THE UNIVERSITY OF CHICAGO

DETERMINANTS OF MICROBIAL COLONIZATION AND RESILIENCE IN THE HUMAN GUT

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON MICROBIOLOGY

BY
ANDREA ROSE WATSON

CHICAGO, ILLINOIS
AUGUST 2021

"There is nothing like looking, if you want to find something. You certainly usually find something, if you look, but it is not always quite the something you were after."

— J.R.R. Tolkien, The Hobbit, or There and Back Again

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 2 supplementary tables are accessible via doi:10.6084/m9.figshare.14138405.

ACKNOWLEDGMENTS

I am deeply thankful for the mentorship of my supervisor, Meren. His faith in me, while perplexing at times, sustained me through the most daunting challenges of graduate school. It is hard to fear failure with a supervisor so devoted to his students' best interests.

I am also grateful to the members of my doctoral thesis committee—Eugene Chang, Bana Jabri, Sam Light, and Howard Shuman—for their support, feedback, and for teaching me to view my work through a multi-disciplinary lens.

This dissertation was enriched by my closest collaborators: Florian Trigodet, Jessika Füssel, Iva Veseli, Christopher Quince, and Thomas Louie. I thank them for believing in this project and for their vital contributions.

My friends and family carried me through the last five years. To my parents, Kevin, Meghan, Taryn, Jay, and others: Thank you. You inspire me and make my life full.

ABSTRACT

Fecal microbiota transplantation (FMT), the process of transferring stool from a donor to a recipient's gastrointestinal tract, collides donor and recipient gut microbial communities chaotically in an already perturbed gut ecosystem. This provides ideal conditions to observe microbial colonization, succession, and competition in the human gut. In this dissertation, I use FMT as a model system in which to study the features of microbes that successfully colonize the human gut ecosystem. I employ genome-resolved metagenomics, high resolution metagenomic read-recruitment analyses, ecological theory, and publicly available data to interrogate the complex ecological and evolutionary events set in motion by FMT. I reveal the ecological forces driving microbial colonization after FMT, identify key metabolic pathways and functions associated with colonization and resilience in perturbed gut environments, and observe rapid genomic structural alterations to the same microbial populations within weeks of colonizing different hosts. Finally, I demonstrate the strength of the computational approaches used in this dissertation to evaluate, inform, and expand the context of bench lab studies through three interdisciplinary side projects. Ultimately, this dissertation bridges a gap between 'omics studies and hypothesis testing by providing targets for mechanistic studies in real world systems to untangle the forces driving microbial gut colonization.

CHAPTER 1

INTRODUCTION

1.1 The human microbiome

Recent estimates suggest that a person is composed, in equal numbers, of both human and bacterial cells [Sender et al., 2016, Messer et al., 2017]. The microbial component of a human, and the many niches that those microorganisms occupy, make up the human microbiome [Berg et al., 2020]. Established at birth, or debatably earlier [Rackaityte et al., 2020], the human microbiome is shaped by genetics [Goodrich et al., 2014, Blekhman et al., 2015], delivery method [Dominguez-Bello et al., 2010], geography [Yatsunenko et al., 2012], age [Yatsunenko et al., 2012, Claesson et al., 2011], environment [Turnbaugh et al., 2010], diet [David et al., 2014b], lifestyle [Jha et al., 2018], medication [Maier et al., 2018, Becattini et al., 2016] and illness [Fan and Pedersen, 2021], and can change on a day to day basis [David et al., 2014a, Priya and Blekhman, 2019]. This dynamic ecosystem of microbes influences maturation and regulation of the human adaptive immune system [Round and Mazmanian, 2009] and provides essential functions for human well-being [Lloyd-Price et al., 2016]. The largest concentration of bacteria in the human microbiome reside in the gastrointestinal tract, where it balances pro- and anti-inflammatory immune responses to maintain immune homeostasis [Round and Mazmanian, 2009]. Disruptions to the microbiome, termed “dysbiosis”, are generally characterized by loss of commensal organisms and a reduction in diversity [Levy et al., 2017]. Such dysbiotic states of the gut microbiome have been associated with a wide range of gastrointestinal, immunological, metabolic, and neurodevelopmental diseases and disorders, including antibiotic refractory *Clostridioides difficile* infection (CDI), inflammatory bowel disease (IBD), irritable bowel syndrome, obesity [Ley et al., 2006], asthma [Fujimura and Lynch, 2015], diabetes [Larsen et al., 2010], depression [Marin et al., 2017], and autism [Wang et al.,

2011, de Vos and de Vos, 2012, Lynch and Pedersen, 2016, Durack and Lynch, 2019]. In light of the myriad associations between the gut microbiota and such diseases, therapeutic modulation of the gut microbial ecosystem has gained significant attention in recent years as one of the major objectives of human microbiome research [Holmes et al., 2012, Vieira et al., 2016, Kim et al., 2019, Quigley and Gajula, 2020].

1.2 Therapeutic modulation of the gut microbiome

During the last century of modern medicine the most common approach to therapeutic modulation of the gut microbiome has been subtractive: the use of antibiotics to kill pathogenic bacteria infecting the human gastrointestinal tract. While antibiotics are essential and life-saving tools of medicine, their overuse can lead to not only antibiotic resistance, but also the loss of beneficial microbes from the gut ecosystem, dysbiosis, and potentially severe disease such as chronic, recurrent CDI: an urgent public health threat with significant morbidity and mortality [Dethlefsen and Relman, 2011, Guh et al., 2020, Laxminarayan et al., 2013, Lessa et al., 2015, Lloyd-Price et al., 2016]. Therefore, therapeutic modulation of the microbiome in the current paradigm of microbiome-awareness has turned to additive approaches, where beneficial microbes are introduced into a patient's gastrointestinal tract [Quigley and Gajula, 2020].

With the exception of CDI, additive microbial therapeutics target complex, non-communicable intestinal diseases that are not usually associated with traditional microbial pathogens [Chow et al., 2011]. The search for therapeutics to restore the gut microbial ecosystem to a state of homeostasis focuses on the use of (1) probiotics and live biotherapeutic products (LBPs, ingestion of live bacteria of selected strains) [Elmer et al., 1996], (2) prebiotics (non-digested food components that stimulate the growth of particular bacteria) [Gibson and Roberfroid, 1995], (3) synbiotics (a combination of probiotics and complementary prebiotics) [Gibson and Roberfroid, 1995], and (4) fecal microbiota transplantation (FMT,

the transfer of stool from a donor into a recipient's gastrointestinal tract) [Eiseman et al., 1958]. Despite the increasing use of these microbial therapies, a complete understanding of the safety, efficacy, and mechanism of these approaches has yet to emerge [Solari et al., 2014, Quera et al., 2014, Suez et al., 2018, Bafeta et al., 2019, DeFilipp et al., 2019, Brüssow, 2019]. The mechanism of action of even the most clinically successful additive microbial therapeutics remains unknown. Even FMT, which has a staggering 90% success rate for the treatment of CDI and has achieved the most clinical success of any additive approach to gut microbiome modulation, has an unknown mechanism of action [van Nood et al., 2013, Cammarota et al., 2015, Kelly et al., 2016, D Goldenberg and Merrick, 2021].

1.3 Fecal microbiota transplantation

FMT-derived benefits are typically attributed to the transfer of live donor microbes from the donor's stool into the recipient's GI tract [Khoruts and Sadowsky, 2016]. Indeed, previous studies have shown the engraftment of donor microbes in recipients after FMT [Li et al., 2016, Smillie et al., 2018, Kumar et al., 2017, Jouhten et al., 2020, Podlesny and Florian Fricke, 2020, Aggarwala et al., 2020], and the potential mechanisms by which those microbes may then alleviate CDI are numerous. Such mechanisms include competition for nutrients between transplanted microbes and *C. difficile*, production of toxins by transplanted microbes which directly kill *C. difficile* cells, microbial-mediated conversion of primary bile acids to secondary bile acids which inhibit *C. difficile* growth, and microbial stimulation of the host mucosal immune system and subsequent gut barrier repair. However, the contribution of these different potential mechanisms in infection clearance is not known even for the treatment of a relatively simple disease with a known cause such as CDI [Khoruts and Sadowsky, 2016, D Goldenberg and Merrick, 2021]. the assumption that the therapeutic success of FMT requires transfer of microbes from donor to recip-

ients is also in question, with one case series showing that the transfer of sterile fecal filtrate samples may also cure CDI [Ott et al., 2017], while another study suggests that engraftment of donor strains into recipient guts can predict positive therapeutic outcomes [Aggarwala et al., 2020]. Furthermore, FMT has had only limited success in the treatment of more complex GI diseases than CDI [Xu et al., 2015, Wang et al., 2019, Knox et al., 2019], as have efforts to distill the microbial component of FMT into defined bacterial consortia for treatment of CDI or more complex diseases [Besselink et al., 2008, Allen et al., 2013, Cuello-Garcia et al., 2015, Olek et al., 2017, Dubberke et al., 2018]. One critical gap in knowledge that must be filled in order to develop more effective, defined additive microbial therapeutics to treat complex diseases is an improved understanding of microbial ecology in the human gut. In order to effectively manipulate the composition and function of gut microbial communities, we need an improved understanding of microbial colonization, succession, competition, and evolution in the human gut [Costello et al., 2012, Messer et al., 2017].

1.4 An ecological framework for microbiome community assembly

The ability of a microbial population to colonize and persist in the adult gut ecosystem depends on multiple interconnected ecological forces [Vellend, 2010, Costello et al., 2012]. Selection can be observed in the human gut ecosystem through associations between microbial community composition and niche-modifying factors such as host genetics, physiology, lifestyle, diet, and medication use [David et al., 2014a, Kurilshikov et al., 2017, Kolde et al., 2018, Rothschild et al., 2018]. These deterministic factors likely influence not only the resident microbial community composition, but also which exogenous microbes are able to colonize [Vellend, 2010]. Dispersal, the order and timing of population arrival into the gut, may also affect which populations are able to colonize through priority effects [Fukami, 2015, Martínez et al., 2018], where earlier colonizers of the human gut

may preempt the niches of later colonizers and prevent their colonization or modify niches they do not occupy in such a way that those niches become more or less suitable for later populations to colonize [Fukami, 2015]. Diversification, the genetic divergence of subpopulations, may amplify priority effects if an earlier colonizer diversifies to occupy new niches through adaptive radiation, pre-empting the niches of later colonizers and increasing colonization resistance [Vellend, 2010, Foster et al., 2017]. At the same time, priority effects can limit the diversification of later colonizers that are constrained by the presence of pre-existing community members [Fukami et al., 2007]. Finally, drift, the stochastic fluctuation in population relative abundances, is a neutral ecological force that affects community composition [Vellend, 2010]. Drift may lead to the exclusion of low abundance populations from a community, and therefore puts low abundance populations at higher risk of exclusion rather than colonization unless they exhibit a competitive advantage [Costello et al., 2012]. Thus, microbial community assembly and evolution in a complex ecosystem like the human gut is driven by deterministic, historical, and neutral factors that govern colonization, succession, and competition through the forces of selection, dispersal, diversification, and drift. Disentangling this multidimensional web of interconnected ecological forces to understand the microbial ecology of community assembly necessitates the use of well-controlled and robust experimental systems that allow for the observation of these dynamics as they unfold *in situ*.

1.5 FMT as a model system to study human gut microbial ecology

The complexity of the gut microbiota provides several challenges for study. Difficulties include the diversity of human lifestyles [David et al., 2014a,b, Lloyd-Price et al., 2019] and the limited utility of model systems to make robust causal inferences for human diseases [Walter et al., 2020]. To investigate functional and metabolic determinants of fitness in the gut environment researchers have employed genome-scale metabolic models [van der

Ark et al., 2017], germ-free mice conventionalized with individual microbial taxa [Lee et al., 2013] or consortia of human microbial isolates [Feng et al., 2020], and human dietary supplements [Martínez et al., 2010]. However, an ideal model system to study microbial ecology of the human gut microbiota exists in plain sight: FMT. While the mechanism of FMT remains unknown and its mixed efficacy raises many questions, FMT enables tracking of individual populations over time and investigation of microbial responses to abruptly changing environmental conditions in a natural system. FMT complements simulations or laboratory models of ecological perturbation by colliding two distinct microbial ecosystems, which offers a powerful framework to study fundamental questions of gut microbial ecology, including the determinants of microbial colonization, succession, and resilience [Schmidt et al., 2018]. The accessibility of human stool samples and the ability to modify the human gut through FMT, as well as the relative stability of the human gut community to perturbation, make FMT an ideal model system. So far FMT has been used to investigate the long-term survival of donor microbes in recipients [Jalanka et al., 2016, Broecker et al., 2016, Moss et al., 2017], characterize correlations between colonization efficiency and taxonomy [Li et al., 2016, Smillie et al., 2018, Podlesny and Florian Fricke, 2020], and discuss the impact of ecological processes that influence colonization outcomes [Lee et al., 2017, Smillie et al., 2018, Podlesny and Florian Fricke, 2020]. However, to fulfill the potential of FMT as a model system for the study of microbial ecology in the human gastrointestinal tract, it is necessary to use methods that can identify and track fundamental units of microbial life as they are transferred from donors to recipients.

1.6 Challenges in tracking fundamental units of microbial life

A complete understanding of microbial ecology through molecular sequencing requires a consensus on the level of resolution to determine cohesive units that can ideally serve as a proxy for taxonomically and functionally homogeneous microbial populations so the

environmental change can be associated with changes in microbial communities. High-throughput sequencing of 16S ribosomal RNA gene amplicons has been one of the most popular approaches to study the microbial ecology of the human gut [Knight et al., 2017]. Modern algorithms that are sensitive even to single-nucleotide differences can partition 16S rRNA gene data into highly-resolved units (e.g., amplicon sequence variants) that are homogeneous at the sequence-level [Eren et al., 2014, Callahan et al., 2017, Amir et al., 2017]. However, the utility of even the most highly-resolved units that can be generated from the amplicon data is limited to track individual microbial populations since microbes with highly divergent genomes and physiologies can carry identical 16S rRNA genes [Jaspers and Overmann, 2004]. Shotgun metagenomics offers an alternative to amplicon sequencing [Quince et al., 2017b], and especially genome-resolved investigations of metagenomes can give access to much higher levels of resolution to track individual microbial populations [Chen et al., 2020]. However, recent metagenomics studies reveal that even microbes that are highly similar given their entire gene content can present distinct ecological behaviors [Zhao et al., 2019, Jiang et al., 2019, Zeevi et al., 2019]. Indeed, subtle genetic variation among microbes created by horizontal gene transfer events or single nucleotide variants have been demonstrated to delineate near-identical microbes into ecologically distinct groups within complex habitats [Denef, 2019, Zeevi et al., 2019, Delmont et al., 2019]. Such genomic alterations that influence functional phenotypes can occur on short evolutionary time scales that cannot be resolved by tracking changes in the genes conserved between environmental populations and existing reference genomes [Ahmed et al., 2008]. Therefore, relying on marker gene sequences or metagenomic read recruitment through reference genomes to track functionally distinct populations [Denef, 2019], especially in natural habitats, may not reveal ecologically relevant distinctions.

1.7 High-resolution genome-resolved metagenomics

Improving our understanding of the fate of donor microbes in recipient guts following additive therapeutic strategies requires long-term tracking of donor microbes with approaches that can resolve subtle changes accurately in diverse hosts. Cultivation and whole-genome sequencing of microbes isolated from the stool samples of FMT donors and FMT recipients allows for the highest resolution level of comparison between strains, allowing for confident determination of donor strain engraftment and observation of very subtle genomic changes [Zhao et al., 2019, Jouhten et al., 2020]. Metagenomic sequencing, the sequencing of the entire DNA content of an environmental sample, is an alternative to cultivation that allows one to track the detection of donor strain genomes in FMT recipients through read recruitment, or mapping, of recipient stool sample metagenome sequences to donor microbial genomes [Aggarwala et al., 2020]. However, these approaches still require cultivation of donor microbes, which is low-throughput and limits study scope to those microbes which can be cultivated under standard laboratory conditions. An alternative approach to cultivation and whole-genome sequencing of donor microbes in genome-resolved metagenomics: where the sequencing reads from donor sample metagenomes are assembled together into longer contiguous sequences, or contigs, which are then binned into microbial genomes based on shared characteristics such as differential coverage across all metagenomic samples and tetranucleotide frequency (Figure 1.1). The quality and completion of these metagenome-assembled genomes (MAGs) can then be measured by the detection and copy-number of single-copy core genes, genes that are common to almost all microbial genomes, and typically appear only once in a genome. High quality MAGs, also referred to as population genomes, provide access to the near-complete genomic sequence of diverse donor microbial populations without the need for cultivation and its potential biases towards more abundant microbes which grow quickly on traditional media. MAGs can then be tracked in recip-

ient samples through metagenomic read-recruitment [Lee et al., 2017]. However, while metagenomic read recruitment can detect the presence of a recipient population with a genome similar enough to a donor population that their short reads map to the same genomic reference, mapping alone does not indicate that the same strain is present in both donor and recipient samples. Single-nucleotide variants, nucleotides in which mapped reads differ from the genomic reference, have been analyzed to detect strains present in different metagenomes, but such approaches have only been applied to unbinned contigs or reference genomes from existing databases, and therefore do not have access to the entire genomic content of the precise environmental populations of interest including their mobile elements, unique polymorphisms, and accessory genes [Li et al., 2016, Smillie et al., 2018, Podlesny and Florian Fricke, 2020]. Therefore a gap remains to be filled: genome-resolved metagenomics with analysis of read-recruitment patterns and single nucleotide variants to achieve strain-resolved observations of the subtle and dynamic microbial ecology of colonization and resilience in the human gut.

1.8 Thesis topics

In this thesis, I use genome-resolved metagenomics, read-recruitment analyses, publicly available data, and ecological theory to identify the determinants of microbial colonization and resilience in the human gut, and to observe the subtle but ecologically significant intra-population dynamics that occur when FMT donor and recipient microbial communities collide. I then demonstrate the power of metagenomics approaches to evaluate, inform, and expand upon bench lab studies through three additional projects.

In Chapter 2, I introduce a longitudinal FMT study of 2 donors and 5 recipients of each donor's stool. I use stool sample metagenomes from this study for genome-resolved metagenomics, resulting in the assembly and binning of 311 novel, high-quality, donor-derived MAGs. I track the detection of these MAGs in donor and recipient samples through

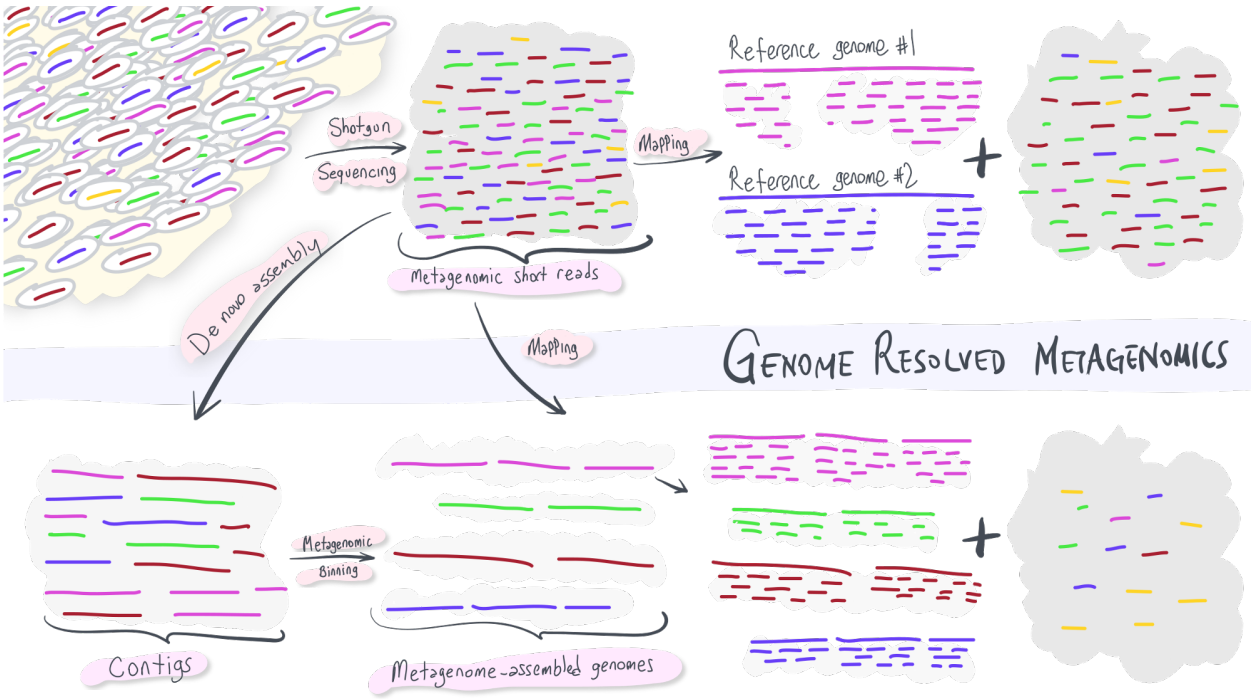


Figure 1.1: Schematic illustrating the difference between recruitment of metagenomic short reads to reference genomes (top), and genome-resolved metagenomics (bottom) [Eren].

metagenomic read-recruitment to determine which donor microbial populations colonized which FMT recipients. This allows me to show that microbial colonization of the recipient gut after FMT is dependent on adaptive, rather than neutral, ecological forces. I then identify metabolic pathways enriched in the superior colonizers, most of which are involved in biosynthesis of essential metabolites, and show that these same metabolic modules are also enriched in MAGs derived from inflamed gut metagenomes. These results indicate that metabolic competence is associated with both colonization and fitness in the gut ecosystem, and that microbes associated with gut inflammation may not be causative of disease, but be selected for under disease states due to their resilience to stress.

In Chapter 3, I build upon the read-recruitment results from Chapter 2 to observe changes in the genomic structure of donor and recipient gut microbial populations as they adapt to the perturbation caused by FMT. I demonstrate that a donor microbial population

that does not colonize a recipient long-term still transfers a prophage to a similar preexisting population in that recipient's gut, and I show multiple instances of donor microbial populations which experienced differential genomic structural rearrangements and mutagenesis events upon colonization of different recipient's gut environments. These results highlight the dynamic and diverse structural alterations that can occur differentially in the same microbial genome under similar environmental conditions, and show that even transient microbial colonizers of the human gut can contribute to the gene content of existing microbial communities.

In Chapter 4, I conclude this dissertation with three examples of how metagenomics analyses can serve bench lab studies. I use metagenomic assembly and read-recruitment to 1) evaluate the completion of a mouse gut isolate collection, 2) identify cultivation conditions supportive of the growth of microbial populations of interest from a previous study, and 3) place a mechanistic study into a global framework to demonstrate its relevance to human health on multiple continents. The findings from these interdisciplinary studies highlight the enormous potential of combining classical microbiology expertise with computational analysis of large 'omics datasets.

CHAPTER 2

METABOLIC COMPETENCY DRIVES MICROBIAL COLONIZATION AND RESILIENCE IN HEALTH AND DISEASE

This chapter is derived from the following publication:

Andrea R Watson, Jessika Fuessel, Iva Veseli, Johanna Zaal DeLongchamp, Marisela Silva, Florian Trigodet, Karen Lolans, Alon Shaiber, Emily Fogarty, Christopher Quince, Michael K Yu, Arda Soylev, Hilary G Morrison, Sonny T M Lee, David T Rubin, Bana Jabri, Thomas Louie, and A Murat Eren. Adaptive ecological processes and metabolicin-dependence drive microbial colonization and resilience in the human gut. *bioRxiv*, March 2021.

2.1 Introduction

The human gut microbiome is associated with a wide range of diseases and disorders [Almeida et al., 2020, Durack and Lynch, 2019, Lynch and Pedersen, 2016]. However, mechanistic underpinnings of these associations have been difficult to resolve in part due to the diversity of human lifestyles [David et al., 2014a] and the limited utility of model systems to make robust causal inferences for microbially mediated human diseases [Walter et al., 2020].

