analysis, and often allows insights into the main factors, as well as more subtle ones, structuring the data to be obtained much more rapidly.

2.1.2 EMPEROR

EMPeror is a thoroughly tested and open-source software package with an interactive user interface and hardware-accelerated graphics, implemented with HyperText Markup Language, version 5 (HTML5), Web Graphics Library (WebGL), Javascript and Python, and tightly integrated with QIIME [23] and PyCogent [111]. EMPeror's command line interface accepts QIIME principal coordinates files and metadata mapping files, and produces an interactive 3D visualization that can be delivered in the context of a web page independent of the command line tool. As an example of EMPeror's ability to deal with continuous variables (time, alpha diversity, pH) that are part of the metadata, these factors can be integrated as an explicit axis in the plot, lines connecting subsequent points of single trajectories (treatments, subjects, sites, etc) or using a colormap to have each sample's color be a function of its position in the gradient. The main features that EMPeror provides are:

1. easily change visibility features of data points in the plots based on metadata;

2. can be easily embedded into other tools, such as Evident[2] as a reusable visualization component;

3. scale to thousands of points with minimal load times (seconds versus many minutes in KiNG);

---

[2]http://github.com/qiime/evident

4. ability to display auxiliary data to increase the understanding of the intrinsic data patterns; these include: biplots [81], procrustes analysis [167], and jackknifed beta diversity plots [123].

To illustrate the effectiveness of EMPeror, we show the combination of [24, 34, 83, 83], see Table 2.1, as generated with the QIIME web application[3]. This combination represents 5740 samples (spheres), and 120 columns of metadata [4]. In KiNG, the resulting files for both the discrete and gradient coloring result in a size of 1.85 GB, but in EMPeror only 26 MB[4], meaning only 1.3% of the original size, see Figure 2.1. Additionally, we can easily view the intrinsic age patterns within the data, Figure 2.1 B, both panels. EMPeror installation instructions can be found in the online documentation[5].

2.1.3 Conclusions

EMPeror provides a user-friendly interface and set of tools for visualizing large numbers of microbial community samples associated with increasingly extensive metadata, and interactively manipulating these data sets to add auxiliary data and visualization techniques. Additionally, it contains several user interface features, which enable straightforward modifications and customization of perceptible aspects in the plot plus the incorporation of statistical techniques, which also help increase the ease and speed of exploratory analysis. We believe that EMPeror

---

[3]http://www.microbio.me/qiime/
[4]ftp://thebeast.colorado.edu/pub/emperor_files/
[5]http://qiime.org/emperor/installation_index.html

Table 2.1: Studies used to create Figure 2.1

Title	General description	Collected samples	References
Moving pictures of the human microbiome	Samples from two subjects are collected for up to 15 months in three body sites (oral, skin and gut)	1964	[24, 55]
Bacterial community variation in human body habitats across space and time	Samples from healthy adult human samples from eight subjects of up to 27 body sites	585	[34]
Structure function and diversity of the healthy human microbiome	Samples from 242 healthy adult human samples from up to eighteen different body sites	3131	[83]
Succession of microbial consortia in the developing infant gut microbiome	Gut samples collected biweekly from an infant through the first 2.5 years of life	60	[83]

will have a large impact on the field, especially for large-scale environmental sampling projects such as the Earth Microbiome Project [64], and large-scale clinical projects such as the Human Microbiome Project [83].

2.1.4 Acknowledgments

2.2 Bringing the Dynamic Microbiome to Life with Animations

Our bodies and natural environment contain complex microbial communities, colloquially termed microbiomes. We previously created a web-based application, EMPeror, for visualizing ordinations derived from comparisons of these microbiome communities. We have now improved EMPeror to create interactive animations that connect successive samples to highlight patterns over time.

2.2.1 Introduction

Recent developments in high throughput DNA sequencing, improvements in molecular methods, and the continual increase in computational power have enabled microbiome researchers to test ever-more sophisticated hypotheses and study designs. Just a decade ago, a microbiome study with a few dozen samples was considered large. However, new technologies now enable studies with thousands of samples that can explore differences at the scale of population or environmentals.

One iconic result of these studies is the presentation of aggregated microbiomes from different human body sites within a single plot, where samples from equivalent body sites across many different individuals all cluster together (for an example, see Figure 2.2). This body site clustering was initially observed in a study of 9 people [34], and then replicated across hundreds of people in the Human Microbiome Project (HMP) [83]. As useful as these static snapshots have been in

examining the variation in diversity across populations, they do not reveal temporal dynamics, which may stem from intrinsic properties of the ecosystem such as inter-species interactions or external drivers such as drugs or diet (in humans) or envionmental perturbations. As seen in macroscopic ecology, changes in the relative abundance of organisms can occur on short time scales. Given that population doubling times of some bacteria can be less than 20 minutes, these time scales can be potentially very short for commensal communities. A single gram of stool can contain over one billion microbial cells, 1000 times the density of microbes in the ocean surface. These microbes are represented by hundreds of species, thousands of strains, and an estimated 3.3 million microbial genes [185]. Individual species and strains compete for resources (e.g., fiber), which can impose limitations on organism populations as the supply and demand of resources are dynamic overtime. Additionally, there are time-dependent bacteriophages which target and kill specific bacteria, or in some cases, promote Horizontal Gene Transfer (HGT), which can alter the strains (e.g., antibiotic resistance genes are commonly transferred via HGT). Furthermore, many selective pressures act on the microbiome. For example, the human immune system acts as a regulatory force for the constituent organisms. This includes the innate and adaptive immune system, but also the variety of white blood cells and their antimicrobial glycoproteins. Additionally, microbial cells, many of which are capable of highly optimized metabolic processes, are subjected to a constantly changing set of compounds from the host's diet, as well as an intermittent exposure to antibiotics and other xenobiotics (in the case

of diseased subjects, antibiotic use can lead to dysbiotic microbiomes, especially when taken frequently and at high doses).

The emergence of quantitative microbial ecology dynamics will be important for strengthening our basic scientific understanding of the microbial world. Additionally, it will also quickly become an essential tool for medicine. Currently, when antibiotics or other pharmaceuticals are given, the doctor does not know the state of the microbiome or how it will be altered. Being able to read out and predict changes in the microbiome will enable a field of precision medicine, for restoring the altered ecology to a healthy one. Similarly, there is great hope that designer pre- and probiotics will be powerful tools for 'gardening' one's microbiome back to health. However, until more quantitative and large-scale trials have been performed with longitudinal tracking of the changes in the microbiome ecology, the precise impact of these tools will remain, at best, a conjecture.

With the improved understanding of human microbial dynamics, the conceptual framework of medicine will continue to experience big changes. As recently described [35] many fundamental concepts from ecological theory are directly applicable to the host-microbiome ecology. However, these topics have not yet been widely incorporated as part of general medical training. Concepts of succession and community assembly in previously unoccupied habitats (infant growth of microbial diversity), assembly after disturbance (recovery from antibiotics) and blooming of invasive species (food poisoning by pathogens). Each of these topics will be elucidated in the coming decade by analyzing time series of microbial ecologies, es-

pecially in testings with defined perturbations, and integrated with metabolomics, metatranscriptomics, metaproteomics, and host biomarkers.

2.2.2 Ordinations

Datasets that have been obtained from DNA sequencing often have very high dimensionality, up to thousands of dimensions for each sample. Given the high dimensionality of microbiome studies, the use of dimensionality reduction techniques, otherwise known as ordinations, have been widely used in the field (some examples include Principal Component Analysis (PCA), Principal Coordinates Analysis (PCoA), Non-metric Multidimensional Scaling (NMDS), and Correspondence Analysis (CA)). These techniques are analogous to modern photography: photographs capture 2D images of a 3D environment, and although they do not capture all of the information about the landscape, they can provide enough information to differentiate and identify objects. Similarly, ordination methods are valuable tools for capturing snapshots of these microbial populations and better understanding the major driving factors affecting these populations.

Although it is clear that the frontier of microbiome research is shifting to understanding microbiome dynamics, it remains a challenge to 'see' the dynamics in these increasingly large datasets. Traditional time-series analysis techniques were developed around densely sampled and evenly collected variables, restrictions that are often impossible in clinical and ecological settings. Beyond these technical limitations with sampling, the complex dynamics of most microbial ecology

datasets violate assumptions of periodicity and stationarity, used in many statistical methods for time-series analysis. For an expanded discussion and review of this topic, see [52].

A few pioneering studies have been published for microbiome time series, but these visualizations were typically performed using custom 'one-off' software that is generally harder to modify, maintain and reuse. However, these examples showed the promise of visualizing microbiome dynamics. For instance [56], described in detail the per-subject alpha and beta diversity variations, as well as the individual variability observed in the time point to time point comparisons. Another remarkable example presented by [36] examined pre- and post-dietary intervention comparisons at the microbiome and gene expression level. Similarly, [141] used the data from an infant's first 27 weeks of life to show that you can visualize the assembly of gut communities and contextualize it with the data of healthy human adults of the HMP. This analysis shows a step-by-step transition of the infant's communities as they transition into one of a healthy adult. [24] showcased the variability in the three body sites of the two surveyed subjects. More importantly, it gradually painted a picture, sample by sample, of the overall stability and dynamics present in the dataset.

2.2.3   EMPeror

EMPeror is a response to the many requirements that are specifically relevant to high throughput microbiome studies. The core idea behind this software

is to allow users to rapidly explore a dataset, traversing hundreds of sample meta-data categories and minimizing the necessity of custom plotting code. All of this can occur without requiring a complex software installation or domain-specific knowledge.

EMPeror [236] is a web-enabled scatter plot viewer for microbiome analysis that was originally made available in 2013 as part of the QIIME pipeline [23], and developed for the Earth Microbiome Project[6]. To use EMPeror, two files need to be provided, the first and most important is the sample metadata file, which contains covariate information about each sample, and the second file is a matrix specifying coordinate positions for each sample along a set of axes. In general, for microbiome studies, the coordinate positions are derived using PCoA. Besides allowing the easy exploration of high dimensional color ordination spaces (such as PCA or PCoA), EMPeror supports more specialized techniques, ranging from displaying jackknifed-resampled data to assess statistical confidence in clustering, to displaying procrustes analysis plots to determine the goodness of fit between two independent datatypes, and even being able to visualize biplots. As of version 0.9.5, users can now create animations inside the user interface without the necessity of any additional computation.

Like most functionality in EMPeror, animations themselves are made possible by selecting two metadata categories, a *trajectory* category and a *gradient* category. The first category determines how the samples are grouped together

(e.g. individual), and the second category describes the order of the grouped samples. As readers might suspect, the ordering is only unambiguous when the values in the *gradient category* are numeric; therefore, we recommend that only these data types are used. Along with these metadata selectors, EMPeror displays a *play*, a *pause*, and a *rewind* button. These control the state of the traces being presented on screen.

With this mapping, samples that belong to the same *trajectory* are subsequently connected by an animated trace (sample by sample). The traces are created by linearly interpolating a series of points between two samples. The number of points present in this interpolation depends on the absolute difference between the two samples in the *gradient*, meaning that samples that were collected further apart along a gradient (e.g., if some samples are daily and some are weekly) will take longer to be connected in the animation than samples that were collected closer in the *gradient*. This allows heterogeneous sampling schemes, common in clinical trials or when some samples are lost or destroyed during processing, to be accommodated in the same animation.

Importantly, these animations do not interfere with the interactivity of a plot. The rest of the controllers (used to manipulate the rotation, zooming, coloring, size, opacity, etc.) remain available and can be used interactively to create a narrative that is more relevant for the data. Lastly, EMPeror is a stand-alone application, that can be easily integrated into a software workflow, for instance using Jupyter notebooks [182].

48

2.2.4   Example

To showcase the utility of this particular type of visualization, we will reproduce the supplementary video presented in 2015 [244], where four *Clostridium difficile* infection (CDI) patients are surveyed before and after a Fecal Microbiota Transplant (FMT) to ameliorate the consequences of this enteric infection.

To create this visualization, we must combine this indiviual dataset with the 16S rRNA amplicon samples from the HMP [83], which includes thousands of microbiome samples collected from 242 healthy individuals. These HMP samples create a visual 'map' of normal oral, skin, vagina, and stool microbiomes and provide context for the CDI samples. To combine these studies, we used Qiita[7], an open-source platform for analysis and visualization of microbiome datasets. Qiita allows users to combine samples, perform the necessary computations for PCoA (among other analyses), and allows for visualizing the result in EMPeror.

For this example, we selected the number of days since the FMT procedure as the *gradient* category and we selected the subject as the *trajectory* category. By animating these traces, we first observe that the original patients with CDI had gut microbiomes that did not resemble healthy adult fecal communities. Immediately after the FMT, we observe a rapid transition into a health state, which can be inferred from the movement of the sample locations by the end of the trajectory to the normal stool region defined by the HMP (Figure 2.2). Critically, these transitions were associated with the patient's immediate recovery from the illness.

---

[7]https://qiita.ucsd.edu
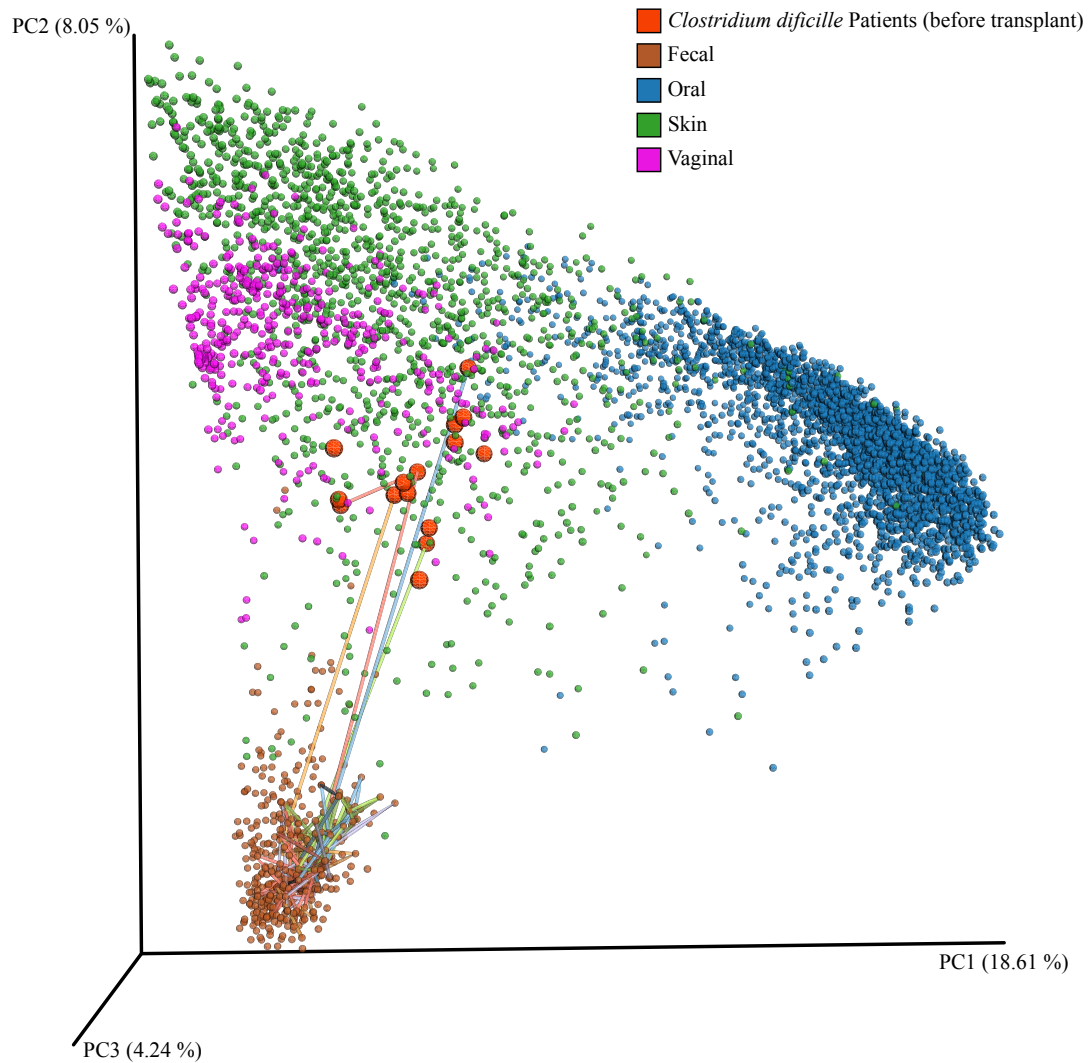
Figure 2.2: **Case study: remission of Clostridium difficile Infection.** *Principal Coordinates Analysis Plot of unweighted UniFrac distances.* After receiving the fecal transplant, the patients undergo a dramatic transition in their gut microbiomes. Samples from healthy individuals are colored: blue for oral, green for skin, pink for vaginal and brown for fecal communities. Patient samples (preceding the transplant) are colored in orange, the traces indicate the trajectory they follow over time (only four subjects receive a transplant). Access this interactive animation using your web browser: http://emperor.microbio.me/animation/

2.2.5 Limitations

The objective of ordinations is to find the dimensions that best describe as much as possible the variation in a dataset, and therefore reduce the amount of information that we have to interpret. This property is most often useful for its ability to distinguish between groups of samples. However, it also makes this technique sensible to outliers. Consequently, it is important to remember that the variation explained in a statistical model does not necessarily correspond to biological importance. In fact, some methods of measuring distance between samples and ordination techniques for reducing dimensionality can explain much of the variation in a given dataset, but fail to reveal biologically important patterns [122]. In addition, PCoA is susceptible to an artifact called the horseshoe effect, where the shared absence of taxa in the middle of the gradient cause the ends to appear similar and wrap around (other techniques such as Non-metric Multidimensional Scaling are susceptible to a related problem called the arch effect, where the second axis is a distorted reflection of the first), and very sparse data matrices where most samples are highly dissimilar to one another can produce artifacts resembling spikes at 90 degree angles to one another, Consequently, choosing the wrong distance metric or dimensionality reduction method for a dataset can result in visual artifacts that render those visualizations unusable. These problems are less frequent, but still possible, when phylogenetic distance metrics such as UniFrac [139] are used. Animating the transitions between samples provides a useful visual aid.

However, the observed patterns of variation still need to be asessed using descriptive statistics. Differences among sets of community samples can be assessed using the Analysis of Similarities (ANOSIM) or a Permutational Multivariate Analysis of Variance (PERMANOVA). These statistical tests compare the groups using the distance matrix to provide a statistic that measures whether sample category labels tend to group similar samples together more often than a random shuffling of the labels. Alternatively, feature-level differences can be assessed using compositional analyses, correlation networks or discriminant features between two groups (using supervised and unsupervised machine learning techniques).

In our experience, the interpretability of animated ordinations is most commonly affected in two cases. The first case can be observed when an inappropriate frame of reference is used, or when only the animated traces are presented. For the example above, the HMP data provides a relevant context due to the types of samples being characterized and the size of the study. However, a reference frame that included only oral samples, or that contained only a single stool sample, would not provide nearly as relevant a data frame for display, and the resulting animations would be difficult to interpret. The second case occurs when many trajectories are animated simultaneously, making the ordination busy and increasing the difficulty of correctly identifying the patterns of variation that are most biologically relevant. This limitation can be overcome by choosing more similar sets of samples to display, such as patients with more consistently defined clinical phenotypes.

2.2.6   Conclusions

Microbiome studies are steadily increasing in size and complexity, challenging our ability to unravel the structure of the ecological interactions occurring at the microbial scale. As researchers, the software we use greatly impacts our ability to ask questions of the data. Platforms like the Jupyter Notebook have greatly facilitated reproducibility of analyses. The success of the Notebook has been in part, due to building on top of an open and common web infrastructure, the same that Emperor has adopted. Interactive animations like the example presented above, can be directly shared with anyone that has access to a modern web browser, whether it be in a personal computer or in a mobile device.

Sample information (also known as sample metadata) is indispensable to data reuse and necessary for the creation of complex visuals, like in Figure 1. This task was partly possible by leveraging the standardization of sample metadata enforced in Qiita. Without standardization, future generations must undertake the burden of post-hoc metadata curation, a process that is deeply frustrating and impossible to automate.

Finally, although EMPeror provides a handful of ways to interrogate a dataset, we urge the community to create other open source interactive tools that democratize robust, reproducible and interactive analysis techniques. These tools should not be usable by only a few experts, but instead must target broader audiences that can benefit from these technological advances.

2.2.7 Availability

The source code and documentation for version 0.9.60 can be found online[8]. Additionally a tutorial (including the data) to generate Figure 2.2 can also be found online[9].

2.2.8 Acknowledgments

This section, in full, is a reprint of the material as it appears in "Bringing the Dynamic Microbiome to Life with Animations". Y. Vázquez-Baeza, A. Gonzalez, L. Smarr, D. McDonald, J. T. Morton, J. A. Navas-Molina, R. Knight. *Cell Host and Microbe*, 11, 2017. The dissertation author was the primary investigator and author of this paper.

---

[8]http://biocore.github.io/emperor/
[9]http://biocore.github.io/emperor/build/html/tutorials/animations.html

Chapter 3

Inflammatory bowel disease in Dogs and Humans

Inflammatory Bowel Disease (IBD) is a family of conditions divided into two major categories, Ulcerative Colitis (UC) and Crohn's Disease (CD), each of which can be further divided into other subtypes specifying location, age of diagnosis, severity, and behaviour of the disease [197]. In all cases IBD is associated with diarrhea, inflammation, and flaring episodes of exacerbated discomfort. In general, diagnostic methods rely on questionnaires, direct examinations of affected intestinal tissue, or quantification of calprotectin [211].

In this chapter, we introduce work motivated by our recent efforts at characterizing the microbiome in new-onset and treatment-naive Crohn's disease subjects [61]. Specifically, we focus on comparing the microbiomes of IBD-affected humans and dogs. As a result, we further expanded the repertoire of features that we know

and associate with IBD in dogs [164], and through a meta-analysis contextualize them with human fecal samples [61].

As we did before [61], we create a dysbiosis network[1] to statistically infer the bacterial genera associated with the disease. This representation allows us to directly compare microbial groups across datasets, and explore their relationship to each other. We find that the human and dog dysbiosis networks are overlapping, both in structure and in the microbes associated with IBD and non-IBD communities. However, we uncover a taxon that shifts from being non IBD-associated in humans to being IBD-associated in dogs.

Chapter 3.1 appeared in the journal *Nature Microbiology, 2016.* As the lead contributor of this project, I co-wrote the text, generated the main figures, processed, analyzed, interpreted, and deposited the data into a public repository.

---

[1] A dysbiosis network is a graph constructed using a correlation matrix as a weighted adjacency matrix.

3.1 Dog and human inflammatory bowel disease rely on overlapping yet distinct dysbiosis networks

Inflammatory Bowel Disease (IBD) is an autoimmune condition that is difficult to diagnose, and where animal models of this disease have questionable human relevance [93]. Here we show that the dysbiosis network underlying IBD in dogs differs from that in humans, with some bacteria such as *Fusobacterium* switching roles between the two species (as *Bacteroides fragilis* switches roles between humans and mice) [61]. For example a dysbiosis index trained on humans fails when applied to dogs, but a dog-specific dysbiosis index achieves high correlations with the overall dog microbial community diversity patterns. In addition, a random forests classifier trained on dog-specific samples achieves high discriminatory power, even when using stool samples rather than the mucosal biopsies required for high discriminatory power in humans [61]. These relationships were not detected in previously published dog IBD datasets due to their limited sample size and statistical power [84]. Taken together, these results reveal the need to train host-specific dysbiosis networks and point the way towards a generalized understanding of IBD across different mammalian models.

3.1.1   Introduction

Dogs are commonly used as large animal models for drug discovery and safety assessment. The usefulness dogs as a model for IBD is yet unexplored, though not unreasonable, as dogs have been useful for studying spontaneously occurring disorders similar to those affecting people [79]. For example, dogs live in a close relationship with and share an environment with their owners, and are, therefore, frequently exposed to similar environmental factors, including enteropathogens and toxins. It is well recognized that dogs and humans suffer from similar spontaneous and lifestyle associated diseases such as obesity, allergies, diabetes mellitus, and cancer, and are often treated with similar antibiotics and drugs. IBD in humans is a chronic autoimmune disease of multifactorial aetiology and has limited treatment options. Similarly, canine idiopathic IBD is a commonly observed chronic inflammatory enteropathy that occurs spontaneously due to similar multifactorial aetiology, namely due to an interplay between an aberrant host immune system, genetics, environmental factors and gut microbiota [93]. Common clinical signs are vomiting, diarrhea, and weight loss. Histological evaluation of intestinal biopsies reveal diffuse or multi-focal inflammatory cell infiltration (most commonly lymphoplasmacytic, followed by eosinophilic and neutrophilic infiltration), with concurrent changes in mucosal architecture (e.g., villus atrophy and fusion) [38]. Enteric protein loss may be observed in severe cases. Also, in a subset of dogs, invasive and adherent *E. coli* have been described, and these

share common features with strains isolated from humans with Crohn's disease [210]. Clinical signs may be controlled by single or combination therapy including dietary modifications, antibiotics, and immunosuppressants. However, clinical relapse occurs frequently, and life-long therapy may be needed. Previous small-scale studies have revealed dysbiosis in the small and large intestine of dogs with IBD, with some changes in bacterial taxa similar to those observed in humans with IBD [84]. For example, in both humans and dogs with IBD, increases in Proteobacteria, specifically Enterobacteriaceae [221], and decreases in Firmicutes, including *Faecalibacterium* and *Blautia* [222] have been reported. However, no detailed studies comparing the changes in gut microbiota between humans and canines with IBD have been reported to date. The aim of this study was to describe in detail the microbiome changes in a large group of dogs with IBD, compare these to the microbiome of humans with IBD, assess host similarities and differences, and to train a dysbiosis index composed of non-IBD and IBD associated bacteria.

3.1.2   Results and Discussion

As expected based on results from human studies [57, 61], but not previously clearly established in canine studies [164, 222], dog IBD cases and controls differ substantially in both microbial community diversity and structure (Figure 3.1.2). IBD dogs had a significantly lower alpha diversity (Mann-Whitney test, p=0.003) compared to non-IBD affected dogs (Figure 3.1.2A), however alpha diversity in this population did not correlate with age, fat intake, weight, or protein intake

(Supplementary Figure 1[2]), nor with body condition scores ($p > 0.05$, see supplemental methods). Clear separation (using Permutational Multivariate Analysis of Variance (PERMANOVA) grouping samples by disease status, $p=0.001$ see supplemental methods) between IBD dogs and controls was observed using unweighted UniFrac (Figure 3.1.2B), and the biplot shows the most abundant taxa driving these overall patterns. Consistent with [164]; antibiotic history did not reveal differences within IBD affected dogs (PERMANOVA $p=0.501$, abx=35, no abx=12), on the contrary a significant effect was observed in non-IBD dogs (PERMANOVA $p=0.01$; abx=8, no abx=77). Finally when combined, the disease effect was significantly stronger (PERMANOVA $p=0.001$) than the history of antibiotic usage (pseudo-F on IBD groups 1.99 and pseudo-F based on antibiotics groups 9.46; $1.99 < 9.46$).

A random forests classifier [21] trained on the dog data achieved an Area Under the Curve (AUC) of 0.93 (Figure 3.1.2C, see discriminant Operational Taxonomic Units (OTUs) in Supplementary Table 1[3]), demonstrating excellent classification accuracy compared to human stool samples where the AUC was only 0.63 using a much larger training set [61], and achieved only AUC = 0.86 even using mucosal biopsies. Consequently, in dogs, but not in humans, high classifier accuracy is achievable for IBD using only stool samples.

Encouraged by these results, we tested whether a dysbiosis (see supplemen-

---

[2]https://images.nature.com/original/nature-assets/nmicrobiol/2016/nmicrobiol2016177/extref/nmicrobiol2016177-s1.pdf

[3]https://images.nature.com/original/nature-assets/nmicrobiol/2016/nmicrobiol2016177/extref/nmicrobiol2016177-s1.pdf
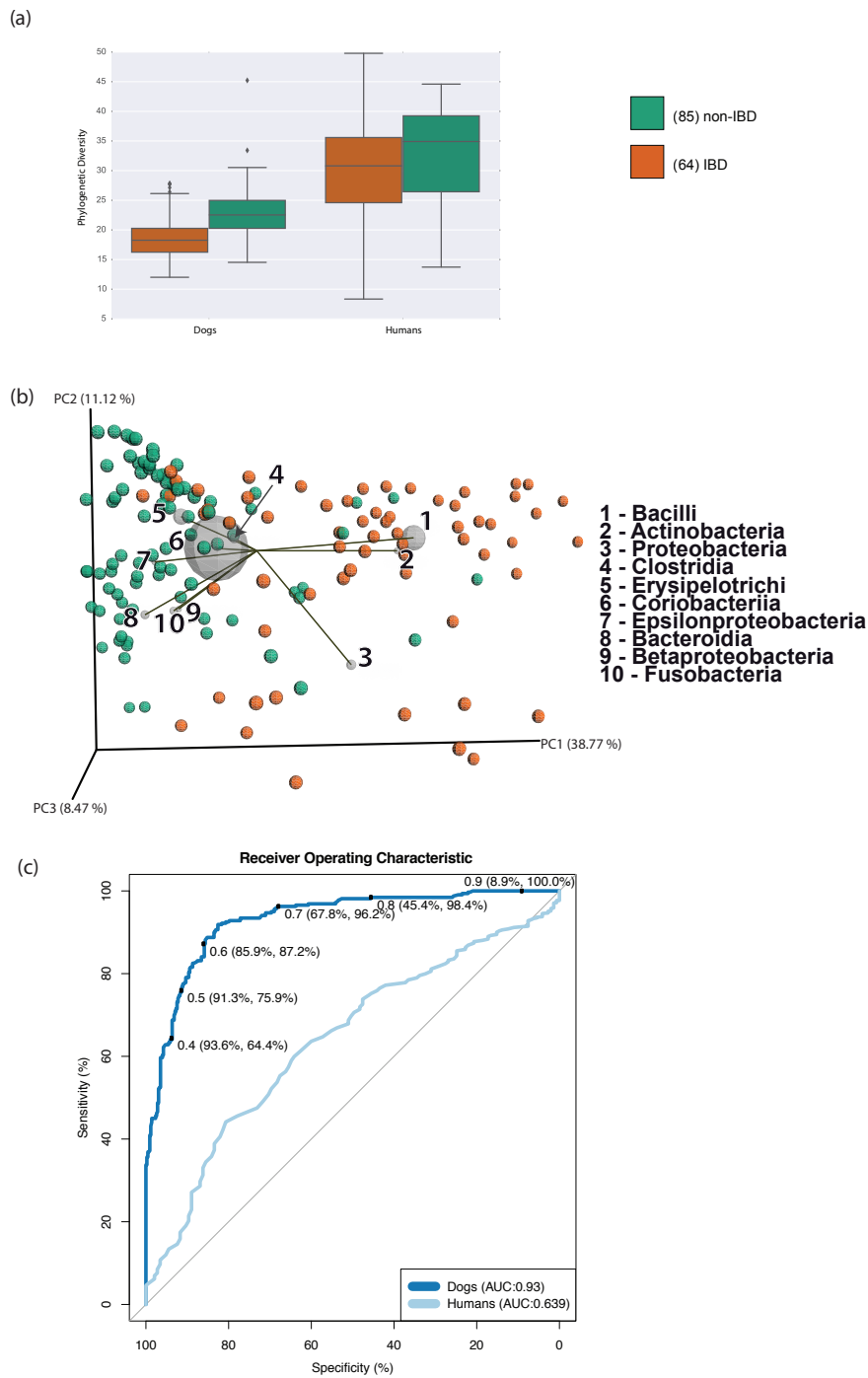
Figure 3.1: **Diversity overview and a comparison against humans.** (a) Comparison between disease states and species of Faith's phylogenetic diversity (whiskers extend for 1.5 times the IQR past the low and high quartiles). (b) Weighted UniFrac beta diversity biplot, for the dog samples colored by disease status. (c) Performance comparison of a random forest classifier trained to separate the healthy from the diseased samples in both humans and dogs.

tal methods) network trained on stool samples alone could be used to interpret the pattern of IBD in dogs, and whether this pattern overlapped the human network. We previously found that in humans, substantially better correlation networks could be achieved from the mucosal biopsies, because many taxa contributing to these networks were not seen in stool. Accordingly, we used the techniques described [61] to generate correlation networks and a dysbiosis index for dog samples (see supplemental methods for more information). Using the network together with the biplots (Figure 3.1.2 B) we see that Gammaproteobacteria (specifically *Enterobacteriaceae*) were significantly associated with IBD, whereas various Firmicutes such as Clostridium and *Ruminococcus* were associated with non-IBD samples. When these features are compared with the discriminant OTUs, obtained from the random forests classifier, we observe a general overlap of the lineages (see Supplementary Table 1 and Supplementary Table 2[4]). However we also see OTUs that are not highlighted by the correlation network. Specifically Erysipelotrichaceae *Allobaculum* and Lachnospiraceae *Blautia producta*, this is likely because these do not consistently co-occur with or co-exclude other taxa.