Inflammatory bowel disease (IBD), a group of increasingly common intestinal disorders that cause inflammation of the gastrointestinal tract [Baumgart and Carding, 2007], has been a model to study human diseases associated with the gut microbiota [Schirmer et al., 2019]. The pathogenesis of IBD is attributed in part to the gut microbiome [Plichta et al., 2019], yet the microbial ecology of IBD-associated dysbiosis remains a puzzle. Despite marked changes in gut microbial community composition in IBD [Ott et al., 2004, Sokol and Seksik, 2010, Joossens et al., 2011], the microbiota associated with the dis-

ease lacks traditional pathogens [Chow et al., 2011], and microbes that are found in IBD typically also occur in healthy individuals [Clooney et al., 2021], which complicates the search for robust functional or taxonomic markers of health and disease states [Lloyd-Price et al., 2019]. One of the hallmarks of IBD is reduced microbial diversity during episodes of inflammation, when the gut environment is often dominated by microbes that typically occur in lower abundances prior to inflammation [Vineis et al., 2016]. The sudden increase in the relative abundance of microbes that are common to healthy individuals suggests that the harsh conditions of IBD likely act as an ecological filter that prevents the persistence of low-fitness populations. Yet, in the absence of a complete understanding of the functional drivers of microbial colonization in this habitat, critical insights into the metabolic requirements of survival in IBD remains elusive.

Understanding the determinants of microbial colonization has been one of the fundamental aims of gut microbial ecology [Costello et al., 2012, Messer et al., 2017]. To overcome the difficulties of conducting well-controlled studies with humans, researchers have studied the determinants of microbial colonization of the gut in model systems, such as germ-free mice conventionalized with individual taxa [Lee et al., 2013] or a consortium of human microbial isolates [Feng et al., 2020]. Despite their utility for hypothesis testing, simpler models do not capture the complex ecological interactions fostered by natural systems and thus the insights they yield do not always translate to human gut microbial ecology [Ley et al., 2006, Finucane et al., 2014]. Between the extremes of well-controlled but simple mouse models and complex yet uncontrolled human populations, there exists a middleground that provides a window into the microbial ecology of complex human systems through a controlled perturbation: human fecal microbiota transplantation (FMT), the transfer of stool from a donor into a recipient's gastrointestinal tract [Eiseman et al., 1958].

FMT complements laboratory models of environmental perturbation by colliding two

distinct microbial ecosystems, and thus offers a powerful framework to study fundamental questions of microbial ecology, including the determinants of microbial succession and resilience [Schmidt et al., 2018]. Here we use FMT as an *in natura* experimental model to investigate the ecological and functional determinants of successful microbial colonization of the human gut at the level of individual populations. Our findings suggest that adaptive ecological forces are key drivers of colonization outcomes after FMT, reveal taxonomy-independent metabolic determinants of fitness in the human gut, and demonstrate that similar ecological principles determine resilience of microbes upon colonizing a new environment and under inflammatory conditions.

2.2 Results and Discussion

Our study includes 109 gut metagenomes (Table 2.1) from two healthy FMT donors (A and B) and 10 FMT recipients (five recipients per donor) who had multiply recurrent *Clostridioides difficile* infection (CDI) and received vancomycin for a minimum of 10 days to attain resolution of diarrheal illness prior to FMT. On the last day of vancomycin treatment, a baseline fecal sample was collected from each recipient, and their bowel contents were evacuated immediately prior to FMT. Recipients did not take any antibiotics on the day of transplant, or during the post-FMT sampling period (Figure 2.6). We also collected 24 Donor A samples over a period of 636 days and 15 Donor B samples over a period of 532 days to establish an understanding of the long-term microbial population dynamics within each donor microbiota. We also collected 5 to 9 samples from each recipient up to 336 days post-FMT. Deep sequencing of donor and recipient metagenomes using Illumina paired-end (2x150) technology resulted in a total of 7.7 billion sequences with an average of 71 million reads per metagenome (Figure 2.1, Table 2.1, Table 2.2). We employed genome-resolved metagenomics, pangenomics, and microbial population genetics for an in-depth characterization of donor and recipient gut microbiota using these data, and we

leveraged publicly available gut metagenomes to benchmark our observations.

2.2.1 Many but not all donor microbes colonized recipients and persisted long-term

We first characterized the taxonomic composition of each donor and recipient sample by aligning metagenomic short reads to reference genomes in the NCBI's RefSeq database (Table 2.2). The phylum-level microbial community composition of both donors reflected those observed in healthy individuals in North America [Human Microbiome Project Consortium, 2012]: a large representation of Firmicutes and Bacteroidetes, and other taxa with relatively lower relative abundances, including Actinobacteria, Verrucomicrobia, and Proteobacteria (Figure 2.1, Table 2.2). In contrast, the vast majority of the recipient pre-FMT samples were dominated by Proteobacteria, a phylum that typically undergoes a drastic expansion in individuals treated with vancomycin [Isaac et al., 2017]. After the FMT, we observed a dramatic shift in recipient taxonomic profiles (Table 2.2, Figure 2.7), a widely documented hallmark of this procedure [Khoruts et al., 2010, Grehan et al., 2010, Shahinas et al., 2012]. Nearly all recipient samples post-FMT were dominated by Bacteroidetes and Firmicutes as well as Actinobacteria and Verrucomicrobia in lower abundances, resembling qualitatively, but not quantitatively, the taxonomic profiles of their donors (Table 2.2). For example, even though the median relative abundance of Bacteroidetes populations were 5% and 17% in donors A and B, their relative abundance in recipients post-FMT increased to 33% and 45%, respectively (Figure 2.1, Table 2.2). A single genus, Bacteroides, made up 76% and 82% of the Bacteroidetes populations in the recipients of Donor A and B, respectively (Table 2.2). The success of the donor Bacteroides populations in recipients upon FMT is not surprising given the ubiquity of this genus across human populations throughout the globe [Wexler and Goodman, 2017] and the ability of its members to survive substantial levels of stress [Swidsinski et al., 2005,

Vineis et al., 2016]. This result suggests that FMT outcomes in our dataset are unlikely to be random, and the study design and resulting dataset offers a framework to study ecological principles of the human gut microbiome.

Next, we assembled short metagenomic reads into contiguous segments of DNA (contigs). Co-assemblies of 24 Donor A and 15 Donor B metagenomes independently resulted in 53,891 and 54,311 contigs that were longer than 2,500 nucleotides, and described 0.70 and 0.79 million genes occurring in 179 and 248 genomes, as estimated by the mode of the frequency of bacterial single-copy core genes (Table 2.2). One way to characterize how well a given assembly describes the DNA content of a given metagenome is to calculate the percentage of reads it recruits from the metagenome through read mapping. Donor contigs recruited on average 80.8% of metagenomic reads from donor metagenomes. In contrast, they recruited 43.4% of reads on average from pre-FMT recipient metagenomes. This number increased to 80.2% for recipient metagenomes post-FMT (Figure 2.1), and the donor contigs continued to represent 76.8% of the recipient metagenomes on average even after a year post-FMT (Table 2.2). These read recruitment results suggest that members of the donor microbiota successfully established in recipient guts upon FMT and largely persisted until the end of the sampling period.

Compared to metagenomic short reads, assembled contigs provide a larger genetic context to study microbial metagenomes. However, a sole focus on contigs may yield misleading results [Kowarsky et al., 2017] that can be ameliorated by reconstructing microbial genomes from metagenomic assemblies [Chen et al., 2020]. We reconstructed genomes from co-assembled donor metagenomes by grouping contigs into putative bins based on sequence composition and differential coverage signal as previously described [Sharon et al., 2013, Lee et al., 2017]. We retained bins that were at least 70% complete and had no more than 10% redundancy as predicted by bacterial single-copy core genes [Bowers et al., 2017, Chen et al., 2020] and manually refined them to improve their quality follow-

ing previously described approaches [Delmont et al., 2018, Shaiber et al., 2020]. Our binning resulted in a final list of 128 metagenome-assembled genomes (MAGs) for Donor A and 183 MAGs for Donor B that included members of Firmicutes (n=265), Bacteroidetes (n=20), Actinobacteria (n=14), Proteobacteria (n=7), Verrucomicrobia (n=2), Cyanobacteria (n=2), and Patescibacteria (n=1) (Table 2.3). The taxonomy of donor-derived genomes largely reflected the taxonomic composition of donor metagenomes as predicted by short reads (Figure 2.1, Table 2.2, Table 2.3). While only 20 genomes (mostly of Bacteroides and Alistipes) explained the entirety of the Bacteroidetes group, we recovered 265 MAGs that represented lower abundance but diverse populations of Firmicutes (Figure 2.1, Table 2.2, Table 2.3). We found no difference between the delivery method of FMT for the recipients of donor A, where, on average 45% and 43% of donor genomes emerged in recipients who received donor stool through colonoscopy (n=3) versus pills (n=2), respectively. However, there was an increase in the efficiency of pills for donor B, where on average 25% and 54% of donor genomes emerged in recipients who received donor stool through colonoscopy (n=2) versus pill (n=3) (Figure 2.8).

Reconstructing genomes gave us access to microbial populations in metagenomes through metagenomic read recruitment strategies and enabled us to characterize (1) population-level microbial colonization dynamics before and after FMT using donor and recipient metagenomes and (2) the distribution of each donor population across geographically distributed humans using 1,984 publicly available human gut metagenomes (Table 2.4). As expected, we detected each donor population in at least one donor metagenome (see Methods for 'detection' criteria). Yet, only 16% of Donor A populations were detected in every Donor A sample, and only 44% of Donor B MAGs were detected in every Donor B sample (Figure 2.1, Table 2.3), in agreement with the previously documented dynamism of gut microbial community composition over time [David et al., 2014a]. A marked increase in the detection of donor populations in recipients after FMT echoed

the general pattern of transfer suggested by the short-read taxonomy (Figure 2.1): while only 38% of Donor A and 54% of Donor B populations were detected in at least one recipient pre-FMT, these percentages increased to 96% and 96% post-FMT (Table 2.3). Not every donor population colonized each recipient, but colonization events did not appear to be random: while some donor populations colonized all recipients, others colonized none (Figure 2.1), providing us with an opportunity to resolve colonization events and quantify colonization success for each donor population in our dataset.

Resolving colonization events accurately is a challenging task as multiple factors may influence the ability to determine colonization outcomes unambiguously. These factors include (1) the inability to detect low-abundance populations, (2) inaccurate characterization of transient populations observed immediately after FMT as successful colonization events, (3) the reliance on relative abundance of populations to define colonization events when abundance estimates from stool do not always reflect the abundance of organisms in the GI tract [Yasuda et al., 2015, Sheth et al., 2019], and (4) the failure to distinguish between colonization by a donor population or emergence of a pre-FMT recipient population after FMT (where a low-abundance recipient population that is closely related to one or more donor populations becomes abundant after FMT and is mistaken as a bona fide colonization event). To mitigate these factors, we have (1) employed deep-sequencing of our metagenomes which averaged 71 million reads per sample, (2) implemented a longitudinal sampling strategy, that spanned 376 days on average, to observe donor populations in our recipients long after the FMT, (3) leveraged a 'detection' metric to define colonization events by presence/absence of populations rather than abundance, and (4) employed microbial population genetics to identify and resolve origins of subpopulations. We also developed an analytical approach (Figure 2.9) to determine whether a given donor population has colonized a given recipient based on the detection of donor subpopulations in the transplant sample, in the recipient pre-FMT, and in the recipient post-FMT (see

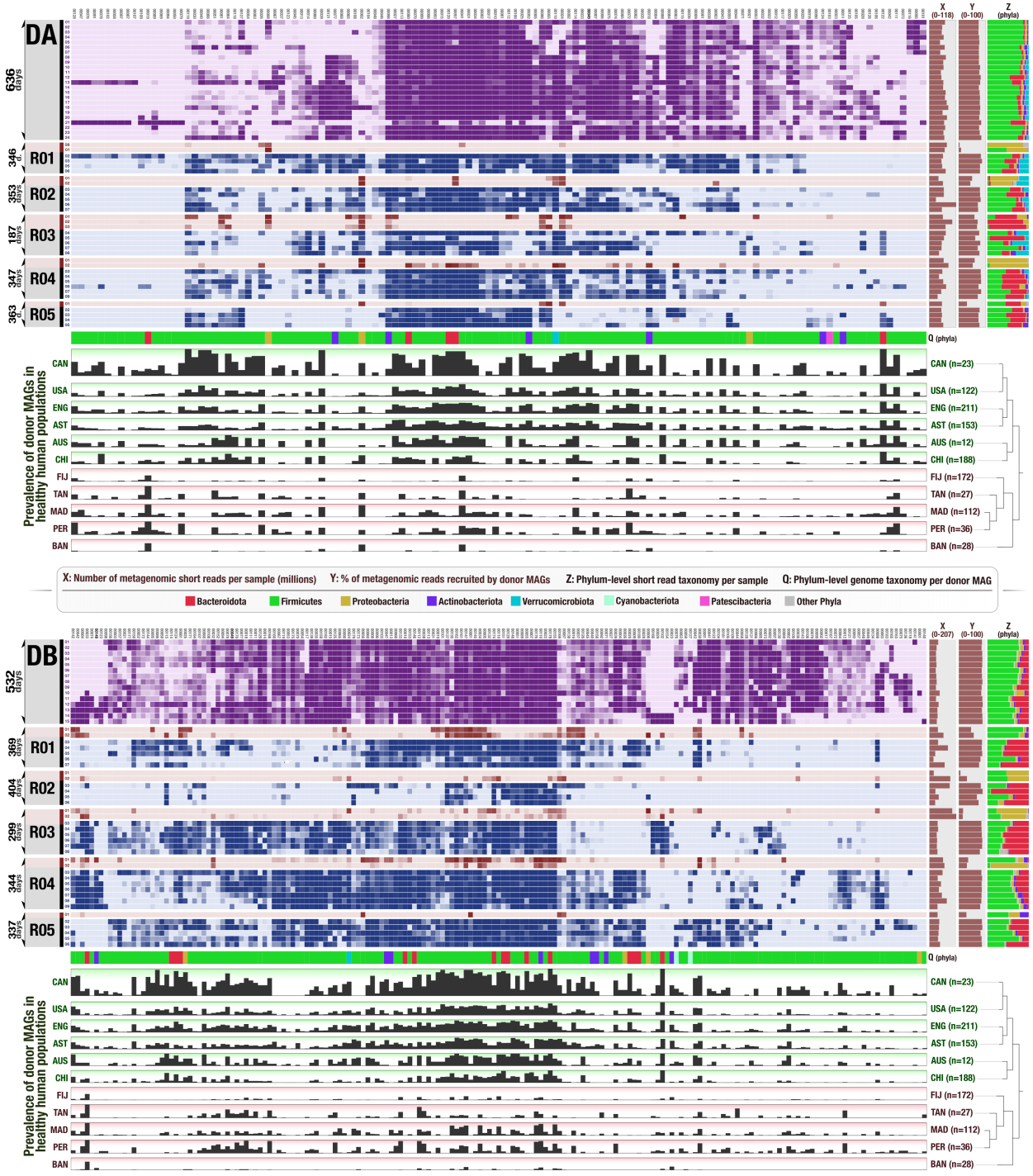


Figure 2.1: FMT donor genomes across recipients and publicly available gut metagenomes.

Figure 2.1 continued: In both heat maps each column represents a donor genome and each row represents a metagenome, and each data point represents the detection of a given genome in a given metagenome. Purple rows represent donor metagenomes which cover 636 days for Donor A and 532 days for Donor B. Each recipient metagenome is colored red for pre-FMT samples and blue for post-FMT samples. The three rightmost columns display for each metagenome (X) the number of metagenomic short reads in millions, (Y) the percent of metagenomic short reads recruited by genomes, and (Z) the taxonomic composition of metagenomes (based on metagenomic short reads) at the phylum level. The row Q provides the phylum-level taxonomy for each donor genome. Finally, the 11 bottom rows under each heat map show the fraction of healthy adult metagenomes from 11 different countries in which a given donor genome is detected (if a genome is detected in every individual from a country it is represented with a full bar). The dendrograms on the right-hand side of these layers organize countries based on the detection patterns of genomes (Euclidean distance and Ward clustering). Red and green shades represent the two main clusters that emerge from this analysis, where green layers are industrialized countries in which donor genomes are highly prevalent and red layers are less industrialized countries where the prevalence of donor genomes is low.

Materials and Methods, Table 2.5). To determine colonization outcomes, we analyzed 640 genome/recipient pairs for Donor A (128 donor genomes in 5 recipients) and identified 99 successful colonization events, 38 failed colonization events, and 503 ambiguous colonization events (Table 2.6). For Donor B, we analyzed 915 genome/recipient pairs (183 donor genomes in 5 recipients) and identified 106 successful colonization events, 109 failed colonization events, and 700 ambiguous colonization events (Table 2.6). Our stringent criteria (see Materials and Methods, Figure 2.9) classified the vast majority of all genome/recipient pairs as ambiguous colonization events. Nevertheless, due to the relatively large number of donor MAGs and FMT recipients in our study, we were left with 352 MAG/recipient pairs with unambiguous phenotypes for downstream analyses.

2.2.2 Adaptive ecological forces are the primary drivers of microbial colonization

The ability of a microbial population to colonize and persist in a complex ecosystem is influenced by both neutral and adaptive forces [Maignien et al., 2014]. Although which of these is the major driver of successful colonization of the human gut remains unclear [Smillie et al., 2018]. In the context of FMT, previous studies have suggested neutral processes to determine colonization success based on the abundance of a microbial population in a donor stool sample [Smillie et al., 2018, Podlesny and Florian Fricke, 2020]. Indeed, ecological drift may have a significant role in a system dominated by neutral processes, where low-abundance donor populations in the transplant would be less likely to be observed in recipients. In contrast, if the system is dominated by adaptive forces, colonization success would be a function of the population fitness in the recipient environment, rather than its abundance in the transplant.

To investigate the impact of neutral versus adaptive processes on colonization in our dataset we first asked whether the prevalence of a donor population in healthy human gut metagenomes, which we define here as a measure of its fitness, was associated with the detection of the same population in donor or recipient metagenomes. Within both FMT cohorts, the mean detection of each population in recipients post-FMT had a stronger association with population fitness than mean detection in donor samples (Figure 2.2a). The fitness of donor A populations explained 4.2% of the variation in mean detection of those populations in donor samples (R^2=0.042, p=0.021) and 19% of variation in mean detection in recipient post-FMT samples (R^2=0.19, p=2.7e-07), an increase of approximately 4.5-fold (Figure 2.2a). Similarly, Donor B population fitness explained 7.3% of the variation in mean detection in donor samples (R^2=0.073, p=2.1e-04), and 36% of the variation in mean detection in recipient post-FMT samples (R^2=0.36, p=4.5e-19), an increase of approximately 5-fold (Figure 2.2a). This suggests that fitness is a better predictor of

colonization outcome than it is of the detection of a population in the donor, suggesting that adaptive forces are likely at play. But detecting a donor population in a recipient post-FMT metagenome through metagenomic read recruitment does not prove colonization, since donor genomes can recruit reads from recipient populations that are closely related (i.e., strain variants) and that were low abundance prior to FMT. Single-nucleotide variants in read recruitment results, however, can reveal such cases [Denef, 2019] and quantify their dynamics [Quince et al., 2017a]. Thus, we developed an improved model that took into consideration the presence and absence of distinct subpopulations in our data and their origins (Figure 2.9). We then used this model to test if colonization success was correlated with population fitness or population dose, which we define here as the relative abundance of a given population in the transplanted donor stool sample. For Donor A populations, colonization outcome was significantly correlated with both dose (Wald test, AUC=0.73, p=7.7e-05) and fitness (Wald test, AUC=0.76, p=6.3e-06) (Figure 2.2b,c). But combining both measures as predictive variables did not substantially improve the performance of our colonization model (AUC=0.82) (Figure 2.2c). This was likely due to the small, but significant, correlation between dose and fitness in Donor A MAGs (R^2=0.053, p=0.0070) (Figure 2.2d). When the fitness of a microbial population is reflected in its relative abundance, the effect of fitness on colonization outcome may be masked by an apparent dose effect. In contrast to Donor A, the fitness of Donor B populations and their relative abundance in Donor B samples were not correlated (R^2=0.0012, p=0.61) (Figure 2.2d), providing us with an ideal case to analyze these two factors independently. Indeed, there was no correlation between dose of a microbial population in Donor B transplant samples and colonization outcome in recipients post-FMT (Wald test, AUC=0.56, p=0.09). Instead, we found a significant correlation between the fitness of each population and the colonization outcome (Wald test, AUC=0.70, p=9.0e-07) (Figure 2.2c).

Taken together, our findings suggest that fitness of a microbial population as measured by its prevalence across global gut metagenomes can predict its colonization success better than its abundance in the donor stool sample, giving credence to the role of adaptive rather than neutral ecological processes in colonization. This finding contrasts with previous studies which suggested that the abundance of a given population in the donor sample was an important determinant of colonization [Smillie et al., 2018, Podlesny and Florian Fricke, 2020]. However, these analyses included many recipient samples collected less than one week after FMT and it is likely that their observations were influenced by the presence of transient populations. Indeed, samples collected immediately after FMT are more likely to inflate the number of colonization events, whereas longitudinal sampling over a longer time course can distinguish transient populations from those that successfully colonized the recipients. We cannot definitively test this hypothesis as we sampled most of our recipients a week after FMT. Still, on average 12% of the donor populations detected in our recipients a week after FMT were no longer detected after a month (Figure 2.1, Table 2.3). Overall, our stringent criteria to determine colonization outcome and the extended post-FMT sampling period likely enabled us to study the long-term engraftment of successful and potentially low-abundance colonizers, instead of high-abundance transient populations that may be dominant directly after FMT.

Accurately distinguishing the role of dose versus fitness in colonization success is further compounded by the fact that microbial populations that are prevalent across human populations may also tend to be more abundant. This is well illustrated by Donor A. Fortunately, the abundant populations in Donor B did not reflect prevalent microbes in healthy adult guts, which demonstrated the importance of fitness as a determinant of colonization success compared to dose without the confounding effect of a correlation between fitness and dose. Thus, it is a theoretical possibility that colonization success is purely driven by adaptive forces and is not influenced by dose, at all. However, while our data assign a

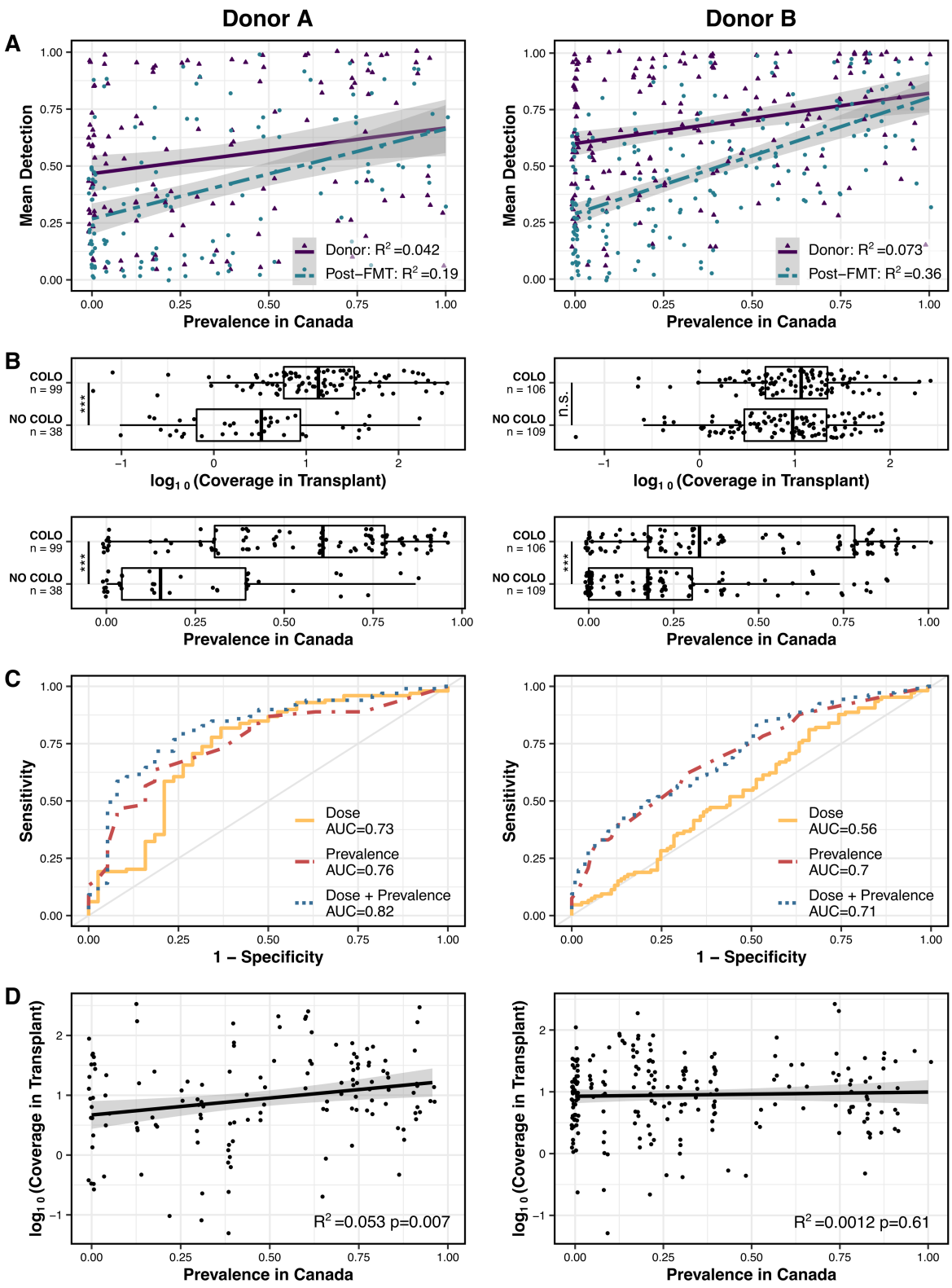


Figure 2.2: Relationships between dose, prevalence, and colonization outcome.

Figure 2.2 continued: Left: Donor A. Right: Donor B. a) Linear regression models of mean detection of each MAG in either donor or recipient post-FMT samples as a function of prevalence. b) Colonization outcome of MAG/recipient pairs as a function of MAG dose or MAG prevalence. Significance calculated by Wald test. c) Receiver operator curves (ROCs) for logistic regression models of colonization. d) Linear regression models of dose as a function of prevalence.

larger role to adaptive forces with confidence, a more accurate determination of the proportional influence of adaptive versus neutral processes in colonization requires a much larger dataset.

2.2.3 Colonizer and resilient microbes are enriched in metabolic pathways for the biosynthesis of essential organic compounds

Fitness in a specific environment is conferred to an organism by a combination of functional traits. In the human gut, such traits drive microbial community succession and structure as a response to changing host diet and lifestyle [Koenig et al., 2011, Rothschild et al., 2018]. Priority effects influence which microbes are able to colonize at different stages of ecosystem assembly. Before FMT, when a recipient's gut is depleted through the use of broad-spectrum antibiotics such as vancomycin, one could argue that the lack of typical microbes associated with healthy gut ecosystems may reduce intra-microbe resource competition and facilitate the colonization and primary succession of low-fitness microbes. However, it is also possible that in some cases a dearth of healthy gut-associated microbes prior to FMT may make colonization more challenging, as some such microbes may be keystone members of the ecosystem that provide resources or otherwise modify the environment in a way that is beneficial to secondary successors. It is therefore impossible to state broadly if the pre-FMT recipient gut environment is more or less hospitable to incoming microbes than a healthy adult gut ecosystem, as this could vary widely depending on the functional traits and niches occupied by both resident and colonizing

microbial populations. Our observation that suggests a primary role of adaptive ecological processes in colonization outcome implies that within our study, the conditions of the recipient pre-FMT gut environment are sufficiently stressful to incoming microbial populations that only a specific subset of high-fitness microbes are able to successfully colonize. We therefore next sought to identify genetic determinants of colonization. For this, we leveraged our access to donor microbial population genomes and global metagenomes to investigate whether a functional enrichment analysis could reveal predictors of success independent of taxonomy.

To generate metabolic insights into colonization success we divided our donor populations into 'high-fitness' and 'low-fitness' groups by considering both their prevalence in FMT recipients and prevalence across global gut metagenomes (Materials and Methods). The 'high-fitness' group included the microbial populations that colonized or persisted in all FMT recipients and were the most prevalent in gut metagenomes from Canada. We assumed that they represented a set of highly fit microbial populations as (1) they were able to colonize human gut environments systematically, (2) they persisted in these environments long-term regardless of the host genetics or lifestyle, and (3) they were prevalent in gut metagenomes outside of our study. In comparison, the 'low-fitness' group comprised microbial populations that failed to colonize or persist in at least three FMT recipients. These populations were nevertheless viable gut microbes as not only our long-term sampling of the donors systematically identified them but also, they sporadically colonized some FMT recipients. Yet, unlike those in the high-fitness group, the distribution patterns of low-fitness populations were sparse, not only within our cohort, but also within publicly available metagenomes. In fact, low-fitness populations were less prevalent than high-fitness genomes in each of the 17 different countries we queried, and in countries including United States, Canada, Austria, China, England, and Australia, we detected high-fitness populations in 5 times more people than low-fitness genomes

in the same country (Figure 2.1, Table 2.3). Overall, we conservatively categorized 20 populations in each group for downstream analyses (Table 2.7). All populations in the low-fitness group resolved to Firmicutes. The high-fitness group was also dominated by Firmicutes (15 of 20) but it also included four Bacteroidetes and one Actinobacteria (Table 2.7). Genome completion estimates did not differ between high and low-fitness groups (Wilcoxon rank sum test, p=0.42) and averaged to 91% and 93%, respectively. However, genome sizes between the two groups differed dramatically (p=2.9e-06), where high-fitness group genomes averaged to 2.8 Mbp while low-fitness group genomes averaged to 1.6 Mbp. These results suggest that the length difference between genomes in high and low-fitness groups is likely to have biological relevance. Indeed, we found a very high correspondence between the lengths of our MAGs and their best matching reference genomes in the GTDB (r=0.88, p=5e-14) (Table 2.7).

Our metabolic enrichment analysis revealed 33 KEGG pathway modules, each containing genes that form a functional unit in a metabolic pathway. Every module that was enriched differentially between these two groups was enriched in the high-fitness group. The lack of any enriched modules in the low-fitness group is in line with the reduction in genome lengths in the low-fitness group and further suggests that the reduction is associated with the absence of metabolic modules. Of all enriched modules, 79% were modules related to biosynthesis, which indicates an overrepresentation of biosynthetic capabilities in the high-fitness group as KEGG modules for biosynthesis only make up 55% of all KEGG modules (Figure 2.3, Table 2.7). Of the 33 enriched modules, 48.5% were associated with amino acid metabolism, 21.2% with vitamin and cofactor metabolism, 18.2% with carbohydrate metabolism, 6% with lipid metabolism and 3% with energy metabolism (Table 2.7). Metabolic modules that were enriched in the high-fitness group included the biosynthesis of seven of the nine essential amino acids, indicating the importance of metabolic competency to synthesize high-demand compounds as a factor increas-

ing fitness in colonizing new gut environments (Table 2.7). This is further supported by the enrichment of biosynthesis pathways for the essential cofactor vitamin B12 (cobalamin), which occurred in 67.5% of the high-fitness populations and only 12.5% of the low-fitness group (Table 2.7). Vitamin B12 is structurally highly complex and costly to produce, requiring expression of more than 30 genes that are exclusively encoded by bacteria and archaea [Martens et al., 2002]. Thus, the competitive advantages conferred by metabolic autonomy appear to outweigh the additional costs. In addition to the biosynthesis of tetrahydrofolate, riboflavin, and cobalamin, the high-fitness group had a larger representation of biosynthetic modules for vitamins including biotin, pantothenate, folate, and thiamine (Table 2.7), micronutrients that are equally important in bacterial and human metabolism and are shown to play important roles in mediating host-microbe interactions [Biesalski, 2016]. Interestingly, enriched metabolic modules in our analysis partially overlap with those that Feng et al. identified as the determinants of microbial fitness using metatranscriptomics and a germ-free mouse model conventionalized with microbial isolates of human origin [Feng et al., 2020].

Even though enriched metabolic modules occurred mostly in high-fitness populations, we did find some of these modules in the low-fitness group as well (Table 2.7), but their distribution was not uniform as they primarily occurred only in a subset of genomes that resolved to Firmicutes (Figure 2.3). We then sought to identify whether the levels of completion of these modules that occurred in both groups were identical. For this, we matched six low-fitness genomes that encoded modules enriched in high-fitness group genomes to six high-fitness genomes from the same phylum (marked as HF and LF subgroups in Figure 2.3). Bacterial single-copy core genes estimated that genomes in both subgroups were highly complete with a slight increase in average completion of low-fitness genomes (93.7%) compared to high-fitness genomes (90.1%). Despite the higher estimated genome completion for low-fitness populations, estimated metabolic module com-

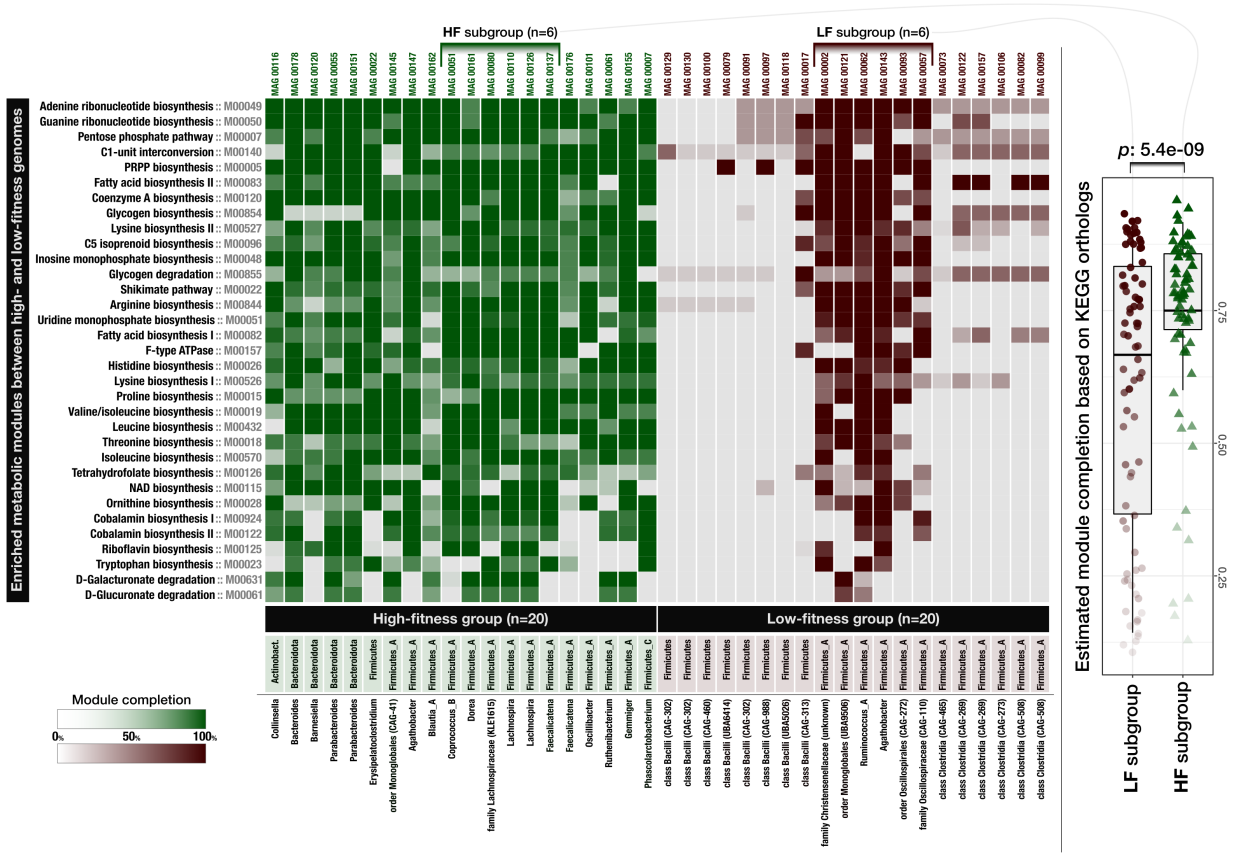


Figure 2.3: Distribution of metabolic modules across low and high-fitness genomes. Each data point in this heat map shows the level of completion of a given metabolic module (rows) in a given genome (columns). The box-plot on the right-side compares a subset of low-fitness (LF) and high-fitness (HF) genomes, where each data point is the level of completion of a given metabolic module in a genome and shows a statistically significant difference between the overall completion of metabolic modules between these subgroups (Wilcoxon rank sum test, p=5.4e-09).

pletion values were significantly lower in the low-fitness group (Wilcoxon rank sum test with continuity correction, V=958, p=5e-09) (Figure 2.3, Table 2.7). This indicates that even when modules that are associated with high-fitness were detected in low-fitness genomes, they were systematically missing genes and were less complete than the same modules in high-fitness genomes.

2.2.4 While gut microbial ecosystems of healthy individuals include both low- and high-fitness microbes, IBD primarily selects for high-fitness populations

Our results so far show that while the healthy donor environment could support both high-fitness and low-fitness populations (Figure 2.1, Table 2.3), challenging microbes to colonize a new environment or to withstand massive ecosystem perturbation during FMT selects for high-fitness populations (Figure 2.3, Table 2.7), suggesting that metabolic competence is a more critical determinant of fitness during stress than during homeostasis. Based on these observations, it is conceivable to hypothesize that (1) a gut environment in homeostasis will support a range of microbial populations with a wide spectrum of metabolic competency, and (2) a gut environment under stress will select for high metabolic competency in microbial populations.

To test these hypotheses, we compared genomes reconstructed from a cohort of healthy individuals [Pasolli et al., 2019] to genomes reconstructed from individuals who were diagnosed with inflammatory bowel disease (IBD). Our IBD dataset was composed of two cohorts: a set of patients with pouchitis [Vineis et al., 2016], a form of IBD with similar pathology to ulcerative colitis [De Preter et al., 2009], and a set of pediatric Crohn's disease patients [Quince et al., 2015]. The number of genomes per individual and the average level of genome completeness per group were similar between healthy individuals and those with IBD: overall, our analysis compared 264 genomes from 22 healthy individuals with an average completion of 90.4%, 44 genomes from 4 pouchitis patients with an average completion of 89.2% and 256 genomes from 12 Crohn's disease patients with an average completion of 94.1% (Table 2.8). Intriguingly, similar to the length differences between genomes of high-fitness and low-fitness populations (2.8 Mbp versus 1.6 Mbp on average), microbial populations associated with IBD patients had larger genomes

compared to healthy people and averaged to 3.0 Mbp versus 2.6 Mbp, respectively (Table 2.8). This suggests that despite the comparable levels of completion of microbial genomes from the healthy cohort, these genomes tended to be smaller in size compared to those reconstructed from individuals with IBD.

Next, we asked whether the completion of those metabolic modules associated with colonization success and resilience differed between the genomes reconstructed from healthy and IBD individuals. The level of completion of the 33 metabolic modules were almost identical between high-fitness genomes and genomes from IBD patients (Wilcoxon rank sum test, p=0.5), but genomes from healthy individuals were significantly less complete compared to high-fitness genomes (Wilcoxon rank sum test, $p < 1e\text{-}07$) as well as genomes from IBD patients (Wilcoxon rank sum test, $p < 1e\text{-}07$) (Figure 2.4, Table 2.8). Metabolic modules with the largest differences in completion between genomes from healthy and IBD individuals included biosynthesis of cobalamin, arginine, ornithine, tryptophan, isoleucine as well as the Shikimate pathway (Figure 2.4, Table 2.8), a seven step metabolic route bacteria use for the biosynthesis of aromatic amino acids (phenylalanine, tyrosine, and tryptophan) [Herrmann and Weaver, 1999].

Our findings show that the same set of key metabolic modules that distinguish high-fitness and low-fitness populations in FMT were also differentially associated with populations that occurred in healthy individuals compared to IBD patients. In particular, while healthy individuals seem to harbor microbes with a broad range of metabolic competency, individuals who suffer from two different forms of IBD appear to harbor organisms with higher metabolic autonomy. It is conceivable that a stable gut microbial ecosystem is more likely to support low-fitness populations through metabolic cross-feeding, where vitamins, amino acids, and nucleotides are exchanged between microbes [D'Souza et al., 2018]. In contrast, host-mediated environmental stress in IBD likely disrupts such interactions and creates an ecological filter that selects for metabolic competence, which subsequently

leads to loss of diversity and the dominance of organisms with large genomes that are not necessarily abundant in states of homeostasis.

These observations have implications on our understanding of the hallmarks of healthy gut environments from an ecological point of view. Defining the 'healthy gut microbiome' has been a major goal of human gut microbiome research [Bäckhed et al., 2012], and remains elusive [Eisenstein, 2020]. Despite comprehensive investigations that considered core microbial taxa [Arumugam et al., 2011, Lloyd-Price et al., 2016] or guilds of microbes that represent coherent functional groups [Wu et al., 2021], the search for 'biomarkers' of healthy microbiomes is ongoing [McBurney et al., 2019]. Given our data we hypothesize that one of the defining features of a healthy gut environment is its ability to support a diverse community of microbes with a broad spectrum of metabolic competence, where both low-fitness and high-fitness populations live in a coherent ecosystem. Conversely, an enrichment of metabolically competent high-fitness populations would likely indicate the presence of environmental stress. Our analyses demonstrate that this is a quantifiable feature of microbial communities through genome-resolved metagenomic surveys. Our analyses have limitations. For instance, metabolic insights in our study have been limited to genomic potential and have considered only well-known metabolic pathways, which, given the extent of the unknown coding space in microbial genomes [Vanni et al., 2020], are likely far from complete. As a result, the disproportional enrichment of biosynthetic modules in high-fitness genomes indicates that the ability to synthesize essential biological compounds is necessary but likely insufficient to survive environmental stress in the gut. Nevertheless, the finding that the same metabolic modules that promote colonization success after FMT are also the hallmarks of fitness in IBD suggests the presence of ecological principles that are shared between these systems and warrants deeper investigation.

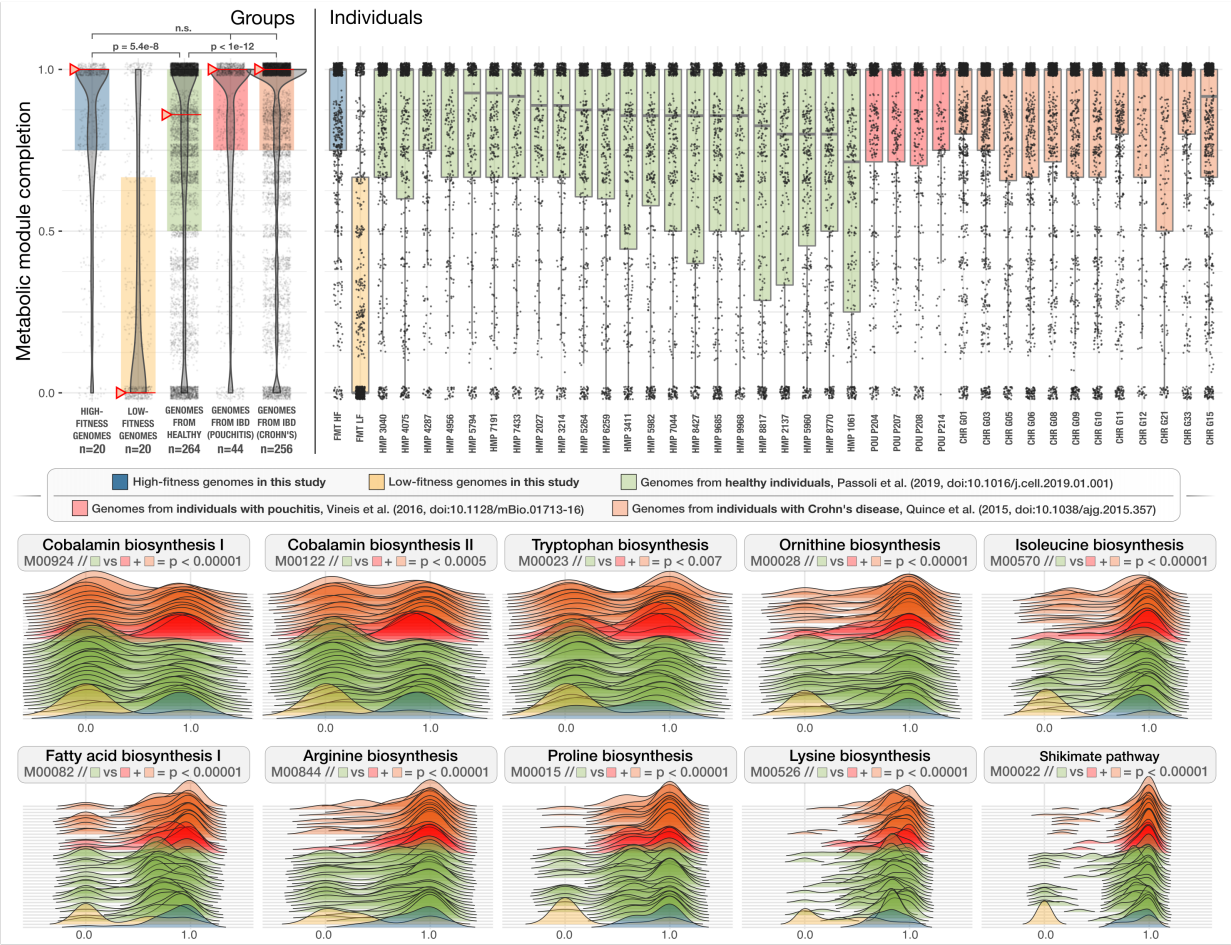


Figure 2.4: Distribution of metabolic modules in genomes reconstructed from healthy individuals and individuals with IBD. The top panel shows the metabolic module completion values for (1) high- and (2) low-fitness donor genomes identified in this study (blue and yellow), (3) genomes from healthy individuals (green), and (4) genomes from individuals with pouchitis (red) and Crohn's disease (orange). Red whiskers in group averages indicate the median. Next to group averages shown the distribution of metabolic modules for each individual. Each dot in a given box-plot represents one of 33 metabolic modules that were enriched in high-fitness FMT donor populations and the y-axis indicates its estimated completion. In the bottom panel the completion values for 10 of the 33 pathways demonstrated as ridge-line plots. Each plot represents a single metabolic module where each layer corresponds to an individual, and the shape of the layer represents the completion of a given metabolic module across all genomes reconstructed from that individual.

2.2.5 Subtle differences in key functions distinguish populations of the same genus with differential colonization success

While adaptive processes that favor metabolic independence explain the determinants of colonization and resilience for distantly related taxa, metabolic features that promote high-fitness at this broad level may not explain differences in fitness between more closely related taxa, such as distinct species within a single genus, which are likely to have similar metabolic capabilities [Martiny et al., 2013] due to unifying ecological traits in higher ranks of taxonomy [Philippot et al., 2010]. We finally investigated whether we could identify determinants of fitness across metabolically similar populations with different levels of success in colonizing unrelated individuals.