The human dysbiosis index failed to negatively correlate with alpha diversity in dog samples (Figure 3.1.2 A.1 and A.2 correlation, A.3 Principal Coordinates Analysis (PCoA) plot), while a dog-specific dysbiosis index showed a statistically significant negative correlation in the same samples (Figure 3.1.2 B.1 and 2 B.2

---

[4]https://images.nature.com/original/nature-assets/nmicrobiol/2016/nmicrobiol2016177/extref/nmicrobiol2016177-s1.pdf

correlation, C.3 PCoA plot, see for a reference the two groups in Figure 3.1.2 D). Additionally we tested the index with previous data [164], and although the sample size is limited, we observed similar patterns (see Supplementary Figure 2[4]). Similarly to humans, the dysbiosis index in dogs is negatively correlated with phylogenetic diversity (r = -0.45, p < 0.001). However, the list of 'non-IBD' (co-occurring in non-IBD samples) and 'IBD' (co-occurring in IBD samples) bacteria only partially overlaps between host species (see Supplementary Table 2[4]). Comparing the correlation networks of the taxa for humans (as described in [61]) and the network generated for the dog data (Figure 3.1.2 C) revealed overlapping, and discordant taxa. In particular, *Fusobacterium* appears to be associated with IBD [61] and colorectal cancer [27] in humans but with non-IBD dog samples. Of note, we previously observed high levels of *Fusobacterium* sp. in dogs [220] but also carnivores of multiple species [132, 167], and noted higher levels of *Fusobacterium* in dogs with more access to the outdoors [214], which may correlate with a wide intake of other immunomodulatory environmental bacteria. Given the limited adaptation (approximately 8.9 Mya when humans split from Gorillas [80]) of the human lineage (historically omnivorous [184]) to a carnivorous diet, as compared to the base of the Carnivora 40-45 Mya [249], it is possible that this taxon has yet to be incorporated into the non-IBD portion of the network in humans. Nonetheless, it is important to remember this is only a statistical association, and further research would need to be developed to properly validate this. Consistent between human and canine networks and former studies were findings in regards to decreased *Fae-*

calibacterium and increased *Escherichia coli* in IBD [222], and these taxa seem to be important in this disease across animal hosts. Other taxa (*Enterococcus* and *Allobaculum*) from the canine network that were associated with IBD or non-IBD are generally consistent with previous results based on either small scale studies or targeted Polymerase Chain Reaction (PCR) [164], but additional taxa were discovered here, such as *Butyricoccus* which was associated with non-IBD dogs.

To measure the relative effect size of host-species and IBD, we combined humans and dogs into a single PCoA plot, marked by clinical status (Supplementary Figure 4[5]). We demonstrate that at a microbial level the disease effect is smaller than the host effect (human vs. dog, Supplementary Figure 4 A[5]). Similarly the disease effect was weaker than the species effect when analyzing PICRUSt [128] metagenome prediction data (PERMANOVA test grouping samples by disease status and by host-species using a binary Jaccard matrix, p=0.001, pseudo-F by disease 14.61, pseudo-F by species 52.59, Supplementary Figure 4 B[6]), indicating that at both the compositional and predicted functional level, species is a more significant influence on the microbial community than disease (see supplemental methods and Supplementary Figure 3[6]). In both dogs and humans, predicted pathways were relatively similar across IBD and non-IBD samples, with the most abundant pathways across both groups in both species including 'housekeeping' pathways such as transporters, ABC transporters, DNA repair and recombination

---

[5]https://images.nature.com/original/nature-assets/nmicrobiol/2016/nmicrobiol2016177/extref/nmicrobiol2016177-s1.pdf

[6]https://images.nature.com/original/nature-assets/nmicrobiol/2016/nmicrobiol2016177/extref/nmicrobiol2016177-s1.pdf
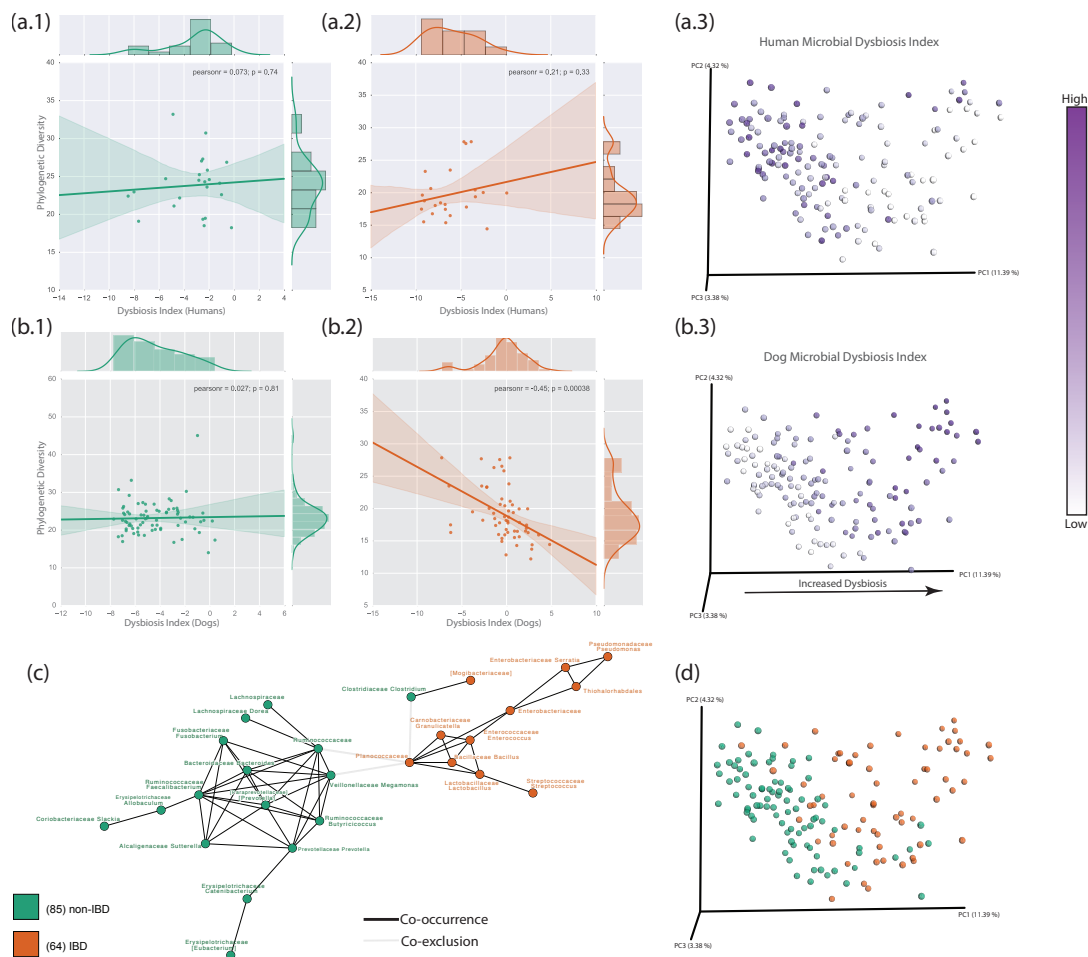
Figure 3.2: **Human and dog dysbiosis index.** (a.1), (a.2) Human dysbiosis index describing the dog samples grouped by disease status, (a.3) weighted UniFrac PCoA plot colored by dysbiosis index. (b.1) and (b.2) Dog dysbiosis index describing the dog samples grouped by disease state, (b.3) weighted UniFrac PCoA plot colored by dysbiosis index. (c) Correlation network used to determine the dysbiosis index in dogs, colored by the bacteria associated with IBD and non-IBD samples, co-exclusion edges are showed in dotted lines and co-occurrence edges are showed in solid lines. (d) Unweighted UniFrac PCoA plot of the dog data. Figures A.1, A.2, B.1 and B.2 all display 95% confidence intervals.

proteins, ribosome, purine metabolism, transcription factors, peptidases, pyrimidine metabolism, and chromosome (Supplementary Figure 5[6]). Although these most abundant pathways were not significantly associated with health or disease, the abundance of several lower abundance pathways was significantly different across IBD and non-IBD samples in dogs, see Supplementary Table 3[6]. No pathways were significantly higher in between disease statuses in humans.

Taken together, these results have important implications for translational medicine and for understanding IBD in dogs. Also, the major functional gene content was conserved across non-IBD and IBD humans and dogs. While some significant predicted functions were identified between non-IBD and IBD dogs, the human sample population did not have enough non-IBD individuals analyzed for proper statistical power. This together with previous work suggests similar functional changes within the microbiota of dogs and humans with IBD. Previous studies have already shown that some treatment approaches to IBD that target the microbiome are conserved across dogs and humans, such as antibiotics, dietary modulation, and probiotics. For example, specific probiotic therapy has been shown to have similar effects on fecal and mucosa-adherent microbiota and host immune response (i.e., increase of tight junction proteins, increase in beneficial mucosa-adherent bacteria) in humans, dogs, and rat models of colitis [148, 194, 230]. Our study also revealed that the dysbiosis networks clearly differ in some key bacterial groups. Better understanding of the similarities in these microbial networks and functional changes may extend the abilities to test therapeutic approaches across

multiple host species.

Data Processing

Demultiplexed and quality controlled sequences were then clustered against the Greengenes [160] database (release 13_8) using the closed reference OTU picking protocol [46] as implemented in Quantitative Insights into Microbial Ecology (QIIME) [23] 1.9.0; these processing steps were performed using the default parameters. The OTU table used for primary analysis was filtered to include only samples from subjects that presented inflammatory bowel disease or that were healthy controls. Blank samples, and subjects with diarrhea were removed. Finally the table was rarefied to normalize for sampling effort [247] at 15,000 sequences per sample, this cutoff was selected as having the best trade-off between sequences and samples per disease status category.

Samples from [164] were processed with the same pipeline as described above, except these were rarefied at 4,500 sequences per sample.

To address the differences in sequencing length, the combination of samples with the human dataset [61] was preceded by trimming the sequences to an even 100 nucleotides per sequence, the rest of the pipeline was performed as described above. The same OTU picking protocol and reference database were used, therefore allowing the combination of the studies using the OTU tables. We only used the fecal samples from the human dataset.

Statistical analysis

Jupyter Notebooks [182] with the analysis of this dataset are available on online[7]. Briefly, alpha and beta diversity calculations of weighted and unweighted UniFrac [139] were performed using QIIME version 1.9.1. To assess statistical significance in these comparisons we used PERMANOVA, grouping the samples by disease status (healthy vs IBD), p=0.001 in both cases; pseudo-F statistics were 9.46 for unweighted UniFrac and 39.65 for weighted UniFrac. Receiver Operating Characteristic (ROC) curves and feature importance were calculated using Caret [124] and hack_ml [8].

Statistical significance between diversity distributions were calculated using Mann-Whitney's test, linear regressions and correlation coefficients were calculated using NumPy 1.10.4 [232] and SciPy version 0.17, and visualized using Seaborn 0.7.0 [241]. Principal coordinates analysis plots were created using Emperor 0.9.51-dev [236].

Metagenome predictions were performed using the Galaxy implementation of PiCRUST version 1.9.0. Significant differences in predicted KEGG Pathways between healthy and IBD groups were calculated using a Kruskal-Wallis test with multiple corrections (FDR and Bonferroni corrected p-values). The 16S data is comprised of 85 healthy dogs, 64 IBD-affected dogs, 29 healthy humans and 450 humans with IBD. On the other hand the PICRUSt predicted data is comprised of

---

[7]https://www.github.com/ElDeveloper/dogs
[8]https://github.com/RNAer/hack_ml

85 healthy dogs, 42 IBD-affected dogs, 23 healthy humans and 344 humans with IBD.

The Jaccard distance was used to compare the PICRUSt predicted data. This metric compares the samples on a presence/absence basis, and is not concerned with the similarity in the abundances of the OTUs but rather with the overlap in their presence. Each value is computed as one minus the ratio of shared OTUs to total OTUs present in the two samples.

To assess the quality of the predictions in the dog samples, we computed the 'nearest sequenced taxon index' (see Supplementary Figure S3[9]), and verified that the samples had acceptable values (below 0.15).

Dysbiosis Network

The dysbiosis network was calculated as described in [61], using CCREPE [202] and visualized using Cytoscape [207]. The index is calculated by taking the log transform of the abundance of the ratio of IBD-associated microbes to non-IBD associated microbes as determined by the correlation network.

This network is created by first scoring the co-occurrence and co-exclusion patterns in the samples. CCREPE uses a compositionally adjusted version of the checkerboard score [219]; the results are filtered to remove non-statistically significant relationships, and to preserve the largest connected component only. We represent the results as a graph where vertices are microbes and edges are

---

[9]https://images.nature.com/original/nature-assets/nmicrobiol/2016/nmicrobiol2016177/extref/nmicrobiol2016177-s1.pdf

interaction types. Vertices are of one of two classes (IBD-associated and Healthy-associated) as determined by the class where they were dominantly abundant, and edges between vertices have a weight given by their adjusted checkerboard score (negative values represent co-exclusions, and positive values represent co-occurrences). The specifics of this processing are described in the supplemental Jupyter notebooks.

A description of the DNA analysis and sample processing is provided in Appendix A

Acknowledgments

Chapter 4

Dynamic features of inflammatory

bowel disease

In recent years, clinical research in IBD has benefited from multi-'omic (genomic [58], metabolomic [92], and metagenomic [113]) characterizations of the condition. Each of these approaches describes a new component of relationships that, one by one, seem to uncover the processes regulating the inflammation and well-being of the affected hosts. Nonetheless, longitudinal descriptions (and as a consequence representations) of IBD were largely unexplored.

This chapter introduces two pioneering longitudinal studies of IBD. One where the sampling is sparse (every three months) and another where the sampling is dense (as much as every day). Both cohorts present increased rates of microbial variation. Specifically, we see that this appears to be different for the different subtypes of IBD. In order to visualize the increased variation, we rely on the

techniques introduced in Chapter 2, and on the definition of a reference plane that acts as a proxy for healthy variation.

Motivated by these findings, we use an classification model to determine the benefits of increased longitudinal sampling. For both (independently collected and sequenced) datasets, we observe that including more than one sample per subject increases the performance of a classifier. This property is only possible when we transform the original representation of the data and create per-subject consensus features based on multiple timepoints. Briefly, these features aggregate multiple samples together and exploit the increased instability as an informative classification feature.

Although we exercise and test this method with human-gut data, the approach is not restricted in any way to this environment. This approach could potentially benefit other applications, specially where longitudinal descriptors might be predictive of a state of interest.

Chapter 4.1 appeared in *Nature Microbiology, 2017*. My role in this project was to act as the lead analyst of the project, I developed a collection of new algorithms to represent the data, produced a series of visualizations used in the paper, co-wrote the text, and interpreted the results. Chapter 4.2 appeared in the journal *Gut, 2017*. As the lead contributor of this project, I co-wrote the text, generated the main figures, wrote the software used for analysis, interpreted the results, and deposited the data into a public repository.

4.1 Dynamics of the human gut microbiome in Inflammatory Bowel Disease

Inflammatory bowel disease (IBD) is characterized by flares of inflammation with periodic need for increased medication and sometimes even surgery. Inflammatory Bowel Disease (IBD) etiology is partly attributed to a deregulated immune response to gut microbiome dysbiosis. Cross-sectional studies have revealed microbial signatures for different IBD diseases, including Ulcerative Colitis (UC), Colonic Crohn's Disease (CCD), and Ileal CD (ICD). Although IBD is dynamic, microbiome studies have primarily focused on single timepoints or few individuals. Here we dissect the long-term dynamic behavior of the gut microbiome in IBD and differentiate this from normal variation. Microbiomes of v subjects fluctuate more than healthy individuals, based on deviation from a newly-defined Healthy Plane (HP). ICD subjects deviated most from the HP, especially subjects with surgical resection. Intriguingly, the microbiomes of some IBD subjects periodically visited the HP then deviated away from it. Inflammation was not directly correlated with distance to the healthy plane, but there was some correlation between observed dramatic fluctuations in the gut microbiome and intensified medication due to a flare of the disease. These results help guide therapies that will re-direct the gut microbiome towards a healthy state and maintain remission in IBD.

4.1.1 Results and Discussion

Both the state and the dynamics of the human gut microbiome in healthy individuals are highly personalized [34, 45, 56, 83, 151, 185, 83, 263]. Although cross sectional studies have revealed dysbiosis of the gut microbiome in IBD [61, 188, 213, 252, 253], little is known about the individual nature of microbiome dynamics in IBD, beyond a study of 3 UC patients before and after ileostomy, and two small studies of IBD patients in remission or during changes in disease activity [150, 254, 260]. Here we studied the long-term dynamics of the gut microbiome from an IBD cohort of 128 individuals (49 Crohn's Disease (CD), 60 UC, 4 Lymphocytic Colitis (LC), 15 Collagenous Colitis (CC)) and 9 Healthy Controls (HC). We sampled at three-month intervals, collecting 110 samples per individual for a total of 683 samples (Supplementary Table 1[1]). The microbiome composition in each sample was determined by sequencing the V4 region of the 16S rRNA gene for a total of 248 million 16S rRNA gene amplicons. To determine links between the gut microbiome and clinical factors, we collected clinical data, including fecal calprotectin (f-calprotectin) concentration and surgical resection status. To control sampling bias, we restricted our statistical analyses of volatility to a subset of the cohort that had sequence data from the first four time points and that had matching f-calprotectin concentrations; yielding 276 samples from 69 patients

---

[1]https://images.nature.com/original/nature-assets/nmicrobiol/2017/nmicrobiol20174/extref/nmicrobiol20174-s1.pdf

(Supplementary Table 1[1], Supplementary Dataset 1[2]); results were similar when all subjects were considered). We also included patient Genetic Load Scores (GLS) based on 163 known IBD risk loci for 29 patients to assess potential links between the host genetics, IBD and the microbiome [94].

As expected from previous work [61, 253], we found that HC and IBD subtypes formed distinct clusters by Principal Coordinates Analysis (PCoA) of unweighted UniFrac distances, with ICD patients least similar to healthy controls (Supplementary Figure 1[1]), ADONIS stratified by time point, $p < 0.001$). As in previous studies [188, 253, 254], we found differences in alpha and beta diversity of the microbiome according to IBD subtype and between subtypes and healthy controls and identified several families that correlated with health or disease state, e.g. *Enterobacteriaceae* with ICD and *Ruminococcaceae* with HC (Supplementary Figure 1-2[3]). Individual taxa that are differentially abundant between IBD subtypes compared to HC are listed in Table 4.1 (DESeq2, log2 fold change). The IBD microbiomes contained significantly lower abundances of putative beneficial Operational Taxonomic Units (OTUs) present in HC, as previously reported [61, 188, 213, 252, 253, 254], including *Prevotella copri* and the butyrate-producing bacterium *Faecalibacterium prauznitzii* (Table 4.1; Supplementary Figure 3[3]).

From the animated ordination of the samples (Supplementary Video 1[4]),

---

[2]https://images.nature.com/original/nature-assets/nmicrobiol/2017/nmicrobiol20174/extref/nmicrobiol20174-s2.txt

[3]https://images.nature.com/original/nature-assets/nmicrobiol/2017/nmicrobiol20174/extref/nmicrobiol20174-s1.pdf

[4]https://images.nature.com/original/nature-assets/nmicrobiol/2017/nmicrobiol20174/extref/nmicrobiol20174-s3.mov

Table 4.1: Differential abundance in specific taxa according to disease phenotype comparisons (DESeq2)

Groups compared	BaseMean	log2FoldChange	padj	Taxonomic annotation
ICD-r over ICD-nr	13,32	-7,05	0,0000000	*Faecalibacterium prausnitzii*
	4,08	-5,57	0,0000011	Lachnospiraceae
	3,32	-5,08	0,0000001	Ruminococcaceae
	6,49	-5,40	0,0000009	Ruminococcaceae
	19,85	-6,11	0,0000005	Ruminococcaceae
	94,87	-5,87	0,0000001	Ruminococcaceae
	53,59	-8,10	0,0000012	*Ruminococcaceae Ruminococcus*
	72,13	-5,14	0,0000241	Clostridiales
ICD-r over HC	14,91	7,20	0,0000000	Alteromonadales [Chromatiaceae]
	13,32	-7,22	0,0000000	*Faecalibacterium prausnitzii*
	2,94	-5,34	0,0000007	Ruminococcaceae
	2,33	-5,62	0,0000000	Clostridiales
	10,41	-7,47	0,0000000	Lachnospiraceae
	15,12	-7,62	0,0000001	Lachnospiraceae *Coprococcus*
	4,13	-8,43	0,0000000	Lachnospiraceae
	19,85	-7,15	0,0000000	Ruminococcaceae
	5,43	-8,72	0,0000000	Ruminococcaceae
	7,16	-6,98	0,0000151	Clostridiales
	3,78	-6,69	0,0000249	Clostridiales
	2,10	-6,53	0,0000000	Ruminococcaceae
	5,76	-7,85	0,0000000	Ruminococcaceae
	53,59	-8,64	0,0000000	Ruminococcaceae *Ruminococcus*
	72,13	-5,71	0,0000000	Clostridiales
	121,13	-9,95	0,0000000	*Prevotella copri*
	5,87	-7,58	0,0000000	*Methanobrevibacter*
ICD-nr over HC	14,91	6,47	0,0000185	Alteromonadales [Chromatiaceae]
	2,94	-7,00	0,0000756	Ruminococcaceae
	5,43	-8,17	0,0000682	Ruminococcaceae
	121,13	-7,82	0,0000185	*Prevotella copri*
CCD over HC	5,43	-8,65	0,0000000	Ruminococcaceae
	121,13	-7,94	0,0000000	*Prevotella copri*
UC over HC	6,53	6,32	0,0000978	*Alistipes massiliensis*

Criteria for inclusion: BaseMean > 1 and padj < 0.0001. Brackets indicate putative taxonomy based upon phylogenetic placement as given in the Greengenes taxonomy. BaseMean is the mean of normalized counts for all samples. padj is the BenjaminiHochberg adjusted p-value.

we observed that although microbiome samples from the healthy individuals varied over time, they were restricted to a small volume of the ordination space. In contrast, IBD subtypes traversed far more of the total volume, sporadically visiting the area where healthy samples resided. To summarize these dynamics, we identified a 'healthy plane' (hereafter referred to as HP). Briefly, this plane is calculated in a space derived from PCoA of unweighted UniFrac distances of healthy subjects (Supplementary Video 1[4]). We constructed a model using the samples from the HC patients, and fit them to a two-dimensional plane embedded in a three-dimensional space using the least squares method (Figure 4.1). The plane is then restricted to only span the three-dimensional ranges of the HC samples. This plane was used as a proxy to represent the normal microbial variation within healthy subjects and to summarize the abnormal, intermittent dysbiosis associated with IBD.

The procedure was as follows: let S be a set of n samples $s_1, s_2, \ldots s_n$ corresponding to a group of trajectories, each trajectory pertaining to a subject with at least four samples collected at distinct points in time. Each sample is represented as a three-dimensional vector corresponding to sample coordinates in ordinated space i.e. $s_1 = (x_1, y_1, z_1); s_2 = (x_2, y_2, z_2); \ldots s_n = (x_n, y_n, z_n)$. We fit a linear model to S by the least squares method to obtain coefficients for the equation of a three dimensional surface T, next we restricted a segment of this surface to the ranges given by $[min_x(S), max_x(S)]$ and $[min_y(S), max_y(S)]$, we defined this to be a plane representative of S or P_S. When S is the set of samples from healthy

77
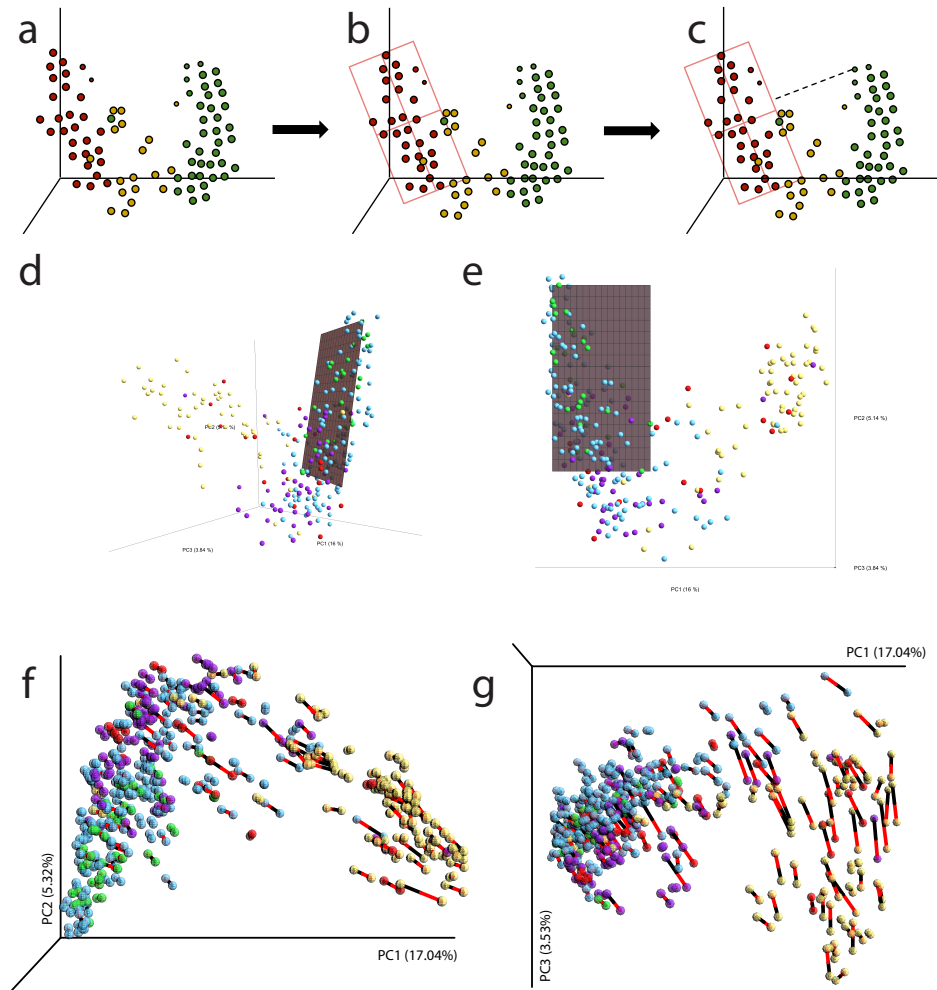
Figure 4.1: **Defining a healthy plane.** Diagram summarizing the procedure for creating a representative plane for a group of samples S: **a**, sample selection, **b**, model fitting and **c**, distance calculations for all samples. The healthy plane is then located in UniFrac space by **d**, fitting a line to the major axis of the points, and **e**, defining a least-squares fit to identify a plane that minimizes the sum of squares of distances to the nearest point on the plane. **f**, Verification that the position of the healthy plane is not driven by proteobacteria-dominated outliers: Procrustes Analysis comparing original samples and those with Proteobacteria removed. A vector connects each original sample (red) with the same samples after Proteobacteria have been omitted (black). $p < 0.001$, $M^2 = 0.018$, 999 permutations. **g**, The short length of most vectors indicates that the relative composition of most samples does not change when proteobacteria are filtered out.

subjects in an ordination space, we refer to P_S as the HP. Finally, we defined d_k to be the Euclidean distance from a sample k to the nearest point lying on P_S. After measuring d_k for all samples in our study, we grouped samples according to their diagnosis, and compared the distributions of distances. Figure 4.1 a-c demonstrates this procedure, and Figure 4.1 d-e demonstrates the placement of the HP in the context of our full IBD dataset. Samples from the HP are co-located with one another, while many samples from IBD patients are further away. To investigate whether this effect is due to 'outlier' groups of samples that are dominated by taxa that are typically rare in healthy individuals, we excluded all Proteobacteria from the dataset, and compared the location of samples in unweighted UniFrac space with and without Proteobacteria using Procrustes analysis. As shown in Figure 1f-g, the omission of Proteobacteria is significant (p < 0.01), but this effect is largest in the already dysbiotic ICD-r patients and aligns with PC3. The effect on healthy controls is minimal, and healthy samples are still located only near the HP (Figure 4.1 f-g).

Calculation of the mean Euclidean distance from each sample to the HP revealed that all subtypes of IBD significantly deviated from the HP (Generalized Linear Mixed effects model (GLM), all p < 0.00268, Figure 4.2a). Samples from patients with CCD and UC were closer to the HP, and some samples did not differ significantly from healthy controls, whereas ICD samples were the most distant from the HP (Supplementary Video 1[5]; Figure 4.2). The highest volatility was

---

[5]https://images.nature.com/original/nature-assets/nmicrobiol/2017/nmicrobiol20174/

observed for ICD patients that had previously undergone ileocecal resection (ICD-r), followed by ICD patients without surgery (ICD-nr) based on UniFrac distances between successive samples (Figure 2b). ICD-r patients also had low gut microbial richness (Supplementary Figure 2[6]). Ileocecal resection is a major modifier of intestinal physiology, and the observed pronounced volatility in ICD-r patients might be partly explained by removal of the ileocecal valve per se. Ileal inflammation may also influence bile salt uptake and, consequently, colonic transit time and microbiome volatility. Interestingly, several IBD patients had complex trajectories that sporadically moved to and from the HP (Figure 4.2c, Supplementary Video 1[5]).