Members of the genus *Bifidobacterium* have long been used as probiotics [Gomes and Malcata, 1999] and are prevalent occupants of the healthy human gut microbiota [Arboleya et al., 2016]. In our dataset, *Bifidobacterium* was the second most abundant genus (14.1%) after Bacteroides (15.8%) in Donor A, from whom we reconstructed three MAGs over 98% completion that resolved to three distinct species in this genus: *B. longum* (DA_MAG_00052), *B. adolescentis* subsp. *adolescentis* (DA_MAG_00018), and *B. animalis* subsp. *lactis* (DA_MAG_00011, Table 2.3). While each of these *Bifidobacterium* populations occurred in Donor A metagenomes in a relatively stable fashion, they showed vastly different colonization efficiency upon FMT (Figure 2.5), enabling us to investigate determinants of colonization among closely related taxa.

In contrast to the *B. longum* and *B. adolescentis* subsp. *adolescentis* (henceforth *B. adolescentis*) populations that colonized most recipients, *B. animalis* subsp. *lactis* (henceforth *B. lactis*) did not seem to have colonized any of our recipients (Table 2.3). Overall, we were able to detect *B. longum*, *B. adolescentis*, and *B. lactis* populations in 83%, 75%, and 4% of all post-FMT recipient metagenomes, respectively (Figure 2.5). Most strikingly, patterns of colonization that emerged from the analysis of FMT recipients reflected

those seen in publicly available gut metagenomes from Canada, where *B. longum*, *B. adolescentis*, and *B. lactis* populations occurred in 74%, 39%, and 13% of the population, demonstrating a positive relationship (Pearson's correlation of 0.9, n.s.) between the colonization efficiency upon FMT and the fitness of these populations. Furthermore, the gut metagenomes from 17 countries confirmed the substantially reduced fitness of *B. lactis* globally (Table 2.9). Interestingly, the *B. lactis* MAG we reconstructed from Donor A was virtually identical (with over 99.99% sequence identity over 99.82% alignment, Table 2.9) to most *B. lactis* strains that are widely used as probiotics [Jungersen et al., 2014], revealing a disagreement between the preferences of commercial microbial therapeutics and human gut microbial ecology.

To identify factors that may explain the differences in colonization success between *B. longum*, *B. adolescentis*, and *B. lactis*, we created a collection of *Bifidobacterium* genomes that, in addition to the three metagenome-assembled genomes we reconstructed, included 31 complete genomes obtained from the NCBI (within-group and across-groups average gANI of 98.9% and 77.3%, respectively) (Table 2.9). All three groups of *Bifidobacterium* genomes encoded the majority of the metabolic pathways associated with the high-fitness group (63% ± 5%). However, missing pathways were not uniformly distributed across three: *B. lactis* lacked the largest fraction of these pathways (42%) compared to the more prevalent *B. adolescentis* (36%) and *B. longum* (33%) (Table 2.9). *B. longum* and *B. adolescentis* carried the complete tetrahydrofolate (vitamin B9) biosynthesis pathway in agreement with previous metabolic descriptions of *Bifidobacterium* [D'Aimmo et al., 2012, Sugahara et al., 2015] which qualifies this group as attractive probiotics [Strozzi and Mogna, 2008, Pompei et al., 2007], but this pathway was absent in *B. lactis* genomes. We also found that *B. longum* and *B. adolescentis* genomes encoded histidine biosynthesis which *B. lactis* lacked (Table 2.9). Finally, the average genome lengths of *B. longum* (2.31 Mbp) and *B. adolescentis* (2.18 Mbp) were longer than the average genome length of *B.*

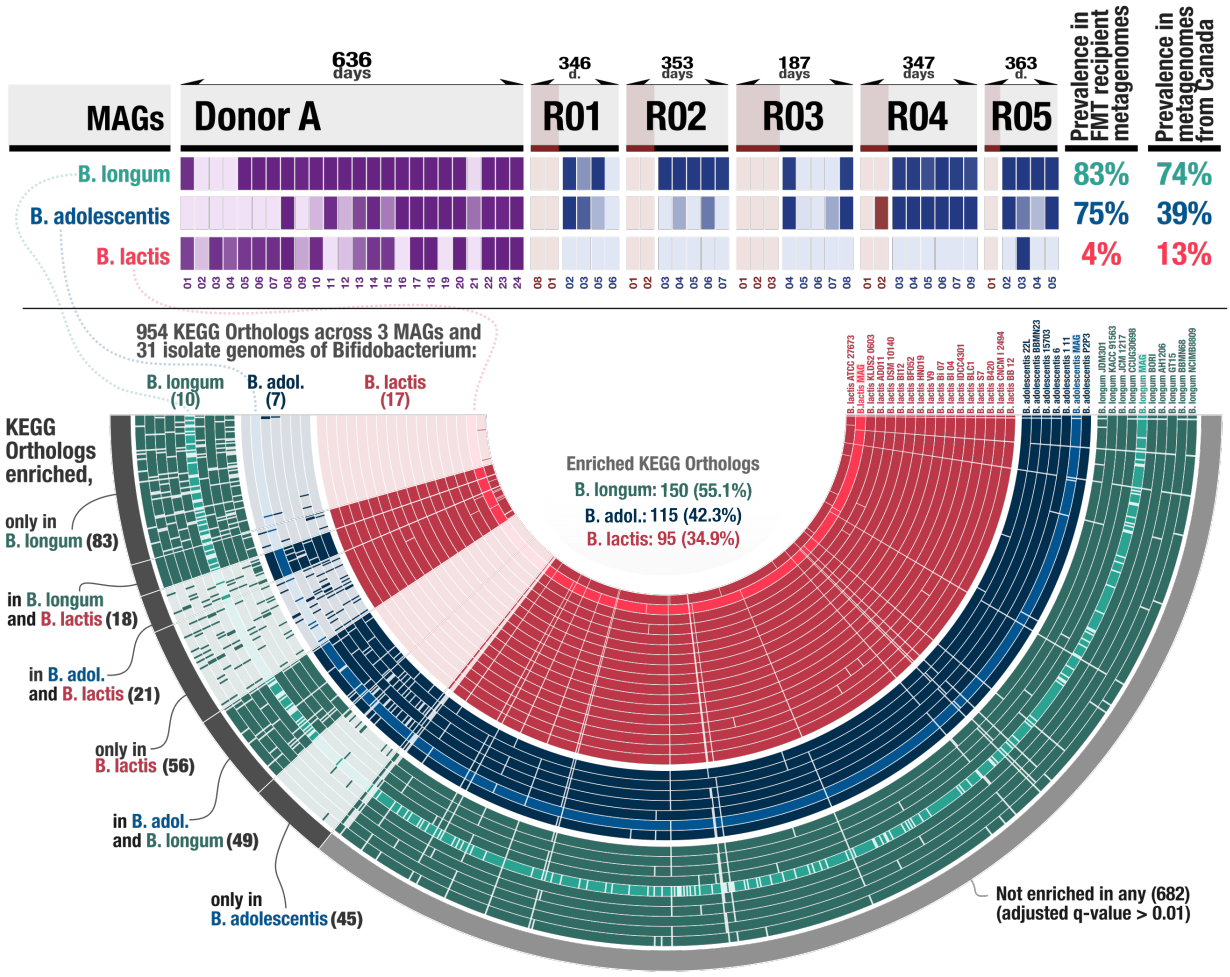


Figure 2.5: Characteristics of three *Bifidobacterium* species. Top panel shows the distribution of Donor A MAGs that represent three distinct *Bifidobacterium* populations across donor and recipient metagenomes before and after FMT. The last two columns in this panel show the prevalence of these populations in post-FMT metagenomes, and publicly available gut metagenomes from Canada. The panel below displays the distribution of KEGG orthologs across the three *Bifidobacterium* MAGs along with 31 high-quality isolate genomes from the NCBI. Each item shown in concentric circles represents a single function assigned by the database of KEGG Orthologs, and each layer is a distinct genome. The intensity of color indicates the presence of a given function in a given genome. The most outer circle marks groups of functions that are enriched in various groups of *Bifidobacterium* genomes as well as those functions that are not enriched in any group as they are either in all genomes, or only a very small number of them.

lactis (1.94 Mbp), which reflects the pattern we observed previously where high-fitness populations tended to have larger genomes. In summary, even though all *Bifidobacterium* genomes in our pangenome had a higher metabolic overlap with one another compared to high-fitness and low-fitness genomes we have previously studied, the reduced fitness of *B. lactis* compared to *B. longum* and *B. adolescentis* could still be explained by the absence of a small number of metabolic competencies associated with the high-fitness group genomes.

Next, we focused on the enrichment of individual functions across the three groups of genomes using gene annotations from KOfam profiles [Aramaki et al., 2020] from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa and Goto, 2000] and Clusters of Orthologous Groups (COGs) from the NCBI [Galperin et al., 2021]. Of all 954 unique KOfams found in our *Bifidobacterium* pangenome, 272 functions were not common to all genomes but statistically enriched in either one or two groups. Our analysis of these accessory functions showed that *B. longum* encoded 150 (55.5%), *B. adolescentis* encoded 115 (42.3%), and *B. lactis* encoded 95 (34.9%) of all accessory functions that were statistically enriched (Figure 2.5, Table 2.9). The same analysis with 1,286 unique COGs confirmed these observations: of all 353 COGs enriched in any group, *B. longum* encoded 212 (60.1%), *B. adolescentis* encoded 172 (48.7%), and *B. lactis* encoded 118 (33.4%) (Table 2.9). Overall, these results reveal a striking correlation between the number of accessory functions associated with *B. longum*, *B. adolescentis*, and *B. lactis*, and echo the absence of metabolic pathways in *B. lactis* even at the level of accessory gene functions, explaining their differential ability to colonize new individuals and distribution in global human gut metagenomes.

We finally investigated the contents of the differentially occurring accessory functions to speculate on whether they could be related to differences in fitness. For instance, in contrast to all *B. longum* and *B. adolescentis* in the *Bifidobacterium* pangenome, none

of the *B. lactis* genomes encoded a phosphoenolpyruvate phosphotransferase (PEP-PTS) system specific for the uptake of β-glucoside (Table 2.9). As the genus *Bifidobacterium* is characterized by a large array of genes associated with carbohydrate uptake and metabolism [Ventura et al., 2009, Schell et al., 2002, Kleerebezem and Vaughan, 2009], *B. lactis* represents a notable exception with a lower number of genes associated with carbohydrate metabolism, fewer genes encoding carbohydrate-specific ABC transporters, and the absence of phosphoenolpyruvate-phosphotransferase (PEP-PTS) systems [Barrangou et al., 2009]. The absence of any other PEP-PTS system in *B. longum* and *B. adolescentis* may indicate the catabolic niche occupied by these populations in the human gut that is shaped by their extensive capacity for uptake and metabolism of plant derived glycosides [Chien et al., 2006, Schell et al., 2002]. Additional functions that exclusively occurred in *B. adolescentis* and *B. longum* genomes included two multidrug resistance pumps of the 'multidrug and toxin extrusion' (MATE) type, three transporters of the major facilitator superfamily (MFS) involved in bile acid tolerance and macrolide efflux, two bile acid:natrium ion symporters, and one proton/chloride ion antiporter conferring acid tolerance (Table 2.9). The drug defense mechanisms may act to protect these populations during periods of inflammation and drug administration, but may also be beneficial with regard to the common ingestion of antibiotics through various food products [Kirchhelle, 2018]. These results show that in the microbial fitness landscape of the human gut, where the determinants of success across distantly related taxa are primarily defined by the presence or absence of a large number of metabolic pathways, there exists smaller niches equally accessible to closely related organisms with similar metabolic potential, among which success can be speculated by subtle differences in key functions.

2.3 Conclusions

Our study points to adaptive ecological processes as primary determinants of both long-term colonization after FMT and microbial fitness in the human gut environment through metabolic competency as conferred by biosynthesis of nucleotides, amino acids, and essential micronutrients. Even when we found these metabolic modules in low-fitness populations, they were systematically less complete compared to their high-fitness counterparts. Our findings suggest that in a healthy gut environment high- and low-fitness populations co-occur in harmony, with their differential fitness indiscernible by taxonomy or relative abundance. However, transfer to a new gut environment through FMT, or host-mediated stress through IBD, initiates an ecological filter that selects for high-fitness populations that can self-sustain. This model offers a null hypothesis to explain how low-abundance members of healthy gut environments can come to dominate the gut microbiota under stressful conditions, while not being causally associated with disease states. If the association between particular microbial taxa and disease is solely driven by their superior metabolic competence, microbial therapies that aim to treat complex diseases by adding microbes associated with healthy individuals will be unlikely to compete with the adaptive processes that regulate complex gut microbial ecosystems.

2.4 Materials and Methods

2.4.1 Sample collection and storage.

We used a subset of individuals who participated in a randomized clinical trial [Kao et al., 2017] and conducted a longitudinal FMT study of two human cohorts (DA and DB), each consisting of one FMT donor and 5 FMT recipients of that donor's stool. All recipients received vancomycin for a minimum of 10 days pre-FMT at a dose of 125 mg four times

daily. Three DA and two DB recipients received FMT via pill, and two DA and three DB recipients received FMT via colonoscopy. All recipients had recurrent *C. difficile* infection before FMT, and two DA recipients and 1 DB recipient were also diagnosed with ulcerative colitis (UC). 24 stool samples were collected from the DA donor over a period of 636 days, and 15 stool samples were collected from the DB donor over a period of 532 days. Between 5 and 9 stool samples were collected from each recipient over periods of 187 to 404 days, with at least one sample collected pre-FMT and 2 samples collected post-FMT. This gave us a total of 109 stool samples from all donors and recipients. Samples were stored at -80℃. (Figure 2.6, Table 2.1)

2.4.2 Metagenomic short-read sequencing.

We extracted the genomic DNA from frozen samples according to the centrifugation protocol outlined in MoBio PowerSoil kit with the following modifications: cell lysis was performed using a GenoGrinder to physically lyse the samples in the MoBio Bead Plates and Solution (5-10 min). After final precipitation, the DNA samples were resuspended in TE buffer and stored at -20℃ until further analysis. Sample DNA concentrations were determined by PicoGreen assay. DNA was sheared to 400 bp using the Covaris S2 acoustic platform and libraries were constructed using the Nugen Ovation Ultralow kit. The products were visualized on an Agilent Tapestation 4200 and size-selected using BluePippin (Sage Biosciences). The final library pool was quantified with the Kapa Biosystems qPCR protocol and sequenced on the Illumina NextSeq500 in a 2 × 150 paired-end sequencing run using dedicated read indexing.

2.4.3 'Omics workflows.

Whenever applicable, we automated and scaled our 'omics analyses using the bioinformatics workflows implemented by the program 'anvi-run-workflow' [Shaiber et al., 2020] in anvi'o [Eren et al., 2015, 2021]. Anvi'o workflows implement numerous steps of bioinformatics tasks including short-read quality filtering, assembly, gene calling, functional annotation, hidden Markov model search, metagenomic read-recruitment, metagenomic binning, pangenomics, and phylogenomics. Workflows use Snakemake [Köster and Rahmann, 2012] and a tutorial is available at the URL https://merenlab.org/anvio-workflows/. The following sections detail these steps.

2.4.4 Taxonomic composition of metagenomes based on short reads.

We used Kraken2 v2.0.8-beta [Wood et al., 2019] with the NCBI's RefSeq bacterial, archaeal, viral and viral neighbours genome databases to calculate the taxonomic composition within short-read metagenomes. Assembly of metagenomic short reads. To minimize the impact of random sequencing errors in our downstream analyses, we used the program 'iu-filter-quality-minoche' to process short metagenomic reads, which is implemented in illumina-utils v2.11 [Eren et al., 2013b] and removes low-quality reads according to the criteria outlined by Minoche et al. [Minoche et al., 2011]. IDBA_UD v1.1.2 [Peng et al., 2012] assembled quality-filtered short reads into longer contiguous sequences (contigs), although we needed to recompile IDBA_UD with a modified header file so it could process 150bp paired-end reads.

2.4.5 Processing of contigs.

We use the following strategies to process both sequences we obtained from our assemblies and those we obtained from reference genomes. Briefly, we used (1) 'anvi-gen-

contigs-database‘ on contigs to compute k-mer frequencies and identify open reading frames (ORFs) using Prodigal v2.6.3 [Hyatt et al., 2010], (2) ‘anvi-run-hmms‘ to identify sets of bacterial [Campbell et al., 2013] and archaeal [Rinke et al., 2013] single-copy core genes using HMMER v3.2.1 [Eddy, 2011], (3) ‘anvi-run-ncbi-cogs‘ to annotate ORFs with functions from the NCBI's Clusters of Orthologous Groups (COGs) [Tatusov et al., 2003], and (4) ‘anvi-run-kegg-kofams‘ to annotate ORFs with functions from the KOfam HMM database of KEGG orthologs (KOs) [Aramaki et al., 2020, Kanehisa and Goto, 2000]. To predict the approximate number of genomes in metagenomic assemblies we used the program ‘anvi-display-contigs-stats‘, which calculates the mode of the frequency of single-copy core genes as described previously [Delmont and Eren, 2016].

2.4.6 Metagenomic read recruitment, reconstructing genomes from metagenomes, determination of genome taxonomy and ANI.

We recruited metagenomic short reads to contigs using Bowtie2 v2.3.5 [Langmead and Salzberg, 2012] and converted resulting SAM files to BAM files using samtools v1.9 [Li et al., 2009]. We profiled the resulting BAM files using the program ‘anvi-profile‘ with the flag ‘–min-contig-length‘ set to 2500 to eliminate shorter sequences to minimize noise. Once we have read recruitment results from each metagenome is profiled to store contig coverages into single anvi'o profile databases, ‘anvi-merge‘ combined all profiles into an anvi'o merged profile for downstream visualization, binning, and statistical analyses. We then used ‘anvi-cluster-contigs‘ to group contigs into 100 initial bins using CONCOCT v1.1.0 [Alneberg et al., 2014], ‘anvi-refine‘ to manually curate initial bins with conflation error based on tetranucleotide frequency and differential coverage signal across all samples, and ‘anvi-summarize‘ to report final summary statistics for each gene, contig, and bin. We used the program ‘anvi-rename-bins‘ to identify bins that were more than 70% complete and less than 10% redundant, and store

them in a new collection as metagenome-assembled genomes (MAG), discarding lower quality bins from downstream analyses. GTBD-tk v0.3.2 [Chaumeil et al., 2019] assigned taxonomy to each of our MAG using GTDB r89 [Parks et al., 2018], but to assign species- and subspecies-level taxonomy for ‘DA_MAG_00057‘, ‘DA_MAG_00011‘, ‘DA_MAG_00052‘ and ‘DA_MAG_00018‘, we used ‘anvi-get-sequences-for-hmm-hits‘ to recover DNA sequences for bacterial single-copy core genes that encode ribosomal proteins, and searched them in the NCBI’s nucleotide collection (nt) database using BLAST [Altschul et al., 1990]. Finally, the program ‘anvi-compute-genome-similarity‘ calculated pairwise genomic average nucleotide identity (gANI) of our genomes using PyANI v0.2.9 [Pritchard et al., 2016].

2.4.7 Criteria for MAG detection in metagenomes.

Using mean coverage to assess the occurrence of populations in a given sample based on metagenomic read recruitment can yield misleading insights since this strategy cannot accurately distinguish reference sequences that represent very low-abundance environmental populations from those sequences that do not represent an environmental population in a sample yet still recruit reads from non-target populations due to the presence of conserved genomic regions. Thus, we relied upon the ‘detection’ metric, which is a measure of the proportion of the nucleotides in a given sequence that are covered by at least one short read, and considered a population was detected in a metagenome if anvi’o reported a detection value of at least 0.25 for its genome (whether it was a metagenome-assembled or isolate genome). Values of detection in metagenomic read recruitment results often follow a bimodal distribution for populations that are present and absent (see Figure 2.7 in ref. [Utter et al., 2020]), thus 0.25 is an appropriate cutoff to eliminate false-positive signal in read recruitment results for populations that are absent.

2.4.8 Identification of MAGs that represent multiple subpopulations.

To identify subpopulations of MAGs in metagenomes, we used the anvi'o command 'anvi-gen-variability-profile' with the '–quince-mode' flag which exported single-nucleotide variant (SNV) information for all MAGs after read recruitment. We then used DESMAN v2.1.1 [Quince et al., 2017a] to analyze SNVs to determine the number and distribution of subpopulations represented by a single genome. To account for non-specific mapping that can inflate the number of estimated subpopulations, we removed any subpopulation that made up less than 1% of the entire population explained by a single MAG. To account for noise due to low-coverage, we only investigated subpopulation for MAGs for which the mean non-outlier coverage of single-copy core genes was at least 10X.

2.4.9 Criteria for colonization of a recipient by a MAG.

We developed a method to determine whether or not a MAG successfully colonized a recipient, and applied this method to each MAG and each recipient within a cohort. In order to confidently assign colonization or non-colonization phenotypes to each MAG/recipient pair, we required that the MAG be detected in the donor sample used for transplant into the recipient. If these criteria were met, we then determined whether the MAG was detected in any post-FMT recipient sample taken more than 7 days after transplant. If not, the MAG/recipient pair was considered a non-colonization event. If the MAG was detected in the recipient greater than 7 days post-FMT, we used subpopulation information to determine if any subpopulation present in the donor and absent in the recipient pre-FMT was detected in the recipient more than 7 days post-FMT. If this was the case, we considered this to represent a colonization event. See Figure 2.9 for a complete outline of all possible cases.

2.4.10 Determination of dose and fitness for MAGs.

We defined population dose as the second and third quartile mean coverage of a population in the transplanted stool sample. We defined fitness as the prevalence of a population in 23 healthy adult gut metagenomes (see Materials and Methods: Criteria for MAG detection in metagenomes) from Canada, the same country in which the FMTs were performed.

2.4.11 Regression analysis.

To examine the association between dose and/or prevalence with colonization outcome, we built binomial logistic regression models using the R stats ‘glm‘ function. We used the R stats ‘predict‘ function and the R pROC ‘roc‘ function to evaluate our models by creating receiver operating characteristic (ROC) curves and calculating the area under the ROC curve (AUC). To determine the correlation between dose and prevalence, we performed linear regression using the R stats ‘lm‘ function. We used the R tidyverse package, including ggplot2, to visualize boxplots, scatterplots, and ROC curves.

2.4.12 Pangenomic analysis and gANI.

We used anvi’o to compute and visualize pangenomes of MAGs and reference genomes. We stored all processed MAG and reference genome contigs (see Contig processing methods section) in an anvi’o database using the command ‘anvi-gen-genomes-storage‘. To create the pangenomes, we then passed that database to the command ‘anvi-pan-genome‘ which used NCBI’s BLAST [Altschul et al., 1990] to quantify gene similarity within and between genomes and the Markov Cluster algorithm (MCL) [Enright et al., 2002] to cluster groups of similar genes. We set the ‘anvi-pan-genome‘ ‘–min-occurrence‘ flag to 2 to remove gene clusters only present in one genome (singletons), and visualized pangenomes using ‘anvi-display-pan‘.

2.4.13 Phylogenomic tree construction.

To concatenate and align amino acid sequences of 46 single-copy core [Campbell et al., 2013] ribosomal proteins that were present in all of our *Bifidobacterium* MAGs and reference genomes, we ran the anvi'o command 'anvi-get-sequences-for-hmm-hits' with the '–return-best-hit', '–get-aa-sequence' and '–concatenate' flags, and the '–align-with' flag set to 'muscle' to use MUSCLE v3.8.1551 [Edgar, 2004] for alignment. We then ran 'anvi-gen-phylogenomic-tree' with default parameters to compute a phylogenomic tree using FastTree 2.1 [Price et al., 2010].

2.4.14 Analysis of metabolic modules and enrichment.

We calculated the level of completeness for a given KEGG module [Kanehisa et al., 2014, 2017] in our genomes using the program 'anvi-estimate-metabolism', which leveraged previous annotation of genes with KEGG orthologs (KOs) (see the section 'Processing of contigs'). Then, the program 'anvi-compute-functional-enrichment' determined whether a given metabolic module was enriched in based on the output from 'anvi-estimate-metabolism'. The URL https://merenlab.org/m/anvi-estimate-metabolism serves a tutorial for this program which details the modes of usage and output file formats. The statistical approach for enrichment analysis is defined elsewhere [Shaiber et al., 2020], but briefly it computes enrichment scores for functions (or metabolic modules) within groups by fitting a binomial generalized linear model (GLM) to the occurrence of each function or complete metabolic module in each group, and then computing a Rao test statistic, uncorrected p-values, and corrected q-values. We considered any function or metabolic module with a q-value less than 0.05 to be 'enriched' in its associated group if it was also at least 75% complete and in at least 50% of the group members. To display the distribution of individual KEGG orthologs across genomes and order them based on their enrichment scores and group affiliations we used the program 'anvi-display-functions'.

2.4.15 Determination of high-fitness and low-fitness MAGs for metabolic enrichment analysis.

We classified MAGs as high-fitness if, in all 5 recipients, they were detected in the donor sample used for transplantation as well as the recipient more than 7 days post-FMT. We classified low-fitness MAGs as those that, in at least 3 recipients, were detected in the donor sample used for FMT but were not detected in the recipient at least 7 days post-FMT. We reduced the number of high-fitness MAGs to be the same as the number of low-fitness MAGs for metabolic enrichment analysis by selecting only the high-fitness MAGs which were the most prevalent in the Canadian gut metagenomes.

2.4.16 Ordination plots.

We used the R vegan v2.4-2 package 'metaMDS' function to perform nonmetric multi-dimensional scaling (NMDS) with Horn-Morisita dissimilarity distance to compare taxonomic composition between donor, recipient, and global metagenomes. We visualized ordination plots using R ggplot2.

2.5 Code and Data Availability

Raw sequencing data for donor and recipient metagenomes are stored under the NCBI BioProject PRJNA701961 (see Table 2.1 for accession numbers for each sample). The URL https://merenlab.org/data/fmt-gut-colonization provides a reproducible bioinformatics workflow and gives access to ad hoc scripts, usage instructions, and intermediate data objects to reproduce findings in our study.

2.6 Supplementary Figures

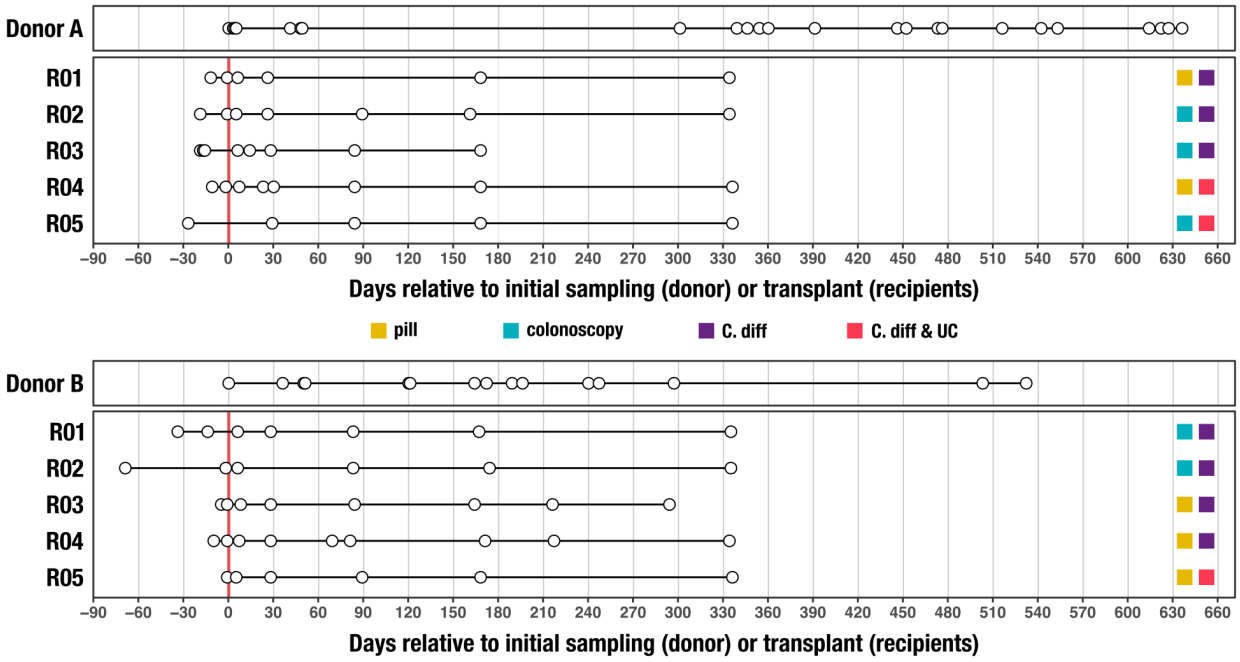


Figure 2.6: Timeline of stool samples collected from FMT study. Each circle represents a stool sample collected from either an FMT donor or FMT recipient. The thicker, red vertical line at day 0 represents the FMT event for each recipient. FMT method (pill or colonoscopy) and FMT recipient health and disease state (C. diff - chronic recurrent *Clostridioides difficile* infection, UC - ulcerative colitis) are indicated on the right.

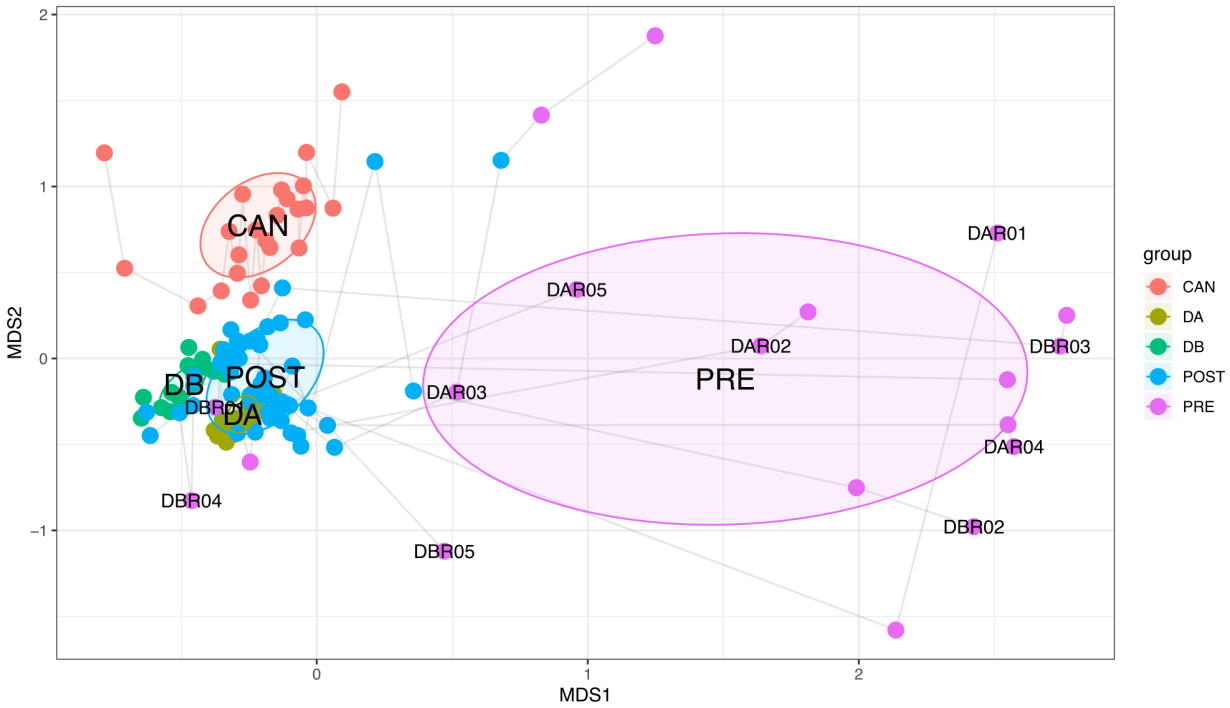


Figure 2.7: Nonmetric multidimensional scaling (NMDS) ordination of the taxonomic composition of donor, recipient, and Canadian gut metagenomes at the genus level based on Morisita-Horn dissimilarity. Samples from the same participant are joined by lines with the earliest time point labeled. CAN: Canadian gut metagenomes, DA: donor A, DB: donor B, POST: recipients post-FMT, PRE: recipients pre-FMT.

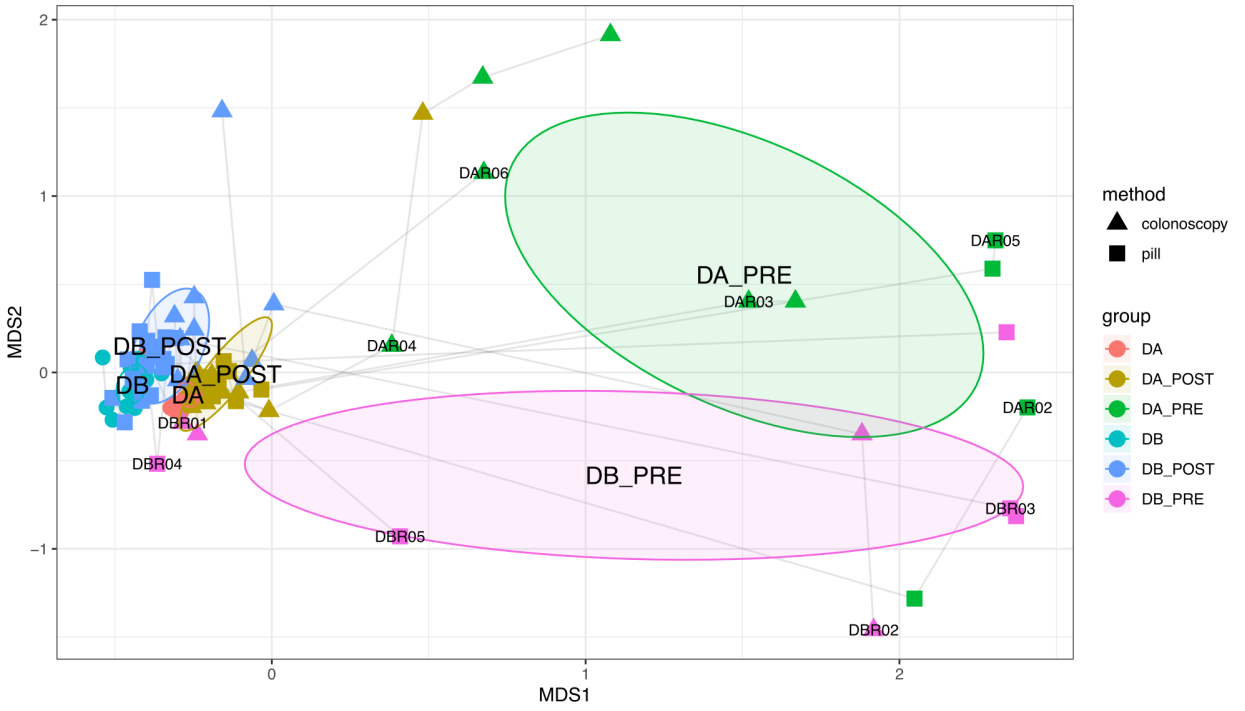


Figure 2.8: Nonmetric multidimensional scaling (NMDS) ordination of the taxonomic composition of the donor and recipient metagenomes at genus level based on Morisita-Horn dissimilarity. Samples from the same participant are joined by lines with the earliest time point labeled. DA_POST: donor A recipients post-FMT, DA_PRE: donor A recipients pre-FMT, DA: donor A, DB_POST: donor B recipients post-FMT, DB_PRE: donor B recipients pre-FMT, DB: donor B.

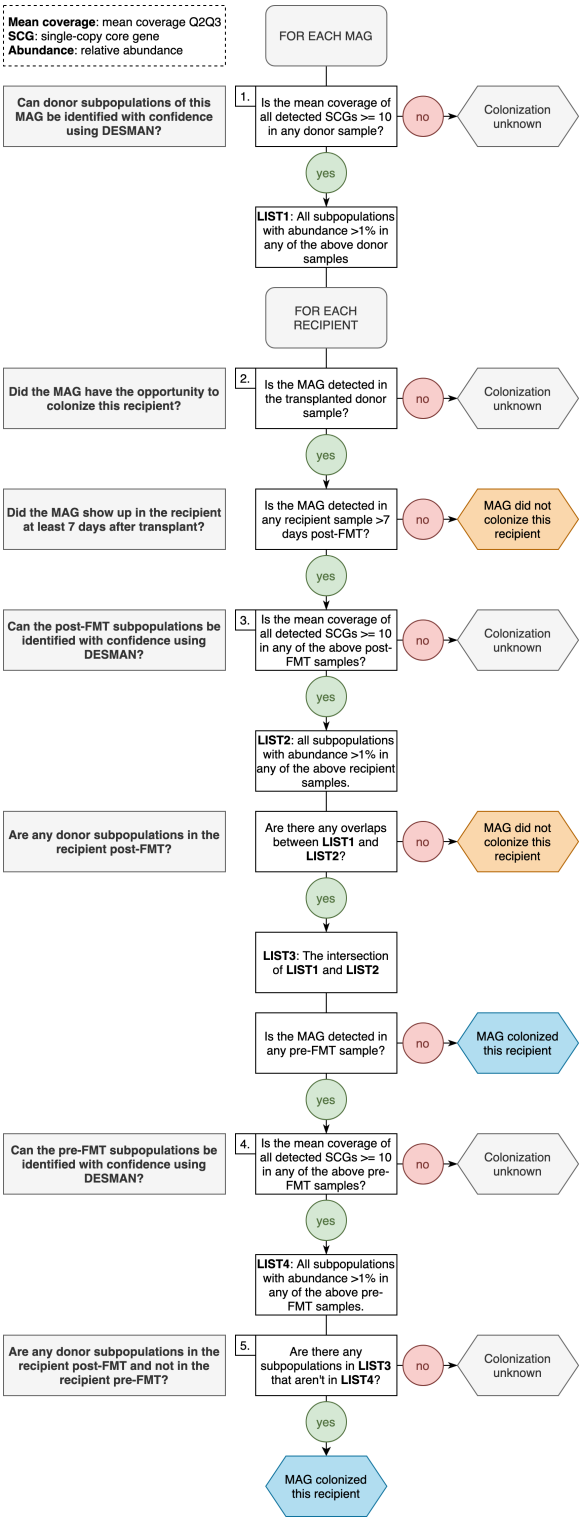


Figure 2.9: A flowchart outlining our method to assign successful colonization, failed colonization, or undetermined colonization phenotypes to donor-derived populations in the recipients of that donor's stool.

2.7 Supplementary Tables

Chapter 2 supplementary tables are accessible via doi:10.6084/m9.figshare.14138405.

Table 2.1: Description of FMT study and stool samples collected. a) Description of FMT donor stool samples and SRA accession numbers. b) Description of FMT recipient samples and SRA accession numbers. c) Description of transplantation events.

Table 2.2: Description of FMT metagenomes and co-assemblies. a) Metagenome SRA accession numbers and number of metagenomic short-reads sequenced and mapped to co-assemblies and MAGs. b) Phylum level taxonomic composition of metagenomes. c) Genus level taxonomic composition of metagenomes. d) Summary statistics for contigs from metagenome co-assemblies.

Table 2.3: Description of MAGs. a) Summary statistics and taxonomic assignments for MAGs. b) and c) Detection of Donor A and Donor B MAGs in FMT metagenomes, respectively. d) and e) Detection of Donor A and Donor B MAGs in global gut metagenomes, respectively. f) and g) Detection summary statistics of Donor A and Donor B MAGs in global gut metagenomes, respectively. h) and i) Mean non-outlier coverage of Donor A and Donor B MAG single-copy core genes in FMT metagenomes.

Table 2.4: Accession numbers of gut metagenomes from 17 countries.

Table 2.5: MAG subpopulation information. a) and b) Number of Donor A and Donor B MAG subpopulations detected in FMT metagenomes, respectively. c) and d) Subpopulation composition of Donor A and Donor B MAGs in FMT metagenomes, respectively.

Table 2.6: MAG/recipient pair colonization outcomes and MAG mean coverage in the 2nd and 3rd quartiles in stool samples used for transplantation.

Table 2.7: Description of high vs. low-fitness populations. a) Taxonomic assignments and genome size estimates for high- and low-fitness populations. b) KEGG module completeness information for high- and low-fitness populations. c) Raw KEGG module enrichment information for high- and low-fitness populations. d) KEGG module enrichment and categorical information for the 33 modules enriched in high-fitness populations. e) and f) Completeness information for the 33 modules enriched in high-fitness populations in all high- and low-fitness populations.

Table 2.8: Description of genomes from healthy individuals and individuals with IBD. a) List of genomes from healthy individuals and individuals with IBD. b) Module completion values across genomes.

Table 2.9: *Bifidobacterium* functional analysis. a) Accession numbers for *Bifidobacterium* reference genomes. b) Summary statistics for *Bifidobacterium* MAGs and reference genomes. c) Prevalence of *Bifidobacterium* MAGs in global gut metagenomes. d) gANI percent identity between *Bifidobacterium* genomes. e) gANI percent alignment coverage between *Bifidobacterium* genomes. f) KOfams enriched in different *Bifidobacterium* species. g) KOfam presence and absence in *Bifidobacterium* genomes. h) COG functions enriched in different *Bifidobacterium* species. i) COG function presence and absence in *Bifidobacterium* genomes. j) KEGG modules enriched in different *Bifidobacterium* species. k) KEGG module completeness in *Bifidobacterium* genomes.

CHAPTER 3

GENOMIC DYNAMISM OF CLOSELY RELATED MICROBIAL POPULATIONS IN UNRELATED INDIVIDUALS

3.1 Introduction

The identification of fundamental units of microbial life is a central goal of microbial ecology [Koeppel et al., 2008, Jain et al., 2018]. Existing technical frameworks to study the ecology of microbes obstruct this goal, as they typically consider individual taxa to be functionally conserved entities across different environments, when in reality members of the same microbial population can display distinct phenotypes under highly similar, or even identical, environmental conditions [Lidstrom and Konopka, 2010, Schreiber and Ackermann, 2020]. Subtle genomic modifications such as single nucleotide variants [Zhao et al., 2019] or genomic rearrangements that do not alter gene content can dramatically change ecological behaviour of a subset of a microbial population within the same environment [Trzilova and Tamayo, 2021, Jiang et al., 2019]. Even without alteration of the genome, viral infection can dramatically alter the metabolic activities and resource utilization of individual microbes in a single population [Howard-Varona et al., 2020]. In different environments, epitranscriptomic processes can influence the proteome of nearly identical microbes [Schwartz et al., 2018], and members of a microbial population within a microenvironment can produce vastly different metabolites as a function of spatial organization [Geier et al., 2020]. Such heritable and non-heritable functional alterations can occur over short evolutionary time scales [Ahmed et al., 2008, Höfer and Jäschke, 2018], and obfuscate associations between microbial taxa and their ecological or environmental effects, especially within complex ecosystems. Assignment of specific phenotypes or environmental effects to functionally distinct microbial populations has been particularly challenging in human gut microbiome research [Fischbach, 2018].

Rapid genomic alterations and within-population heterogeneity in different environments present major challenges to the identification of members of the gut microbiota associated with health and disease states, or bacterial strains that are suitable for inclusion in additive microbial therapeutics. To address this challenge, we must improve our understanding of microbial lifestyles in the complex and variable human gut ecosystem by studying very closely related microbes in different gut environments to establish deeper insights into microbial behavior [Fischbach, 2018]. The lack of complex systems that are suitable to investigate the fate of identical populations in different gut environments is partially resolved by FMT, which enables tracking of individual populations over time and investigation of microbial responses to abruptly changing environmental conditions in a natural human gut system.

Understanding the fate of donor microbes in recipient guts following additive microbial therapeutic strategies like FMT requires long-term tracking of donor microbes with approaches that can accurately resolve subtle changes in different host ecosystems. Analysis of marker gene sequences or metagenomic read recruitment to reference genomes cannot reveal the full extent of within-population variation that occurs in complex, naturally occurring microbial ecosystems. Genome-resolved metagenomics provides an alternative strategy that allows for the study of precise environmental populations of interest and their near-complete genomic context. Here we use MAGs and read-recruitment results from a previously described genome-resolved metagenomics FMT study [Watson et al., 2021] and additional long-read sequencing to investigate the ecology and evolution of donor populations as they colonize different recipients. Our analyses reveal mobile genetic elements that transfer between donor and recipient members of the same population upon collision in the recipient gut, genomic structural variation in response to new host environments that affects gene expression without altering gene content, and events of rapid evolution that occur differently in the same donor population upon colonization of

different individuals.

3.2 Results and Discussion

3.2.1 Competition and mobile element transfer between two highly similar Alistipes shahii *subpopulations*

A small fraction of the FMT donor populations from [Watson et al., 2021] also occurred in recipient pre-FMT samples, which enabled us to investigate interactions between highly similar but distinct subpopulations to resolve the genomic basis of colonization outcome. We identified one such *Alistipes shahii* population, which occurred both in donor A and in recipient 3 pre-FMT. We improved this MAG using contig extension to track even subtle differences in donor and recipient subpopulations with confidence. Donor and recipient derived *A. shahii* MAGs had a gANI of > 98%.

We tracked both subpopulations through single-nucleotide variants and the differential presence/absence of genes and mobile elements (Figure 3.1). Immediately post-FMT, the recipient's native *A. shahii* subpopulation became undetectable, and the donor subpopulation accounted for 100% of the population (Figure 3.1). However, over time the recipient's original subpopulation re-emerged, until it nearly eclipsed the donor subpopulation at 334 days post-FMT. Interestingly, during this time a prophage was transferred from the donor subpopulation to the recipient subpopulation. Our data show consistent coverage and lack of SNVs in the prophage region in donor samples and the absence of the prophage region in the recipient pre-FMT samples. At 334 days post-FMT, the *A. shahii* population in the recipient gut contains the prophage region in combination with SNVs characteristic of the recipient's native subpopulation (Figure 3.1). This indicates that a novel *A. shahii* population emerged in the recipient nearly one year post-FMT that was no longer identical to either the donor or native recipient subpopulations.

These results show that colonization outcome is decided at even finer taxonomic resolution than previously anticipated and even microbial subpopulations that fail to colonize recipients long-term may influence the genetic repertoire of resident populations through the transfer of mobile genetic elements. Such interactions may facilitate adaptive processes of microbial populations and result in the emergence of novel subpopulations. In our study, the transfer of a prophage from a donor to a recipient subpopulation could have negated the competitive advantage that the donor population initially had over the recipient population. This would then allow the recipient's original subpopulation, which was likely better adapted to life in the recipient gut aside from the presence of the newly introduced prophage, to take over the niche once more.

The close similarity between the donor and recipient *A. shahii* subpopulations likely indicates that they occupy the same niche in the gut environment and share the same major characteristics such as a fermentative metabolism, bile acid resistance and hydrolysis of tryptophan to indole [Parker et al., 2020]. Our results show that a foreign subpopulation can outcompete a resident subpopulation in the human gut due to genomic differences unresolvable at the species level. Previous studies have shown that the presence of a single prophage is able to confer a competitive advantage among strains of *Bacteroides fragilis* in culture conditions (Zhao et al., 2019). While the prophage we identified may confer a similar advantage in *A. shahii*, 40/45 of the genes encoded by the prophage are of unknown function and we cannot relate any fitness advantage offered by the prophage to a specific function. Members of the genus *Alistipes* have been inconsistently characterized as beneficial or pathogenic [Parker et al., 2020]. Our findings indicate that fitness of a microbe in the gut environment may be conveyed by yet uncharacterized traits that decide the fate of a microbial community in the gut environment and hence may impact its interactions with its host. Thus, beneficial or adverse effects associated with specific microbial populations may not always be resolvable on the species level and may fluently

change in relation to subpopulation level competition or, as observed here, the emergence of novel subpopulations as a result of prophage infection.