Although our study represents the largest longitudinal analysis of the IBD microbiome to date, the small number of patients in specific subgroups limited some statistical comparisons. Furthermore, the number of healthy controls was lower than the number of IBD patients. To address this limitation of unequal sampling of healthy individuals and IBD patients, we compared the volatility of our healthy controls to healthy participants in two published studies, the Student Microbiome Project (SMP) and the Moving Pictures (MP) datasets [24, 56]. There was less variability over time in healthy individuals across all three cohorts compared to those with IBD (Figure 2b). This result emphasizes that IBD is characterized by volatile dysbiosis not found in healthy people, and confirms earlier preliminary

extref/nmicrobiol20174-s3.mov

[6]https://images.nature.com/original/nature-assets/nmicrobiol/2017/nmicrobiol20174/extref/nmicrobiol20174-s1.pdf
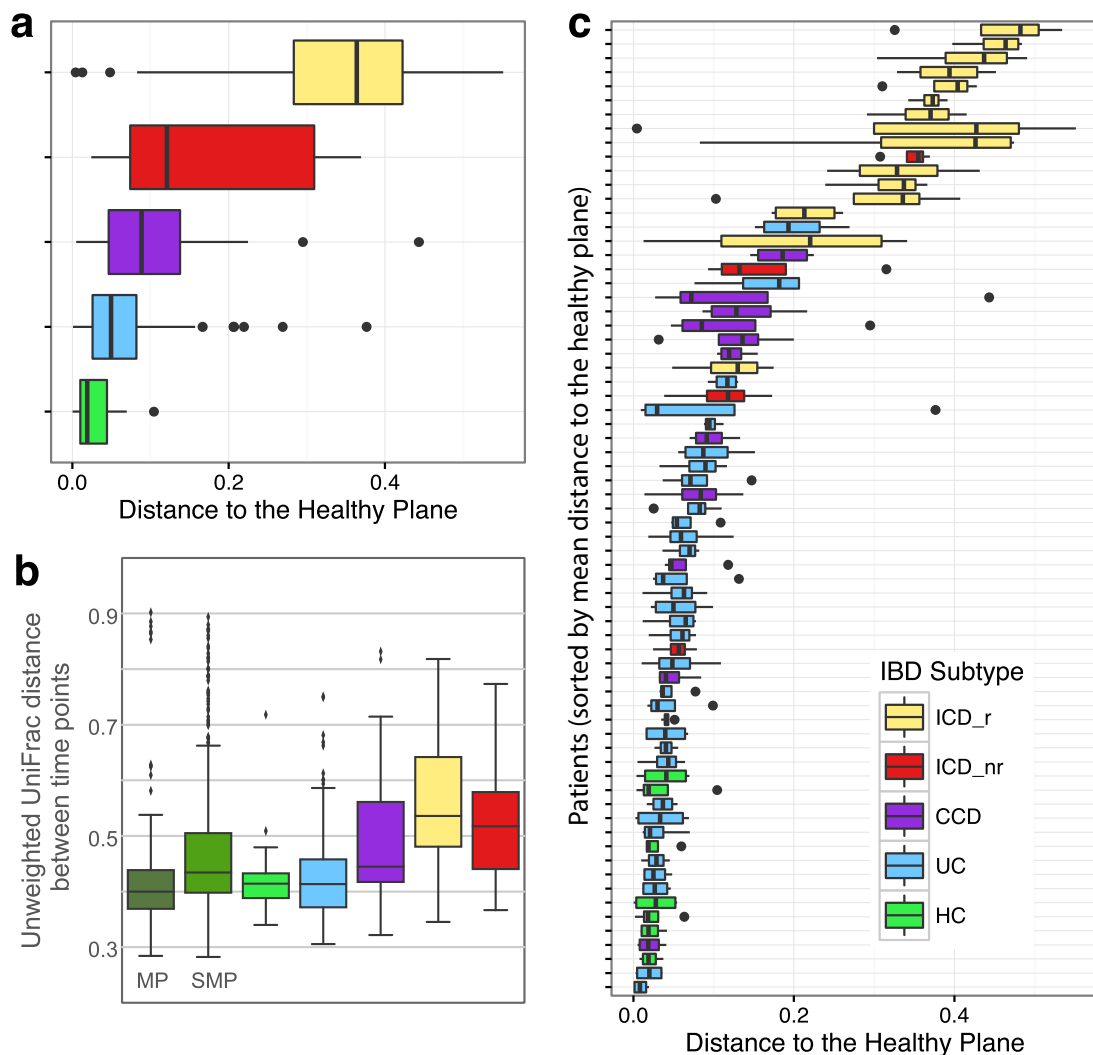
Figure 4.2: **The gut microbiomes of different IBD subtypes display different distributions relative to a healthy plane (HP). a**, Median distances from HP for each IBD subtype. All IBD subtypes were significantly different from healthy controls (GLM, all $p < 0.00261$). b, UniFrac distances between subsequent samples. c, Distance to HP for each individual patient. HP was defined using data shown in Supplemental Video 1[5]. See Supplementary Table 1[6] for composition of downstream analysis cohort. Boxes show interquartile range (IQR). Whiskers denote the lowest and highest values within 2.5 IQR of the median. Circles represent outliers.

results and meta-analysis of much smaller studies [147, 150, 254, 260].

To extend our understanding of the mechanisms underlying the microbiome dynamics, we explored the correlation between the dynamics and inflammatory activity in each sample using f-calprotectin > 150 μg/g as a surrogate for inflammatory activity. The concentration of f-calprotectin in stool samples has previously been correlated with endoscopic and histopathologic activity, and is used in daily clinical practice because it is non-invasive [131]. We observed that concentrations of f-calprotectin were higher in all IBD subtypes than in healthy controls (Figure 4.3). However, we did not observe a significant correlation between f-calprotectin and distance from the HP (Figure 4.3, GLM p = 0.275). Although recent microarray analyses of the gut microbiome in a cohort of anti-TNF treated pediatric IBD patients[116] and experiments with gnotobiotic fecal transplants suggest that microbial composition and function are causally associated with inflammatory activity [192], our differential abundance testing revealed only weak trends and no specific OTUs that varied significantly with active inflammation, using f-calprotectin > 150 μg/g as cut-off. However, the use of f-calprotectin as a proxy for inflammatory activity might have introduced bias, because f-calprotectin is a less accurate marker of ileal than colonic inflammation [211] and ICD patients displayed the greatest distance from the HP.

Examples of the microbiome dynamics for one representative HC and one from each IBD subgroup are shown in addition to changes in f-calprotectin concentrations, distance to HP, and Shannon diversity in Figure 4.4 (all individual

profiles are shown in Supplementary Figure 3[7]), illustrating the more stable dynamics over time for HC and UC compared to the other clinical phenotypes of IBD, with the most fluctuations occurring for patient 69 that had undergone surgical resection. The f-calprotectin levels were low and relatively stable for the HC compared to the IBD patient (Figure 4.4). In these examples, there were also substantial fluctuations in diversity for the HC, ICD-r, and CCD by contrast to UC and ICD-nr patients.

We further explored the individual dynamics of the gut microbiome in IBD patients who experienced increased clinical disease activity, according to the physician's global assessment (Supplementary Figure 4[7]). Recently, the short-term (6 weeks) dynamics of the microbiome in pediatric patients with active IBD treated with anti-TNF indicated that initiation of medical treatment changes the microbial composition at the genus level [116]. Because we had few anti-TNF exposed patients, and our patients received a course of corticosteroids at flare as first line therapy, we explored how corticosteroid administration influenced microbiome dynamics. Our data demonstrate that change in medication influenced the volatility of the microbiome (Supplementary Figure 4[8]). Patients receiving a course of oral corticosteroids (n=7) had more microbiome fluctuations than patients on stable medication (n=49), based on calculations of unweighted UniFrac distance between time points (Wilcoxon Signed-rank test; p=0.04). Our dynamic model

---

[7]https://images.nature.com/original/nature-assets/nmicrobiol/2017/nmicrobiol20174/extref/nmicrobiol20174-s1.pdf

[8]https://images.nature.com/original/nature-assets/nmicrobiol/2017/nmicrobiol20174/extref/nmicrobiol20174-s1.pdf
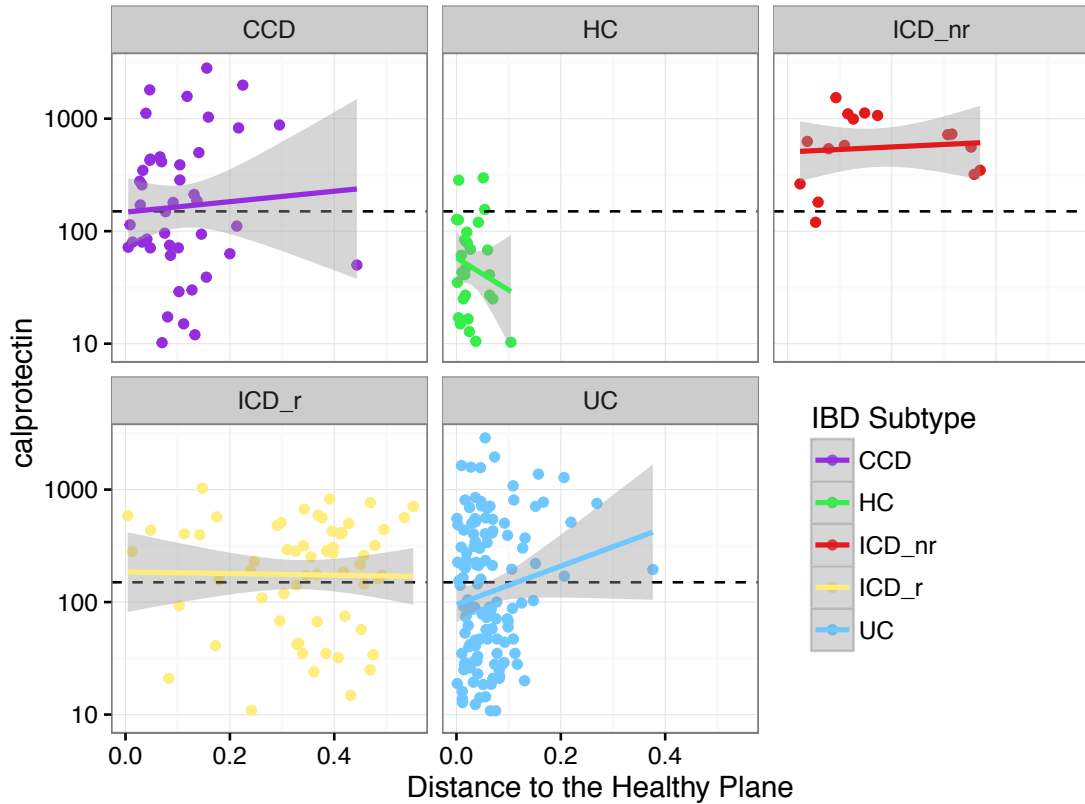
Figure 4.3: **Correlation between fecal calprotectin concentrations and distance to a defined healthy plane (HP) in 3D ordination space.** Data represent a correlation of f-calprotectin levels and distance to the here defined healthy plane in 3D ordination space (see Supplementary Video 1[5]) for each individual and time point for different inflammatory bowel disease (IBD) subtypes. To compare the relationship between f-calprotectin and the healthy plane, a generalized linear mixed effects model was fit, with a conditional Gamma distribution, using f-calprotectin and disease type as fixed effects and including a random subject effect; f-calprotectin was not significant (p = 0.27501).
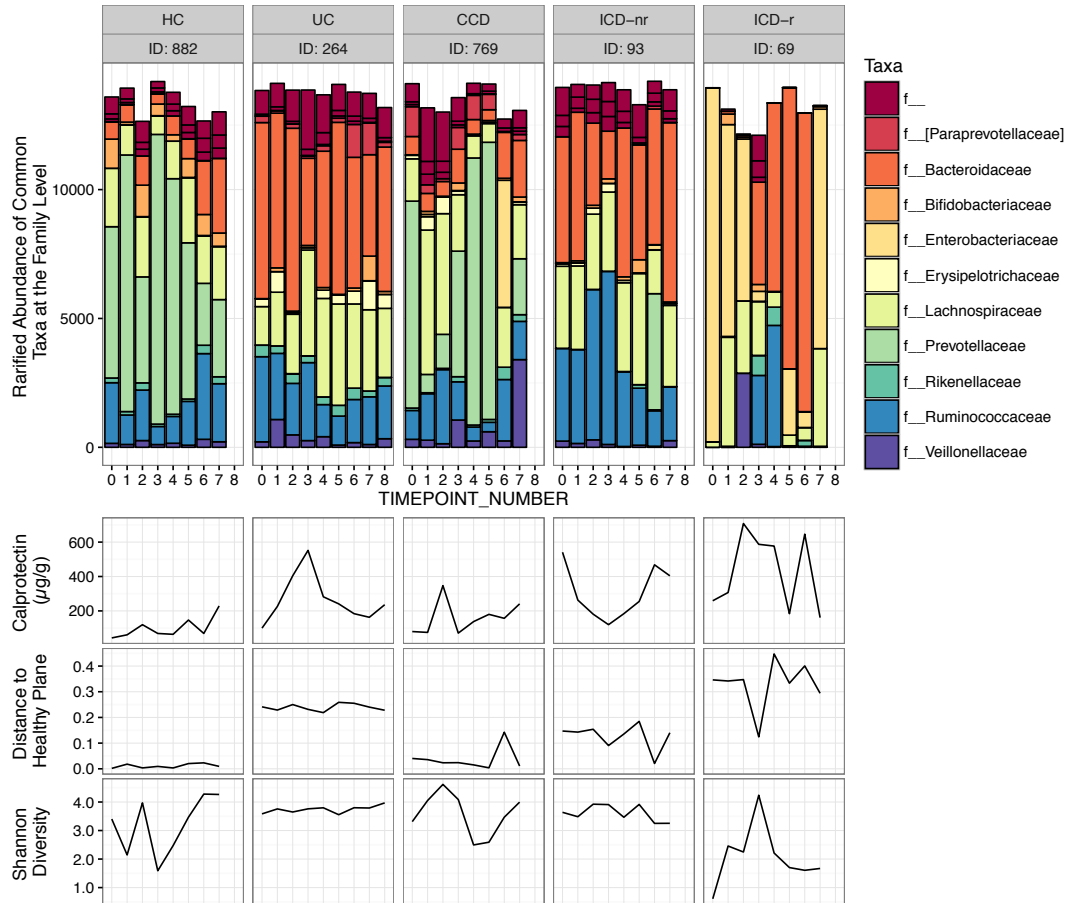
Figure 4.4: **Microbiome dynamics of selected individuals from each IBD subtype and a healthy control.** From each IBD subtype and healthy control group, representative individuals sampled over the most time points and having complete clinical and sequence data were selected. Data represent f-calprotectin values, distance to the healthy plane, and Shannon diversity and rarified abundances of most common taxa at the family level. Note that taxa unclassified at the family level are represented in the f_' category.

suggests that beyond the association with IBD subtype and the weak correlation with inflammation, the dynamics of the microbiome composition are influenced by changes in medication. The extent to which other factors, such as dietary changes and smoking, may have influenced the observed volatility remains speculative, because the collected information was insufficiently detailed to include these factors as covariates.

To evaluate the microbiome as a predictive tool, we combined the microbial and clinical data and used a supervised learning Random Forests model to predict IBD subtypes [21, 138]. To avoid overfitting, our models were built using OTU abundances from the first time points only, along with clinical metadata (Body Mass Index (BMI), f-calprotectin concentrations, sex, and Distance to the HP). Accuracy was evaluated using the remaining three time points, which were not used to train the model. Using this model, the IBD subtypes were discriminated from healthy controls and correctly predicted for 66.6% of samples (Supplementary Table 2[9]), consistent with the findings reported in Gevers et al. (2014). Feature importance scores from this model revealed several potential microbial indicators of IBD subtypes, including OTUs matching to *Lachnospira*, *Clostridium*, *Oscillospira*, and many unidentified *Ruminococcaceae* (Supplementary Table 3[9]). Intriguingly, the accuracy of the model increased slightly if f-calprotectin concentrations were omitted, but decreased by at least 10% if the distance to the HP was removed

---

[9]https://images.nature.com/original/nature-assets/nmicrobiol/2017/nmicrobiol20174/extref/nmicrobiol20174-s1.pdf

(Supplementary Tables 2 and 3[9]), suggesting that the HP is a more important factor in the model. Comparable levels of accuracy were previously achieved using rectal samples [61], but here we show that the same can be achieved with fecal samples, which are easier to collect. When immunochip data were included for a subset of 29 IBD individuals and the Random Forests model was repeated, the samples were still classified into the four IBD subtypes (UC, CCD, ICD-r, ICD-nr) (Supplementary Tables 2 and 3[9]). While distance to the HP remained as the single most important feature for classification,GLS were more predictive than sex, f-calprotectin, or BMI when included in the model (Supplementary Table 2[9]). However, including GLS only increased the overall accuracy of the model by about 2%, demonstrating the predictive potential of the microbiome. Interestingly, in a recent study of obesity covering 339,224 individuals, the 97 risk loci for obesity accounted for only 2.7% of BMI variation [137], whereas the microbiome classified lean from obese individuals with 90% accuracy [115], providing precedent for the predictive value of the gut microbiome over human genetics in chronic disease.

In summary, by analyzing fecal samples collected every 3 months from a large IBD cohort, we determined the long-term volatility of the gut microbiome in IBD. Our data revealed that although the microbiome of healthy individuals varied it was only within a newly defined HP, whereas there was considerable volatility away from the HP for several of the IBD cohorts. Devising improved methods to detect the healthy state with non-invasive sampling, to predict when the healthy state will be departed, and to sustain the microbiome in this healthy state by

87

erecting barriers that prevent the slip back into dysbiosis, will be an important focus of future work.

4.1.2 Methods

Cohort demographics, sample collection and sample processing details can be found in Appendix B.1.

Phylogenetic Analysis

Sequence data were processed using QIIME 1.9.0-dev through the online platform QIITA[10][23]. Four HiSeq lanes of data were demultiplexed with default quality filtering settings and subsequently combined, resulting in 248,547,926 total sequences. These sequences were clustered using SortMeRNA at 97% identity against the Greengenes rRNA reference database May 2013 release [118, 160]. Sequences that failed to match the database were discarded. 237,653,256, or approximately 95.6% of the sequences, clustered against the Greengenes reference dataset. Even sampling was performed at 14,553 sequences per sample for beta diversity and supervised learning analyses. The beta diversity principal coordinates plot of unweighted UniFrac distances was constructed using the same rarefied OTU table, and visualized in Emperor [139, 236]. Clinical matched metadata, including f-calprotectin concentrations, were included when available. From each patient, the first four microbiome samples with matched f-calprotectin concentrations were

---

[10]https://qiita.ucsd.edu

sub-selected for use in downstream analysis.

Statistical Analysis.

The R package phyloseq was used to import and graph data, while the packages DESeq2, randomForest, and vegan were used to perform differential abundance testing and supervised learning [89, 134, 138, 162, 174]. Statistical significance of unweighted UniFrac distance matrices comparing healthy controls and IBD subtypes was assessed using the ADONIS test. The variation of the microbial community over time was calculated with vector lengths produced by summing the total distance between each subject's time points over the first three PCoA axes of unweighted UniFrac space. The Random Forests model was constructed using the first time point from each patient in the downstream analysis cohort and prediction accuracy was measured using the subsequent three time points (CCD = 11 patients, ICD without resection (ICD-nr) = 4 patients, ICD with resection (ICD-r) = 15 patients, UC = 30 patients, and HC = 7 patients). The following metadata categories used as features for supervised learning: BMI, f-calprotectin (continuous), sex, and Distance to the Healthy Plane. For classification on the subset of samples with immunochip data, the above process was used while adding GLS. We used a dataset of 29 IBD samples (25 CD, 4 UC) with available immunochip data to estimate a GLS for each sample. In particular, GLS was calculated summing the number (counts) of risk alleles (0,1,2) at lead Single Nucleotide Polymorphism (SNP) from each IBD risk locus (n = 163) according to Jostins et al. [94].

After performing principal coordinates analysis of the unweighted UniFrac distance matrix, the samples of HC were used to fit using the least squares method on the first three principal coordinates. The distance from each sample to this plane was measured and added to the sample metadata both for weighted and unweighted UniFrac. To compare distribution of samples relative to this healthy plane, a generalized linear mixed effects model was fit, with a conditional Gamma distribution, using disease type as a fixed effect and including a random subject effect.

To assess the variation between samples in an orderly manner (as described by the time point), we measured the UniFrac (weighted and unweighted) distance between samples that occur sequentially for any given subject; samples representing the final collection point do not have a value in these columns, as there is no subsequent sample to compare to.

We examined the power of differential abundance tests for the 16S data. DESeq2 [138] assumes that counts can be modeled as a negative binomial distribution with a mean parameter, allowing for size factors, and a dispersion parameter. The test for differential abundances fits a generalized linear model with a negative binomial family and a log link function.

A power analysis was conducted by using samples from the downstream analysis cohort (Supplementary Table 4[11]). The power of the differential abun-

[11]https://images.nature.com/original/nature-assets/nmicrobiol/2017/nmicrobiol20174/extref/nmicrobiol20174-s1.pdf

dance test is dependent on the sample size of groups, difference in mean counts, type 1 error rate, and the dispersion value. Here we used a conventional type 1 error rate of 0.05 and assumed that we have two groups with sample sizes of n1 = 25 and n2 = 80. A total of 1000 OTUs were randomly selected from the data and dispersion parameters were estimated for each OTU. Data was then simulated from a negative binomial distribution, as specified by DESeq2, for each estimated dispersion value with means giving 2 fold, 1.5 fold, and 1.25 fold differences. For each mean fold difference value and dispersion value, a total of 5000 data simulations were done and a Wald test for difference in means using a generalized linear model was conducted. Based on these simulations, where data is generated with true differences, the fraction of times that the null hypothesis is correctly rejected (the power), was calculated and our comparisons were well powered given our subset sample sizes (Supplementary Table 4[12]).

Code Availability

Our analysis methods make use of standard, open source software. R software packages are available on CRAN[13], bioconductor[14], and GitHub[15]. Python software is available in the bioconda and biocore Conda channels, and is maintain on GitHub[16].

---

[12]https://images.nature.com/original/nature-assets/nmicrobiol/2017/nmicrobiol20174/extref/nmicrobiol20174-s1.pdf

[13]cran.r-project.org

[14]bioconductor.org

[15]github.com/joey711/phyloseq

[16]github.com/biocore,github.com/ElDeveloper/reference-plane

4.1.3 Acknowledgments

4.2 Guiding longitudinal sampling in inflamma-
tory bowel diseases cohorts

The volatile microbial signatures of Crohn's Disease (CD) patients greatly hinder our ability to classify healthy from affected subjects using 16S rRNA profiles from stool. The recent work by Pascal et al. [180] overcomes this issue and other complications [196], producing a decision tree that classifies subjects with CD, Ulcerative Colitis (UC), irritable bowel syndrome and anorexia. Although the authors note that both subtypes of Inflammatory Bowel Disease (IBD), particularly CD, have increased microbial community instability, this information is not used as a feature to improve classifier accuracy. Could microbiome instability become actionable by creating a new classifier that benefits from repeated measurements? If so, how many samples per individual are needed to assess instability?

We collected daily stool samples for up to six weeks from 19 CD subjects and 12 controls (see the analysis notebook for cohort description[17]) over 2 separate periods of 2 or 4 weeks spread over 2 and 5 months, for a total of 960 samples. We believe this is the most densely sampled longitudinal study of CD; previous studies collected samples every 1-3 months [72, 180]. Our cohort shows decreased alpha diversity and increased stability, as previously reported in CD and other subtypes of IBD [58, 61, 72, 180]. We also noted that subjects who underwent resection have lower alpha diversity than other CD-affected subjects (see analysis notebooks[17]).

---

[17]https://github.com/knightlab-analyses/longitudinal-ibd

A critical experimental design question for clinical studies is whether a finite budget should best be spent collecting samples from more patients, or collecting more serial samples from each patient? Therefore, we created a Random Forests [21] model based on per-subject aggregation of longitudinal data for alpha diversity [49], beta diversity [139] and abundances of two phylogenetic factors found to be associated with CD in ileal biopsies [61, 240] (Figure 4.5). With one sample per subject, our model performs worse than a classifier that uses microbial relative abundances at a single time point, but when more samples per-subject are added, the classifier outperforms that approach and results previously only attained with biopsy samples [61]. Furthermore, we replicate this observation with a different cohort (Table 4.2).

Novel analyses aggregating features over time and combining both alpha and beta diversity over time using our intensive daily sampling demonstrates that the main benefits are already obtained by collecting between 3-5 fecal specimens, and no additional benefits are obtained beyond 7 serial samples. Similar results are found for monthly sampling. These results highlight the importance of treating CD as a volatile, time-varying condition, even during clinical remission, but provide hope to clinicians in that a relatively small number of samples yields large additional benefits, facilitating patient compliance. This information can be used to design collection of fecal samples for a large prospective cohort of CD patients for longitudinal studies of host-microbial interactions over time.

The methods demonstrated here have not previously been used for micro-
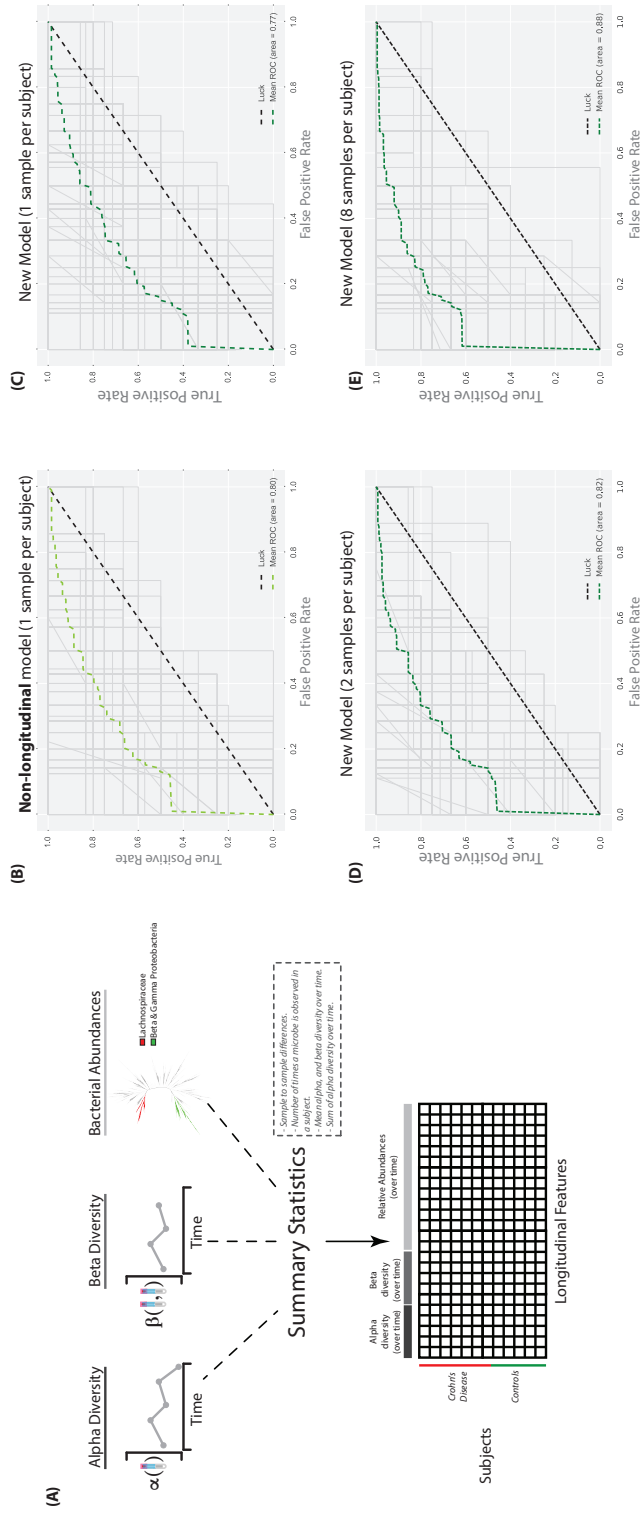
Figure 4.5: Diagram for the model creation, and comparison of four receiver operating characteristic (ROC) curves. (A) Diagram describing the origin for the classifying features. (B) ROC curve for a model that relies on relative abundances and one sample per subject (as used in previous publications). (C-D) ROC curve for our new model at 1, 2 and 8 samples per subject. The gray lines represent the performance at each of the 100 iterations. The dotted black diagonal line represents the performance of a classifier that guesses the labels at random.

Table 4.2: Performance summary of the classifier at increased samples per subject for this cohort (daily samples) and a previously published cohort. The area under the curve (AUC) summarizes the performance; closer to 1 is better, of the model trained on the different sample sizes as described by the other columns. The row marked with represents the performance of a classifier that relies on non-longitudinal relative abundances only.

AUC	Samples per Subject	Controls	Crohn's Disease	Sampling
0.80	1	12	19	
0.77	1	12	19	
0.82	2	12	19	
0.85	3	12	18	
0.86	4	12	18	
0.87	5	12	18	
0.87	6	12	18	Daily samples
0.88	7	12	18	
0.88	8	12	18	
0.87	9	12	18	
0.87	10	12	17	
0.86	11	12	16	
0.80	1	9	19	
0.80	1	9	19	
0.83	2	9	15	Monthly samples [72]
0.86	3	8	14	
0.92	4	8	12	

96

biome analyses but have been used for other engineering applications, for example in production lines to predict product specification outcomes in a steel manufacturer's facility [30]. We expect the results to generalize in other systems, including other gastrointestinal and hepatic disorders, where dynamic features of the microbiome, host gene expression, or other accessible features can act as indicators of underlying dysbiotic states.

4.2.1   Methods

Data Processing

Demultiplexed sequences were filtered using Quantitative Insights into Microbial Ecologys (QIIMEs) [23] default parameters for quality control using (version 1.8.0-dev). To account for the fact that a subset of the sequences were processed using a MiSeq instrument and the rest using a HiSeq instrument, the resulting reads were trimmed to an even length of 99 nucleotides [25, 136]. The resulting sequences were clustered using the closed-reference Operational Taxonomic Unit (OTU)-picking protocol (utilizing UCLUST as the clustering algorithm [46]) against the Greengenes database (release 13_8) [160]. Distance matrices were calculated using the UniFrac [139] distance metric. Bimodality tests were performed using Hartigan's dip test [78]. To account for uneven sampling efforts in the samples, the resulting OTU table was rarefied at 7,400 sequences per sample. The number was selected so the main category of interest (health status of the patient)

had a balanced number of classes.

Classifier

To assess the benefits of increased number of samples per subject, we used a Random Forests [21] model and we created the features following these steps. First we randomly selected N samples for each subject. For this set of N samples, we created features based on alpha diversity, beta diversity and microbial relative abundances. We then split the dataset into training and testing subsets and evaluated the classifier's sensitivity and specificity. Finally, we repeated the selection of N random samples for M iterations, and summarized the results using a Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). After the M iterations, we increased N by one and repeated the procedure until we reached N'. The result is a total of $N' - N$ ROC curves and AUC scores (one for each number of samples per subject).