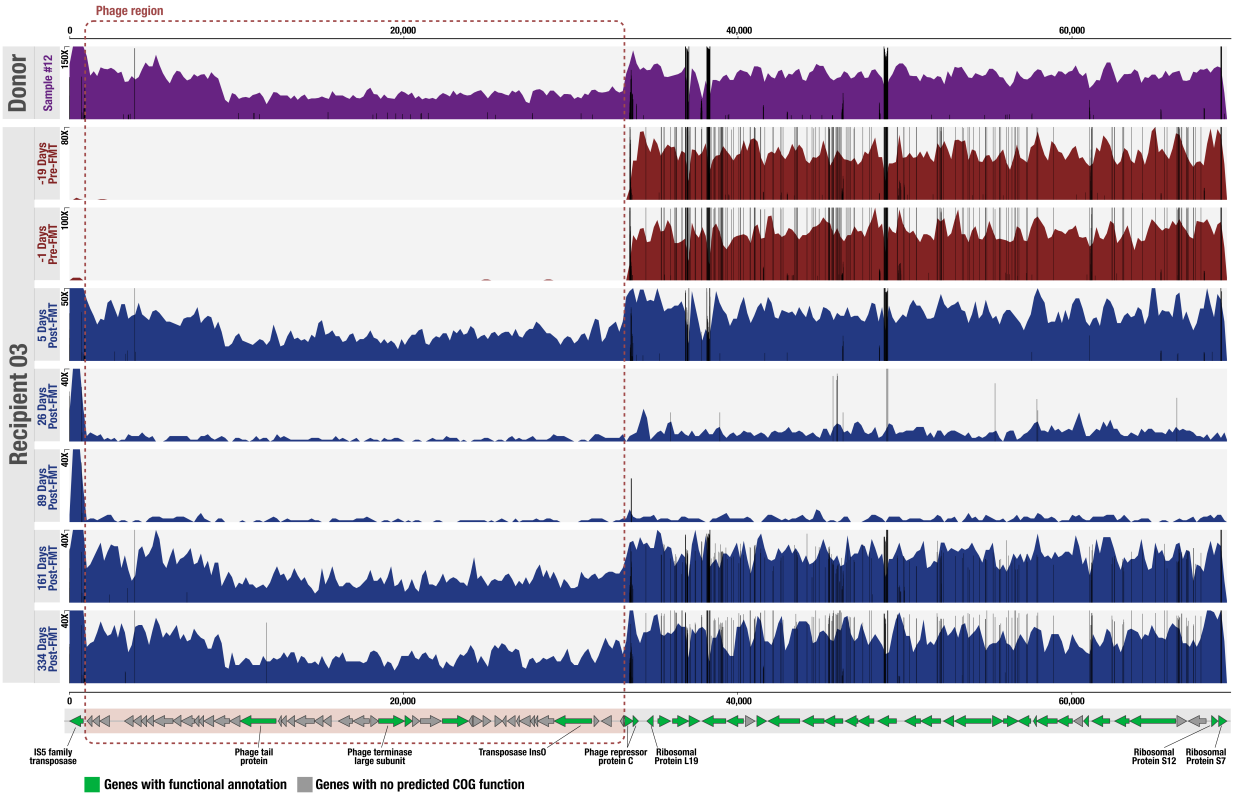


Figure 3.1: Read recruitment of donor and recipient metagenomes to an *A. shahii* contig containing a prophage region. Coverage values after metagenomic short-read recruitment are shown for each nucleotide position in a 69,293bp *A. shahii* MAG contig. Black vertical lines indicate nucleotide positions with a minimum coverage of 10x where at least one mapped read contained a single-nucleotide variant (SNV) relative to the reference sequence. The height of the vertical lines as a fraction of the total height of each row represents the fraction of mapped reads containing a SNV at that position.

3.2.2 Recipient-specific invertase-mediated structural alterations to Akkermansia muciniphila *genome rapidly emerge after FMT*

Bacterial phase variation is the rapid and reversible alteration of genome structure to control the expression of one or more genes. Bacterial phase variation can occur through invertible promoters, where a promoter falls within an invertible genomic sequence, or

inverton, that can be inverted by an invertase enzyme targeting flanking inverted repeat sequences. Such inversions change the orientation of the promoter so that in one orientation a downstream gene is expressed, and in the opposite orientation the same gene is not, essentially creating an ON/OFF switch for gene expression and allowing for the rapid adaptation to abrupt changes in environmental conditions [Jiang et al., 2019, Trzilova and Tamayo, 2021]. Invertible promoters are widely distributed amongst most bacterial phyla and predominantly appear in host associated bacteria regulating bacterial cell surface products or antibiotic resistance genes [Jiang et al., 2019]. We investigated bacterial phase variation through invertible promoters in an FMT donor-derived *Akkermansia muciniphila* population that successfully colonized three recipients [Watson et al., 2021]. To characterize all invertons in the exact *A. muciniphila* genome of interest, we used long-read metagenomic sequencing of recipient stool samples to circularize the existing MAG.

A. muciniphila is a mucin degrading gram negative Verrucomicrobia commonly found in the human gut [Collado et al., 2007], with reported vancomycin resistance [Dubourg et al., 2013] that explains the very high coverage observed in recipient 3 of donor A's stool pre-FMT (mean coverage of 1220x in the recipient 3 before FMT) (see Chapter 2 Code and Data Availability). We identified all putative inversions with PhaseFinder [Jiang et al., 2019] and computed the relative proportion of each inversion's orientation using metagenomic short-read sequences and a novel application of oligotyping [Eren et al., 2013a]. Our method identified the same promoter and inverted repeat motifs as previously described in *A. muciniphila* by [Jiang et al., 2019] (Figure 3.4), but identified a larger number of informative short-reads for each inverton than PhaseFinder (Table 3.1).

We identified 31 possible invertons in *A. muciniphila* with PhaseFinder, but only 13 changed phase over time (Figure 3.5) within FMT donor A and the three FMT recipients colonized by this population. As anticipated, the proportion of inversions in the first recipient sample after FMT matched the proportion of inversions the donor sample used for

FMT, for all recipients (Figure 3.2). However, the proportion of some inversions changed dramatically within weeks after colonizing a new host. We inspected the genomic context of each inversion and inferred the promoter phase (ON or OFF) when possible, based on the orientation of downstream genes. Three promoter inversions were associated with glycosyltransferases genes, involved in cell wall biosynthesis. Of these invertons, two (Inv 1 and Inv 2) had synchronized proportions of ON/OFF orientation in the donor, recipient 2, and recipient 3, but not in recipient 4. In contrast, the third inversion regulating a glycosyltransferase (Inv 3) consistently showed a higher proportion of ON orientation when Inv 1 and 2 were in the OFF orientation in the population.

Glycosyltransferases play a key role in bacterial adaptation to the environment [Coyne et al., 2008, Porter et al., 2017], and the associations of glycosyltransferases and other predicted transmembrane proteins with invertons in our data supports cell surface modification as a keystone of *A. muciniphila* adaptation to a new host following FMT.

A previous study proposed that a single invertase regulates all *A. muciniphila* invertons [Jiang et al., 2019], however; our data show invertons in the same organism changing in orientation independently. This could be explained by a basal rate of invertase activity and adaptive ecological processes selecting for specific inversion orientations. Our results show that bacterial adaptation and evolution through phase variation can occur rapidly upon colonization of a new gut environment, and that the same population of bacteria transplanted into different recipient gut environments will adapt differently. Not only do our analyses reveal changes in inverton orientation when a donor microbe colonizes a new host, but they also reveal dramatic, non-random changes in inverton orientation within the donor over time, indicating that variation in conditions within a single individual's gut environment is sufficient to select for specific promoter orientations. We also show that glycosyltransferases and other transmembrane proteins were affected by inverton regulation, indicating that cell wall and cell surface molecules are likely drivers of

this rapid adaptation to changing gut environmental conditions. Ultimately, these findings reveal that genomic structural rearrangements that do not alter gene content are likely drivers of microbial adaptation to changing environmental conditions in the human gut, yet are overlooked by the majority of human microbiome studies.

3.2.3 Diversity-generating retroelements introduce substantial genomic dynamism within a single population across different individuals

Diversity-generating retroelements (DGRs) are genetic mechanisms that diversify the nucleotide composition of a given gene at defined locations [Liu et al., 2002]. This template-dependent, reverse transcriptase-mediated 'mutagenic homing' process can yield a remarkable number of protein variants within a single population, enabling rapid adaptation to changing environmental conditions [Medhekar and Miller, 2007] across broad lineages of bacteria and archaea [Guo et al., 2014, Paul et al., 2017] and its activity can be measured through metagenomics [Yan et al., 2019, Benler et al., 2018]. By re-analyzing publicly available human gut metagenomes, Roux et al. showed that obese patient gut microbiomes had high DGR activity after a period of dramatic diet and lifestyle change, compared to the lower DGR activity seen in the gut microbiomes of individuals with more consistent diets [Roux et al., 2020], indicating that DGR activity may be higher after gut microbiome perturbation caused by diet and lifestyle change. Here we leveraged the unique opportunity our dataset enables in studying the genetic dynamism of a single microbial population in different individuals by means of diversity-generating retroelements (DGRs).

In our study, a *Bacteroides fragilis* population of Donor A (represented by MAG80) that successfully colonized and remained abundant in four FMT recipients for over 300 days, contained two DGR sites that were were more than 600,000 nts apart from each other and located within prophage regions. Both DGRs targeted genes of unknown function.

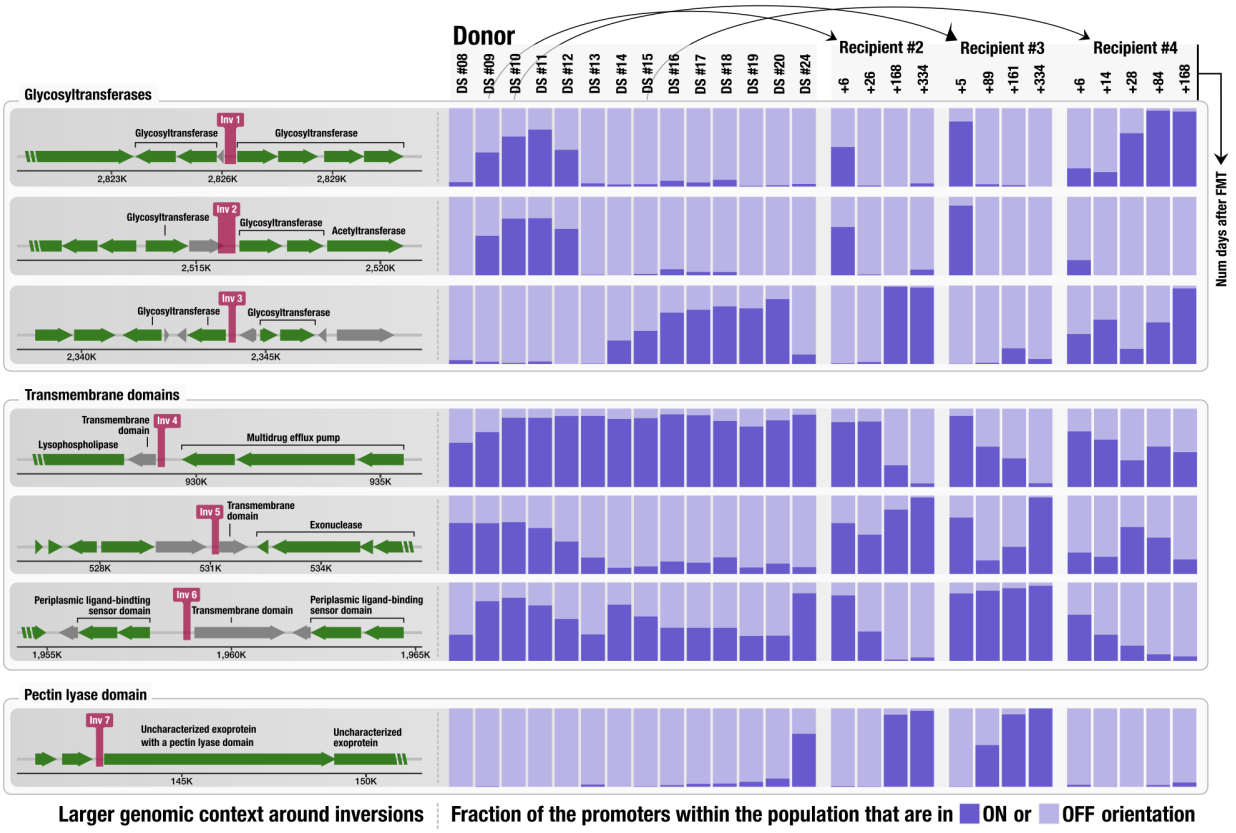


Figure 3.2: Selected *A. muciniphila* promoter inversion orientations in FMT study metagenomes. We identified promoter inversions in *A. muciniphila* associated with genes encoding glycosyltransferases, transmembrane domains, and pectin lyase domains. The genomic context for each invertible promoter is shown on the left: genes with COG and/or KEGG functional annotations are green, and genes without functional annotations are grey. The transmembrane and pectin lyase domains were inferred using InterProScan. For each invertible promoter, we computed the relative proportion of the "ON" and "OFF" orientation in each metagenome (purple stacked bars). Arrows indicate the donor samples used for each recipient's FMT.

This population was not detected in any of these recipients pre-FMT and enabled us to characterize DGR activity within a single individual for about a year and compare changes in diversity per site across DGR sites and different individuals.

To characterize within-population diversity of DGR-targeted variable regions (VRs) in each recipient, we first attempted to use single-nucleotide variants in our metagenomic read recruitment results. However, the use of SNVs proved to be a significant challenge due to the decreasing identities of individual short reads to the VR region which led to significant decreases in coverage, especially in metagenomes with increased DGR activity (Figure 3.3). To establish a more accurate description of sequence diversity in VRs, we developed a method for targeted identification of short metagenomic reads without mapping (see Methods) and resolving their diversity through oligotyping [Eren et al., 2013a].

The oligotype profiles of metagenomic short reads matching to VRs in donor MAG80 revealed relatively stable profiles for both DGR sites in donor metagenomes over 400 days (Figure 3.3), confirming previous observations that suggested low DGR activity in unperturbed gut environments [Roux et al., 2020]. Expectedly, the VRs in recipient samples collected immediately after FMT showed similar oligotype profiles to the donor (Figure 3.3). However, diverging patterns of variation emerged within less than two months in half of the recipients (Figure 3.3). Patterns of variation these data reveal include (1) population-level dynamics that include instances of population bottlenecks (VR #1, recipient 3, day 89), (2) hypermutagenic explosions (VR #1, recipient 3, between days 116 and 334), (3) gradual increase in and eventual dominance of an initially absent oligotype (VR #2, recipient #6, purple oligotype), (4) sudden turnover of dominant oligotypes (VR #1, recipient 6, the dominant oligotypes between days 29 and 84 differ at 16 of 18 positions), (5) oligotype fixation (VR #2, donor, red oligotype), and (6) stable equilibrium of non-fixed oligotypes (VR #1, recipient 5, green, orange, and blue oligotypes). There are numerous hypotheses capable of explaining each of these dynamics that are all underpinned by

drift, purifying selection, or adaptive evolution of the prophage and/or its host. For example, the rise in frequency of the purple oligotype in VR #2 of recipient 6 may be caused by conferring a selective advantage, hitchhiking on a positively selected genetic background, or simply due to drift. In general, our data alone cannot distinguish what we consider to be equally valid and likely causal models. But we found particularly interesting that while VRs in both DGR sites showed relatively stable profiles in recipients 2 and 5, recipient 3 and 6 showed dramatic and non-identical shifts from the donor profiles. VR #1 and VR #2 were more than 600,000 nts apart from each other, yet in recipients 3 and 6 we observed the overall level of divergence of VR #1 relative to the donor tended to match that of VR #2 (Figure 3.3), suggesting that DGR activity in a single population may be synchronized through global processes.

Overall, these data enable a direct observation of a single microbial population that genetically diverges from its ancestral donor population in a host-dependent manner. Distinguishing many evolutionary hypotheses that may explain both the observed emergence of hypervariability and the maintenance of stability in DGR-mediated variable sites will require careful consideration of phage life cycle and additional transcriptomic insights to survey DGR expression. Nevertheless, these data illustrate that (1) the proteome of the same microbial population can differentially diverge from its ancestral population within weeks in the human gut, and (2) the rich diversity of ways in which this can happen.

3.3 Conclusions

Our study reveals within-population genomic dynamism with functional consequences that occur differentially in different individual's gut environments over short time scales. Our findings demonstrate that even when microbes fail to colonize a recipient's gut, their transient presence can have a dramatic impact on the gene content of native members of the recipient gut microbiota. When a population does colonize, our data show that genomic

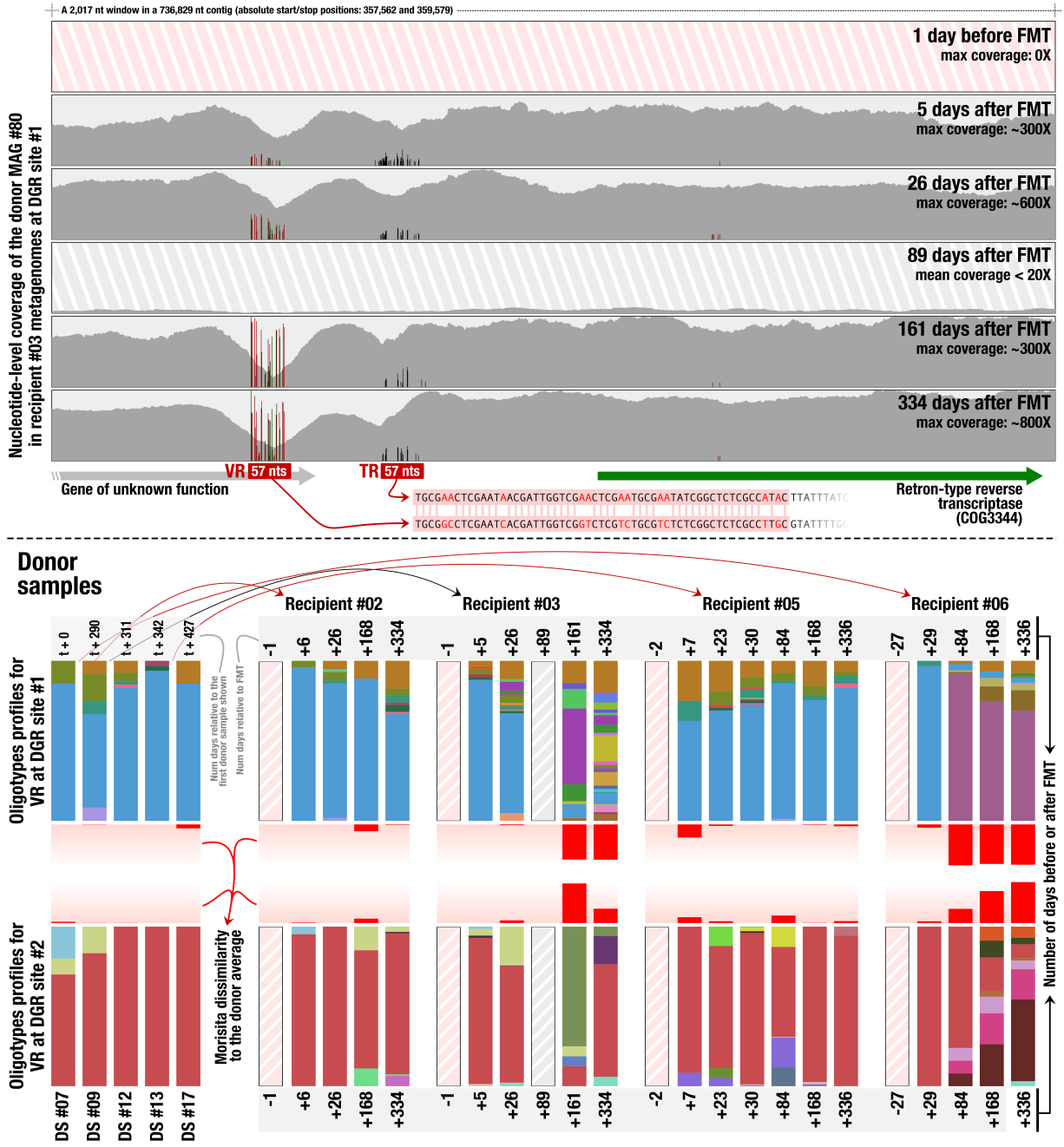


Figure 3.3: Diversity generating retroelement oligotypes in *B. fragilis* following FMT.

Figure 3.3 continued: Diversity generating retroelements (DGRs) consist of a retron-type reverse transcriptase (RT) found downstream of a target gene which contains a variable region (VR). That variable region matches to a template region (TR) found in the intergenic space between the target gene and the RT gene. The top panel shows the metagenomic coverage of a DGR from a *B. fragilis* MAG in recipient 3 of donor A. Vertical bars indicate single nucleotide variations coming from the short-read mapping and reflect the variability induced by the DGR activity. The VR and TR sequences only differ in adenosine positions in the TR as shown in the alignment. The bottom panel shows the oligotyping profiles for the VR region of two DGRs identified in the same *B. fragilis* MAG. Each colored bar represents a unique VR sequence identified from the metagenomics short-read using oligtyping; and bar size corresponds to relative proportion in each sample. Arrows indicate the donor samples used for each recipient's FMT. Red bars quantify the divergence of oligotype profiles from the donor average using Morisita-Horn diversity index.

adaptation to the new gut environment can occur on the order of weeks, and leads to different functional consequences in different host environments.

Taken together, these results reveal the impact of colliding donor and recipient microbial communities on the genomic content of the microbiome. Both donor and recipient microbial populations can undergo structural rearrangements and rapid genetic divergence that could only be observed in this study through systematic analysis of metagenomic read-recruitment patterns. These findings are particularly relevant to the evaluation of the safety and efficacy of additive microbial therapeutics, where divergent evolution of a population after colonization may have important functional consequences that could alter the efficacy or safety of a probiotic or LBP strain.

3.4 Materials and Methods

3.4.1 Metagenomic long-read sequencing.

We performed long-read sequencing on three samples from donor A recipient 3. We used a modified Qiagen Genomic Tip 20/G (Qiagen) kit to extract high molecular weight DNA prior to long-read sequencing. We centrifuged 250 µg of frozen stool with 1.8 ml

of Phosphate-Buffered Saline (PBS, pH 7.4) for 2 min at 645 x g and centrifuged the supernatant for 1 min at 7168 x g. We resuspended the pellet in 1 ml of PBS (pH 7.4) and repeated the two centrifugations. We then followed the manufacturer's protocol "Preparation Gram-negative and some Gram-positive Bacterial Samples" with the addition of 9 µl lysostaphin (4000 U/mL, Sigma-Aldrich) and 45 µl mutanolysin (10 KU/mL Sigma-Aldrich) in the lytic cocktail. We used 48 µl of extracted DNA for the MinION's library preparation using the Oxford Nanopore Technology (ONT) Ligation Sequencing Kit (SQK-LSK109). We performed DNA fragmentation with a 22G needle (10 passes) and DNA repair with the NEBNext FFPE DNA Repair Mix (NEB M6630) using 0.45X AMPure clean-ups steps prior and post DNA repair. We loaded the libraries on R9.4 flow cells (one flow cell per sample) according to ONT instructions and set the runtime to 72 hours. A technical failure stopped a sequencing run (sample DA_R03_CDI_C_03_POST) that was later resumed with no issues. We used guppy v2.3.1 for the basecalling and quality filtering with a minimum quality score of 7.

3.4.2 Metagenomic assembly of long reads.

We assembled long-read metagenomes using Flye v2.6 [Kolmogorov et al., 2019] with the –meta argument followed by Pilon v1.23 [Walker et al., 2014] correction using the associated sample's short-reads. We identified circular contigs with Flye [Kolmogorov et al., 2019].

3.4.3 Contig elongation.

We used PriceTI v1.2 [Ruby et al., 2013] to extend the *A. shahii* MAG (DA_MAG_00057) contigs with short reads from DA donor sample 12 (DA_D_12). We ran PriceTI with the "-icf" flag set to use one additional step, one cycle per step, and a 3x multiplier for quality

scores, the “-nc” flag set to 3 in order to perform 3 cycles of extension, and the “-fp” flag set to a total read and amplicon size of 600. We then recruited short reads from all DA cohort samples to the extended contigs and manually refined a new *A. shahii* MAG (see Chapter 2 Materials and Methods: Read recruitment and binning). We then re-mapped metagenomes to the newly refined MAG, but used competitive read-recruitment where the reference sequence encompassed the MAG and most contigs from the DA assembly, to avoid non-specific mapping. Before mapping, we removed the original *A. shahii* MAG contigs from the DA assembly, as well as any contigs which aligned to the extended *A. shahii* MAG with a percent identity of 95% or greater, and an alignment length of at least 25% of the contig.

3.4.4 Promoter inversions in Akkermansia muciniphila*.*

We identified three *Akkermansia* circular genomes (one per sample) and used pyANI v0.2.9 [Pritchard et al., 2016] to compute their average nucleotide identity. We used both COG [Tatusov et al., 2003] and KEGG [Kanehisa et al., 2016] for gene functional annotation.

We used PhaseFinder [Jiang et al., 2019] (flank size of 200 bp) to identify putative promoter inversions in the circular genome of *A. muciniphila* assembled from the sample DA_R03_CDI_C_03_POST. We selected for inversions with 10 reads supporting both orientation in at least one sample, and a ratio >10% of reverse orientation in at least one sample (n = 29). PhaseFinder uses short reads mapping to estimate the relative proportion of invertons in the reverse and forward orientation in a given sample, but the number of short reads reported were very low compared to the associated MAG mean coverage per sample. We used another approach to quantify the inversions: we identified all short reads matching to 6 bp of genomic context followed by the inverted repeat and at least 20 bp into the inversion (that can be in two orientations). We then use oligotyping

[Eren et al., 2013a] to obtain the relative proportion of forward and reverse orientation. We eventually kept 13 inversions based on their orientation dynamic. We identified putative promoter sequences in a similar fashion as by [Jiang et al., 2019]: we used MEME [Bailey et al., 2009] to find shared motifs between the 13 inversions and the upstream region of the three 16S rRNA genes. We inferred the promoter phase (ON or OFF) based on the direction of the putative downstream regulated gene for each inversion when possible, while assuming that the 16S rRNA gene promoter was ON.

3.4.5 Identification and analysis of diversity generating retroelements in Bacteroides fragilis*.*

We screened reference bacterial genomes (available through the NCBI accession ID PRJNA28331) generated by the Human Microbiome Project [Ribeiro et al., 2012, Human Microbiome Jumpstart Reference Strains Consortium et al., 2010] for diagnostic features that are common to all known DGRs, including reverse transcriptase, and VR-TR near-repeats as described in [Paul et al., 2017]. In a given contig, we identified reverse transcriptase-coding genes using pHMMER v3.1 [Finn et al., 2011] and used a sliding window approach to search for near-repeats in close proximity to the RT (*i.e.* +/- 5-kbp).

The DGR region was first identified in a large contig (365kbp) in a *Bacteroides fragilis* MAG from a donor sample (DA_D_08) single assembly. We then proceed to improve the donor co-assembled MAG affiliated to *B. fragilis* (DA_MAG_000080) using a combination of short reads from different assemblies and long reads. In brief, we used Unicycler [Wick et al., 2017] with the short reads mapping to the *B. fragilis* MAG (1) from the co-assembly, and (2) from the DA_D_08 single assembly. We included the MinION long-reads from the donor's A recipient 3 sample (DA_R03_CDI_C_03_POST) mapping to an OPERA-MS [Bertrand et al., 2019] assembly cluster affiliated to *B. fragilis* (as it did colonize recipient 3). The resulting MAG was composed of only 27 contigs. Due to the de-

creasing mappability of short reads into the DGR region as a result of the high-rate of sequence evolution, the standard metagenomic read recruitment approach using reference genomes is not typically sufficient to study the heterogeneity of these hypervariable genomic regions. Hence, to characterize the diversity of the DGR region we first determined a set of 20nt sequences that were identical to the conserved 5'-flanking downstream positions of the variable region of interest. Among these *in silico* primer sequences, a sequence derived from *Bacteroides dorei* CL02T12C06, 5'-ACTACTGGGCAGCTGGTTCT-3', NZ_JH724134, uniquely and identically matched to the reassembled *Bacteroides fragilis* MAG that was transferred from donor A to multiple recipients. We identified the next one, 5'-GTGTTACGTGGTGTCCTGTTC-3', through a manual inspection of the genome (we saw the retron type RT, and dramatic drop in coverage). Then, we identified all short reads in each metagenome that contains this sequence from each individual, and used oligotyping to explain the diversity of resulting sequences.

3.5 Supplementary Figures

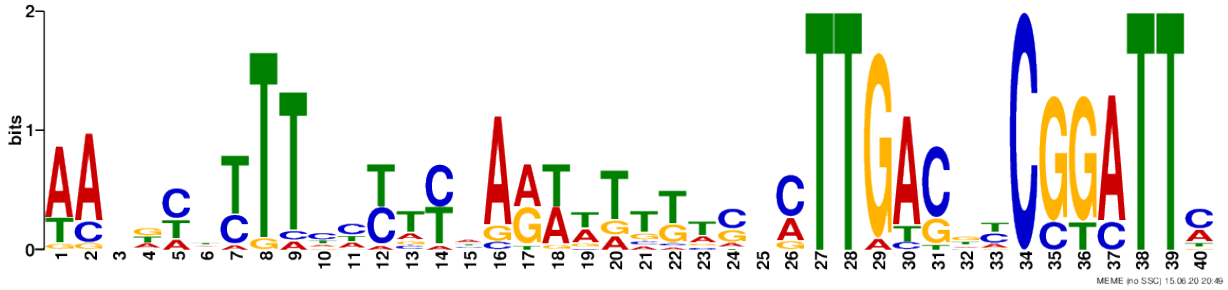


Figure 3.4: *A. muciniphila* promoter motif sequence logo.

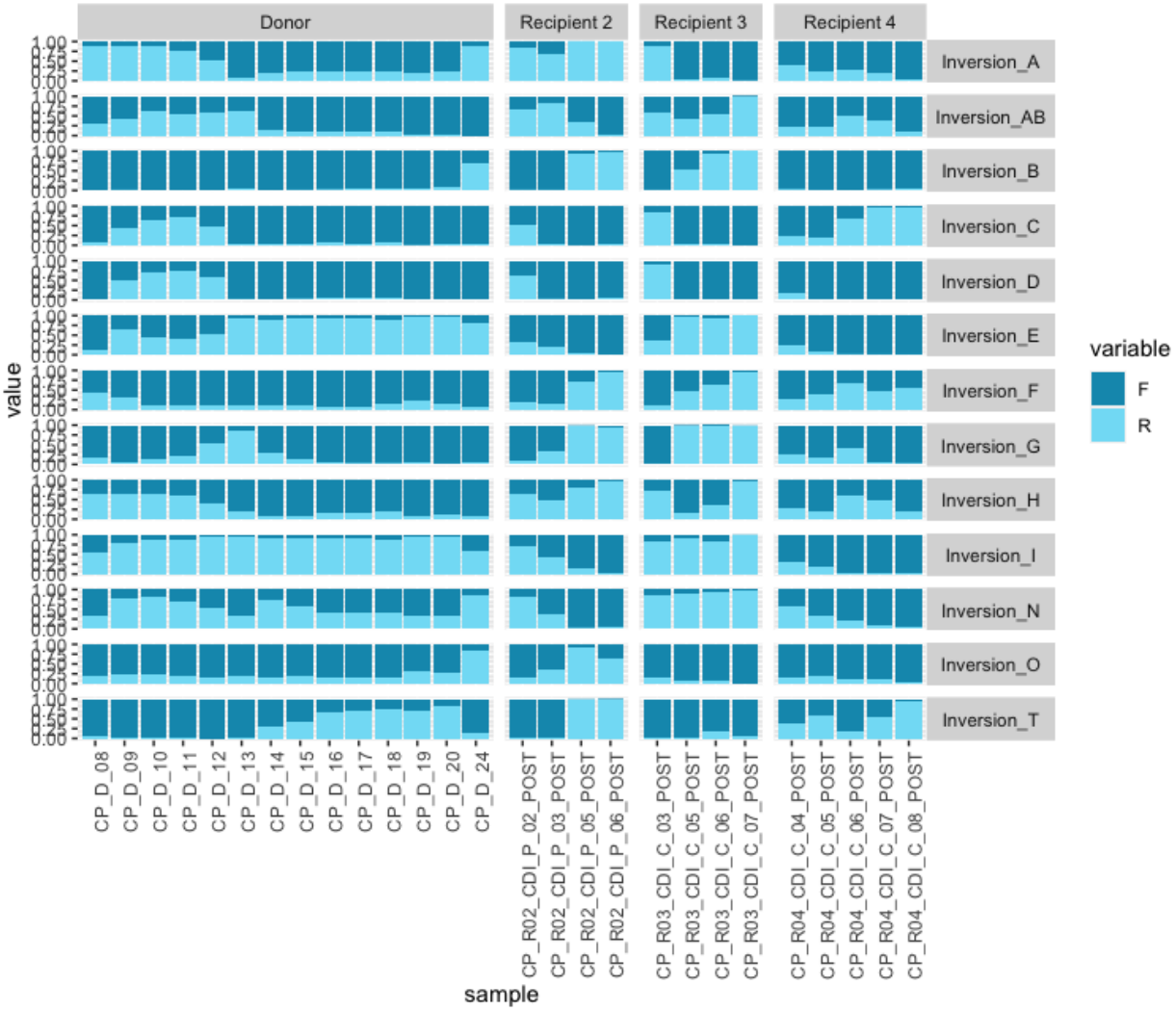


Figure 3.5: The orientations of the 13 active *A. muciniphila* promoter inversions in FMT study metagenomes.

3.6 Supplementary Tables

Sample	Mean Coverage	Oligotyping		PhaseFinder	
		Forward	Reverse	Forward	Reverse
DA_D_01	17.5	17	4	0	0
DA_D_03	0.6	1	0	0	0
DA_D_08	38.6	5	35	0	0
DA_D_09	144.8	13	116	0	0
DA_D_10	222.6	30	223	2	0
DA_D_11	84.6	21	68	0	0
DA_D_12	117.5	58	63	0	2
DA_D_13	216.1	205	22	52	2
DA_D_14	112.4	88	24	24	8
DA_D_15	255.2	239	71	90	14
DA_D_16	252.7	228	70	56	20
DA_D_17	316.4	267	75	64	18
DA_D_18	175.2	137	46	46	14
DA_D_19	95.5	68	18	42	2
DA_D_20	247.7	191	65	50	20
DA_D_22	2.3	0	2	0	0
DA_D_23	2.5	0	4	0	0
DA_D_24	131.1	16	135	6	46
DA_R02_CDI_C_02_PRE	450.6	0	2	0	0
DA_R02_CDI_C_03_POST	341.5	33	308	4	26
DA_R02_CDI_C_04_POST	6.2	4	8	0	0
DA_R02_CDI_C_05_POST	174.3	151	9	6	0
DA_R02_CDI_C_06_POST	372	306	19	38	2
DA_R02_CDI_C_07_POST	103	96	1	4	0

Table 3.1: The number of short-reads identified by PhaseFinder and oligotyping for *Akkermansia muciniphila* inverton A in FMT study metagenomes.

CHAPTER 4

ADVANCING BENCH LAB SCIENCE WITH METAGENOMICS

4.1 Introduction

The experience I developed analyzing large microbial 'omics datasets, and my background in microbiology, led me to identify new applications of 'omics tools to areas of microbiology outside of FMT. Metagenomic datasets, including the wealth of publicly available metagenomes, can be leveraged to put bench lab discoveries into a larger, and even global, environmental context. Through interdisciplinary collaborations and small projects, I used metagenomics to compare a mouse gut microbial isolate collection to the environmental mouse gut community, determine which culture conditions were able to support microbial populations of interest, and show the global relevance of a mechanistic study. This chapter describes these three studies and demonstrates the power of metagenomics to evaluate, inform, and expand the context of wet lab research.

4.2 Results and Discussion

4.2.1 A stool-derived mouse gut isolate collection differs at the phylum level from the gastrointestinal community

While the human gut microbiome has been associated with a myriad of health and disease states, the majority of these associations exist without established causality or known mechanism [Fischbach, 2018, Lynch et al., 2019, Walter et al., 2020]. Mechanistic studies are required to move beyond correlation and achieve an understanding of which components of the human gut microbiome are responsible for different host disease phenotypes, and would accelerate therapeutics targeting microbiome-mediated diseases [Fischbach,

2018]. Model systems provide a means to carry out well-controlled mechanistic studies of microbial communities.

The mouse model is the most frequently used model system for studying the human gut microbiota. Use of this model system usually involves transplant of human-derived gut microbial communities into germ-free mice, who were raised under sterile conditions and lack their own microbiome, to create human-microbiota associated (HMA) mice. However, not all human-derived microbial taxa are able to successfully colonize the murine gut, and as a result HMA mouse gut microbial communities look substantially different than those that seeded them [Zhang et al., 2017, Staley et al., 2017, Walter et al., 2020]. Additionally, germ-free mice may not replicate the host conditions that contribute to dysbiosis or disease state in humans, but do have their own unique developmental deficiencies which may lead to very different host-microbe interactions [Chung et al., 2012, Marcobal et al., 2013, Arrieta et al., 2016, Walter et al., 2020]. Despite evidence showing the inability of an HMA mouse to accurately model the human gut ecosystem, conclusions from HMA mouse studies are still often extrapolated to apply to human health [Walter et al., 2020].

In order to take full advantage of mouse model systems to study the human microbiota, we require a greater understanding of the native mouse gut microbial ecosystem. Identifying the differences and similarities between human and mouse microbiota would allow for more nuanced extrapolation of mouse model conclusions to human health. In addition, studies of the native mouse microbiome would allow for mechanistic insights into microbial colonization and fitness within a complex, natural gut ecosystem. Mechanistic studies benefit greatly from isolated culture collections of bacteria, and while many such collections exist for the human gut, there are relatively fewer isolates available which are derived from the murine gut microbiota. Establishing mouse gut isolate collections would be greatly beneficial to mechanistic studies of the mouse gut microbiota.

Existing mouse gut bacterial isolate collections include the Mouse Intestinal Bacte-

rial Collection (miBC) [Lagkouvardos et al., 2016] and the mouse gut microbial biobank (mGMB) [Liu et al., 2020]. However, recent analyses show that less than 10% of MAGs derived from the mouse gut microbiome belonged to species with a cultivated representative [Hitch et al., 2021]. While this number may be artificially deflated due to contamination of automatically binned MAGs making it difficult to assign them taxonomy [Chen et al., 2020], only 36% of 16S rRNA amplicon sequencing surveys of the mouse gut had cultivated species representatives for more than 50% of their sequences in 2020 [Hitch et al., 2021]. The missing cultivated representatives for mouse microbiota indicate that further cultivation efforts are required. Known gaps in the culture collection may be due to cultivation bias, where some taxa grow more easily under currently available laboratory conditions [Watterson et al., 2020]. It is therefore important to be able to evaluate any isolate collection for cultivation bias, and to see how well it represents the natural microbiota ecosystem from which it was derived.

We used metagenomics to compare genomes from a mouse gut isolate collection to environmental populations from the same environment to evaluate the completion of the isolate collection and identify any key gaps. We used metagenomic assembly, read-recruitment, and phylogenetics to create a phylogenetic tree of all Ribosomal L16 protein sequences from isolate genomes and environmental microbial populations present in gut metagenomes of the same mice from which the culture collection was derived. We found that some environmentally abundant phyla were underrepresented in the culture collection, and other phyla not detected in the metagenomes were overrepresented in the collection, indicating significant cultivation bias.

To investigate a collaborator's mouse gut isolate genome collection in the context of gut microbial metagenomes, we used a combination of metagenomic read recruitment and phylogenomic analysis strategies. First, we used hidden Markov models distributed by the Pfam protein families database [El-Gebali et al., 2019] to identify ribosomal proteins in

isolate genomes and metagenomic assemblies. From the list of suitable candidate genes for downstream analyses we chose Ribosomal L16, a ribosomal protein that typically occurs only once in each bacterial genome [Campbell et al., 2013], since it was the most frequently identified ribosomal protein in the assembly. The alignment and subsequent phylogenetic analysis of Ribosomal L16 amino acid sequences enabled an organization of all isolate genomes and gut microbial populations found in metagenomes in a single evolutionary context (Figure 4.1). Then, we identified the coverage of each Ribosomal L16 gene from isolate genomes and metagenomic contigs across each gut metagenome (see Methods for details) to estimate the relative abundance of populations to which each gene belongs (Figure 4.1).

While some cultivars represented microbial clades that were abundant in mouse gut metagenomes, we were unable to detect others. Similarly, some clades that were abundant and prevalent across metagenomes did not have a representative in our culture collection. Isolates for Proteobacteria (n=6) and Actinobacterota (n=9) in the cultivar collection remained below the detection limit of our sequencing effort in mice gut metagenomes, with the exception of *Bifidobacterium pseudolongum* isolate NM87 A27A and *Parasutterella* sp. isolate NM82 D38, which were covered 0.13X to 0.53X in three metagenomes (Figure 4.1). The 46 Ribosomal L16 genes recovered from mouse gut metagenomes did not include any that resolved to these phyla, either. Similarly, the Bacilli class of the Firmicutes phylum was well represented within the cultivar collection with 10 isolates, but also had no representative Ribosomal L16 genes within the metagenomes. However, 5 of the 10 Bacilli Ribosomal L16 genes had low coverage in at least one metagenome. In contrast, the Clostridia class of the Firmicutes phylum was represented in both the cultivar collection (n=5) and in the metagenomes (n=9). However, the Clostridia from the cultivar collection recruited no reads from any metagenomes, even though microbial populations within the same family, Lachnospiraceae, were abundant within several different mice in

our cohort. The phylum Bacteroidota was also well-detected both in the culture collection and metagenomes (Figure 4.1). Many of the isolate genomes in this phylum were detected in metagenomes in high relative abundances. While we were able to detect members of phyla Deferribacterota and Campylobacterota in metagenomes, neither of these phyla were represented in the cultivar collection. The relative abundance of Defferibacterota was moderate in multiple metagenomes (D10, D4, and D6), and Campylobacter occurred in moderate to high abundance in all metagenomes from diseased mice but was absent from healthy mouse metagenomes.

To summarize, the culture collection had relatively good coverage of the most abundant Bacteroidota, but otherwise was very lacking in representation of environmental Campylobacter and Defferobacter phyla. While the isolate collection contained several Firmicutes strains, these genomes were much less environmentally abundant across mouse gut samples than the environmental Firmicutes populations. The culture collection also contained many Actinobacteria and Proteobacteria strains, however; these phyla were almost completely undetected in metagenomes. While the absence of detection in metagenomes does not mean that these strains were not present in the environment, it indicates that if they were present, then these populations were relatively low in abundance compared to the higher coverage populations. Overall, this culture collection appears to have severe cultivation bias and bears very little resemblance to the microbial composition of the environments from which the cultivars were isolated.

These results may appear disappointing because they bring to attention the shortcomings of this particular culture collection. However, this approach allows for the identification of gaps in the cultivar collection to focus further cultivation efforts. Gaps in the collection could also be further explored by binning population genomes of the unrepresented members from metagenomes to access their genomic context and potentially learn more about how to cultivate them [Baart et al., 2007, Carini et al., 2013]. These

results also demonstrate our novel approach to compare isolate genomes with metagenomic populations without losing complex populations from the analyses that may not be easily or accurately binned into MAGs. This new approach can place cultivars isolated in the wet lab into the much larger environmental context accessed through metagenomics.

4.2.2 Metagenomics-informed media selection

The strong bias and limited representation seen in the mouse gut isolate collection from the previous section demonstrates the challenge of cultivating a breadth of gastrointestinal organisms. While a large number of human gut isolates have been characterized by several large cultivation efforts [HMP, etc], an estimated 70% of known genomes associated with human gut bacteria still do not have cultured representatives. Several "most wanted" lists have been created for isolation and full-genome sequencing of microbial taxa of particular relevance to human health, but traditional cultivation methods where samples are plated onto solid agar and individual colonies are picked and characterized are very labor intensive and time consuming. While new high-throughput cultivation methods have been developed, the throughput of isolation does not matter when organisms of interest simply are not supported on the culture media being used.

Genomic analysis and metabolic network modelling have been used to inform minimal media requirements for some fastidious microbes. However, these microbes were successfully cultivated prior to these analyses, albeit on more complex and undefined media, and their genomes were fully sequenced [Baart et al., 2007, Carini et al., 2013]. Such techniques are unlikely to be successful in determining the appropriate cultivation conditions for microbes which have never been cultivated under any conditions and whose genomes are largely incomplete or unknown, and this approach also requires in depth analyses that are difficult to scale up.

Metagenomic sequencing of the communities grown on different culture media, and

Figure 4.1: Phylogenetic tree of ribosomal L16 gene sequences from cultivar genomes and mouse gut metagenomes.

Figure 4.1 continued: D02, D04, D06, and D10 are metagenomes from mice that developed colitis, and M01, M02, M03, and M04 are metagenomes from healthy mice.

recruitment of those reads to population genomes of interest, provides a high-throughput way of identifying whether a precise environmental population of interest can be supported by the culture conditions. We mapped metagenomic sequences of DNA extracted from solid agar plate scrapings onto population genomes that were binned and characterized by a previous FMT study [Lee et al., 2017] to identify which, if any, of our five media were able to support the populations of interest. Lee et al. 2017 identified four main groups of microbial populations: group I which colonized both recipients of the FMT donor's stool, group II which colonized only recipient 1, group III which colonized only recipient 2, and group IV which failed to colonize both recipients. Our read-recruitment analysis revealed that only group I, group III, and ungrouped microbial populations had grown to the level of detection on any of the media used (Fig 4.2). The group with the most cultivated representatives was group I, with 14 out of 20 populations cultivated. GMM was able to support the growth of the most populations (n=23), followed by RFM (n=19), LKV (n=12), KV (n=10), and YCFA (n=4). These results demonstrate the ability of metagenomic read recruitment of plate metagenomes to genomes of interest to determine growth media before beginning arduous large-scale cultivation efforts to isolate populations of interest.

Overall, we were able to determine the media and cultivation conditions that could support the growth of different microbial populations and taxa from an FMT study. Our study demonstrates how metagenomics can inform cultivation efforts to isolate precise environmental populations of interest, and bridge the gap between computationally derived hypotheses and mechanistic studies.