The features are created based on longitudinal patterns of the different data types (alpha diversity, beta diversity and relative abundances). In all cases, we relied on the sample collection date to order the data and measured a series of summary statistics. For alpha diversity and the microbial relative abundances, the per-sample summaries were ordered and treated as vectors. For beta diversity, we considered only the distances between ordered and subsequent timepoints and treated this as a vector. Features were extracted from these vectors as described in the TSFRESH package [30], for example by counting the number of samples where

an individual OTU is not zero, to include the mean of alpha or beta diversity distances, the mean rate of change of the alpha diversity vector, etc.

In the case of the microbial abundances, we performed a feature selection step using phylofactor [240]. Phylofactorization was performed by maximizing the F-statistic from logistic regression on disease state in patients with CD. Over 200 factors were found as significant predictors of CD in the terminal ileum, even when controlling for multiple comparisons by a Bonferroni correction. Two factors identified large clades (>100 OTUs) used for feature selection. Factor 3 identified a monophyletic clade of 518 OTUs in the Lachnospiraceae family which decreases in the terminal ileum of patients with CD relative to healthy patients, and factor 26 identified a monophyletic clade containing 737 OTUs of the *Gammaproteobacteria* and *Betaproteobacteria* classes which increases in relative abundance in the terminal ileum of patients with CD relative to healthy patients (Supplementary Figure 1[18]). The dataset used for feature selection [61], was not used during the training or testing stages of the classifier construction.

Jupyter notebooks and source code describing all the analyses in this paper can be found online[19].

Cohort demographics, sample collection and sample processing details can be found in Appendix B.2.

---

[18]https://static-content.springer.com/esm/art%3A10.1186%2Fs40168-017-0269-3/MediaObjects/40168_2017_269_MOESM2_ESM.tif

[19]https://github.com/knightlab-analyses/longitudinal-ibd

4.2.2 Acknowledgments

Chapter 5

Restoring a lost ecosystem

Clostridium difficile infection (CDI) is a hospital-acquired infection resulting from antibiotic administration for an unrelated condition. Paradoxically, the treatment for CDI is to prescribe more antibiotics. This treatment generally translatesto depleted microbial diversity. As a consequence, resources that could be consumed by other communities are now available for *Clostridium difficile* to thrive on. More recently CDI has been treated through the administration of Fecal Microbiota Transplants (FMTs), with a success rate above 90% [233]. A FMT consists of reintroducing bacteria into the Gastro Intestinal (GI) tract of an affected patient. Frequently, after a day, the symptoms of CDI disappear and the recipients recover. In this chapter, we examine two cohorts and focus on the short and long-term changes in the gut microbiota after a FMT.

In the first cohort, relying on an animated β-diversity plot (as introduced in Chapter 2.2), we visualize the immediate changes in the gut microbiome after

a FMT. In addition, we also visualize the stability of the microbiome after the transplant and note that the only point of instability occurs immediately after the FMT.

The second cohort is composed of patients suffering from CDI and in some cases with underlying Inflammatory Bowel Disease (IBD). We find that, while both groups recover from the CDI, the micro-ecological effects of FMT appear to be dampened in subjects with underlying IBD. Overall, individual phylogenetic diversity is lower, and the number of taxa differentially present before and after the transplant is also smaller (when compared to the non-IBD counterparts).

Chapter 5.1 appeared in the journal *Microbiome, 2015.* As the lead analyst in this project, I contributed to the text, analyzed and interpreted the data, wrote software used for the analysis, and generated the main visuals. Chapter 5.2 appered in the journal *Microbiome, 2017.* As the co-lead contributor to this project, I co-wrote the text, generated the figures, analyzed, interpreted, and deposited the data into a public repository.

5.1 Dynamic Changes in Short- and Long-Term Bacterial Composition Following Fecal Microbiota Transplantation for Recurrent *Clostridium difficile* Infection

FMT is an effective treatment for recurrent CDI that often fails standard antibiotic therapy. Despite its widespread recent use, however, little is known about the stability of the fecal microbiota following FMT. Here we report on short- and long-term changes, and provide kinetic visualization of fecal microbiota composition in patients with multiply recurrent CDI (R-CDI) that were refractory to antibiotic therapy and treated using FMT. Fecal samples were collected from four patients before and up to 151 days after FMT, with daily collections until 28 days and weekly collections until 84 days post-FMT. The composition of fecal bacteria was characterized using high throughput 16S rRNA gene sequence analysis, and compared to microbiota across body sites in the Human Microbiome Project (HMP) database, and visualized in a movie-like, kinetic format. FMT resulted in rapid normalization of bacterial fecal sample composition from a markedly dysbiotic state to one representative of normal fecal microbiota. While the microbiome appeared most similar to the donor implant material one day post-FMT, the composition diverged variably at later time points. The donor microbiota composition also varied over time. However, both post-FMT and donor samples

103

remained within the larger cloud of fecal microbiota characterized as healthy by the HMP. Dynamic behavior is an intrinsic property of normal fecal microbiota, and should be accounted for in comparing microbial communities among normal individuals and those with disease states. This also suggests that more frequent sample analyses are needed in order to properly assess success of FMT procedures.

5.1.1 Introduction

FMT has emerged in recent years as a highly effective treatment for refractory CDI that cannot be cured with antibiotics alone [107]. The procedure leads to prompt engraftment of donor microbiota, attainment of donor-like bacterial diversity, and normalization of the overall microbial community structure [59, 76, 105, 204, 206, 233, 245]. However, existing data characterizing long-term stability of engrafted microbiota are limited. One recent study suggests the microbiota of patients after FMT may not fully recover until 16 weeks after the procedure [215]. This type of analysis, however, is complicated by the fact that the microbial communities are intrinsically dynamic, and affected by daily fluctuations in the host's diet, activities, and health [37, 40, 114]. In addition, multiple fixed host factors, such as different states of immune competence, genetics, or gastrointestinal anatomy, likely also affect the composition, stability, or resilience of colonic microbiota [15, 95, 120, 140, 143]. Therefore, it is unclear whether divergence in post-FMT microbiota from that of donor implant material represents continued recovery, or whether these temporal changes are a general characteristic

of host-associated gut microbiota in a changing host environment.

Here we describe both short- and long-term dynamic changes of fecal bacterial composition in four patients following FMT. All patients received microbiota from the same pre-qualified donor according to the standardized FMT protocol we described previously [75]. Three patients received freshly prepared microbiota and one patient received microbiota that had previously been frozen. We compared pre- and post-FMT fecal microbial communities from these patients, as well as pre-FMT communities from 10 additional patients with multiply R-CDI, to the sequences of normal subjects described in the Human Microbiome Project [227]. In addition, we compared temporal changes in fecal bacterial composition in recipients following FMT to temporal changes observed within samples from the donor.

5.1.2   Materials and Methods

A description of sample collection and processing is provided in Appendix C.1.

Sequence processing and analysis

Sequence data was processed and analyzed using Quantitative Insights into Microbial Ecology (QIIME) [23] according to the Illumina demultiplexing and processing protocol [25] and current quality-filtering recommendations [20], using the 1.8.0 pipeline and the default parameters in split_libraries_fastq.py. After quality control and demultiplexing, we picked close references at 97% similarity against

the 97% similarity Greengenes database [160] version 13_8. All further analyses were performed at a rarefied depth of 5,000 reads/sample. EMPeror [236] was used for data visualization of BIOM-format [159] Operational Taxonomic Unit (OTU) tables. OTU analyses were performed by clustering at the 97% level with UCLUST [46], and data were integrated with the HMP dataset according to the protocols used for similar previous meta-analyses [120, 55]. Sequences were analyzed by using both weighted and unweighted UniFrac [139], followed by principal coordinate analysis [23]. Data were visualized using Phinch. The Phinch program provides an easy-to-use, browser-based, platform to visualize contingency tables along with their sample metadata (Bik et al., manuscript in preparation[1]).

Analysis of microbiome stability and centrality

For each set of post-transplant patient samples we assessed the similarity of that set to the set of reference samples from the donor (2,000 reads/sample). To reduce noise and compare patient samples along only relevant dimensions in UniFrac distance space, we applied Principal Coordinates Analysis (PCoA) to the unweighted UniFrac distance matrix containing only the post-transplant and donor samples for that donor-patient pair, then recalculated the distances using only the first n principal coordinates axes required to explain at least 80% of the variation in the distance matrix. An 80% cutoff was chosen to balance bias and overfitting. Distances were recalculated using Euclidean distances between points in PCoA

---

[1]https://github.com/PitchInteractiveInc/Phinch

space in order to convert PCoA coordinates to a distance matrix. The empirical p-values for the 'normality' were obtained by comparing the mean distance between patient and donor samples to the histogram of within-donor distances (generated using all samples from a given donor by enumerating the pairwise distance between those samples). The empirical p-values for the 'dynamic range' (stability) were obtained by comparing the mean distance within patient samples to the histogram of within-donor distances. These analyses were also performed using alternative parameters including, weighted UniFrac, Jensen-Shannon, root Jensen-Shannon, and Bray-Curtis.

5.1.3 Results

Bacterial composition of fecal samples from patients with R-CDI becomes healthy and donor-like following FMT

Four patients (CD1-CD4) with R-CDI were treated with FMT using material obtained from a single donor but from different time points, and fecal samples were collected from these patients before and after the procedure as well as from the donor at times of donation. Bacterial communities from these fecal samples were characterized by sequencing the V4 region of the 16S rRNA gene. Following trimming and quality filtering from a total of 12,536,492 sequences, we randomly subsampled to 5,000 sequences/sample in order to normalize read depth across all samples. All further analyses were performed using this rarefied read depth.

To better understand changes in bacterial communities following FMT, we compared the bacterial composition of patient fecal samples to those of microbial communities from various body sites from the 252 healthy individuals characterized in the HMP [227] (Figure 5.1) using unweighted UniFrac [139] followed by PCoA [23] (see Movie Supplement[2]). The composition of pre-FMT fecal samples from patients CD1-CD4 and 10 additional patients with R-CDI was distinct from both fecal samples from healthy individuals and microbial communities at other body sites, including mouth, vagina, and skin, demonstrating severe alterations in pre-FMT communities compared to healthy fecal communities as has been previously shown [76, 233]. In contrast, microbial communities from the donor fell within the range of healthy fecal samples. Using an animated visualization of FMT-associated changes in patients' fecal microbial communities, we observed rapid and dramatic shifts after FMT towards the communities found in the feces of healthy individuals and of the original donor (see Movie Supplement[3]).

Fecal microbial communities remain dynamic following FMT

To more closely examine temporal changes in recipient fecal samples following FMT, we analyzed fecal microbial communities from patients CD1-CD4 and donor, as well as from 10 additional donor samples, using weighted and unweighted UniFrac [139] followed by PCoA [23]. This analysis demonstrated that fecal bac-

---

[2]https://static-content.springer.com/esm/art%3A10.1186%2Fs40168-015-0070-0/MediaObjects/40168_2015_70_MOESM1_ESM.mp4

[3]https://static-content.springer.com/esm/art%3A10.1186%2Fs40168-015-0070-0/MediaObjects/40168_2015_70_MOESM1_ESM.mp4
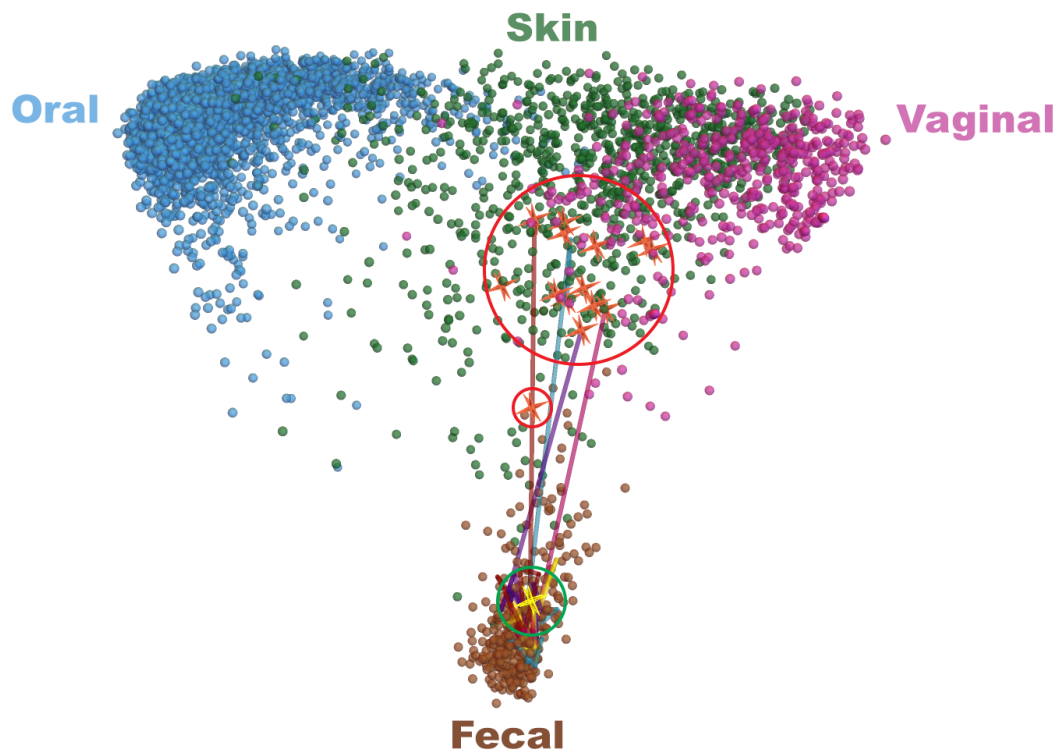
Figure 5.1: Fecal bacterial communities of recurrent CDI patients shift towards HMP fecal bacterial communities after FMT. Red circles = pre-FMT patient samples; green circle = post-FMT patient samples; blue line = trajectory of patient fecal communities after FMT.

terial communities continued to undergo compositional fluctuation following FMT (Figure 5.2A and S1; individuals OTUs listed in Table S1).

To determine whether this dynamic range of post-FMT microbial composition fits within the range seen across healthy individuals, we also compared communities in our samples to those in the HMP via weighted UniFrac and PCoA (Figure 5.2B). Again, fecal microbial communities prior to FMT were highly distinct from healthy fecal microbial communities, and following the procedure these communities more closely resembled those of healthy individuals. Similar to the comparison with donor communities above, fecal microbial communities of R-CDI patients following FMT shifted within the cluster of communities from healthy individuals.

Rapid and substantial changes to *Enterobacteriales* in feces following FMT

While overall fecal microbial communities were dramatically altered following FMT, we also examined the effects of the procedure on the abundance and dynamics of individual bacterial taxa within the four original CDI patients. As shown previously [59, 76, 105, 204, 206, 233, 245], the relative abundance of bacterial phyla in patient fecal samples shifted substantially following FMT, with relative decreases in *Proteobacteria* and relative increases in *Bacteroidetes* and *Firmicutes* (Figure 5.3). These *Proteobacteria* are primarily the order *Enterobacteriales*, which were also substantially decreased in relative abundance following
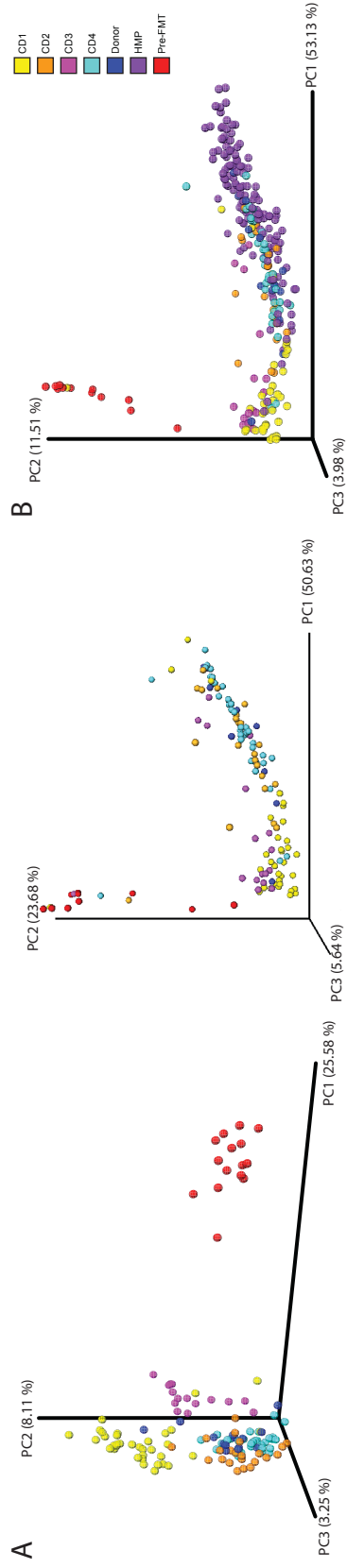
Figure 5.2: Microbial communities shift following FMT. A) Unweighted (left) and weighted (right) UniFrac analyses followed by principal component analysis of bacterial communities of recurrent CDI patient fecal samples before (red) and after FMT and donor samples (blue). B) Weighted UniFrac analysis followed by principal component analysis of bacterial communities of patients before (red) and after FMT versus HMP fecal communities (purple). PC: principal component. Percentages represent percent variability explained by each principal component. Se key at right for colors associated with samples before FMT (Pre-FMT), from HMP and donor, and from patients after FMT (CD1-CD4).

FMT (Figure 5.4A).

We focused on these changes by examining the relative abundance of *Enterobacteriales* alone in each patient before and after FMT. The relative abundance of this taxon ranged from 44 to 82% in all four patient samples prior to FMT, and rapidly dropped to undetectable levels within one week after the procedure. Moreover, abundance of this taxon remained low at 26 days after FMT, the latest time point shared by all four patients (Figure 5.4A), although other members of the *Proteobacteria* remain detectable if decreased in relative abundance (Figure 5.3). In addition, we generated individual value control charts based on the average abundance of this taxon in R-CDI patients. Compared to relative abundance, these control charts displayed the expected variation of the abundance of *Enterobacteriales* in these fecal samples. In all patients, the abundance of *Enterobacteriales* was above the expected variation (i.e. more than three standard deviations above the mean relative abundance [Standard Upper Control Limit (UCL)] of this order across all samples) prior to FMT, and rapidly fell below the upper control limit within 1-2 days after the procedure (Fig 5.4B). These results suggest that the relative abundance of *Enterobacteriales* significantly decreased in all patients soon after FMT to levels similar to donor samples, and remained within a statistically expected range for the duration of sample collection (up to 151 days post-FMT).
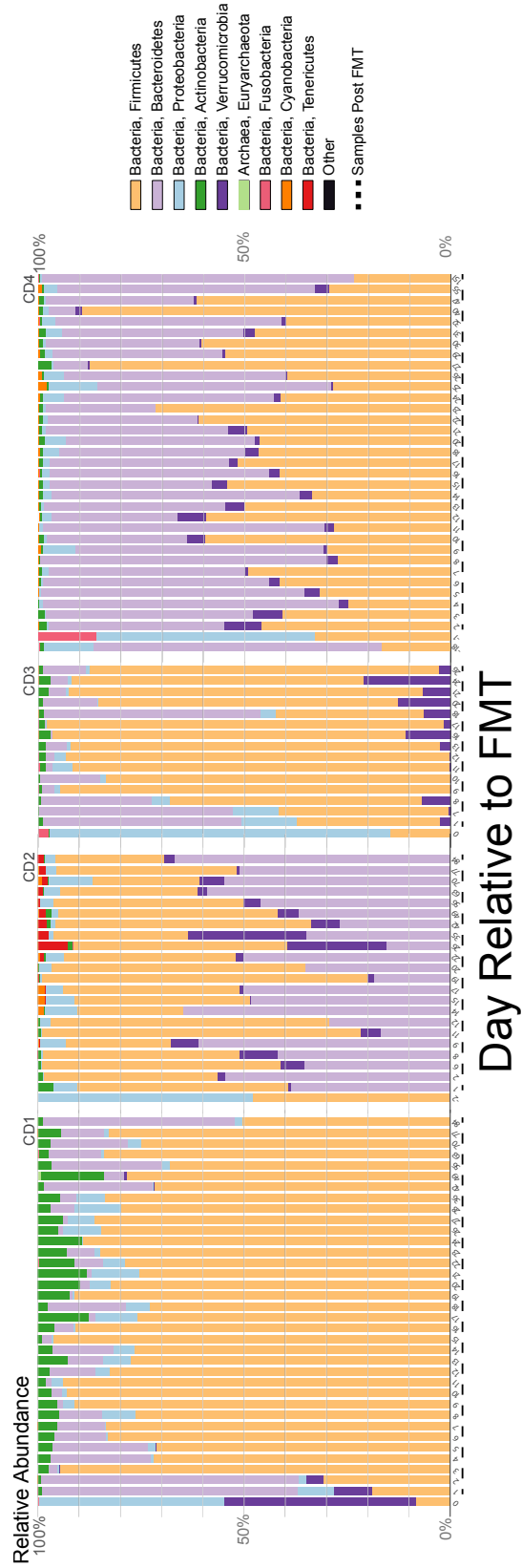
Figure 5.3: Changes in fecal microbial communities after FMT. Relative abundance of sequences classified to the level of bacterial phyla before and after FMT in patient fecal samples. Samples after FMT indicated with dashed line. See key at right.
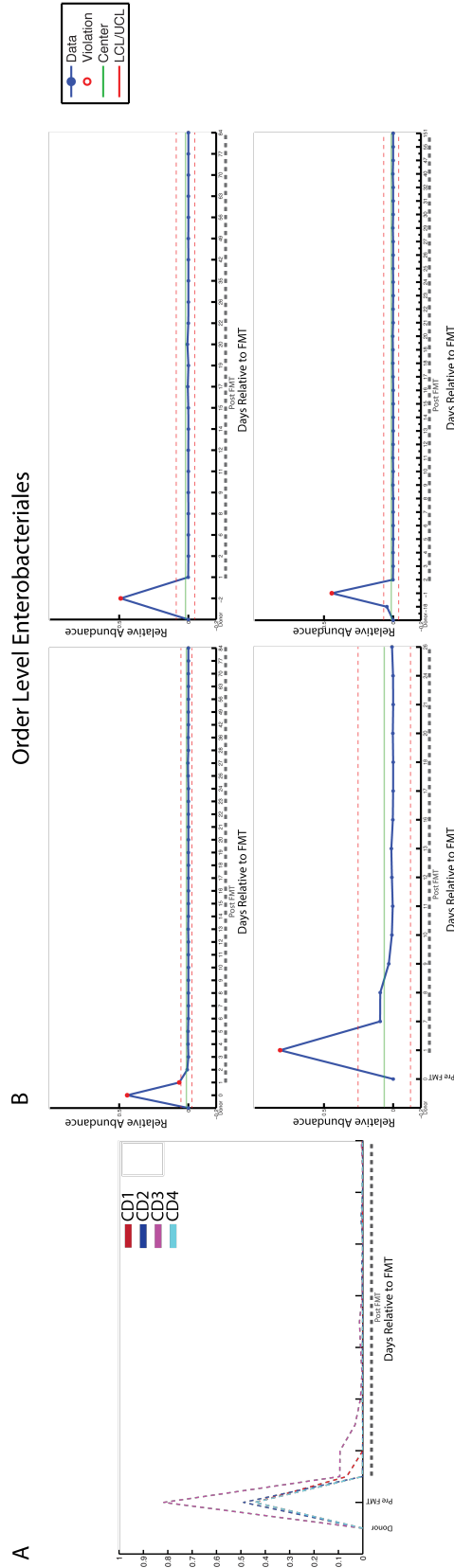
Figure 5.4: Changes in the order *Enterobacteriales* after FMT. A) Relative abundance of *Enterobacteriales* in donor and patient samples before and after FMT in samples common across all patients. B) Control charts of relative abundance of *Enterobacteriales* in donor (leftmost sample) and patient samples before and after FMT. Top left: patient CD1, top right: patient CD2, bottom left: patient CD3, bottom right: patient CD4. LCL: lower control limit; UCL: upper control limit; Center: mean relative abundance in all samples. LCL and UCL represent three standard deviations in relative abundance below and above the mean, respectively. Dashed lines indicate samples after FMT.

Post-FMT communities are initially similar to donor samples but can later diverge

Next, we compared fecal microbial communities within each patient over time to that of the initial donor sample. We generated heat maps based on Pearson correlations between every sample within a given patient set, including respective donor samples and samples from 10 additional pre-FMT patients (Figure 5.5A). This analysis revealed that while microbiota in samples from patients after FMT quickly became similar to microbiota in donor samples, the similarity of samples taken at later time points after FMT fluctuated.

To further investigate how fecal microbial communities in these patients correlate to donor communities, we examined Pearson and Spearman correlations between donor and patient samples, which were common to each patient (pre-FMT samples and those up to 26 days post-FMT; Figure 5B and C and Figure S2). While fecal microbial communities from patients before FMT were highly distinct from those in the donor, fecal microbial communities from samples one day after the procedure were highly correlated to donor communities via both Pearson and Spearman analyses in all patients. After the initial time point after FMT, the Pearson correlation values of patient to donor samples were highly variable within and across patients, although Spearman correlations remained high for three patients. To examine whether this variation is similar in healthy individuals, we determined Pearson and Spearman correlations within the four donor samples used
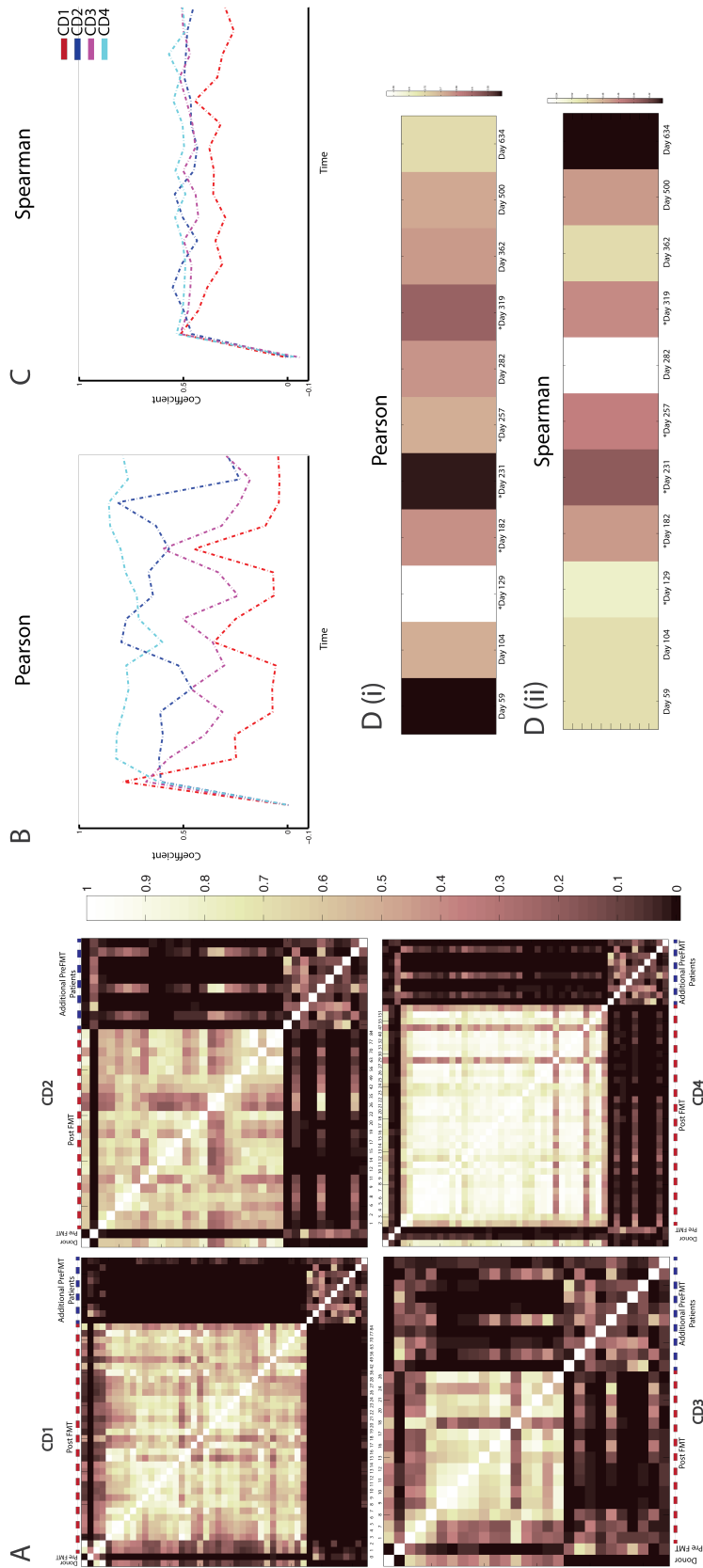
Figure 5.5: Pearson and Spearman correlations between fecal communities before and after FMT. A) Heat map of Pearson correlation values between each sample within each patient set, corresponding donor, and 10 additional pre-FMT patient samples (far right). B) Pearson correlation values between donor sample and each patient sample. C) Spearman correlations between donor sample and each patient sample. D) Heat maps of Pearson (i) and Spearman (ii) correlation values between earliest donor sample and eleven subsequent samples; days represent collection time of each sample versus earliest donor sample. CD1-4: patients 1-4. Dashed lines indicate samples after FMT.

in FMT, as well as eight additional donor samples from the same individual as a control. Results of this analysis revealed that donor microbiota also changed over time (Figure 5D). These findings suggested that the level of variability seen across patient post-FMT fecal microbial communities was within the range of normal microbiota behavior in a healthy individual.

Normalization and dynamic range of post-FMT patient fecal microbial communities are similar to donor communities

Because of the observed variability in later post-FMT patient fecal communities relative to single donor communities, we compared the communities of these patient samples to an expanded set of 17 samples taken from the same donor. We generated two metrics to evaluate the relationships between these communities: normalization and dynamic range (stability). Normalization refers to the mean between-sample distance for each set of patient samples versus the set of donor samples, while dynamic range is the mean distance between each sample within a single patient set. Effectively, the normality of a post-FMT patient sample set is a measure of how similar it is to the donor (healthy) sample set, while dynamic range is a measure of variability within a given patient sample set. We found that neither the normalization nor the dynamic range of any post-FMT patient sample set was significantly different than the donor set following analysis using unweighted UniFrac (Table 5.1). This suggested that although fecal microbial communities of patients post-FMT do not remain identical to the donor,

they nonetheless fall within expected parameters relative to the healthy donor. Similar results were obtained when these analyses were repeated with other parameters, including weighted UniFrac, Jensen-Shannon, and root Jensen-Shannon, and Bray-Curtis (data not shown).

Table 5.1: P-values of normalization and dynamic range of patient samples sets versus donor set.

Patient	CD1	CD2	CD3	CD4
Normalization	0.154	0.429	0.165	0.484
Dynamic Range	0.484	0.429	0.308	0.473

5.1.4   Discussion

It is now well understood that the fecal microbiota change substantially following FMT, typically shifting to fecal microbial communities more similar to those of the donor after transplant [59, 76, 105, 204, 206, 233, 245]. Here we show that these communities shift away from a dysbiotic state towards a composition that is representative of fecal microbial communities from hundreds of healthy individuals, collected in the HMP [227]. Similarly to previous studies [59, 76, 204, 233, 245], the dysbiotic state in these patients with multiply R-CDI is characterized by a large expansion of *Proteobacteria* (primarily members of the order Enterobacteriales, which contains the family *Enterobacteriaceae*), and FMT is associated with reemergence of dominance by members of the *Bacteroidetes* and *Firmicutes* phyla.

Analysis of multiple donor and post-FMT samples demonstrates the dynamic behavior of fecal microbial communities over time. Both donor and recipient samples are characterized by highly dynamic shifts that nonetheless remain within the compositional range of normal fecal microbiota. This observation is consistent with known rapid responsiveness of the fecal microbiome to environmental inputs, such as dietary variations [37], and drifts in microbiota composition over time in healthy individuals [50].

The dynamic nature of intestinal microbiota is an intrinsic property, which should be taken into account when considering how therapeutic interventions, including FMT, impact its composition over time. In long-term post-FMT follow-up, Song and colleagues also noted dynamic changes in the fecal microbiome of R-CDI patients up to 16 weeks post-FMT [215]. These investigators concluded that the fecal microbiome of post-FMT patients did not fully recover over this time, despite clinical recovery. Indeed, we observed divergence of microbiome in some of the patients away from the original implanted material over time. However, analysis of multiple donor samples showed that this movement is within the same dynamic range observed in the donor's fecal microbiome. We therefore conclude that the dynamic behavior of microbiota needs to be taken into account in making comparisons between individuals, and should become an integral part of analysis of the success of FMT.

Three of the recipients in this study received freshly prepared microbiota, while one received frozen/thawed preparation. Use of frozen microbiota prepa-

rations is increasing in clinical practice [261], and its equivalency has not been rigorously established in randomized clinical trials. The ability to store microbiota allows the most up-to-date testing of the donor and fecal material for infectious pathogens, as some of the current tests may take several weeks to complete. Therefore, ability to preserve donor microbiota long-term is critical for its development as a therapeutic agent in clinical practice. Our results here, although limited in the number of patients, demonstrate indistinguishable behavior of fresh and frozen/thawed microbiota preparation.

The patients in this study did not have any significant gastrointestinal co-morbidities. However, a significant proportion of patients with R-CDI have underlying inflammatory bowel disease, take potent immunosuppressive medications, or have multiple other medical problems [75, 24]. The importance of these host factors in contributing to microbiota behavior is currently unknown, but is a subject of great interest [61]. Understanding these influences will require analysis of multiple samples. Recently, Fuentes and colleagues [59] reported that some specific microbial groups and interactive networks are likely to be very important for the maintenance of microbiota in healthy individuals. However, although there is a great deal of effort focused on discovery of compositional differences in microbiota between normal subjects and individuals with different gastrointestinal and medical conditions, the dynamic behavior of fecal microbiota constitutes another dimension that may distinguish these cases. Thus, predictors of stable or dysbiotic intestinal microflora may also change over time. Further detailed studies of dy-

namic behavior of post-FMT microbiota may improve our understanding of causal connections between microbial communities and different disease states.

Acknowledgments

5.2 Changes in Microbial Ecology after Fecal Microbiota Transplantation for recurrent *C. difficile* Infection Affected by Underlying Inflammatory Bowel Disease

Background

Gut microbiota play a key role in maintaining homeostasis in the human gut. Alterations in the gut microbial ecosystem predispose to CDI, and gut inflammatory disorders such as IBD. FMT from a healthy donor can restore gut microbial diversity and pathogen colonization resistance; consequently it is now being investigated for its ability to improve inflammatory gut conditions such as IBD. In this study, we investigated changes in gut microbiota following FMT in 38 patients with CDI with or without underlying IBD.

Results

There was a significant change in gut microbial composition towards the donor microbiota, and an overall increase in microbial diversity consistent with previous studies after FMT. FMT was successful in treating CDI using a diverse set of donors, and varying degrees of donor stool engraftment suggesting that donor type and degree of engraftment are not drivers of a successful FMT treatment of CDI. However, patients with underlying IBD experienced an increased number of

CDI relapses (during a 24-month follow-up), and a decreased growth of new taxa, as compared to the subjects without IBD, note that the test used presents some limitations in sample size and statistical assumptions (see methods). Moreover, the need for IBD therapy did not change following FMT. These results underscore the importance of the existing gut microbial landscape as a decisive factor to successfully treat CDI, and potentially for improvement of the underlying pathophysiology in IBD.

Conclusions

FMT leads to a significant change in microbial diversity in patients with recurrent CDI and complete resolution of symptoms. Stool donor type (related or unrelated) and degree of engraftment are not key for successful treatment of CDI by FMT. However, CDI patients with IBD have higher proportion of the original community after FMT and lack of improvement of their IBD symptoms and increased episodes of CDI on long-term follow-up.

5.2.1 Background

Gut microbiota play a key role in maintaining homeostatic host functions and deleterious shifts in the gut microbial ecosystem, often referred to as dysbiosis, are associated with CDI, IBD and other systemic inflammatory conditions [103]. A diverse gut microbial community confers colonization resistance against pathogens such as *C. difficile*, and disruption of a diverse community structure from antibi-

otics, comorbidities, altered gastrointestinal transit or other risk factors can lead to pathogen colonization and infection [101].

With increasing incidence of community and hospital acquired CDI, high rates of recurrent CDI (estimated 20-30% after a first and 50-60% after a third infection), high mortality (29,000 deaths annually) in the United States, and an urgent need for newer non-antibiotic therapies has led to the emergence of microbiome based therapies [130]. FMT in CDI patients restores phylogenetic diversity to levels more typical of a healthy person, with response rates >85% by enema, oral capsule or endoscopic delivery modes [96, 102, 175]. A recent study suggests significantly lower response of CDI to FMT in patients with underlying IBD [106]. We have also previously described a higher rate of recurrence of CDI following FMT in patients with CDI and underlying IBD [104]. It remains unclear if changes in gut microbial ecology play a role in long-term success of FMT in these patients.

FMT has not shown consistent success in treating other diseases associated with microbial dysbiosis such as IBD. Three clinical trials to treat Ulcerative Colitis (UC) with FMT have shown conflicting results and one highlighted the potential role of specific gut microbial members in donor stool in determining success after FMT in UCs [165, 179, 193]. The underlying host or donor factors that may be important for success of FMT in treatment of IBD remain unclear.

In this study, we assessed the effect of donor type (standard donor versus related donor) and changes in gut microbial ecology on response to FMT in R-CDI

with and without underlying IBD as well as clinical response to FMT.

5.2.2   Methods

Patient recruitment, sample collection and clinical analysis can be found in Appendix C.2.

Alpha diversity values were calculated using Faith's phylogenetic diversity [49]. To assess differential abundance between the groups, we used ANCOM [146], as implemented in scikit-bio 0.5.1[4]. This is tested by looking at the individual OTUs across the patient types (with and without underlying IBD); OTUs of the same genus are grouped for displaying purposes. We note that ANCOM makes the statistical assumption that fewer than 25% of taxa change, not met in all these comparisons (before FMT and post FMT communities are expected to be very different [244]).

The donor-plane is created using all the donor samples, and serves as a proxy for where their microbiomes are in the ordination space, and how as time goes by this proximity changes. This procedure was originally presented by Halfvarson et al. [72].

Beta diversity matrices were created using unweighted UniFrac [139], and plotted using Emperor [236] (all other plotting was done using the Seaborn visualization package).

---

[4]http://scikit-bio.org/docs/0.5.1/

Processed tables and sample information can be found in Qiita[5] under study id 10057, alternatively the data can be found under accession number ERP021216 at the European Bioinformatics Institute.

SourceTracker Analysis

To assess the proportion of pre-transplant communities that were retained in the patients' microbiota, we used SourceTracker [112]. The pre-transplant samples and the donor samples were described as *sources*; all the other samples were used as *sinks*. For all samples at day seven and twenty-eight, SourceTracker estimated the proportion of communities that were attributed one of three environments, (1) the donor, (2) the patient pre-transplant, and (3) and unknown community. Using these proportions, we grouped the samples according to their IBD status and compared their distributions using the Mann-Whitney test (as implemented in SciPy 0.15.1[160]).

5.2.3 Results

FMT leads to resolution of CDI

In order to assess gut microbiota changes following FMT, 38 patients with recurrent CDI were enrolled in the study and a fecal sample was obtained prior to transplant, as well as 7 and 28 days post-transplant. Sample handling, donor and recipient sample collection, sample processing and data analyses are detailed

---

[5]https://qiita.ucsd.edu

in supplementary methods. FMT was accomplished by colonoscopy using fresh donor stools from related (n=12) or unrelated (n=26) donors. None of the IBD patients received stool from a related donor. The demographic, disease and treatment characteristics are outlined in Table 5.2. Detailed characteristics of IBD patients are shown in supplementary table 1. Twelve patients (31.6%) had IBD (6 with UCs and 6 with Crohn's disease), with median age 27.6 years (range, 23.3-74.9), and median IBD duration 5 years (range, 2-33). 58.3% percent of patients were on 5-ASA (amino salicylic acid) agents, 50% on biologics and 33.3% on immunomodulators and 58.3% on steroids. Among patients with IBD, at the time of colonoscopy, 2 had normal colonoscopy, 1 had pseudopolyps, 5 had severe pancolitis, 1 had moderate colitis, 1 had mild colitis, 1 had mild procto-sigmoiditis and 1 had moderate ileo-colitis (Supplementary Table 1).

All patients responded to FMT with regards to clinical or microbiologic remission of CDI (negative *C. difficile* testing), 92.1% (n=35) of patient symptoms returned to baseline bowel pattern (as before CDI) and resolution of CDI, 5.3% (n=2, both with IBD) had worsening diarrhea (*C. difficile* negative), and 2.6% (n=1) had new onset constipation after FMT. Upon long-term follow-up of 24 months; 13.2% (n=5/38; of these n=1 within 56 days, n=1 from 56 days to 1 year and n=3 beyond 1 year, Supplementary Table 2) had another episode of CDI and 10.5% (n=4/38) required a second FMT due to multiply R-CDI. One patient with R-CDI was treated with vancomycin. The risk of another episode of CDI after FMT in IBD patients was 25% (n=3/12) compared to 7.7% (n=2/26) in non-IBD

Table 5.2: Clinical Characteristics

		Overall (n=38)	IBD (n=12)	No IBD (n=26)
Age	median	53.1	27.6	58.3
	(range)	(21.9-82.7)	(23.3-74.9)	(21.9-82.7)
Sex distribution (% female)		81.6	66.7	88.5
BMI, kg/m2	median	24.8	25.6	23.8
	(range)	(14.9-39.9)	(18.5-30.3)	(14.9-39.9)
Number prior CDI episodes	median	5	4.5	5
	(range)	(3-13)	(3-7)	(3-13)
Number prior metronidazole courses	median	1	1	1
	(range)	(0-8)	(0-2)	(0-8)
Number prior vancomycin 10-14 day courses	median	2	2	2
	(range)	(0-4)	(0-4)	(0-3)
Number prior vancomycin tapers	median	1	1	1
	(range)	(0-5)	(0-1)	(0-5)
Number prior fidaxomicin courses	median	0	0	0
	(range)	(0-4)	(0-2)	(0-4)
Recurrent CDI after FMT (%)		13.2	25	8.4

patients (p=0.16, chi-square test). Seven of the 12 patients with IBD were on systemic immunosuppression. None of the patients with IBD had improvement in their IBD course after FMT, and none were able to withhold, de-escalate or stop IBD treatment. This is not an unexpected finding as one time FMT would not be expected to alter the disease course in IBD patients.

FMT decreases microbial dysbiosis

FMT led to a significant increase in alpha diversity based on Faith's phylogenetic diversity, Shannon's diversity index and observed species, both at day 7 and day 28 (Mann-Whitney $p<0.05$; Supplementary Figure 1, comparing pre- and post-FMT in patients with CDI with or without underlying IBD). Also, patient's stool closely resembled donor stool, as evidenced by a rapid and sustained change in unweighted and weighted UniFrac-based beta diversity following FMT at day 7 and 28 post-transplant (Figure 5.6A; PERMANOVA $p<0.05$) [139].

To characterize the changes in community composition, we use the Microbial Dysbiosis (MD) index as a reference to describe the dominance of individual taxa (Supplementary Table 3). The MDs index is composed of 18 taxonomic groups, as defined by Gevers et al, with a higher value correlated with greater disease severity in IBD, and lower values associated with healthier states [61]. As CDI is also associated with dysbiosis and inflammation, we wanted to determine the effect of FMT on dysbiosis. The MDs index values were significantly higher in patients with CDI compared to donors (Mann-Whittney's U, $p < 0.05$, Fig-
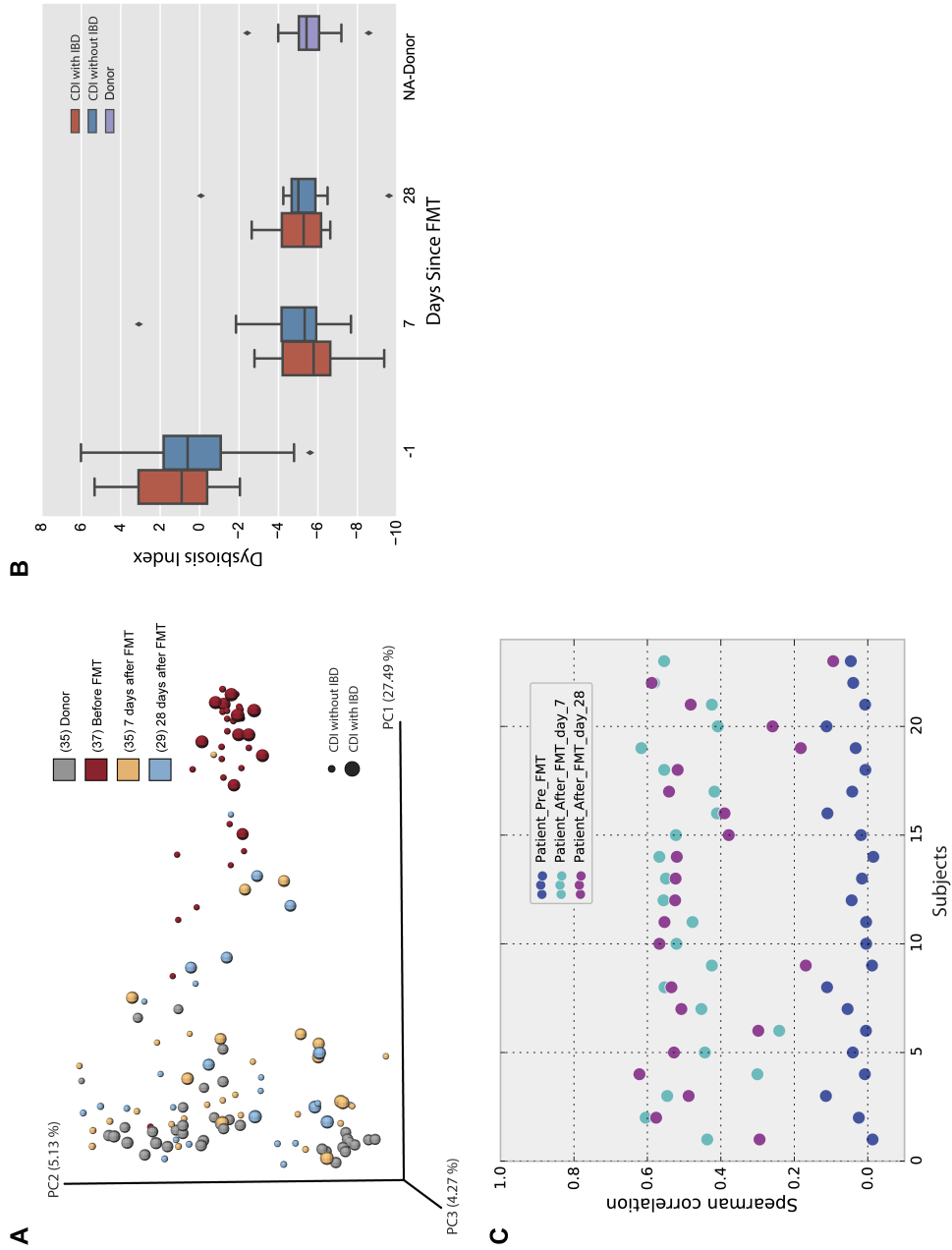
Figure 5.6: (A) Principal Coordinates Analysis of the unweighted UniFrac distances, showing change in the phylogenetic diversity between patients with CDI, 7 and 28 days after fecal microbiota transplant. (B) Change in dysbiosis index following fecal microbiota transplant in patients with CDI with or without IBD. (C) Spearman correlation to donor stool 7 and 28 days following fecal microbiota transplantation.

ure 5.6B). However, on day 7 and 28 after the transplantation, the MDs index values were similar to donors (Mann-Whitney's U p > 0.05, Figure 5.6B) and this change was independent of whether recipients had IBD or not.

In order to determine if the changes seen in our subjects following FMT were similar to other published studies we compared our samples with recently published data from Weingarden et al. 2015 (Supplementary Figure 2A) wherein 4 patients with R-CDI (but not IBD) received FMT from a single donor [244]. Similar to our findings, there was a rapid and sustained change in beta diversity (Supplementary Figure 2A) following FMT and the regression to the donor plane (change in microbial composition to resemble healthy donors) following FMT was remarkably similar in the two studies (Supplementary Figure 2B). In this context, we refer to the donor plane as a proxy to the region in thePCoA: a dimensionality reduction method to visualize beta-diversity distance matrices) space where the donors are located; we do this by fitting a three-dimensional plane (using the least squares method) to the samples from the donors. As the communities change post-FMT, the distance to this plane is reduced.

Clinical response of CDI to FMT is independent of engraftment or donor type but underlying IBD influences changes in gut microbial ecology after FMT

In order to determine if the response of CDI to FMT was dependent on donor stool engraftment, we determined Spearman's correlation coefficient between

fecal microbial communities prior to and 7 and 28 days post-transplant. The fecal microbial communities from patients with CDI were distinct from donor communities prior to transplant (Spearman's r<0.2 for all subjects, Figure 5.6C). Following transplant the communities showed an increase in correlation to donor stool at day 7 (Spearman's r>0.4 for 85% of the subjects, Figure 5.6C) and a spread for all subjects at day 28 ranging from below 0.2 up to 0.6 (Figure 5.6C). Using SourceTracker [112], we found that after FMT, subjects with IBD retained a higher proportion of their original communities (Mann-Whitney $p < 0.05$ at day 7, and $p = 0.06$ at day 28; Figure 5.7A and 5.7B) and a significantly lower proportion of new communities (Mann-Whitney $p < 0.05$ at day 7 and 28), as compared to the patients without IBD. The expansion of new taxa following FMT represents a beneficial ecological change following FMT as seen in patients without IBD, while those with IBD are more prone to revert to the original community structure. Consequently, in patients with IBD we observed a smaller group of taxa that change significantly seven days after FMT. In both groups, *Bacteroides*, and *Faecalibacterium* showed a significant increase in relative abundance, with *Blautia*, only being increased for patients without IBD. Additionally, these patients showed a decrease in relative abundance of *Lactobacillus*, *Veillonella*, *Enterobacter*, *Klebsiella*, *Erwina*, *Proteus*, *Salmonella*, and *Trabulsiella* (Figure 5.7C and 5.7D, ANCOM $p < 0.05$, corrected for multiple comparisons using Bonferroni-Holm's method [146]).

All patients had either clinical or microbiological remission, confirming that initial response of CDI to FMT is not dependent on the degree of donor stool
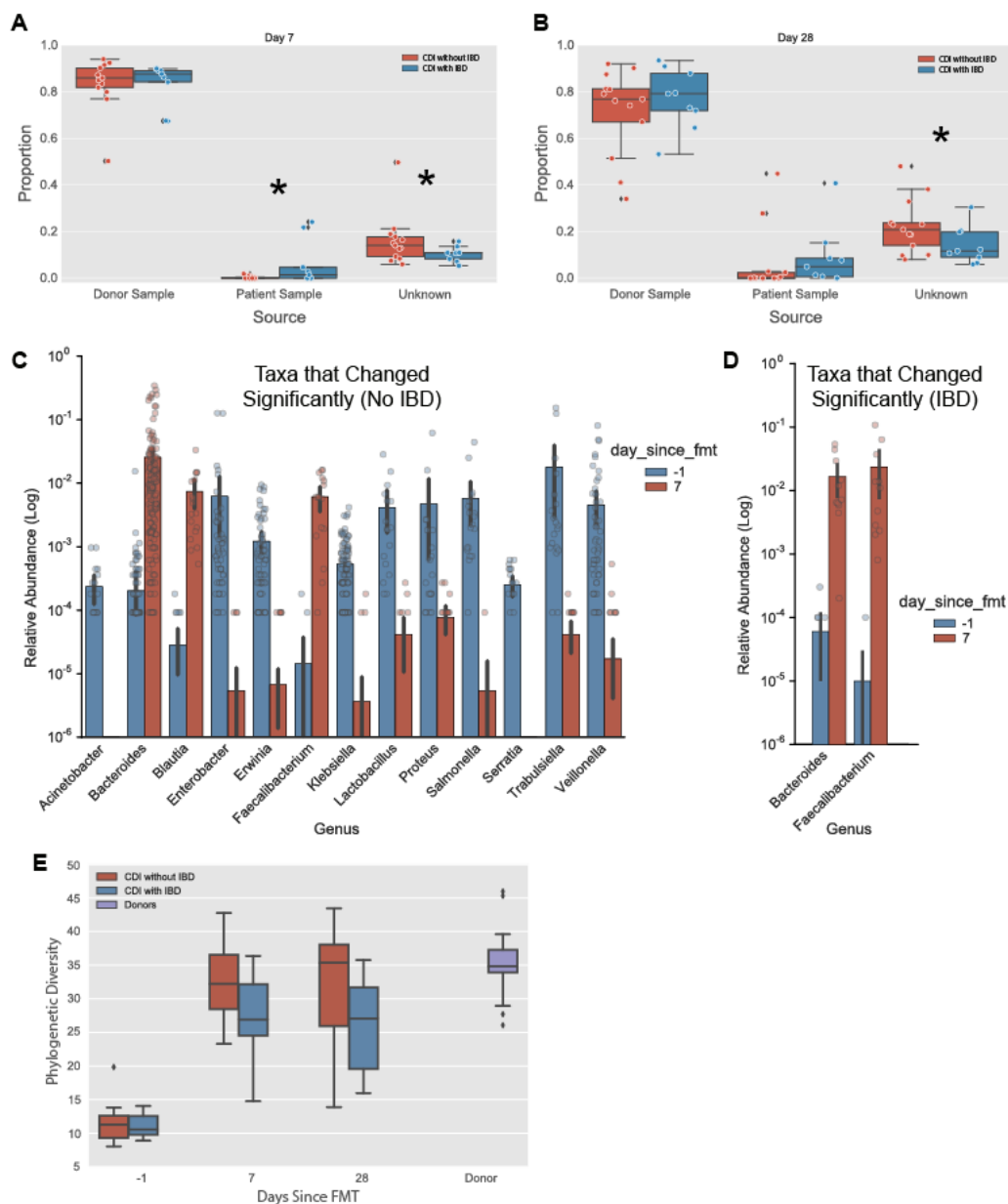
Figure 5.7: (A) and (B) Subjects with IBD retain a higher proportion of their original communities (Mann-Whitney $p < 0.05$ at day 7, and $p = 0.06$ at day 28 and a significantly lower proportion of new communities (Mann-Whitney $p < 0.05$ at day 7 and 28), as compared to the patients without IBD using SourceTracker. (C) Bacterial taxa that change significantly in patients with IBD after FMT (ANCOM $p < 0.05$, corrected for multiple comparisons using Bonferroni-Holm's method). (D) Bacterial taxa that change significantly in patients without IBD after FMT (ANCOM $p < 0.05$, corrected for multiple comparisons using Bonferroni-Holm's method). (E) Change in phylogenetic diversity based alpha diversity 7 and 28 days following fecal microbiota transplant in patients with CDI with and without IBD (Mann-Whitney's U $p < 0.001$).

engraftment. In this small cohort of patients, those with underlying IBD had higher number of late relapses of CDI. We found no significant differences in gut microbiota composition following FMT from standard donors or related donors (Mann-Whitney p> 0.05 at day 7 and 28), suggesting that engraftment of donor stool was independent of donor type. Furthermore as all patients had ongoing clinical remission with microbiological response (if measured), donor type does not appear to affect CDI related clinical response.

Change in bacterial diversity after FMT is dependent on underlying IBD.

IBD disease course, as measured by the need for specific IBD therapies, did not change after FMT, and patients with CDI and underlying IBD retained a higher proportion of the pre-transplant communities and lower proportion of new communities following FMT. Thus, underlying IBD appears to affect the change in gut microbial ecology resulting in a less significant increase in overall diversity. In subjects without IBD, Faith's phylogenetic diversity (which measures the total branch length of a phylogenetic tree that a given sample covers [48]) reached a level comparable to healthy donors (Mann-Whitney's U p ¡ 0.001, Figure 2E). The differences in phylogenetic diversity following FMT between subjects with and without IBD became evident on day 7 and persisted on day 28 (Mann-Whitney, day -1 p = 0.163, day 7 p = 0.0058, and day 27 p = 0.008, Figure 2E). A linear regression of phylogenetic diversity vs MDs index (Supplementary Figure 3) shows a

significantly lower negative correlation between the increase in phylogenetic diversity and the increase of the MDs index in patients with IBD (Pearson's correlation coefficient, IBD R=-0.68, No IBD R=-0.83; p < 0.0001; Supplementary Figure 3) suggesting a lack of recovery of phylogenetic diversity in patients with IBD as the MDs index improves.

5.2.4 Discussion

In this study, we found that gut microbiota diversity changes rapidly following FMT for treatment of CDI and resembles donor microbiota diversity, similar to previous studies. A successful response of CDI to FMT was seen with a diverse group of donors and at levels of engraftment (as measured by correlation to donor stool) varying from 50-94% (at day 7) and 34-93% (at day 28) based on the proportion of communities attributed to the donor following FMT per SourceTracker, suggesting these are not critical factors in determining response. Similarly, a recent study that evaluated pre- and post-FMT (for recurrent CDI) gut microbiome samples from a subset of patients enrolled in a randomized controlled trial [97], compared donor FMT to autologous FMT suggested that complete engraftment of donor bacteria may be not necessary, if functionally critical taxa are present in subjects following initial antibiotic therapy for CDI [218]. This study excluded patients with IBD but was able to compare autologous to donor FMT unlike our study. There was a higher number of R-CDI following FMT in patients with CDI and IBD but this was not statistically significant, likely given the small sample size.

However we have previously reported similar findings in a larger cohort of patients with CDI and IBD [104], where gut microbiota changes were not monitored. Interestingly, in this cohort all patients had an initial clinical or microbiological remission of CDI following FMT and we did not see a difference in initial response reported in a recent study [106], which is also likely due to the smaller sample size of our study and differences in underlying disease characteristics.

We also did not see changes in need for IBD therapy in the subset of patients with IBD underlying CDI. While dynamic variations can be seen in patients following FMT [244], patients with underlying IBD in our study show a higher proportion of the original pre-transplant microbial community and lower recovery of phylogenetic diversity following FMT compared to those without IBD. This lack of beneficial change in microbial ecology may be relevant for long term response of CDI in patients with IBD and the lack of clinical response of IBD to FMT seen in our and previous studies [106]. Future studies designed to study the effect of compositional and functional changes in gut microbiota on clinical outcomes following FMT in patients with IBD will be needed to definitively address the potential importance of changes in microbial ecology, donor selection [165], underlying disease characteristics and multiple-dose FMTs, in correcting the underlying pathophysiology of IBD.

5.2.5 Conclusions

There is a significant increase in microbial diversity in patients with recurrent CDI after FMT. Both, the degree of microbial engraftment or donor type (related or unrelated) are not key for successful treatment of R-CDI by FMT. Compared to CDI patients without IBD, CDI patients with IBD have higher proportion of the original microbial communities after FMT and increased episodes of future CDI on long-term follow-up.

5.2.6 Acknowledgments

This section, in full, is a reprint of the material as it appears in "Changes in microbial ecology after fecal microbiota transplantation for recurrent C. difficile infection affected by underlying inflammatory bowel disease". S. Khanna, Y. Vázquez-Baeza, A. González, S. Weiss, B. Schmidt, D. A. Muiz-Pedrogo, J. F. Rainey 3rd, P. Kammer, H. Nelson, M. Sadowsky, A. Khoruts, S. L. Farrugia, R. Knight, D. S. Pardi, P. C. Kashyap. *Microbiome.* 5, 2017. The dissertation author was the co-primary investigator and author of this paper.

Chapter 6

Future Work

This thesis presents a collection of static and dynamic snapshots of diverse microbial systems. Although our new methods are in no way specific to the human gut, part of our focus has been to bridge the gap between computation and biology, to get us closer towards a future where drug prescription, diet and recreational activities can be further *personalized* – through a layer invisible to the naked eye but perceivable in most respects of life.

It seems impossible to accurately predict the impact of scientific discoveries. Historically, we have seen great evidence of seemingly anachronistic findings (for example neural networks [225]) that later, usually when other technologies catch up, become new fields of research or cornerstones in consumer applications. Therefore, while acknowledging this is a tough problem, I present my personal view on future directions that can build up from the contributions in this thesis. I divide these into three sections: diagnostic methods, treatment and analysis.

6.1 Diagnostic Methods

In Chapter 3 and Chapter 4, we presented examples of biomarkers for IBD and Crohn's Disease (CD) from two different analytical perspectives. First, in a dog cohort, we showed that we can reformulate the dysbiosis index we originally developed for humans [61], and make it specific for a different host-species. Both in humans and dogs, this log-ratio of bacterial groups is associated with decreased phylogenetic diversity, and in humans, it is also associated with increased inflam-

mation. Next, after noticing the increased volatility in the microbiome of subjects with IBD, we benchmarked microbial variability in fecal samples as an effective classifier for disease. By collecting more samples per subject, we can overcome the low classification accuracy of fecal samples.

Although both approaches are encouraging, translating research into consumer-level applications, commonly presents formidable challenges that might only be solvable by future generations. Take for example the Electrocardiograph (ECG), it took almost 125 years, between the first observation of biopotentials[1], and the moment when the first table ECG[2] became commercially available [268]. Even after the tremendous progress, proper validation of computer-generated diagnostics only appeared more than 200 years after the initial discovery [250].

I use the example of the ECG, because much like raw heart biopotentials, (i) microbiome data is plagued with noise, and (ii) determining the appropriate filters and thresholds directly depends on the use-case. However, unlike with the early developments of the ECG, we live in digital and connected world. As such, the following focuses will shorten the time between future discoveries and innovation:

Mechanistic Studies The novelty characteristic of microbiome studies, has produced a large amount of descriptive studies. Contrastingly, mechanistic experiments have lagged in validating several of these findings (likely due to the more expensive requirements). If the goal is to relate the presence or

---

[1]Credited to Luigi Galvani in 1787.

[2]Willem Einthoven's ectrocardiograph was manufactured by the Cambridge Scientific Instrument Company of London in 1911.

abundance of a microbiome feature to a disease state or biochemical process, the underlying methods must be informed by biological inferences in order for these biomarkers to gain credibility.

Open Data Medical research is especially affected by human factors, being able to consistently collect samples from a subject is not always easy or deterministic (think of bowel problems and fecal samples). A powerful practice used to counter underpowered studies is to improve on the current sample size through the reuse of previously published data. This approach is only possible through open resources, like Qiita[3], that make data reuse seamless. Importantly, the work in this thesis was only possible through the reuse of existing datasets (see Chapter 2 through Chapter 5). Although making data openly available is an important first step, proper standardization of processing protocols will also be key to maximizes data re-usability. A remarkable example of this practice is the Earth Microbiome Project (EMP) [63].

6.2   Treatment

Microbiome-based treatments are generally based on the transplantation of microbial communities from a *healthy donor* into an *affected recipient*. Three spearheading examples are: FMTs to treat CDI [233], skin microbiomes transplants to treat atopic dermatitis [169], and (although still in early stages) a capsule full

---

[3]https://qiita.ucsd.edu/

of microbial spores to treat UC by Seres Therapeutics[4].

Although most FMTs succeed at treating CDI, what makes a successful transplant is not clear yet. The number of variables implicated in answering this question is immense. From a computational complexity standpoint, *a priori* determination of whether or not a new community can colonize an existing ecosystem is considered a hard computational problem, belonging to the **#-P** class [88]. As such our focus should be on strengthening our systematic understanding of the transplant, and characterize not only what makes a successful transplant but also what makes a failed one. For example, the impact of the work presented in Chapter 5.2 could have been magnified if we had included subjects for which the FMT failed. With appropriate sample sizes, we could have applied a number of techniques to single out (if any) the common features that lead to a failed transplant. With this knowledge, we could pre-screen *donors* and *recipients* to ensure a successful treatment and avoid unexpected side-effects.

6.3 Analysis

Much as our microbial symbionts depend on the host environment (and vice versa), the development and funding of analytical tools depends on the ever-growing necessity to unravel complex patterns in a number of experimental setups (cross-sectional, longitudinal, multi-'omic, etc). Thus, these tools must take into

---

[4]http://www.serestherapeutics.com

consideration the ability to be flexible, scalable, and when possible interactive.

Novel analytical methods and software infrastructure should be built making scalability a priority. We have seen a steady increase in our capability to generate data, and it is likely that this trend will continue. Emperor (Chapter 2.1) was partly a response to the limitations in existing software, and more recently we had to re-architect the underlying implementation to cope with modern and larger datasets.

Flexibility and compliance with community standards makes software available to a wider audience. Take the count table, often acting as the core data-structure for metabolomic, transcriptomic, proteomic, (and other 'omics) analyses. If the software producing this data complies with a standard, like the Biological Observation Matrix (BIOM)-format, the end user is now free to select from a variety of methods as opposed to being limited to *niche-software*. This idea is being taken a step further with QIIME-2, where a semantic type system defines the methods and visualizations that can be applied to any given dataset (regardless of its biological origin).

Finally, the value of interactively exploring a dataset lies in our ability to quickly test hypotheses and iteratively develop new ones. Future exploratory analysis tools should be developed with interactivity and interoperability in mind. For example, brushing to select a group of samples in a view of data might act as a filter for a different representation. Modern web technologies, and software development frameworks like QIIME-2, will likely be the pioneers of these global

overviews of microbial diversity.

Appendix A

Sample collection and processing for Chapter 3

A.1   Methods

Naturally passed fecal samples were analyzed from 85 healthy dogs and 65 dogs with chronic signs of gastrointestinal disease and confirmed inflammatory changes on histopathology. All dogs participated in different clinical studies and leftover fecal samples were utilized for this study. The protocol for sample collection was approved by the Clinical Research Review Committee of the College of Veterinary Medicine, Texas A&M University (CRRC#09-06).

Dogs with clinical signs of chronic GI disease (i.e., vomiting, diarrhea, anorexia, weight loss, etc.) were diagnosed with idiopathic IBD based on theWorld Small Animal Veterinary Association (WSAVA) criteria: (i) chronic (i.e., > 3

weeks) GI signs; (ii) histopathologic evidence of mucosal inflammation; (iii) inability to document other causes of GI inflammation; (iv) inadequate response to dietary, antibiotic, and anthelmintic therapies, and (v) clinical response to anti-inflammatory or immunosuppressive agents. Histological samples were obtained endoscopically. The clinical status of each dog was evaluated using a published clinical Clinical Canine IBD Activity Index (CIBDAI). Within the IBD dogs, 41 dogs had histological confirmed inflammation in the small intestine, 18 dogs had histological changes in both small intestine and colon, and 5 dogs had only histological changes reported in the colon. Histological changes were predominantly of lymphoplasmacytic infiltrates, with a subset of dogs also showing eosinophilic and/or neutrophilic components. The mean (SD) CIBDAI for IBD dogs was 6.4 (3.1).

Dogs were excluded if they received antibiotics within the past 2 weeks of sample collection. Data on antibiotic history was nevertheless collected: 34/65 dogs with IBD had no history of prior antibiotics administration, while 13 dogs received antibiotics several weeks (>2) or months before sample collection. The remaining 18 dogs in the IBD group had no information about prior antibiotic use. In the healthy group (n=85), 76 dogs had not received any antibiotics, and 9 dogs had a history of antibiotic use, but not within the last 2 weeks of sample collection. No technical replicates were generated in this study.

Sample and animal information (i.e., age, weight, gender, breed, duration of clinical signs, histopathology, antibiotic usage) was obtained from clinical records.

146

Also, if the owner provided the information, the exact diet (trade name and manu-facturer) fed at time of sample collection was recorded in the clinical records, and the dietary macronutrients (protein, fat, and carbohydrate content) were recorded from manufacturer's provided data on the labels.

Body weights ranged from 2.9 to 55 kg (mean 22 kg, SD: 14.9 kg), which was not significantly different from (Mann Whitney test; p=0.087) the healthy dogs (range 0.9 to 50 kg; mean 20.3 kg, SD: 10.7g). Mean age (SD) was 5.4 (3.07) in the IBD group, which was not significantly (Mann Whitney test; p=0.311) different from healthy dogs (4.7, 3.2). There was a wide breed distribution with 37 different breeds in the IBD group and 42 different breeds in the control group. In the IBD group, Yorkshire terrier, German Shepherd dogs and Labrador Retrievers (n=5 each) were most commonly represented.

Body Condition Scores (BCS) were assessed according to the WSAVA cri-teria. BCS is rated in a 9-point scale that visually evaluates a dog's body com-position. This score has been validated against the standard dual-energy X-ray Absorptiometry (DEXA) [157]. For this dataset, the BCS was restricted to a sub-set of the healthy samples, therefore IBD vs. Healthy comparisons could not be made in this case.

A.1.1 DNA Extraction and Sequencing

DNA isolation was performed as described by the EMP Protocol (version 4_13) for 16S rRNA[8]. The full cohort included 192 samples, of which 15 were

removed because those subjects had acute hemorrhagic diarrhea, and had little clinical information available. The remaining 28 samples did not recover enough sequences after quality control including screening for low counts of reads per sample. All samples were sequenced using the Illumina HiSeq platform (2 x 100 nucleotide sequences and an index read).

A.1.2 Accession Numbers

Raw sequences for the dog samples have been deposited to the European Nucleotide Archive (ENA) at the following accession number ERP014919, equivalent processed OTU tables and metadata can be accessed through Qiita[1] under study 833 - 'Dog models of inflammatory bowel disease'.

The data for the human dataset[61] can be found in the ENA at the following accession numbers ERP015241 and ERP015242, equivalent processed OTU tables and metadata can be accessed through Qiita[1] under study identifiers 1939 and 1998 -'The Treatment-Naive Microbiome in New-Onset Crohn's Disease'.

Data for the additional dog study [164] can be found at the Sequence Read Archive (SRA) of the under accession number SRP040310.

---

[1]https://qiita.microbio.me

Appendix B

Sample collection and processing for Chapter 4

B.1 Methods for Chapter 4.1

B.1.1 Cohort Demographics

Patients with CD or UC, the two major forms of IBD, attending the outpatient clinic were consecutively invited to take part. After obtaining written consent, Body Mass Index (BMI) was recorded and patients were asked to provide a fecal sample and to fill in a questionnaire with clinical disease activity, present medication, dietary habits, use of antibiotics and use of Non-steroidal anti-inflammatory drugs (NSAIDs). Disease phenotype was classified according to the Montreal classification [209]. Individuals were then followed prospectively, asked to provide fecal

samples and to fill in the questionnaire every third month for a two-year period. If a patient did not provide a fecal sample at any of the three months periods, a reminder letter was sent. In total 109 patients with IBD (CD; n=49 and UC; n=60) took part. Nine additional individuals with no IBD or any other gastrointestinal conditions were recruited as Healthy Controls (HC) as well as 19 patients with other chronic inflammatory gastrointestinal diseases (4 Lymphocytic Colitis (LC) and 15 Collagenous Colitis (CC)). All 137 individuals were Caucasians and together they provided 683 fecal samples during the two-year period (Supplementary Table 4[1]). The study was approved by the Ethical Committee of the Medical Faculty, Uppsala University (2007/291).

B.1.2   Sample Collection

Fecal samples were self-collected in sterile plastic containers and stored at -80 C until shipping on dry ice and processing.

B.1.3   Fecal calprotectin

To assess the degree of inflammatory activity at the collection of each fecal sample, the concentration of f-calprotectin was assessed by commercially available ELISA, Calprotectin Elisa Buhlmann Laboratories AG, Basel, Switzerland, according to the manufacturer's protocol.

---

[1]https://images.nature.com/original/nature-assets/nmicrobiol/2017/nmicrobiol20174/extref/nmicrobiol20174-s1.pdf

B.1.4 DNA Extraction and Amplification

Genomic DNA was extracted from 0.25 g of fecal material from each sample using the Earth Microbiome DNA extraction protocol [2]. Briefly, DNA was extracted using the 96-well format MoBio Powersoil DNA kit on an EpMotion 5075 robot with vacuum (Eppendorf, Hamburg, Germany). DNA was quantified with the Qubit 2.0 fluorometer (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions.

PCR amplification and library preparation were performed similarly to the protocol described by Caporaso et al. [26]. 515F/806R Illumina primers with unique reverse primer barcodes were used to target the V4 region of the 16S rRNA gene. Samples were amplified in triplicate and cleaned using the MO BIO 69 htp PCR cleanup kit. Each PCR reaction included 1X PCR buffer, 10 μM each forward and reverse primer, 200 M dNTPs, 1 U/ml Taq polymerase, 15 ng template DNA, and PCR grade water, with a total reaction volume of 25 μL. Reactions were kept at 94°C for 3 minutes for denaturation to occur. Amplification was performed by 25 cycles of 94°C for 45s, 58°C for 60s, and 72°C for 90s. The V4 amplicons were sequenced on the Illumina HiSeq 2000 platform, yielding single end, 100 base pair reads. Sequencing and quality assessment were performed at the Yale Center for Genome Analysis.

B.1.5   Data Availability

Microbiome data from this study is available on Qiita under study ID 1629 (https://qiita.ucsd.edu/study/description/1629) and using the EBI accession number ERP020401. Patient clinical information is available on Qiita and in Supplemental Dataset 1[2].

B.2   Methods for Chapter 4.2

B.2.1   Fecal collection

Stool was collected daily using a swab technique [34], which enables the study subject to collect stool samples from the toilet paper. Samples were stored on a -20°C freezer. Fifteen patients collected daily stool for two 2-week periods separated by an interval of 4 weeks, during which no stool was collected. The short Clinical Disease Activity Index (CDAI) was evaluated at entry and at the end of collection period 1 and 2 [226]. The collection periods for the family substudy, which included 4 patients with inactive CD with same exclusion criteria as above and 3 unaffected family members of each CD patient were two separate 4 weeks periods interrupted by a 3 month collection-free interval.

Patients with proven CD diagnosed for at least 3 months and in clinical remission were recruited at the University of North Carolina (see Table 4.2). Serial

---

[2]https://images.nature.com/original/nature-assets/nmicrobiol/2017/nmicrobiol20174/extref/nmicrobiol20174-s2.txt

stool samples as outlined below were submitted for analysis. Exclusion criteria for entry into the study were active CD as defined by a short CDAI score > 150; fistulizing CD, concomitant use of azathioprine (AZA), 6-mercaptopurine (6-MP), methotrexate or anti-TNF agents for less than 3 months; or concomitant use of systemic steroids or budesonide. Steroids or budesonide had to be discontinued at least 8 weeks before inclusion, and local, rectally administered therapies containing 5-ASA (enemas, suppositories) or steroid enemas/foams should have not been used for the previous 4 weeks. Also, NSAIDs were not allowed as regular treatment, which was defined as use for at least 4 days a week each month. Patients were excluded if they were on antibiotic therapy ≥5 days each month or had antibiotic therapy ≥5 days in the previous 24 weeks. No probiotics were allowed in the last 24 weeks before inclusion and patients on long-term therapy with narcotics >2 days weekly were excluded. Further subject exclusions were known Hepatitis B, Hepatitis C or PSC or regular, high dose alcohol consumption (more than seven drinks per week). All trial participants were prohibited from consuming specific diets (e.g. Atkins diet, low carbohydrate diet). In the second subset we also included nonaffected family members to collect serial fecal samples. The requirement for the family subjects was one sibling age ≥ 8 years without known CD or ulcerative colitis and two living parents without CD or ulcerative colitis. The same exclusion criteria for antibiotics, probiotics, narcotics and concomitant diseases as in the main study were applied.

Table B.1: Summary of the demographic data of the 19 CD patients.

Male/Female (n)		8 / 11
Age (years, median; range)		31 (15-51)
Duration of Crohn's disease (years median; range)		9 (0.5-35)
	Ileal	5
Location of Crohn's disease (n)	Ileo-colonic	10
	Colonic	4
History of ileocecal resection		7
History of isolated small intestinal resection		2
	Steroids	0
	Azathioprine/6-MP/methotrexate	8
Concomitant therapies (n)	anti-TNF agent	12
	5-ASA	2
CDAI end of week 2 (mean; SD)		63 (51)
CDAI end of week 4 (mean; SD)		72 (61)

B.2.2   DNA Extraction

Fecal DNA isolation was performed according to the 16S Earth Microbiome Project Protocol [8]. Of the 960 samples analyzed, the initial subset (384 samples) were processed using the Illumina MiSeq platform (150 nucleotide sequences), and for the second subset (576 samples) were processed using the HiSeq platform (100 nucleotide sequences).

B.2.3   Data Availability

The sequences have been deposited on EBI and are available under the following accession number ERP104742. In addition, the processed sequences and sample information can be found in the Qiita[3] database under the study identifier 2538.

---

[3]https://qiita.ucsd.edu/study/description/2538

Appendix C

Sample collection and processing for Chapter 5

C.1 Methods for Chapter 5.1

C.1.1 Patients and donors

All patients suffered from multiply R-CDI refractory to standard antibiotic therapies. A single standard donor was used in the preparation of all fecal microbiota material as described previously [75]. The Institutional Review Board at the University of Minnesota approved prospective collection of fecal specimens and their analysis. All patients satisfied the inclusion criteria for the FMT within our program, which included at least two spontaneous recurrences of CDI within a month of discontinuation of antibiotics and failure of at least one advanced antibi-

otic regimen such as a vancomycin pulse/taper protocol or vancomycin treatment followed by administration of rifaximin or fidaxomicin for 2-3 weeks. The specific clinical characteristics of patients involved in this study are summarized in Supplementary Table 1[1].

C.1.2 Fecal microbiota transplantation

FMT was done using a standardized preparation of concentrated fresh or frozen fecal bacteria via colonoscopy as previously described [75]. All patents were treated with oral vancomycin, 125 mg four times daily, until two days prior to the procedure [75]. The day before the procedure, patients received a polyethylene glycol-based colonoscopy prep (GoLYTELY® or MoviPrep®) to remove residual antibiotics and fecal material. Donor fecal microbiota was placed into the terminal ileum and/or cecum via the biopsy channel of the colonoscope. A total of 17 donor samples from the same individual were used in these studies. The CD1-CD4 donor samples were given to patients CD1-CD4, respectively. Patients CD1, CD3, and CD4 received freshly prepared fecal microbiota, while patient CD2 received a previously frozen preparation of fecal microbiota, all from the same standardized, anonymous donor.

---

[1]https://static-content.springer.com/esm/art%3A10.1186%2Fs40168-015-0070-0/MediaObjects/40168_2015_70_MOESM3_ESM.xlsx

C.1.3 Sample collection

Fecal samples were collected at home by the patients using swabs to sample feces deposited into a toilet hat immediately after production, and stored frozen at approximately -20°C. Samples were subsequently transferred to the laboratory and stored at -80°C until used. Donor samples for DNA extraction were collected during processing of material for FMT, and stored frozen at -80°C until used. Samples from patients CD1- 4 were obtained prior to FMT and between 1 to 151 days post-FMT, with daily collection until day 28, and weekly collection until day 84. Fecal material prior to FMT was obtained from patients CD5-CD14.

C.1.4 DNA extraction

DNA was extracted from donor and recipients' pre- and post-FMT fecal samples using MOBIO PowerSoil DNA extraction kits (MOBIO, Carlsbad, CA), according to the manufacturer's instructions. Fecal DNA concentrations were measured using a QuBit DNA quantification system (Invitrogen, Carlsbad, CA).

C.1.5 PCR amplification

Extracted DNA was amplified using the EMP standard protocols[2] following the recommendations of Caporaso et al. [25]. Briefly, F515/R806 primers were used, with 12-base Golay codes introduced on the 806 end to provide unique sample indices. Cycling and annealing conditions were as previously described [25].

---

[2]http://www.earthmicrobiome.org/

C.1.6 DNA sequencing

DNA sequencing was performed as previously described [25] on an Illumina MiSeq platform using 2 x 150 bp paired-end reads and the Illumina v3 reagent chemistry.

C.2 Methods for Chapter 5.2

C.2.1 Patient selection

Patients undergoing FMT for R-CDI were prospectively recruited in this study. Informed consent was obtained to collect clinical data and stool samples. Data collected included demographics, clinical history, CDI treatment history, co-morbid conditions and response to FMT. A donor fecal sample was collected prior to FMT. Stool samples from the recipients were collected before FMT, and at day 7 and day 28 and were stored at -80°C. The donors were either related (genetically related family members) or unrelated (screened hospital employee volunteer donors or unrelated family members) and a fresh sample was obtained on the day of FMT. All donors underwent extensive screening in accordance with standard practice and guidelines from the Food and Drug Administration [100]. Donor selection criteria and experience from our group have been previously published [224]. The donor stool sample is weighed and divided into 50 grams aliquots. Each aliquot of 50 grams is diluted in normal saline in a 1:5 ratio (50 grams of stool diluted with

250 ml of normal saline) and is placed in the blender bag (a 2 bag system with a semipermeable membrane in the inner bag and the outside bag is plastic). The stool is placed in the inner bag and normal saline is added. The bag is placed in a sealed compartment in the stomacher 400 (Seward) and blended for 60 seconds at 230 rotations per minute. The filtrate is then placed into 50 ml conical tubes using 50 ml pipettes and placed on an ice pack prior to the procedure. R-CDI was defined as another episode of CDI within 56 days after symptom resolution with recurrence of symptoms and a positive stool polymerase chain reaction test. For this study, future *C. difficile* episodes after FMT up to 2 years were captured. These were categorized as up to 56 days, 56 days to 1 year and beyond 1 year.

C.2.2 Sequencing and analytic methods

After fecal DNA isolation (MoBio, Carlsbad, CA fecal DNA kit), amplicons spanning the variable region 4 of bacterial 16S rRNA were generated and sequenced using Illumina MiSeq platform at the Mayo Clinic Medical Genome Facility, Rochester, MN. The 16S rRNA sequencing data from the Illumina runs were quality controlled, trimmed, demultiplexed and assigned to OTUs following the closed reference at 97% similarity (using SortMeRNA as a clustering algorithm [118] protocol against the Greengenes [160] database 13_8 release, as implemented in QIIME 1.9.0 software [23], default parameters were used for all these steps unless otherwise noted. After quality control, 10,583,052 sequences were obtained, for a mean of 76,688 sequences per sample (min: 33,559, max: 154,200).

C.2.3 Clinical Statistics

Statistical analyses for clinical data were performed with JMP version 11.0 (SAS institute, NC). Data analysis included descriptive statistics, t-tests for normally distributed variables, non-parametric tests for skewed variables, chi-square tests and ANOVA tests as applicable. A p-value less than 0.05 was considered statistically significant.

Bibliography

[1] Wolfram mathematica. http://www.wolfram.com/, 2010.

[2] Emp dna extraction protocol. 2011. 00000.

[3] Microsoft excel. http://office.microsoft.com/en-us/excel/, 2011.

[4] Emperor dataset. http://dx.doi.org/10.5524/100068, 2013.

[5] Matlab. http://www.mathworks.com/products/matlab/, 2013.

[6] The r programming language. http://www.r-project.org/, 2013.

[7] Spotfire. http://spotfire.tibco.com/, 2013.

[8] The earth microbiome project standard protocols. http://www.earthmicrobiome.org/emp-standard-protocols/, 2015.

[9] The american gut project. http://americangut.org/, 2017.

[10] N. Alang and C. R. Kelly. Weight gain after fecal microbiota transplantation. *Open Forum Infect Dis*, 2(1):ofv004, 2015.

[11] M. Amieva and Jr. Peek, R. M. Pathobiology of helicobacter pylori-induced gastric cancer. *Gastroenterology*, 150(1):64–78, 2016.

[12] A. F. Andersson, M. Lindberg, H. Jakobsson, F. Backhed, P. Nyren, and L. Engstrand. Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS One*, 3(7):e2836, 2008.

[13] 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. Borruel, 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, H. I. T. Consortium Meta, M. Antolin, F. Artiguenave, H. M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, 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, and P. Bork. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–80, 2011.

[14] M. B. Azad, T. Konya, H. Maughan, D. S. Guttman, C. J. Field, R. S. Chari, M. R. Sears, A. B. Becker, J. A. Scott, A. L. Kozyrskyj, and Child Study Investigators. Gut microbiota of healthy canadian infants: profiles by mode of delivery and infant diet at 4 months. *CMAJ*, 185(5):385–94, 2013.

[15] F. Backhed, R. E. Ley, J. L. Sonnenburg, D. A. Peterson, and J. I. Gordon. Host-bacterial mutualism in the human intestine. *Science*, 307(5717):1915–20, 2005.

[16] A. Bashan, T. E. Gibson, J. Friedman, V. J. Carey, S. T. Weiss, E. L. Hohmann, and Y. Y. Liu. Universality of human microbial dynamics. *Nature*, 534(7606):259–62, 2016.

[17] B. Bjorkholm, C. M. Bok, A. Lundin, J. Rafter, M. L. Hibberd, and S. Pettersson. Intestinal microbiota regulate xenobiotic metabolism in the liver. *PLoS One*, 4(9):e6958, 2009.

[18] J. M. Blair, M. A. Webber, A. J. Baylay, D. O. Ogbolu, and L. J. Piddock. Molecular mechanisms of antibiotic resistance. *Nat Rev Microbiol*, 13(1):42–51, 2015.

[19] M. J. Blaser and S. Falkow. What are the consequences of the disappearing human microbiota? *Nat Rev Microbiol*, 7(12):887–94, 2009.

[20] N. A. Bokulich, S. Subramanian, J. J. Faith, D. Gevers, J. I. Gordon, R. Knight, D. A. Mills, and J. G. Caporaso. Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing. *Nat Methods*, 10(1):57–9, 2013.

[21] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[22] H. P. Browne, S. C. Forster, B. O. Anonye, N. Kumar, B. A. Neville, M. D. Stares, D. Goulding, and T. D. Lawley. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature*, 533(7604):543–6, 2016.

[23] 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, and R. Knight. Qiime allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–6, 2010.

[24] J. G. Caporaso, 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(5):R50, 2011.

[25] J. G. Caporaso, C. L. Lauber, W. A. Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S. M. Owens, J. Betley, L. Fraser, M. Bauer, N. Gormley, J. A. Gilbert, G. Smith, and R. Knight. Ultra-high-throughput microbial community analysis on the illumina hiseq and miseq platforms. *ISME J*, 6(8):1621–4, 2012.

[26] J. Gregory Caporaso, Christian L. Lauber, William A. Walters, Donna Berg-Lyons, Catherine A. Lozupone, Peter J. Turnbaugh, Noah Fierer, and Rob Knight. Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*, 108(Suppl 1):4516–4522, Mar 2011. 01135.

[27] M. Castellarin, R. L. Warren, J. D. Freeman, L. Dreolini, M. Krzywinski, J. Strauss, R. Barnes, P. Watson, E. Allen-Vercoe, R. A. Moore, and R. A. Holt. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res*, 22(2):299–306, 2012.

[28] J. Chase, J. Fouquier, M. Zare, D. L. Sonderegger, R. Knight, S. T. Kelley, J. Siegel, and J. G. Caporaso. Geography and location are the primary drivers of office microbiome composition. *mSystems*, 1(2), 2016.

[29] Vincent B. Chen, Ian W. Davis, and David C. Richardson. King (kinemage, next generation): A versatile interactive molecular and scientific visualization program. *Protein Science*, 18(11):2403–2409, 2009.

[30] M. Christ, Kempa-Liehr, A.W., and M Feindt. Distributed and parallel time series feature extraction for industrial big data applications. 2017.

[31] D. M. Chu, J. Ma, A. L. Prince, K. M. Antony, M. D. Seferovic, and K. M. Aagaard. Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nat Med*, 23(3):314–326, 2017.

[32] A. Clauset, D. B. Larremore, and R. Sinatra. Data-driven predictions in the science of science. *Science*, 355(6324):477–480, 2017.

[33] J. C. Clemente, E. C. Pehrsson, M. J. Blaser, K. Sandhu, Z. Gao, B. Wang, M. Magris, G. Hidalgo, M. Contreras, O. Noya-Alarcon, O. Lander, J. Mc-Donald, M. Cox, J. Walter, P. L. Oh, J. F. Ruiz, S. Rodriguez, N. Shen, S. J. Song, J. Metcalf, R. Knight, G. Dantas, and M. G. Dominguez-Bello. The microbiome of uncontacted amerindians. *Sci Adv*, 1(3), 2015.

[34] E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, and R. Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–7, 2009.

[35] E. K. Costello, K. Stagaman, L. Dethlefsen, B. J. Bohannan, and D. A. Relman. The application of ecological theory toward an understanding of the human microbiome. *Science*, 336(6086):1255–62, 2012.

[36] L. A. David, A. C. Materna, J. Friedman, M. I. Campos-Baptista, M. C. Blackburn, A. Perrotta, S. E. Erdman, and E. J. Alm. Host lifestyle affects human microbiota on daily timescales. *Genome Biol*, 15(7):R89, 2014.

[37] L. A. David, C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma, M. A. Fischbach, S. B. Biddinger, R. J. Dutton, and P. J. Turnbaugh. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–63, 2014.

[38] M. J. Day, T. Bilzer, J. Mansell, B. Wilcock, E. J. Hall, A. Jergens, T. Minami, M. Willard, R. Washabau, and Group World Small Animal Veterinary Association Gastrointestinal Standardization. Histopathological standards for the diagnosis of gastrointestinal inflammation in endoscopic biopsy samples from the dog and cat: a report from the world small animal veterinary association gastrointestinal standardization group. *J Comp Pathol*, 138 Suppl 1:S1–43, 2008.

[39] J. Debelius, S. J. Song, Y. Vazquez-Baeza, Z. Z. Xu, A. Gonzalez, and R. Knight. Tiny microbes, enormous impacts: what matters in gut microbiome studies? *Genome Biol*, 17(1):217, 2016.

[40] L. Dethlefsen and D. A. Relman. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc Natl Acad Sci U S A*, 108 Suppl 1:4554–61, 2011.

[41] A. Dhariwal, J. Chong, S. Habib, I. L. King, L. B. Agellon, and J. Xia. Microbiomeanalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res*, 2017.

[42] R. Diaz Heijtz, S. Wang, F. Anuar, Y. Qian, B. Bjorkholm, A. Samuelsson, M. L. Hibberd, H. Forssberg, and S. Pettersson. Normal gut microbiota modulates brain development and behavior. *Proc Natl Acad Sci U S A*, 108(7):3047–52, 2011.

[43] M. G. Dominguez-Bello, 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(26):11971–5, 2010.

[44] R. R. Dunn, N. Fierer, J. B. Henley, J. W. Leff, and H. L. Menninger. Home life: factors structuring the bacterial diversity found within and between homes. *PLoS One*, 8(5):e64133, 2013.

[45] P. B. Eckburg, 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(5728):1635–8, 2005.

[46] R. C. Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–1, 2010.

[47] E. F. Enright, C. G. Gahan, S. A. Joyce, and B. T. Griffin. The impact of the gut microbiota on drug metabolism and clinical outcome. *Yale J Biol Med*, 89(3):375–382, 2016.

[48] D. P. Faith. Phylogenetic diversity, functional trait diversity and extinction: avoiding tipping points and worst-case losses. *Philos Trans R Soc Lond B Biol Sci*, 370(1662):20140011, 2015.

[49] D. P. Faith and A. M. Baker. Phylogenetic diversity (pd) and biodiversity conservation: some bioinformatics challenges. *Evol Bioinform Online*, 2:121–8, 2006.

[50] J. J. Faith, J. L. Guruge, M. Charbonneau, S. Subramanian, H. Seedorf, A. L. Goodman, J. C. Clemente, R. Knight, A. C. Heath, R. L. Leibel, M. Rosenbaum, and J. I. Gordon. The long-term stability of the human gut microbiota. *Science*, 341(6141):1237439, 2013.

[51] G. Falony, M. Joossens, S. Vieira-Silva, J. Wang, Y. Darzi, K. Faust, A. Kurilshikov, M. J. Bonder, M. Valles-Colomer, D. Vandeputte, R. Y. Tito, S. Chaffron, L. Rymenans, C. Verspecht, L. De Sutter, G. Lima-Mendez, K. D'Hoe, K. Jonckheere, D. Homola, R. Garcia, E. F. Tigchelaar, L. Eeckhaudt, J. Fu, L. Henckaerts, A. Zhernakova, C. Wijmenga, and J. Raes. Population-level analysis of gut microbiome variation. *Science*, 352(6285):560–4, 2016.

[52] K. Faust, L. Lahti, D. Gonze, W. M. de Vos, and J. Raes. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr Opin Microbiol*, 25:56–66, 2015.

[53] C. Faustino, C. Serafim, P. Rijo, and C. P. Reis. Bile acids and bile acid derivatives: use in drug delivery systems and as therapeutic agents. *Expert Opin Drug Deliv*, 13(8):1133–48, 2016.

[54] N. Fierer, C. L. Lauber, N. Zhou, D. McDonald, E. K. Costello, and R. Knight. Forensic identification using skin bacterial communities. *Proc Natl Acad Sci U S A*, 107(14):6477–81, 2010.

[55] S. M. Finegold, S. E. Dowd, V. Gontcharova, C. Liu, K. E. Henley, R. D. Wolcott, E. Youn, P. H. Summanen, D. Granpeesheh, D. Dixon, M. Liu, D. R. Molitoris, and 3rd Green, J. A. Pyrosequencing study of fecal microflora of autistic and control children. *Anaerobe*, 16(4):444–53, 2010.

[56] G. E. Flores, J. G. Caporaso, J. B. Henley, J. R. Rideout, D. Domogala, J. Chase, J. W. Leff, Y. Vazquez-Baeza, A. Gonzalez, R. Knight, R. R. Dunn, and N. Fierer. Temporal variability is a personalized feature of the human microbiome. *Genome Biol*, 15(12):531, 2014.

[57] D. N. Frank, C. E. Robertson, C. M. Hamm, Z. Kpadeh, T. Zhang, H. Chen, W. Zhu, R. B. Sartor, E. C. Boedeker, N. Harpaz, N. R. Pace, and E. Li. Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. *Inflamm Bowel Dis*, 17(1):179–84, 2011.

[58] D. N. Frank, A. L. St Amand, R. A. Feldman, E. C. Boedeker, N. Harpaz, and N. R. Pace. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A*, 104(34):13780–5, 2007.

[59] S. Fuentes, E. van Nood, S. Tims, I. Heikamp-de Jong, C. J. ter Braak, J. J. Keller, E. G. Zoetendal, and W. M. de Vos. Reset of a critically disturbed microbial ecosystem: faecal transplant in recurrent clostridium difficile infection. *ISME J*, 8(8):1621–33, 2014.

[60] A. P. Garcia-Gonzalez, A. D. Ritter, S. Shrestha, E. C. Andersen, L. S. Yilmaz, and A. J. M. Walhout. Bacterial metabolism affects the c. elegans response to cancer chemotherapeutics. *Cell*, 169(3):431–441 e8, 2017.

[61] D. Gevers, S. Kugathasan, L. A. Denson, Y. Vazquez-Baeza, W. Van Treuren, B. Ren, E. Schwager, D. Knights, S. J. Song, M. Yassour, X. C. Morgan, A. D. Kostic, C. Luo, A. Gonzalez, D. McDonald, Y. Haberman, T. Walters, S. Baker, J. Rosh, M. Stephens, M. Heyman, J. Markowitz, R. Baldassano,

A. Griffiths, F. Sylvester, D. Mack, S. Kim, W. Crandall, J. Hyams, C. Huttenhower, R. Knight, and R. J. Xavier. The treatment-naive microbiome in new-onset crohn's disease. *Cell Host Microbe*, 15(3):382–92, 2014.

[62] Sean Gibbons, Claire Duvallet, and Eric J. Alm. Correcting for batch effects in case-control microbiome studies. *bioRxiv*, 2017.

[63] J. A. Gilbert, J. K. Jansson, and R. Knight. The earth microbiome project: successes and aspirations. *BMC Biol*, 12:69, 2014.

[64] J. A. Gilbert, F. Meyer, D. Antonopoulos, P. Balaji, C. T. Brown, N. Desai, J. A. Eisen, D. Evers, D. Field, W. Feng, D. Huson, J. Jansson, R. Knight, J. Knight, E. Kolker, K. Konstantindis, J. Kostka, N. Kyrpides, R. Mackelprang, A. McHardy, C. Quince, J. Raes, A. Sczyrba, A. Shade, and R. Stevens. Meeting report: the terabase metagenomics workshop and the vision of an earth microbiome project. *Stand Genomic Sci*, 3(3):243–8, 2010.

[65] J. A. Gilbert, R. A. Quinn, J. Debelius, Z. Z. Xu, J. Morton, N. Garg, J. K. Jansson, P. C. Dorrestein, and R. Knight. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*, 535(7610):94–103, 2016.

[66] S. R. Gill, 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(5778):1355–9, 2006.

[67] E. M. Glass, J. Wilkening, A. Wilke, D. Antonopoulos, and F. Meyer. Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc*, 2010(1):pdb prot5368, 2010.

[68] A. Gonzalez and R. Knight. Advancing analytical algorithms and pipelines for billions of microbial sequences. *Current opinion in biotechnology*, 23(1):64–71, 2012.

[69] J. C. Gower and P. Legendre. Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3(1):5–48, 1986.

[70] H. J. Haiser, D. B. Gootenberg, K. Chatman, G. Sirasani, E. P. Balskus, and P. J. Turnbaugh. Predicting and manipulating cardiac drug inactivation by the human gut bacterium eggerthella lenta. *Science*, 341(6143):295–8, 2013.

[71] H. J. Haiser, K. L. Seim, E. P. Balskus, and P. J. Turnbaugh. Mechanistic insight into digoxin inactivation by eggerthella lenta augments our understanding of its pharmacokinetics. *Gut Microbes*, 5(2):233–8, 2014.

[72] J. Halfvarson, C. J. Brislawn, R. Lamendella, Y. Vazquez-Baeza, W. A. Walters, L. M. Bramer, M. D'Amato, F. Bonfiglio, D. McDonald, A. Gonzalez, E. E. McClure, M. F. Dunklebarger, R. Knight, and J. K. Jansson. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol*, 2:17004, 2017.

[73] M. Hamady and R. Knight. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res*, 19(7):1141–52, 2009.

[74] M. Hamady, J. J. Walker, J. K. Harris, N. J. Gold, and R. Knight. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods*, 5(3):235–7, 2008.

[75] M. J. Hamilton, A. R. Weingarden, M. J. Sadowsky, and A. Khoruts. Standardized frozen preparation for transplantation of fecal microbiota for recurrent clostridium difficile infection. *Am J Gastroenterol*, 107(5):761–7, 2012.

[76] M. J. Hamilton, A. R. Weingarden, T. Unno, A. Khoruts, and M. J. Sadowsky. High-throughput dna sequence analysis reveals stable engraftment of gut microbiota following transplantation of previously frozen fecal bacteria. *Gut Microbes*, 4(2):125–35, 2013.

[77] Roger Harris. The l4 time-series: the first 20 years. *Journal of Plankton Research*, 32(5):577–583, 2010.

[78] J. A. Hartigan and P. M. Hartigan. The dip test of unimodality. pages 70–84, 1985.

[79] J. J. Hayward, M. G. Castelhano, K. C. Oliveira, E. Corey, C. Balkman, T. L. Baxter, M. L. Casal, S. A. Center, M. Fang, S. J. Garrison, S. E. Kalla, P. Korniliev, M. I. Kotlikoff, N. S. Moise, L. M. Shannon, K. W. Simpson, N. B. Sutter, R. J. Todhunter, and A. R. Boyko. Complex disease and phenotype mapping in the domestic dog. *Nat Commun*, 7:10460, 2016.

[80] S. B. Hedges, J. Marin, M. Suleski, M. Paymer, and S. Kumar. Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol*, 32(4):835–45, 2015.

[81] K. M. Hewitt, F. L. Mannino, A. Gonzalez, J. H. Chase, J. G. Caporaso, R. Knight, and S. T. Kelley. Bacterial diversity in two neonatal intensive care units (nicus). *PLoS One*, 8(1):e54703, 2013.

[82] E. M. Hill-Burns, J. W. Debelius, J. T. Morton, W. T. Wissemann, M. R. Lewis, Z. D. Wallen, S. D. Peddada, S. A. Factor, E. Molho, C. P. Zabetian, R. Knight, and H. Payami. Parkinson's disease and parkinson's

disease medications have distinct signatures of the gut microbiome. *Mov Disord*, 32(5):739–749, 2017.

[83] HMP-Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–14, 2012.

[84] J. B. Honneffer, Y. Minamoto, and J. S. Suchodolski. Microbiota alterations in acute and chronic gastrointestinal inflammation of cats and dogs. *World J Gastroenterol*, 20(44):16489–97, 2014.

[85] M. A. Hullar, A. N. Burnett-Hartman, and J. W. Lampe. Gut microbes, diet, and cancer. *Cancer Treat Res*, 159:377–99, 2014.

[86] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[87] S. M. Huse, D. B. Mark Welch, A. Voorhis, A. Shipunova, H. G. Morrison, A. M. Eren, and M. L. Sogin. Vamps: a website for visualization and analysis of microbial population structures. *BMC Bioinformatics*, 15:41, 2014.

[88] R. Ibsen-Jensen, K. Chatterjee, and M. A. Nowak. Computational complexity of ecological and evolutionary spatial dynamics. *Proc Natl Acad Sci U S A*, 112(51):15636–41, 2015.

[89] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299, Sep 1996. 08748.

[90] D. Ishikawa, T. Sasaki, T. Osada, K. Kuwahara-Arai, K. Haga, T. Shibuya, K. Hiramatsu, and S. Watanabe. Changes in intestinal microbiota following combination therapy with fecal microbial transplantation and antibiotics for ulcerative colitis. *Inflamm Bowel Dis*, 23(1):116–125, 2017.

[91] J. P. Jacobs, M. Goudarzi, N. Singh, M. Tong, I. H. McHardy, P. Ruegger, M. Asadourian, B. H. Moon, A. Ayson, J. Borneman, D. P. McGovern, Jr. Fornace, A. J., J. Braun, and M. Dubinsky. A disease-associated microbial and metabolomics state in relatives of pediatric inflammatory bowel disease patients. *Cell Mol Gastroenterol Hepatol*, 2(6):750–766, 2016.

[92] J. P. Jacobs, M. Goudarzi, N. Singh, M. Tong, I. H. McHardy, P. Ruegger, M. Asadourian, B. H. Moon, A. Ayson, J. Borneman, D. P. McGovern, Jr. Fornace, A. J., J. Braun, and M. Dubinsky. A disease-associated microbial and metabolomics state in relatives of pediatric inflammatory bowel disease patients. *Cell Mol Gastroenterol Hepatol*, 2(6):750–766, 2016.

[93] A. E. Jergens and K. W. Simpson. Inflammatory bowel disease in veterinary medicine. *Front Biosci (Elite Ed)*, 4:1404–19, 2012.

[94] L. Jostins, S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, J. Essers, M. Mitrovic, K. Ning, I. Cleynen, E. Theatre, S. L. Spain, S. Raychaudhuri, P. Goyette, Z. Wei, C. Abraham, J. P. Achkar, T. Ahmad, L. Amininejad, A. N. Ananthakrishnan, V. Andersen, J. M. Andrews, L. Baidoo, T. Balschun, P. A. Bampton, A. Bitton, G. Boucher, S. Brand, C. Buning, A. Cohain, S. Cichon, M. D'Amato, D. De Jong, K. L. Devaney, M. Dubinsky, C. Edwards, D. Ellinghaus, L. R. Ferguson, D. Franchimont, K. Fransen, R. Gearry, M. Georges, C. Gieger, J. Glas, T. Haritunians, A. Hart, C. Hawkey, M. Hedl, X. Hu, T. H. Karlsen, L. Kupcinskas, S. Kugathasan, A. Latiano, D. Laukens, I. C. Lawrance, C. W. Lees, E. Louis, G. Mahy, J. Mansfield, A. R. Morgan, C. Mowat, W. Newman, O. Palmieri, C. Y. Ponsioen, U. Potocnik, N. J. Prescott, M. Regueiro, J. I. Rotter, R. K. Russell, J. D. Sanderson, M. Sans, J. Satsangi, S. Schreiber, L. A. Simms, J. Sventoraityte, S. R. Targan, K. D. Taylor, M. Tremelling, H. W. Verspaget, M. De Vos, C. Wijmenga, D. C. Wilson, J. Winkelmann, R. J. Xavier, S. Zeissig, B. Zhang, C. K. Zhang, H. Zhao, I. B. D. Genetics Consortium International, M. S. Silverberg, V. Annese, H. Hakonarson, S. R. Brant, G. Radford-Smith, C. G. Mathew, J. D. Rioux, EE Schadt, MJ Daly, A Franke, M Parkes, S Vermeire, JC Barrett, and JH Cho. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, Nov 2012. 01571.

[95] S. Kawamoto, M. Maruya, L. M. Kato, W. Suda, K. Atarashi, Y. Doi, Y. Tsutsui, H. Qin, K. Honda, T. Okada, M. Hattori, and S. Fagarasan. Foxp3(+) t cells regulate immunoglobulin a selection and facilitate diversification of bacterial species responsible for immune homeostasis. *Immunity*, 41(1):152–65, 2014.

[96] C. R. Kelly, S. Kahn, P. Kashyap, L. Laine, D. Rubin, A. Atreja, T. Moore, and G. Wu. Update on fecal microbiota transplantation 2015: Indications, methodologies, mechanisms, and outlook. *Gastroenterology*, 149(1):223–37, 2015.

[97] C. R. Kelly, A. Khoruts, C. Staley, M. J. Sadowsky, M. Abd, M. Alani, B. Bakow, P. Curran, J. McKenney, A. Tisch, S. E. Reinert, J. T. Machan, and L. J. Brandt. Effect of fecal microbiota transplantation on recurrence in multiply recurrent clostridium difficile infection: A randomized trial. *Ann Intern Med*, 165(9):609–616, 2016.

[98] S. L. Kelly and D. E. Kelly. Microbial cytochromes p450: biodiversity and biotechnology. where do cytochromes p450 come from, what do they do and what can they do for us? *Philos Trans R Soc Lond B Biol Sci*, 368(1612):20120476, 2013.

[99] A. H. Keshteli, B. Millan, and K. L. Madsen. Pretreatment with antibiotics may enhance the efficacy of fecal microbiota transplantation in ulcerative colitis: a meta-analysis. *Mucosal Immunol*, 10(2):565–566, 2017.

[100] S. Khanna and D. S. Pardi. Clostridium difficile infection: management strategies for a difficult disease. *Therap Adv Gastroenterol*, 7(2):72–86, 2014.

[101] S. Khanna and D. S. Pardi. Clinical implications of antibiotic impact on gastrointestinal microbiota and clostridium difficile infection. *Expert Rev Gastroenterol Hepatol*, pages 1–8, 2016.

[102] S. Khanna, D. S. Pardi, C. R. Kelly, C. S. Kraft, T. Dhere, M. R. Henn, M. J. Lombardo, M. Vulic, T. Ohsumi, J. Winkler, C. Pindar, B. H. McGovern, R. J. Pomerantz, J. G. Aunins, D. N. Cook, and E. L. Hohmann. A novel microbiome therapeutic increases gut microbial diversity and prevents recurrent clostridium difficile infection. *J Infect Dis*, 2016.

[103] S. Khanna and P. K. Tosh. A clinician's primer on the role of the microbiome in human health and disease. *Mayo Clin Proc*, 89(1):107–14, 2014.

[104] S. Khanna, R. Weatherly, P. P. Kammer, and D.S. Pardi. Management and outcomes of patients with failed fecal microbiota transplantation for recurrent clostridium difficile infection. *Am J Gastroenterol*, 110:S580, 2015.

[105] A. Khoruts, J. Dicksved, J. K. Jansson, and M. J. Sadowsky. Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent clostridium difficile-associated diarrhea. *J Clin Gastroenterol*, 44(5):354–60, 2010.

[106] A. Khoruts, K. M. Rank, K. M. Newman, K. Viskocil, B. P. Vaughn, M. J. Hamilton, and M. J. Sadowsky. Inflammatory bowel disease affects the outcome of fecal microbiota transplantation for recurrent clostridium difficile infection. *Clin Gastroenterol Hepatol*, 2016.

[107] A. Khoruts and A. R. Weingarden. Emergence of fecal microbiota transplantation as an approach to repair disrupted microbial gut ecology. *Immunol Lett*, 2014.

[108] D. Kim, L. Song, F. P. Breitwieser, and S. L. Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res*, 26(12):1721–1729, 2016.

[109] I. S. Kim, D. H. Yoo, I. H. Jung, S. Lim, J. J. Jeong, K. A. Kim, O. N. Bae, H. H. Yoo, and D. H. Kim. Reduced metabolic activity of gut microbiota by antibiotics can potentiate the antithrombotic effect of aspirin. *Biochem Pharmacol*, 122:72–79, 2016.

[110] D. Kirpichnikov, S. I. McFarlane, and J. R. Sowers. Metformin: an update. *Ann Intern Med*, 137(1):25–33, 2002.

[111] R. Knight, P. Maxwell, A. Birmingham, J. Carnes, J. G. Caporaso, B. C. Easton, M. Eaton, M. Hamady, H. Lindsay, Z. Liu, C. Lozupone, D. McDonald, M. Robeson, R. Sammut, S. Smit, M. J. Wakefield, J. Widmann, S. Wikman, S. Wilson, H. Ying, and G. A. Huttley. Pycogent: a toolkit for making sense from sequence. *Genome Biol*, 8(8):R171, 2007.

[112] D. Knights, J. Kuczynski, E. S. Charlson, J. Zaneveld, M. C. Mozer, R. G. Collman, F. D. Bushman, R. Knight, and S. T. Kelley. Bayesian community-wide culture-independent microbial source tracking. *Nat Methods*, 8(9):761–3, 2011.

[113] D. Knights, M. S. Silverberg, R. K. Weersma, D. Gevers, G. Dijkstra, H. Huang, A. D. Tyler, S. van Sommeren, F. Imhann, J. M. Stempak, H. Huang, P. Vangay, G. A. Al-Ghalith, C. Russell, J. Sauk, J. Knight, M. J. Daly, C. Huttenhower, and R. J. Xavier. Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Med*, 6(12):107, 2014.

[114] D. Knights, T. L. Ward, C. E. McKinlay, H. Miller, A. Gonzalez, D. McDonald, and R. Knight. Rethinking "enterotype". *Cell Host Microbe*, 16(4):433–7, 2014.

[115] Dan Knights, Laura Wegener Parfrey, Jesse Zaneveld, Catherine Lozupone, and Rob Knight. Human-associated microbial signatures: examining their predictive value. *Cell host & microbe*, 10(4), Oct 2011. 00048.

[116] Kaija-Leena Kolho, Katri Korpela, Tytti Jaakkola, Madharasi V. A. Pichai, Erwin G. Zoetendal, Anne Salonen, and Willem M. de Vos. Fecal microbiota in pediatric inflammatory bowel disease and its relation to inflammation. *The American Journal of Gastroenterology*, 110(6):921–930, Jun 2015. 00024.

[117] M. R. Konikoff and L. A. Denson. Role of fecal calprotectin as a biomarker of intestinal inflammation in inflammatory bowel disease. *Inflamm Bowel Dis*, 12(6):524–34, 2006.

[118] E. Kopylova, L. Noe, and H. Touzet. Sortmerna: fast and accurate filtering of ribosomal rnas in metatranscriptomic data. *Bioinformatics*, 28(24):3211–7, 2012.

[119] T. Korem, D. Zeevi, N. Zmora, O. Weissbrod, N. Bar, M. Lotan-Pompan, T. Avnit-Sagi, N. Kosower, G. Malka, M. Rein, J. Suez, B. Z. Goldberg, A. Weinberger, A. A. Levy, E. Elinav, and E. Segal. Bread affects clinical parameters and induces gut microbiome-associated personal glycemic responses. *Cell Metab*, 25(6):1243–1253 e5, 2017.

[120] O. Koren, J. K. Goodrich, T. C. Cullender, A. Spor, K. Laitinen, H. K. Backhed, A. Gonzalez, J. J. Werner, L. T. Angenent, R. Knight, F. Backhed, E. Isolauri, S. Salminen, and R. E. Ley. Host remodeling of the gut microbiome and metabolic changes during pregnancy. *Cell*, 150(3):470–80, 2012.

[121] O. Koren, D. Knights, A. Gonzalez, L. Waldron, N. Segata, R. Knight, C. Huttenhower, and R. E. Ley. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput Biol*, 9(1):e1002863, 2013.

[122] J. Kuczynski, Z. Liu, C. Lozupone, D. McDonald, N. Fierer, and R. Knight. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Methods*, 7(10):813–9, 2010.

[123] J. Kuczynski, J. Stombaugh, W. A. Walters, A. Gonzalez, J. G. Caporaso, and R. Knight. Using qiime to analyze 16s rrna gene sequences from microbial communities. *Curr Protoc Microbiol*, Chapter 1:Unit 1E 5, 2012.

[124] M. Kuhn. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

[125] P. Kurdi, H. Tanaka, H. W. Van Veen, K. Asano, F. Tomita, and A. Yokota. Cholic acid accumulation and its diminution by short-chain fatty acids in bifidobacteria. *Microbiology*, 149(Pt 8):2031–7, 2003.

[126] Z. D. Kurtz, C. L. Muller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol*, 11(5):e1004226, 2015.

[127] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins,

E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[128] M. G. Langille, J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. Vega Thurber, R. Knight, R. G. Beiko, and C. Huttenhower. Predictive functional profiling of microbial communities using 16s rrna marker gene sequences. *Nat Biotechnol*, 31(9):814–21, 2013.

[129] C. L. Lauber, M. Hamady, R. Knight, and N. Fierer. Pyrosequencing-based assessment of soil ph as a predictor of soil bacterial community structure at the continental scale. *Applied and Environmental Microbiology*, 75(15):5111–20, 2009.

[130] F. C. Lessa, L. G. Winston, L. C. McDonald, and C. difficile Surveillance Team Emerging Infections Program. Burden of clostridium difficile infection in the united states. *N Engl J Med*, 372(24):2369–70, 2015.

[131] James D. Lewis. The utility of biomarkers in the diagnosis and therapy of inflammatory bowel disease. *Gastroenterology*, 140(6):1817–1826.e2, May 2011. 00000.

[132] R. E. Ley, M. Hamady, C. Lozupone, P. J. Turnbaugh, R. R. Ramey, J. S. Bircher, M. L. Schlegel, T. A. Tucker, M. D. Schrenzel, R. Knight, and J. I. Gordon. Evolution of mammals and their gut microbes. *Science*, 320(5883):1647–1651, 2008.

[133] R. E. Ley, C. A. Lozupone, M. Hamady, R. Knight, and J. I. Gordon. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol*, 6(10):776–88, 2008.

[134] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. 03910.

[135] J. Lindenbaum, D. G. Rund, Jr. Butler, V. P., D. Tse-Eng, and J. R. Saha. Inactivation of digoxin by the gut flora: reversal by antibiotic therapy. *N Engl J Med*, 305(14):789–94, 1981.

[136] Z. Liu, T. Z. DeSantis, G. L. Andersen, and R. Knight. Accurate taxonomy assignments from 16s rrna sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res*, 36(18):e120, 2008.

[137] A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, C. Powell, S. Vedantam, M. L. Buchkovich, J. Yang, D. C. Croteau-Chonka,

T. Esko, T. Fall, T. Ferreira, S. Gustafsson, Z. Kutalik, J. Luan, R. Magi, J. C. Randall, T. W. Winkler, A. R. Wood, T. Workalemahu, J. D. Faul, J. A. Smith, J. H. Zhao, W. Zhao, J. Chen, R. Fehrmann, A. K. Hedman, J. Karjalainen, E. M. Schmidt, D. Absher, N. Amin, D. Anderson, M. Beekman, J. L. Bolton, J. L. Bragg-Gresham, S. Buyske, A. Demirkan, G. Deng, G. B. Ehret, B. Feenstra, M. F. Feitosa, K. Fischer, A. Goel, J. Gong, A. U. Jackson, S. Kanoni, M. E. Kleber, K. Kristiansson, U. Lim, V. Lotay, M. Mangino, I. M. Leach, C. Medina-Gomez, S. E. Medland, M. A. Nalls, C. D. Palmer, D. Pasko, S. Pechlivanis, M. J. Peters, I. Prokopenko, D. Shungin, A. Stancakova, R. J. Strawbridge, Y. J. Sung, T. Tanaka, A. Teumer, S. Trompet, S. W. van der Laan, J. van Setten, J. V. Van Vliet-Ostaptchouk, Z. Wang, L. Yengo, W. Zhang, A. Isaacs, E. Albrecht, J. Arnlov, G. M. Arscott, A. P. Attwood, S. Bandinelli, A. Barrett, I. N. Bas, C. Bellis, A. J. Bennett, C. Berne, R. Blagieva, M. Bluher, S. Bohringer, L. L. Bonnycastle, Y. Bottcher, H. A. Boyd, M. Bruinenberg, I. H. Caspersen, Y. I. Chen, R. Clarke, E. W. Daw, A. J. M. de Craen, G. Delgado, and M. Dimitriou. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, Feb 2015. 00449.

[138] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12), 2014. 01282.

[139] C. Lozupone and R. Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*, 71(12):8228–35, 2005.

[140] C. A. Lozupone, M. Li, T. B. Campbell, S. C. Flores, D. Linderman, M. J. Gebert, R. Knight, A. P. Fontenot, and B. E. Palmer. Alterations in the gut microbiota associated with hiv-1 infection. *Cell Host Microbe*, 14(3):329–39, 2013.

[141] C. A. Lozupone, J. Stombaugh, A. Gonzalez, G. Ackermann, D. Wendel, Y. Vazquez-Baeza, J. K. Jansson, J. I. Gordon, and R. Knight. Meta-analyses of studies of the human microbiota. *Genome Res*, 23(10):1704–14, 2013.

[142] L. Lu, Y. Wu, L. Zuo, X. Luo, and P. J. Large. Intestinal microbiome and digoxin inactivation: meal plan for digoxin users? *World J Microbiol Biotechnol*, 30(3):791–9, 2014.

[143] C. Lupp, M. L. Robertson, M. E. Wickham, I. Sekirov, O. L. Champion, E. C. Gaynor, and B. B. Finlay. Host-mediated inflammation disrupts the intestinal microbiota and promotes the overgrowth of enterobacteriaceae. *Cell Host Microbe*, 2(2):119–29, 2007.

[144] J. R. Lupton. Microbial degradation products influence colon cancer risk: the butyrate controversy. *J Nutr*, 134(2):479–82, 2004.

[145] E. A. Magee, C. J. Richardson, R. Hughes, and J. H. Cummings. Contribution of dietary protein to sulfide production in the large intestine: an in vitro and a controlled feeding study in humans. *Am J Clin Nutr*, 72(6):1488–94, 2000.

[146] S. Mandal, W. Van Treuren, R. A. White, M. Eggesbo, R. Knight, and S. D. Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis*, 26:27663, 2015.

[147] Chaysavanh Manichanh, Natalia Borruel, Francesc Casellas, and Francisco Guarner. The gut microbiota in ibd. *Nature Reviews Gastroenterology and Hepatology*, 9(10):599–608, Oct 2012. 00418.

[148] H. E. Mardini and A. Y. Grigorian. Probiotic mix vsl#3 is effective adjunctive therapy for mild to moderately active ulcerative colitis: a meta-analysis. *Inflamm Bowel Dis*, 20(9):1562–7, 2014.

[149] P. Marteau, P. Pochart, J. Dore, C. Bera-Maillet, A. Bernalier, and G. Corthier. Comparative study of bacterial groups within the human cecal and fecal microbiota. *Appl Environ Microbiol*, 67(10):4939–42, 2001.

[150] Cristina Martinez, Maria Antolin, Javier Santos, Antonio Torrejon, Francesc Casellas, Natalia Borruel, Francisco Guarner, and Juan-R. Malagelada. Unstable composition of the fecal microbiota in ulcerative colitis during clinical remission. *The American Journal of Gastroenterology*, 103(3):643–648, Mar 2008. 00114.

[151] Ines Martinez, James M. Lattimer, Kelcie L. Hubach, Jennifer A. Case, Junyi Yang, Casey G. Weber, Julie A. Louk, Devin J. Rose, Gayaneh Kyureghian, Daniel A. Peterson, Haub M., and Walter J. Gut microbiome composition is linked to whole grain-induced immunological improvements. *The ISME Journal*, 7(2):269–280, Feb 2013. 00104.

[152] S. L. Martz, M. Guzman-Rodriguez, S. M. He, C. Noordhof, D. J. Hurlbut, G. B. Gloor, C. Carlucci, S. Weese, E. Allen-Vercoe, J. Sun, E. C. Claud, and E. O. Petrof. A human gut ecosystem protects against c. difficile disease by targeting tcda. *J Gastroenterol*, 52(4):452–465, 2017.

[153] N. M. Maruthur, E. Tseng, S. Hutfless, L. M. Wilson, C. Suarez-Cuervo, Z. Berger, Y. Chu, E. Iyoha, J. B. Segal, and S. Bolen. Diabetes medications as monotherapy or metformin-based combination therapy for type 2 diabetes: A systematic review and meta-analysis. *Ann Intern Med*, 164(11):740–51, 2016.

[154] C. Mataki, B. C. Magnier, S. M. Houten, J. S. Annicotte, C. Argmann, C. Thomas, H. Overmars, W. Kulik, D. Metzger, J. Auwerx, and K. Schoonjans. Compromised intestinal lipid absorption in mice with a liver-specific deficiency of liver receptor homolog 1. *Mol Cell Biol*, 27(23):8330–9, 2007.

[155] Z. Matuskova, E. Anzenbacherova, R. Vecera, H. Tlaskalova-Hogenova, M. Kolar, and P. Anzenbacher. Administration of a probiotic can change drug pharmacokinetics: effect of e. coli nissle 1917 on amidarone absorption in rats. *PLoS One*, 9(2):e87150, 2014.

[156] C. F. Maurice, H. J. Haiser, and P. J. Turnbaugh. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell*, 152(1-2):39–50, 2013.

[157] D. I. Mawby, J. W. Bartges, A. d'Avignon, D. P. Laflamme, T. D. Moyers, and T. Cottrell. Comparison of various methods for estimating body fat in dogs. *J Am Anim Hosp Assoc*, 40(2):109–14, 2004.

[158] L. J. McCreight, C. J. Bailey, and E. R. Pearson. Metformin and the gastrointestinal tract. *Diabetologia*, 59(3):426–35, 2016.

[159] D. McDonald, J. C. Clemente, J. Kuczynski, J. R. Rideout, J. Stombaugh, D. Wendel, A. Wilke, S. Huse, J. Hufnagle, F. Meyer, R. Knight, and J. G. Caporaso. The biological observation matrix (biom) format or: how i learned to stop worrying and love the ome-ome. *Gigascience*, 1(1):7, 2012.

[160] D. McDonald, M. N. Price, J. Goodrich, E. P. Nawrocki, T. Z. DeSantis, A. Probst, G. L. Andersen, R. Knight, and P. Hugenholtz. An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J*, 6(3):610–8, 2012.

[161] S. E. McGarr, J. M. Ridlon, and P. B. Hylemon. Diet, anaerobic bacterial metabolism, and colon cancer: a review of the literature. *J Clin Gastroenterol*, 39(2):98–109, 2005.

[162] Paul J. McMurdie and Susan Holmes. phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, 8(4):e61217, Apr 2013. 00352.

[163] J. L. Metcalf, Z. Z. Xu, S. Weiss, S. Lax, W. Van Treuren, E. R. Hyde, S. J. Song, A. Amir, P. Larsen, N. Sangwan, D. Haarmann, G. C. Humphrey, G. Ackermann, L. R. Thompson, C. Lauber, A. Bibat, C. Nicholas, M. J. Gebert, J. F. Petrosino, S. C. Reed, J. A. Gilbert, A. M. Lynne, S. R. Bucheli, D. O. Carter, and R. Knight. Microbial community assembly and metabolic function during mammalian corpse decomposition. *Science*, 351(6269):158–62, 2016.

[164] Y. Minamoto, C. C. Otoni, S. M. Steelman, O. Buyukleblebici, J. M. Steiner, A. E. Jergens, and J. S. Suchodolski. Alteration of the fecal microbiota and serum metabolite profiles in dogs with idiopathic inflammatory bowel disease. *Gut Microbes*, 6(1):33–47, 2015.

[165] P. Moayyedi, M. G. Surette, P. T. Kim, J. Libertucci, M. Wolfe, C. Onischi, D. Armstrong, J. K. Marshall, Z. Kassam, W. Reinisch, and C. H. Lee. Fecal microbiota transplantation induces remission in patients with active ulcerative colitis in a randomized controlled trial. *Gastroenterology*, 149(1):102–109 e6, 2015.

[166] E. Montassier, G. A. Al-Ghalith, T. Ward, S. Corvec, T. Gastinne, G. Potel, P. Moreau, M. F. de la Cochetiere, E. Batard, and D. Knights. Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection. *Genome Med*, 8(1):49, 2016.

[167] B. D. Muegge, J. Kuczynski, D. Knights, J. C. Clemente, A. Gonzalez, 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(6032):970–4, 2011.

[168] T. Nakatsuji, T. H. Chen, S. Narala, K. A. Chun, A. M. Two, T. Yun, F. Shafiq, P. F. Kotol, A. Bouslimani, A. V. Melnik, H. Latif, J. N. Kim, A. Lockhart, K. Artis, G. David, P. Taylor, J. Streib, P. C. Dorrestein, A. Grier, S. R. Gill, K. Zengler, T. R. Hata, D. Y. Leung, and R. L. Gallo. Antimicrobials from human skin commensal bacteria protect against staphylococcus aureus and are deficient in atopic dermatitis. *Sci Transl Med*, 9(378), 2017.

[169] Teruaki Nakatsuji, Tiffany H. Chen, Saisindhu Narala, Kimberly A. Chun, Aimee M. Two, Tong Yun, Faiza Shafiq, Paul F. Kotol, Amina Bouslimani, Alexey V. Melnik, Haythem Latif, Ji-Nu Kim, Alexandre Lockhart, Keli Artis, Gloria David, Patricia Taylor, Joanne Streib, Pieter C. Dorrestein, Alex Grier, Steven R. Gill, Karsten Zengler, Tissa R. Hata, Donald Y. M. Leung, and Richard L. Gallo. Antimicrobials from human skin commensal bacteria protect against staphylococcus aureus and are deficient in atopic dermatitis. *Science Translational Medicine*, 9(378), 2017.

[170] H. Nakayama, T. Okumichi, S. Nakashima, A. Kimura, M. Ikeda, and H. Kajihara. Papillary adenocarcinoma of the sigmoid colon associated with psammoma bodies and hyaline globules: report of a case. *Jpn J Clin Oncol*, 27(3):193–6, 1997.

[171] D. J. Novo, N. G. Perlmutter, R. H. Hunt, and H. M. Shapiro. Multiparameter flow cytometric analysis of antibiotic effects on membrane potential,

membrane permeability, and bacterial counts of staphylococcus aureus and micrococcus luteus. *Antimicrob Agents Chemother*, 44(4):827–34, 2000.

[172] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner. metaspades: a new versatile metagenomic assembler. *Genome Res*, 27(5):824–834, 2017.

[173] S. I. O'Donoghue, A. C. Gavin, N. Gehlenborg, D. S. Goodsell, J. K. Heriche, C. B. Nielsen, C. North, A. J. Olson, J. B. Procter, D. W. Shattuck, T. Walter, and B. Wong. Visualizing biological data-now and in the future. *Nature Methods*, 7(3 Suppl):S2–4, 2010.

[174] Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, and the Vegan Development Team. *vegan: Community Ecology Package*. Jun 2016. 00000.

[175] R. Orenstein, E. Dubberke, R. Hardi, A. Ray, K. Mullane, D. S. Pardi, M. S. Ramesh, and Punch Cd Investigators. Safety and durability of rbx2660 (microbiota suspension) for recurrent clostridium difficile infection: Results of the punch cd study. *Clin Infect Dis*, 62(5):596–602, 2016.

[176] K. Otani, T. Tanigawa, T. Watanabe, S. Shimada, Y. Nadatani, Y. Nagami, F. Tanaka, N. Kamata, H. Yamagami, M. Shiba, K. Tominaga, Y. Fujiwara, and T. Arakawa. Microbiota plays a key role in non-steroidal anti-inflammatory drug-induced small intestinal damage. *Digestion*, 95(1):22–28, 2017.

[177] N. R. Pace. A molecular view of microbial diversity and the biosphere. *Science*, 276(5313):734–40, 1997.

[178] C. Palmer, E. M. Bik, D. B. DiGiulio, D. A. Relman, and P. O. Brown. Development of the human infant intestinal microbiota. *PLoS Biol*, 5(7):e177, 2007.

[179] S. Paramsothy, M. A. Kamm, N. O. Kaakoush, A. J. Walsh, J. van den Bogaerde, D. Samuel, R. W. Leong, S. Connor, W. Ng, R. Paramsothy, W. Xuan, E. Lin, H. M. Mitchell, and T. J. Borody. Multidonor intensive faecal microbiota transplantation for active ulcerative colitis: a randomised placebo-controlled trial. *Lancet*, 2017.

[180] V. Pascal, M. Pozuelo, N. Borruel, F. Casellas, D. Campos, A. Santiago, X. Martinez, E. Varela, G. Sarrabayrouse, K. Machiels, S. Vermeire, H. Sokol, F. Guarner, and C. Manichanh. A microbial signature for crohn's disease. *Gut*, 66(5):813–822, 2017.

[181] M. A. Peppercorn and P. Goldman. The role of intestinal bacteria in the metabolism of salicylazosulfapyridine. *J Pharmacol Exp Ther*, 181(3):555–62, 1972.

[182] F. Perez and B. E. Granger. Ipython: A system for interactive scientific computing. *Computing in Science & Engineering*, 9(3):21–29, 2007.

[183] J. K. Pickrell and J. K. Pritchard. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet*, 8(11):e1002967, 2012.

[184] S. A. Price, S. S. Hopkins, K. K. Smith, and V. L. Roth. Tempo of trophic evolution and its impact on mammalian diversification. *Proc Natl Acad Sci U S A*, 109(18):7008–12, 2012.

[185] 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, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, H. I. T. Consortium Meta, P. Bork, S. D. Ehrlich, and J. Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.

[186] R. A. Quinn, L. F. Nothias, O. Vining, M. Meehan, E. Esquenazi, and P. C. Dorrestein. Molecular networking as a drug discovery, drug metabolism, and precision medicine strategy. *Trends Pharmacol Sci*, 38(2):143–154, 2017.

[187] F. Rafii and C. E. Cerniglia. Reduction of azo dyes and nitroaromatic compounds by bacterial enzymes from the human intestinal tract. *Environ Health Perspect*, 103 Suppl 5:17–9, 1995.

[188] Sylvie Rajca, Virginie Grondin, Edouard Louis, Gwenola Vernier-Massouille, Jean-Charle Grimaud, Yoram Bouhnik, David Laharie, Jean-Louis Dupas, Helene Pillant, Laurence Picon, M Veyrac, G Flamant, Mand Savoye, R Jian, M Devos, G Paintaud, E Piver, M Allez, JY Mary, H Sokol, JF Colombel, and P Seksik. Alterations in the intestinal microbiome (dysbiosis) as a predictor of relapse after infliximab withdrawal in crohn's disease. *Inflammatory Bowel Diseases*, 20(6):978–986, Jun 2014. 00034.

[189] F. Raymond, A. A. Ouameur, M. Deraspe, N. Iqbal, H. Gingras, B. Dridi, P. Leprohon, P. L. Plante, R. Giroux, E. Berube, J. Frenette, D. K. Boudreau, J. L. Simard, I. Chabot, M. C. Domingo, S. Trottier, M. Boissinot, A. Huletsky, P. H. Roy, M. Ouellette, M. G. Bergeron, and J. Corbeil. The initial

state of the human gut microbiome determines its reshaping by antibiotics. *ISME J*, 10(3):707–20, 2016.

[190] J. Reeder and R. Knight. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat Methods*, 7(9):668–9, 2010.

[191] J. M. Ridlon, D. J. Kang, and P. B. Hylemon. Bile salt biotransformations by human intestinal bacteria. *J Lipid Res*, 47(2):241–59, 2006.

[192] Michelle G Rooks, Patrick Veiga, Leslie H Wardwell-Scott, Timothy Tickle, Nicola Segata, Monia Michaud, Carey Ann Gallini, Chloé Beal, Johan ET van Hylckama-Vlieg, Sonia A Ballal, X Morgan, J Glickman, D Gevers, C Huttenhower, and W Garret. Gut microbiome composition and function in experimental colitis during active disease and treatment-induced remission. *The ISME Journal*, 8(7):1403–1417, Jul 2014. 00067.

[193] N. G. Rossen, S. Fuentes, M. J. van der Spek, J. G. Tijssen, J. H. Hartman, A. Duflou, M. Lowenberg, G. R. van den Brink, E. M. Mathus-Vliegen, W. M. de Vos, E. G. Zoetendal, G. R. D'Haens, and C. Y. Ponsioen. Findings from a randomized controlled trial of fecal transplantation for patients with ulcerative colitis. *Gastroenterology*, 149(1):110–118 e4, 2015.

[194] G. Rossi, G. Pengo, M. Caldin, A. Palumbo Piccionello, J. M. Steiner, N. D. Cohen, A. E. Jergens, and J. S. Suchodolski. Comparison of microbiological, histological, and immunomodulatory parameters in response to treatment with either combination therapy with prednisone and metronidazole or probiotic vsl#3 strains in dogs with idiopathic inflammatory bowel disease. *Plos ONE*, 9(4):e94699, 2014.

[195] R. B. Sartor. Microbial influences in inflammatory bowel diseases. *Gastroenterology*, 134(2):577–94, 2008.

[196] R. B. Sartor and G. D. Wu. Roles for intestinal bacteria, viruses, and fungi in pathogenesis of inflammatory bowel diseases and therapeutic approaches. *Gastroenterology*, 152(2):327–339 e4, 2017.

[197] J. Satsangi, M. S. Silverberg, S. Vermeire, and J. F. Colombel. The montreal classification of inflammatory bowel disease: controversies, consensus, and implications. *Gut*, 55(6):749–53, 2006.

[198] D. C. Savage. Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol*, 31:107–33, 1977.

[199] S. I. Sayin, A. Wahlstrom, J. Felin, S. Jantti, H. U. Marschall, K. Bamberg, B. Angelin, T. Hyotylainen, M. Oresic, and F. Backhed. Gut microbiota regulates bile acid metabolism by reducing the levels of tauro-beta-muricholic acid, a naturally occurring fxr antagonist. *Cell Metab*, 17(2):225–35, 2013.

[200] C. Scarpignato, W. Dolak, A. Lanas, P. Matzneller, C. Renzulli, M. Grimaldi, M. Zeitlinger, and I. Bjarnason. Rifaximin reduces the number and severity of intestinal lesions associated with use of nonsteroidal anti-inflammatory drugs in humans. *Gastroenterology*, 152(5):980–982 e3, 2017.

[201] S. L. Schnorr, M. Candela, S. Rampelli, M. Centanni, C. Consolandi, G. Basaglia, S. Turroni, E. Biagi, C. Peano, M. Severgnini, J. Fiori, R. Gotti, G. De Bellis, D. Luiselli, P. Brigidi, A. Mabulla, F. Marlowe, A. G. Henry, and A. N. Crittenden. Gut microbiome of the hadza hunter-gatherers. *Nat Commun*, 5:3654, 2014.

[202] E Schwager and C Huttenhower. Detecting statistically significant associations between sparse and high dimensional compositional data. *Unpublished*, 2016.

[203] T. A. Scott, L. M. Quintaneiro, P. Norvaisas, P. P. Lui, M. P. Wilson, K. Y. Leung, L. Herrera-Dominguez, S. Sudiwala, A. Pessia, P. T. Clayton, K. Bryson, V. Velagapudi, P. B. Mills, A. Typas, N. D. E. Greene, and F. Cabreiro. Host-microbe co-metabolism dictates cancer drug efficacy in c. elegans. *Cell*, 169(3):442–456 e18, 2017.

[204] A. M. Seekatz, J. Aas, C. E. Gessert, T. A. Rubin, D. M. Saman, J. S. Bakken, and V. B. Young. Recovery of the gut microbiome following fecal microbiota transplantation. *MBio*, 5(3):e00893–14, 2014.

[205] R. Sender, S. Fuchs, and R. Milo. Are we really vastly outnumbered? revisiting the ratio of bacterial to host cells in humans. *Cell*, 164(3):337–40, 2016.

[206] D. Shahinas, M. Silverman, T. Sittler, C. Chiu, P. Kim, E. Allen-Vercoe, S. Weese, A. Wong, D. E. Low, and D. R. Pillai. Toward an understanding of changes in diversity associated with fecal microbiome transplantation based on 16s rrna gene deep sequencing. *MBio*, 3(5), 2012.

[207] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504, 2003.

[208] W. Shen, H. R. Gaskins, and M. K. McIntosh. Influence of dietary fat on intestinal microbes, inflammation, barrier function and metabolic outcomes. *J Nutr Biochem*, 25(3):270–80, 2014.

[209] Mark S. Silverberg, Jack Satsangi, Tariq Ahmad, Ian D. R. Arnott, Charles N. Bernstein, Steven R. Brant, Renzo Caprilli, Jean-Frédéric Colombel, Christoph Gasche, Karel Geboes, D Jewel, A Karban, A Loftus,

EV Jr andPena, RH Riddel, DB Sachar, S Schreiber, AH Steinhart, SR Targan, S Vermeire, and BF Warren. Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: report of a working party of the 2005 montreal world congress of gastroenterology. *Canadian Journal of Gastroenterology = Journal Canadien De Gastroenterologie*, 19 Suppl A:5A–36A, Sep 2005. 00000.

[210] K. W. Simpson, B. Dogan, M. Rishniw, R. E. Goldstein, S. Klaessig, P. L. McDonough, A. J. German, R. M. Yates, D. G. Russell, S. E. Johnson, D. E. Berg, J. Harel, G. Bruant, S. P. McDonough, and Y. H. Schukken. Adherent and invasive escherichia coli is associated with granulomatous colitis in boxer dogs. *Infect Immun*, 74(8):4778–92, 2006.

[211] T. Sipponen, P. Kärkkäinen, E. Savilahti, K.-L. Kolho, H. Nuutinen, U. Turunen, and Färkkilä M. Correlation of faecal calprotectin and lactoferrin with an endoscopic score for crohn's disease and histological findings. *Alimentary Pharmacology & Therapeutics*, 28(10):1221–1229, Nov 2008. 00000.

[212] I. P. Smirnov. [case of multiple race congenital abnormalities]. *Vrach Delo*, 5(5):130–1, 1978.

[213] Harry Sokol, Bénédicte Pigneur, Laurie Watterlot, Omar Lakhdari, Luis G. Bermedez-Humaron, Jean-Jacques Gratadoux, Sébastien Blugeon, Chantal Bridonneau, Jean-Pierre Furet, Gérard Corthier, Corinne Grangette, Nadia Vasquez, Philippe Pochart, Philippe Seksik, and Philippe Langella. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of crohn disease patients. *Proceedings of the National Academy of Sciences of the United States of America*, 105(43):16731–16736, Oct 2008. 01592.

[214] S. J. Song, C. Lauber, E. K. Costello, C. A. Lozupone, G. Humphrey, D. Berg-Lyons, J. G. Caporaso, D. Knights, J. C. Clemente, S. Nakielny, J. I. Gordon, N. Fierer, and R. Knight. Cohabiting family members share microbiota with one another and with their dogs. *Elife*, 2:e00458, 2013.

[215] Y. Song, S. Garg, M. Girotra, C. Maddox, E. C. von Rosenvinge, A. Dutta, S. Dutta, and W. F. Fricke. Microbiota dynamics in patients treated with fecal microbiota transplantation for recurrent clostridium difficile infection. *PLoS One*, 8(11):e81330, 2013.

[216] T. Sousa, R. Paterson, V. Moore, A. Carlsson, B. Abrahamsson, and A. W. Basit. The gastrointestinal microbiota as a site for the biotransformation of drugs. *Int J Pharm*, 363(1-2):1–25, 2008.

[217] P. Spanogiannopoulos, E. N. Bess, R. N. Carmody, and P. J. Turnbaugh. The microbial pharmacists within us: a metagenomic view of xenobiotic metabolism. *Nat Rev Microbiol*, 14(5):273–87, 2016.

[218] C. Staley, C. R. Kelly, L. J. Brandt, A. Khoruts, and M. J. Sadowsky. Complete microbiota engraftment is not essential for recovery from recurrent clostridium difficile infection following fecal microbiota transplantation. *MBio*, 7(6), 2016.

[219] L. Stone and A. Roberts. The checkerboard score and species distributions. *Oecologia*, 85(1):74–79, 1990.

[220] J. S. Suchodolski, J. Camacho, and J. M. Steiner. Analysis of bacterial diversity in the canine duodenum, jejunum, ileum, and colon by comparative 16s rrna gene analysis. *FEMS Microbiol Ecol*, 66(3):567–78, 2008.

[221] J. S. Suchodolski, S. E. Dowd, V. Wilke, J. M. Steiner, and A. E. Jergens. 16s rrna gene pyrosequencing reveals bacterial dysbiosis in the duodenum of dogs with idiopathic inflammatory bowel disease. *Plos ONE*, 7(6):e39333, 2012.

[222] J. S. Suchodolski, M. E. Markel, J. F. Garcia-Mazcorro, S. Unterer, R. M. Heilmann, S. E. Dowd, P. Kachroo, I. Ivanov, Y. Minamoto, E. M. Dillman, J. M. Steiner, A. K. Cook, and L. Toresson. The fecal microbiome in dogs with acute diarrhea and idiopathic inflammatory bowel disease. *Plos ONE*, 7(12):e51907, 2012.

[223] T. Suzuki, A. Masui, J. Nakamura, H. Shiozawa, J. Aoki, H. Nakae, S. Tsuda, J. Imai, O. Hideki, M. Matsushima, T. Mine, A. Tamura, T. Ohtsu, Y. Asami, and A. Takagi. Yogurt containing lactobacillus gasseri mitigates aspirin-induced small bowel injuries: A prospective, randomized, double-blind, placebo-controlled trial. *Digestion*, 95(1):49–54, 2017.

[224] R. Tariq, R. Weatherly, P. Kammer, D. S. Pardi, and S. Khanna. Donor screening experience for fecal microbiota transplantation in patients with recurrent c. difficile infection. *J Clin Gastroenterol*, 2016.

[225] Olivier Temam. The rebirth of neural networks. In *International Symposium on Computer Architecture (ISCA)*, June 2010.

[226] K. Thia, Jr. Faubion, W. A., Jr. Loftus, E. V., T. Persson, A. Persson, and W. J. Sandborn. Short cdai: development and validation of a shortened and simplified crohn's disease activity index. *Inflamm Bowel Dis*, 17(1):105–11, 2011.

[227] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The human microbiome project. *Nature*, 449(7164):804–10, 2007.

[228] P. J. Turnbaugh, 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(6):6ra14, 2009.

[229] S. Turroni, J. Fiori, S. Rampelli, S. L. Schnorr, C. Consolandi, M. Barone, E. Biagi, F. Fanelli, M. Mezzullo, A. N. Crittenden, A. G. Henry, P. Brigidi, and M. Candela. Fecal metabolome of the hadza hunter-gatherers: a host-microbiome integrative view. *Sci Rep*, 6:32826, 2016.

[230] J. M. Uronis, J. C. Arthur, T. Keku, A. Fodor, I. M. Carroll, M. L. Cruz, C. B. Appleyard, and C. Jobin. Gut microbial diversity is reduced by the probiotic vsl#3 and correlates with decreased tnbs-induced colitis. *Inflamm Bowel Dis*, 17(1):289–97, 2011.

[231] P. Van den Abbeele, C. Grootaert, M. Marzorati, S. Possemiers, W. Verstraete, P. Gerard, S. Rabot, A. Bruneau, S. El Aidy, M. Derrien, E. Zoetendal, M. Kleerebezem, H. Smidt, and T. Van de Wiele. Microbial community development in a dynamic gut model is reproducible, colon region specific, and selective for bacteroidetes and clostridium cluster ix. *Appl Environ Microbiol*, 76(15):5237–46, 2010.

[232] S. van der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[233] E. van Nood, A. Vrieze, M. Nieuwdorp, S. Fuentes, E. G. Zoetendal, W. M. de Vos, C. E. Visser, E. J. Kuijper, J. F. Bartelsman, J. G. Tijssen, P. Speelman, M. G. Dijkgraaf, and J. J. Keller. Duodenal infusion of donor feces for recurrent clostridium difficile. *N Engl J Med*, 368(5):407–15, 2013.

[234] Y. Vazquez-Baeza, A. Gonzalez, L. Smarr, D. McDonald, J. T. Morton, J. A. Navas-Molina, and R. Knight. Bringing the dynamic microbiome to life with animations. *Cell Host Microbe*, 21(1):7–10, 2017.

[235] Y. Vazquez-Baeza, E. R. Hyde, J. S. Suchodolski, and R. Knight. Dog and human inflammatory bowel disease rely on overlapping yet distinct dysbiosis networks. *Nat Microbiol*, 1:16177, 2016.

[236] Y. Vazquez-Baeza, M. Pirrung, A. Gonzalez, and R. Knight. Emperor: a tool for visualizing high-throughput microbial community data. *Gigascience*, 2(1):16, 2013.

[237] S. Vega-Lopez, L. M. Ausman, J. L. Griffith, and A. H. Lichtenstein. Interindividual variability and intra-individual reproducibility of glycemic index values for commercial white bread. *Diabetes Care*, 30(6):1412–7, 2007.

[238] R. Vrolix and R. P. Mensink. Variability of the glycemic response to single food products in healthy subjects. *Contemp Clin Trials*, 31(1):5–11, 2010.

[239] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W. T. Liu, M. Crusemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderon, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C. C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C. C. Liaw, Y. L. Yang, H. U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, P. Ca Boya, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. Rodriguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P. M. Allard, P. Phapale, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat Biotechnol*, 34(8):828–37, 2016.

[240] A. D. Washburne, J. D. Silverman, J. W. Leff, D. J. Bennett, J. L. Darcy, S. Mukherjee, N. Fierer, and L. A. David. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*, 5:e2969, 2017.

[241] M Waskom. Seaborn: statistical data visualization. 2012.

[242] M. Watanabe, S. M. Houten, C. Mataki, M. A. Christoffolete, B. W. Kim, H. Sato, N. Messaddeq, J. W. Harney, O. Ezaki, T. Kodama, K. Schoonjans, A. C. Bianco, and J. Auwerx. Bile acids induce energy expenditure by promoting intracellular thyroid hormone activation. *Nature*, 439(7075):484–9, 2006.

[243] J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S.

Moore, J. Laskin, N. Bandeira, and P. C. Dorrestein. Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci U S A*, 109(26):E1743–52, 2012.

[244] A. Weingarden, A. Gonzalez, Y. Vazquez-Baeza, S. Weiss, G. Humphry, D. Berg-Lyons, D. Knights, T. Unno, A. Bobr, J. Kang, A. Khoruts, R. Knight, and M. J. Sadowsky. Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent clostridium difficile infection. *Microbiome*, 3:10, 2015.

[245] A. R. Weingarden, C. Chen, A. Bobr, D. Yao, Y. Lu, V. M. Nelson, M. J. Sadowsky, and A. Khoruts. Microbiota transplantation restores normal fecal bile acid composition in recurrent clostridium difficile infection. *Am J Physiol Gastrointest Liver Physiol*, 306(4):G310–9, 2014.

[246] S. Weiss, W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, A. Birmingham, J. A. Cram, J. A. Fuhrman, J. Raes, F. Sun, J. Zhou, and R. Knight. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J*, 10(7):1669–81, 2016.

[247] S. J. Weiss, Z. Xu, A. Amir, K. Peddada, S.; Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. Vazquez-Baeza, A. Birmingham, and R Knight. Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. *PeerJ Pre-print*, 2015.

[248] J. J. Werner, D. Knights, M. L. Garcia, N. B. Scalfone, S. Smith, K. Yarasheski, T. A. Cummings, A. R. Beers, R. Knight, and L. T. Angenent. Bacterial community structures are unique and resilient in full-scale bioenergy systems. *Proc Natl Acad Sci U S A*, 108(10):4158–63, 2011.

[249] J. J. Wesley-Hunt, G. D.; Flynn. Phylogeny of the carnivora: Basal relationships among the carnivoramorphans, and assessment of the position of 'miacoidea' relative to carnivora. *Journal of Systematic Palaeontology*, 3(1):1–28, 2005.

[250] Jos L. Willems, Cassiano Abreu-Lima, Pierre Arnaud, Jan H. van Bemmel, Christian Brohet, Rosanna Degani, Bernard Denis, Jrgen Gehring, Ian Graham, Gerard van Herpen, Hilario Machado, Peter W. Macfarlane, Jrg Michaelis, Spyridon D. Moulopoulos, Paul Rubel, and Christoph Zywietz. The diagnostic performance of computer programs for the interpretation of electrocardiograms. *New England Journal of Medicine*, 325(25):1767–1773, 1991. PMID: 1834940.

[251] B. P. Willing, J. Dicksved, J. Halfvarson, A. F. Andersson, M. Lucio, Z. Zheng, G. Jarnerot, C. Tysk, J. K. Jansson, and L. Engstrand. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology*, 139(6):1844–1854 e1, 2010.

[252] Ben Willing, Jonas Halfvarson, Johan Dicksved, Magnus Rosenquist, Gunnar Järnerot, Lars Engstrand, Curt Tysk, and Janet K. Jansson. Twin studies reveal specific imbalances in the mucosa-associated microbiota of patients with ileal crohn's disease. *Inflammatory Bowel Diseases*, 15(5):653–660, May 2009. 00298.

[253] Ben P. Willing, Johan Dicksved, Jonas Halfvarson, Anders F. Andersson, Marianna Lucio, Zongli Zheng, Gunnar Järnerot, Curt Tysk, Janet K. Jansson, and Lars Engstrand. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology*, 139(6):1844–1854.e1, Dec 2010. 00448.

[254] Edgar S. Wills, Daisy M. A. E. Jonkers, Paul H. Savelkoul, Ad A. Masclee, Marieke J. Pierik, and John Penders. Fecal microbial composition of ulcerative colitis and crohn's disease patients in remission and subsequent exacerbation. *PLoS ONE*, 9(3), Mar 2014. 00030.

[255] G. D. Wu, J. Chen, C. Hoffmann, K. Bittinger, Y. Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight, R. Sinha, E. Gilroy, K. Gupta, R. Baldassano, L. Nessel, H. Li, F. D. Bushman, and J. D. Lewis. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–8, 2011.

[256] H. Wu, E. Esteve, V. Tremaroli, M. T. Khan, R. Caesar, L. Manneras-Holm, M. Stahlman, L. M. Olsson, M. Serino, M. Planas-Felix, G. Xifra, J. M. Mercader, D. Torrents, R. Burcelin, W. Ricart, R. Perkins, J. M. Fernandez-Real, and F. Backhed. Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat Med*, 2017.

[257] X. Xiao, G. Nakatsu, Y. Jin, S. Wong, J. Yu, and J. Y. Lau. Gut microbiota mediates protection against enteropathy induced by indomethacin. *Sci Rep*, 7:40317, 2017.

[258] T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, and J. I. Gordon. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–7, 2012.

[259] K. Yoshikawa, C. Kurihara, H. Furuhashi, T. Takajo, K. Maruta, Y. Yasutake, H. Sato, K. Narimatsu, Y. Okada, M. Higashiyama, C. Watanabe, S. Komoto, K. Tomita, S. Nagao, S. Miura, H. Tajiri, and R. Hokari. Psychological stress exacerbates nsaid-induced small bowel injury by inducing changes in intestinal microbiota and permeability via glucocorticoid receptor signaling. *J Gastroenterol*, 52(1):61–71, 2017.

[260] Vincent B. Young, Laura H. Raffals, Susan M. Huse, Marius Vital, Dongjuan Dai, Patrick D. Schloss, Jennifer M. Brulc, Dionysios A. Antonopoulos, Rose L. Arrieta, John H. Kwon, GAutham Reddy, Nathaniel Hubert, Sharon Grim, Joseph Vineis, Sushila Dalal, Hilary Morrison, Murat Eren, Folker Meyer, Thomas Schmidt, James Tiedje, Eugene Chang, and Mitch Sogin. Multiphasic analysis of the temporal development of the distal gut microbiota in patients following ileal pouch anal anastomosis. *Microbiome*, 1(1):9, Mar 2013. 00016.

[261] I. Youngster, J. Sauk, C. Pindar, R. G. Wilson, J. L. Kaplan, M. B. Smith, E. J. Alm, D. Gevers, G. H. Russell, and E. L. Hohmann. Fecal microbiota transplant for relapsing clostridium difficile infection using a frozen inoculum from unrelated donors: a randomized, open-label, controlled pilot study. *Clin Infect Dis*, 58(11):1515–22, 2014.

[262] U. M. Zanger and M. Schwab. Cytochrome p450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol Ther*, 138(1):103–41, 2013.

[263] Egija Zaura, Bernd W. Brandt, M. Joost Teixeira de Mattos, Mark J. Buijs, Martien P. M. Caspers, Mamun-Ur Rashid, Andrej Weintraub, Carl Erik Nord, Ann Savell, Yanmin Hu, Anthony Coates, Mike Hubank, David Spratt, Michael Wilson, Bart Keijser, and Wim Crielaard. Same exposure but two radically different responses to antibiotics: Resilience of the salivary microbiome versus long-term microbial shifts in feces. *mBio*, 6(6):e01693–01615, 2015.

[264] D. Zeevi, T. Korem, N. Zmora, D. Israeli, D. Rothschild, A. Weinberger, O. Ben-Yacov, D. Lador, T. Avnit-Sagi, M. Lotan-Pompan, J. Suez, J. A. Mahdi, E. Matot, G. Malka, N. Kosower, M. Rein, G. Zilberman-Schapira, L. Dohnalova, M. Pevsner-Fischer, R. Bikovsky, Z. Halpern, E. Elinav, and E. Segal. Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079–94, 2015.

[265] Z. Zhang, J. Geng, X. Tang, H. Fan, J. Xu, X. Wen, Z. S. Ma, and P. Shi. Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. *ISME J*, 8(4):881–93, 2014.

[266] A. Zhernakova, A. Kurilshikov, M. J. Bonder, E. F. Tigchelaar, M. Schirmer, T. Vatanen, Z. Mujagic, A. V. Vila, G. Falony, S. Vieira-Silva, J. Wang, F. Imhann, E. Brandsma, S. A. Jankipersadsing, M. Joossens, M. C. Cenit, P. Deelen, M. A. Swertz, study LifeLines cohort, R. K. Weersma, E. J. Feskens, M. G. Netea, D. Gevers, D. Jonkers, L. Franke, Y. S. Aulchenko, C. Huttenhower, J. Raes, M. H. Hofker, R. J. Xavier, C. Wijmenga, and J. Fu. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, 352(6285):565–9, 2016.

[267] E. G. Zoetendal, J. Raes, B. van den Bogert, M. Arumugam, C. C. Booijink, F. J. Troost, P. Bork, M. Wels, W. M. de Vos, and M. Kleerebezem. The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. *ISME J*, 6(7):1415–26, 2012.

[268] C. Zywietz. A brief history of electrocardiography - progress through technology. *Published Online by Biosigna, Institute for Biosignal Processing and System Analysis.*