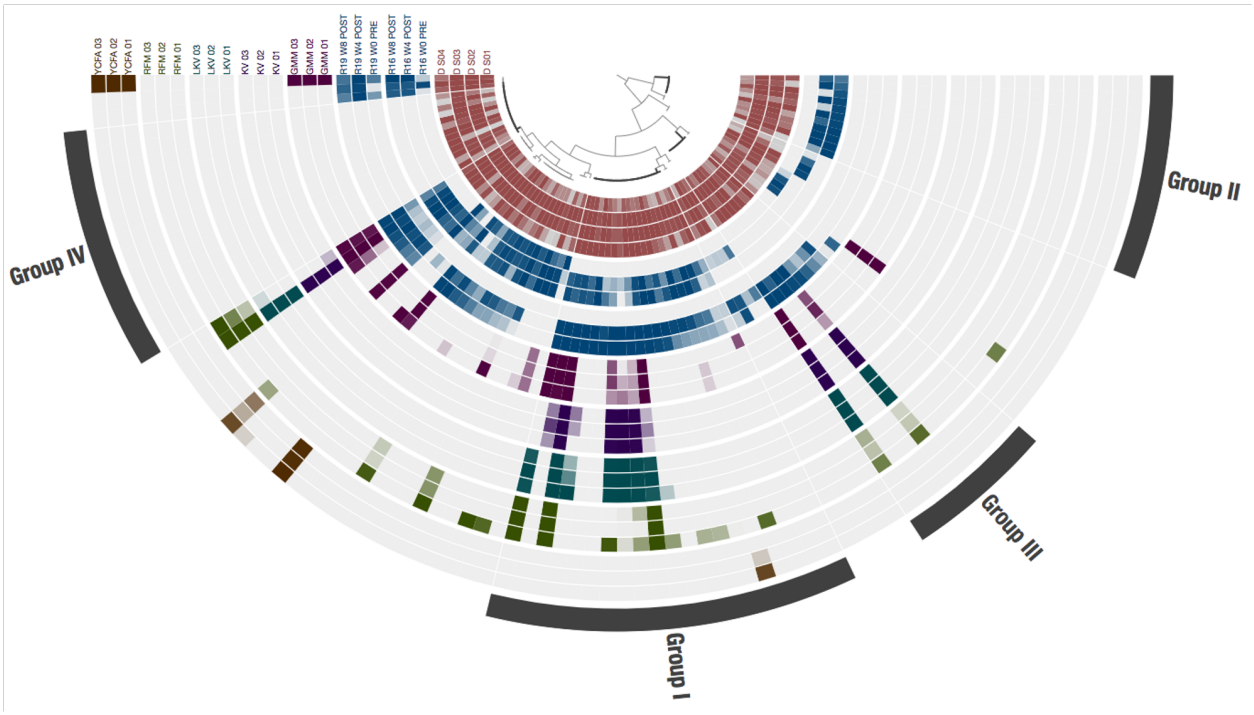


Figure 4.2: Detection of MAGs from Lee et al. 2017 across different types of culture media. In this heatmap, each radii of the semi-circle represents a donor MAG from Lee *et al*. 2017, and each half-ring represents a metagenome. The data points represent the detection of a MAG in a given metagenome. The innermost red and blue rings show the detection of each MAG in the metagenomes from the original study participants. The remaining rings show the detection of each MAG in metagenomes from scraped solid agar plates of different media that were inoculated with the original donor stool. The MAGs are organized by a dendrogram clustering them based on their detection patterns in the metagenomes. The groups from Lee et al. 2017 that each MAG belongs to are marked on the outermost layer.

4.2.3 Ruminococcus gnavus *B-cell superantigens are geographically widespread and part of the core genome*

The gastrointestinal mucosal barrier protects the host from pathogenic and commensal bacteria alike. It consists of a layer of epithelial cells protected by host-secreted mucus, which is primarily composed of mucin glycoproteins, but also contains antimicrobial peptides and secretory intestinal immunoglobulin A (IgA) [Martens et al., 2018]. IgA, the most abundant immunoglobulin isotype in the human body, is secreted into the mucus layer regardless of infection or inflammatory state, where it recognizes microbial and dietary antigens to carry out a variety of protective functions for the host as well as contributing to intestinal homeostasis [Cunningham-Rundles, 2001, Moon et al., 2015, Bunker et al., 2015, Macpherson et al., 2018, Bunker and Bendelac, 2018].

Intestinal secretory IgA protects the host by neutralizing and removing harmful molecules and preventing excessive colonization or translocation of the mucosal barrier by commensal bacteria, and sampling antigens in the gut lumen [Macpherson et al., 2018]. Monoclonal secretory IgA from the intestine are typically polyreactive, targeting multiple distinct microbial taxa *in vivo*, and several different microbial antigens *in vitro* [Bunker et al., 2017]. IgA coating of commensal bacteria during homeostasis has been well-documented, however; IgA also targets pathogens for elimination and its coating of subsets of gut microbiota has been associated with inflammatory conditions [Palm et al., 2014, Kau et al., 2015, Bunker et al., 2015, Viladomiu et al., 2017, Macpherson et al., 2018, Bunker and Bendelac, 2018].

It is unclear how IgA coating affects different gut bacteria, and if commensal microbes may express surface molecules to either evade or encourage targeting by IgA. The bacterial antigens targeted by secretory IgA in the intestine had not been characterized *in vivo* at the time of the collaboration featured in this subsection. Our collaborators identified two superantigen proteins expressed by *Ruminococcus gnavus* that bind to VH3-positive

B cells, likely through the B cell receptor, to stimulate a large B-cell mediated immune response in the gut leading to the expansion of intestinal secretory IgA which then targets and coats R gnavus cells *in vivo* [Bunker et al., 2019].

These findings were highly significant as they described a novel mechanism through which a specific commensal microbiota taxon can interact with the host to stimulate an immunological host phenotype [Bunker et al., 2019]. An estimated 50-60% of B cells and intestinal plasma cells express VH3, indicating that the immune response to *R. gnavus* superantigens could be highly relevant to diseases with B cell and IgA involvement [Benckert et al., 2011, Pauli et al., 2014]. However, while *R. gnavus* is a common gut commensal microbe, the superantigen binding activity characterized in this study was seen only in 4 of the 10 characterized infant stool samples [Bunker et al., 2019]. The relevance of the *R. gnavus* population encoding IbpA and IbpB to human populations on a larger scale remained unknown.

We mapped metagenomic reads from 424 publicly available healthy adult gut metagenomes to the IbpA and IbpB encoding *R. gnavus* genome to determine the global prevalence and distribution of this microbial population. *R. gnavus* and the superantigen genes were most highly detected in metagenomes from China (43%), followed by the United States (42%), and Fiji (7%). *R. gnavus* was not detected in any metagenomes from Tanzanian individuals. Inspecting metagenomic coverage of all genes in the *R. gnavus* genome allowed us to identify core and accessory genes (Figure 4.3). Accessory genes are genes which are not always detected in the environment along with the rest of the genome, whereas core genes are genes that are always detected in the environment along with the genome but are never detected in its absence. Our analysis determined that the *R. gnavus* genes encoding the IbpA and IbpB superantigens are core to this *R. gnavus* population. The prevalence and geographic breadth of *R. gnavus* and its core superantigen genes indicate that the mechanism through which *R. gnavus* stimulates

a disproportionate immune response discovered by our collaborators may represent a common and fundamental mechanism of host-microbe interaction.

We contributed metagenomic analysis to this collaboration using only publicly available and pre existing data, but by doing so were able to greatly expand the context of the findings in this study. We put a mechanistic discovery made in a single lab into a global framework that demonstrated its relevance to human populations across three different continents. Our straightforward approach can be expanded to include even more publicly available gut metagenomes, such as the 1,990 publicly available gut metagenomes used in Chapter 2 of this dissertation. Through this method, wet lab research can be examined at the same scale as the massive amounts of metagenomic data being generated more rapidly than hypotheses can be tested.

4.3 Conclusions

Through the application of metagenomic assembly and read recruitment to three different wet-lab based studies, we were able to 1) evaluate a mouse gut isolate collection to identify a lack of representation of key environmental taxa, 2) inform the media selection for the cultivation of microbial populations of interest from a previous FMT study, and 3) expand the context of a novel mechanism for host-microbe interaction in the human gut to show its relevance on a global scale. These studies exemplify the power of combining wet lab techniques with the large datasets accessible through metagenomics, and emphasize how an interdisciplinary computational and bench work approach can lead to exciting and novel insights into fundamental microbiological problems.

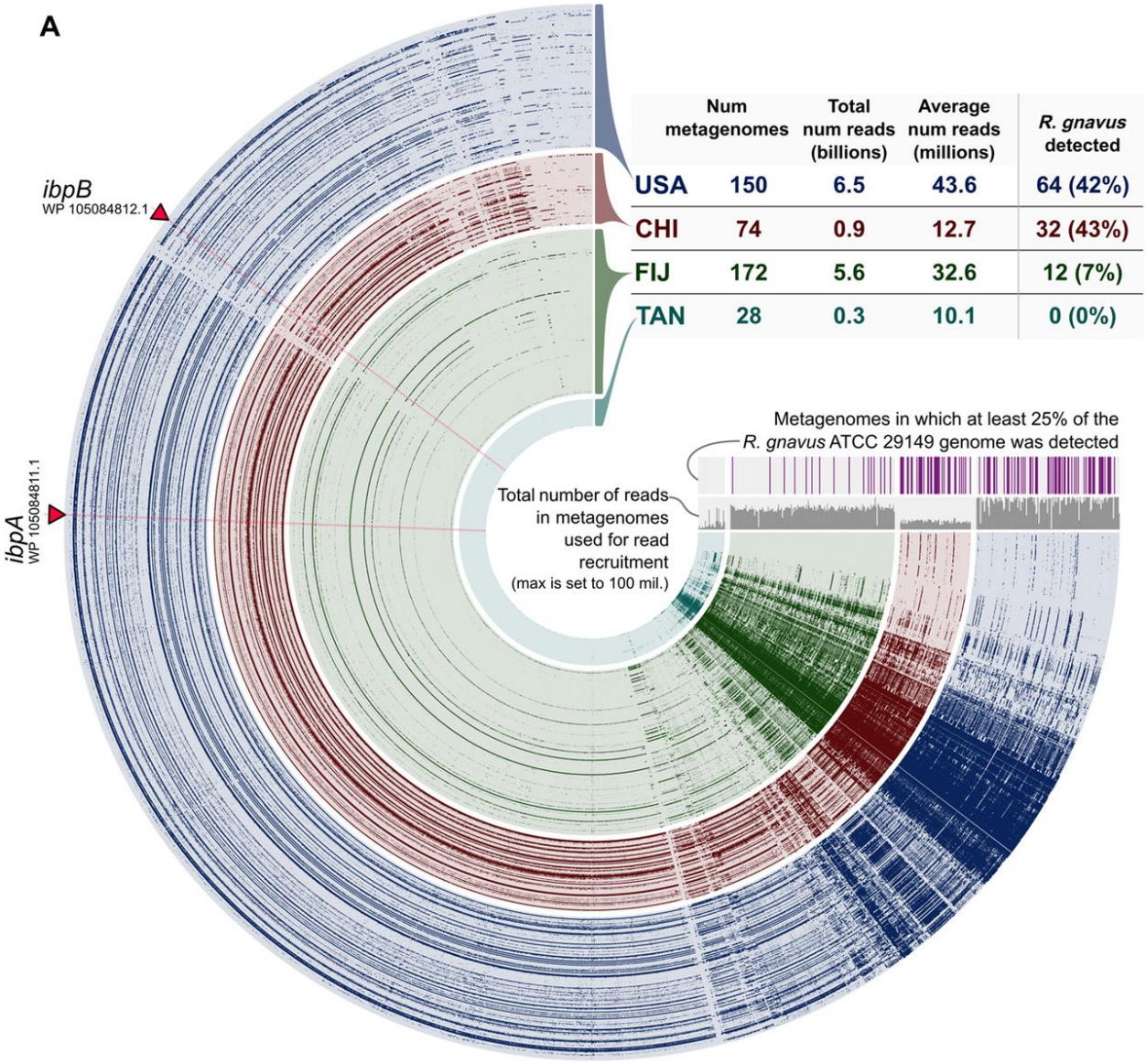


Figure 4.3: Distribution of *Ruminococcus gnavus* and its superantigens across human metagenomes.

Figure 4.3 continued: Dendrogram alignment of the *R. gnavus* ATCC 29149 genome to 424 human metagenomes. Each spoke represents one gene in the *R. gnavus* genome, and each layer represents an individual human metagenome. The two superantigen genes are labeled. Intensity represents coverage of the open reading frame in the metagenome.

4.4 Materials and Methods

4.4.1 Constructing a phylogenetic tree of SCGs from mouse gut microbial isolates and metagenomes

We quality filtered metagenomic short reads according to the criteria outlined by [Minoche et al., 2011] using the illumina-utils [Eren et al., 2013b] library with the program 'iu-filter-quality-minoche' and default parameters. We then performed single assemblies of each metagenome using IDBA-UD [Peng et al., 2012] with default parameters and a minimum contig length of 1,000 nucleotides. We recruited short reads from metagenomes to cultivar genomes and metagenomic assemblies using the anvi'o v6.2 [Eren et al., 2015] metagenomic workflow, which mapped short reads to references with Bowtie2 v2.3.5 [Langmead and Salzberg, 2012], and converted SAM files into BAM files with samtools v1.9[Li et al., 2009].

We processed isolate genome and metagenomic assembled contigs using anvi'o v6.2 to (1) create an anvi'o contigs database using the program 'anvi-gen-contigs-database', during which Prodigal v2.6.3 [Hyatt et al., 2010] identified open reading frames in contigs, (2) identify bacterial single-copy core genes using the program 'anvi-run-hmms', which ran HMMER v3.2.1 [Eddy, 2011], (3) associate ribosomal proteins with taxonomy using the program 'anvi-run-scg-taxonomy', which searches ribosomal proteins against the Genome Taxonomy Database [Parks et al., 2018] using DIAMOND [Buchfink et al., 2015] (https://merenlab.org/2019/10/08/anvio-scg-taxonomy/), (4) process read recruitment results in BAM files per metagenomes using the program 'anvi-profile', and finally (5) recover

coverage values of ribosomal proteins using the program 'anvi-summarize' with the flag '–init-gene-coverages'.

We used the mean non-outlier coverage of ribosomal genes after read recruitment as an approximation of the relative abundance of the microbial population containing that gene. However, because metagenomic reads were mapped to each isolate genome individually and non-competitively, there is a risk that non-specific mapping to isolate genomes may artificially inflate the relative abundance of some isolate genomes, particularly those with multiple distinct but closely related microbial populations present in the metagenome. To account for this, we multiplied the non-outlier coverage of each isolate genome ribosomal by the detection of the entire isolate genome in the metagenome of interest.

We used a collection of 22 bacterial ribosomal genes for constructing phylogenetic trees: Ribosomal_L2, Ribosomal_S2, Ribosomal_S8, Ribosomal_S20p, Ribosomal_L4, Ribosomal_L9_C, ribosomal_L24, Ribosomal_S3_C, Ribosomal_S6, Ribosomal_S7, Ribosomal_S9, Ribosomal_S11, Ribosomal_L1, Ribosomal_L3, Ribosomal_L6, Ribosomal_L13, Ribosomal_L16, Ribosomal_L17, Ribosomal_L20, Ribosomal_L21p, and Ribosomal_L22, and Ribosomal_L27A. We exported the sequences of these genes from all isolated genomes and metagenomic assemblies using the anvi'o program 'anvi-get-sequences-for-hmm-hits', and aligned the sequences of each gene using MUSCLE v3.8.1551. We trimmed alignments using trimAl v1.4 and a gap threshold of 0.5 to remove alignment positions with gaps in greater than 50% of the sequences, and we used the anvi'o program 'anvi-script-reformat-fasta' with the '–max-percentage-gaps 50' parameter to remove sequences which consisted of greater than 50% gaps from our alignments. We manually removed sequences from our alignments that did not align to the appropriate ribosomal genes using blastp against the NCBI protein database. We then computed maximum likelihood phylogenetic trees for each alignment using IQ-TREE v1.6.12

[Nguyen et al., 2015] with the 'WAG' general matrix model [Whelan and Goldman, 2001] and 1,000 ultrafast bootstrap replicates. Phylogenetic trees and associated metadata were visualized using the anvi'o v6.2 program 'anvi-interactive' program.

4.4.2 Read-recruitment of mixed culture metagenomes to MAGs from Lee et al. 2017

See the description of metagenomic read recruitment in Chapter 2 Materials and Methods Metagenomic read recruitment, reconstructing genomes from metagenomes, determination of genome taxonomy and ANI.

4.4.3 Characterizing the occurrence of R. gnavus *across healthy human gut metagenomes*

To estimate the abundance of *R. gnavus* ATCC 29149 genome and its superantigen genes across healthy human gut metagenomes, we used Bowtie2 v2.3.2 [Langmead and Salzberg, 2012] with default parameters to recruit reads from publicly available gut metagenomes from healthy individuals. We used anvi'o v5.1 [Eren et al., 2015, 2021] to profile short metagenomic reads aligned to the *R. gnavus* ATCC 29149 genome, to estimate coverage and detection statistics per metagenome, and to visualize merged profiles in "gene mode" where the distribution of each gene of a genome is shown independently for accurate estimates of gene-level detection. To avoid overestimating "detection" as a result of nonspecific short read recruitment due to genomic regions conserved across multiple populations, we assumed that *R. gnavus* was detected in a given metagenome only if more than 25% of it was covered by at least 1X. We applied the same principle to identify metagenomes in which superantigens were detected.

CHAPTER 5

DISCUSSION

5.1 Summary of Results

In this work I demonstrated that adaptive ecological forces rather than neutral ecological forces drive colonization outcomes after FMT. Having established that colonization outcome is based on fitness in the gut environment, I then identified metabolic pathways and specific functions associated with microbial colonization and resilience in the human gut. Specific pathways implicated in microbial fitness regardless of taxonomy were those for the synthesis of seven out of nine essential amino acids and six out of seven B vitamins. Within the genus *Bifidobacterium*, we further identified a PEP-PTS system, multi-drug efflux pumps, bile acid efflux pumps, and acid tolerance as likely drivers of increased fitness in the gut environment. We also found associations between genome size and global prevalence with resilience in the gut. We found that healthy individual's gut environments contained a relatively equal distribution of microbial populations which did or did not encode these functions, however; in the inflamed gut environment there were many more microbial populations encoding the pathways and functions associated with fitness. This brought to light a novel null hypothesis: microbes associated with inflammatory states in the human gut are not causative of those states, but are being selected for by challenging conditions.

We then showed that gut microbial populations can undergo rapid structural alterations to their genomes after collision of donor and recipient microbial communities in the recipient gut. We observed transfer of a prophage from a transiently colonizing donor population of *A. shahii* to a recipient's native population of the same species after FMT. We also demonstrated that members of a microbial population from one individual's gut undergo structural genomic changes differently upon colonization of different individual's gut

environments through invertible promoters and DGRs. These results illustrate the subtle yet functionally relevant genomic consequences of mixing donor and recipient microbial communities together that are rarely characterized in FMT studies.

Finally, we used metagenomics to evaluate and complement wet lab studies. We found that a murine gut isolate collection was lacking in several key environmental phyla while overrepresenting others, we identified culture media that could support the growth of different microbial populations of interest from a previous study, and we expanded the scale and relevance of a study showing a novel mechanism through which the commensal gut bacteria *R. gnavus* interacts with the human immune system by leveraging publicly available global gut metagenomes.

Collectively, this dissertation provides single-nucleotide level resolution insight into the journey of microbial populations as they are transplanted into a novel environment, using FMT as a model system. Our strategy bridges the gap between 'omics studies and hypothesis testing by computationally identifying targets for mechanistic studies in real world systems to ultimately untangle the ecological and mechanistic determinants of microbial colonization and resilience of the human gut.

5.2 Concluding Remarks and Perspectives

5.2.1 Implications for microbial therapeutics

My dissertation presents a null hypothesis for the association of certain microbes with disease states. This hypothesis states that gut microbial associations with disease may not be causal, but rather the microbes associated with disease are selected for under stressful gut conditions because they are more metabolically independent. We refer to this null hypothesis as "The Dark Knight hypothesis". In the film The Dark Knight the protagonist, Batman, and secondary antagonist, Harvey Dent, both fall off a building while

fighting. Harvey Dent is killed, but Batman survives because he is wearing body armor. After this incident, Batman takes the blame for Harvey Dent's murders so that Harvey Dent can be remembered as the hero and a symbol of hope he once was. In this situation, Batman is a highly resilient individual who is essential to the city of Gotham's wellbeing. However, when he survives tremendous stress and the perceived hero of the city does not, Batman becomes a scapegoat for the tragedy that befell Gotham at the hand's of Harvey Dent. Similarly, microbial populations which are a part of the healthy human gut ecosystem that are selected for after a disease state causes great stress and perturbation to the microbiome, could be mistakenly assumed to have a causal role in the disease.

If specific microbial populations are not causative of disease states, what does this mean for microbial therapeutics? One might think that if a disease is not microbially-mediated, populating the gut ecosystem with microbes associated with health may not be sufficient to treat disease. Even if a disease is caused by a dysbiotic gut microbiome, microbes associated with healthy gut conditions may not be able to colonize and persist under the selective environmental conditions associated with disease states, like inflammation. However, these considerations do not preclude the development of successful probiotics or LBPs. Bacteria can deliver molecules or functions to the gut ecosystem that can mediate disease symptoms, regardless of whether the disease is caused by the microbiome. By carefully selecting microbes that are able to successfully colonize and persist in multiple individuals' guts under stressful conditions, or by identifying traits required for these properties, microbial therapeutics can be developed which can colonize inhospitable gut environments.

My dissertation also raises considerations for the development of safe and effective additive microbial therapeutics in terms of the ecology and evolution of a bacterial strain after it colonizes the gut. My work led to the discovery that members of the same microbial population from the same host pool can evolve rapidly and differently in different

recipients' guts, with likely functional consequences. Rapid and differential adaptation to different individuals' gut environments could theoretically be a positive trait in a probiotic or LBP strain. Such adaptation could increase the likelihood of colonization and therefore the long term efficacy of a therapeutic. However; the functional consequences of such adaptations may interfere with the intended role of a strain and lead to inconsistent therapeutic results across individuals, and in the worst case scenario, these adaptations could lead to harmful impacts in patients. Safety considerations for microbial therapeutics must take into account not only how an individual microbe may adapt to a new host ecosystem, but also how an introduced microbe may modify the genetic content and functioning of native microbial populations. My work demonstrates that even transient microbes, which ultimately fail to colonize a recipient's gut, can transfer mobile elements to pre-existing populations. In theory, therapeutic microbes could be engineered to deliver desirable functions to resident microbes using horizontal gene transfer, but practical concerns include ensuring that potentially harmful functions, like antibiotic resistance, are not transferred to the recipient's microbiota. Development of safe additive microbial therapeutics should therefore screen for invertons, DGRs, mobile elements, and other genomic features that could allow for rapid and inconsistent evolution of a microbial strain across different individuals.

5.2.2 Empowering microbiologists to use computational tools through anvi'o

The findings in this dissertation would not have been possible without the analysis of very large and very complicated microbial 'omics datasets. As a classically trained wet-lab microbiologist, when I began my PhD my computational training was minimal, and my experience with tools for analysis of large 'omics datasets was limited to those that function as black-boxes: pre-defined workflows which require little to no understanding of the

many bioinformatics steps that they carry out. However, over the course of my dissertation work, I was empowered by the use of the modular, well documented, open-source software anvi'o (Figure 5.1). Anvi'o allowed me complete control over all data analysis steps and the freedom to explore my data through helpful interactive visualizations and summary statistics. Most importantly, anvi'o allowed me to ask my own questions of my data, rather than confining my curiosity to the pre-existing workflows that had become standard in my field. Not only was anvi'o essential to the work outlined in this dissertation, but this work also informed the development: there were features that I needed for my project, and I was able to work with anvi'o developers to make them a reality [Eren et al., 2021].

Anvi'o empowered me to make informed decisions about the bioinformatic tools I used in this project from the very beginning. Upon initially receiving the metagenomic sequencing from the longitudinal FMT study that lay the foundation for chapters 2 and 3 of this dissertation, I needed to assemble these sequences into contigs for eventual binning. There are myriad different assembly software to choose from, and I needed a way to decide which to use. I assembled the same metagenomes with several different options, but needed a way to evaluate which assembly was best. So I worked with a developer of anvi'o to implement a new anvi'o tool that could compute and visualize assembly summary statistics for easy comparison (Figure 5.2). Ultimately, my need to choose an assembler resulted in the creation of anvi-display-contigs-stats, which I then used to make an informed decision about the tools I would use going forward [Eren et al., 2021].

The freedom granted by anvi'o to investigate a dataset deeply not only informed which tools I would use to analyze my data, but contributed directly to findings in my dissertation. Having assembled my contigs, I was presented with the challenge of binning them into MAGs. The most common binning method is automatic: software exists that will cluster and bin contigs based on differential coverage and tetranucleotide frequency. However,

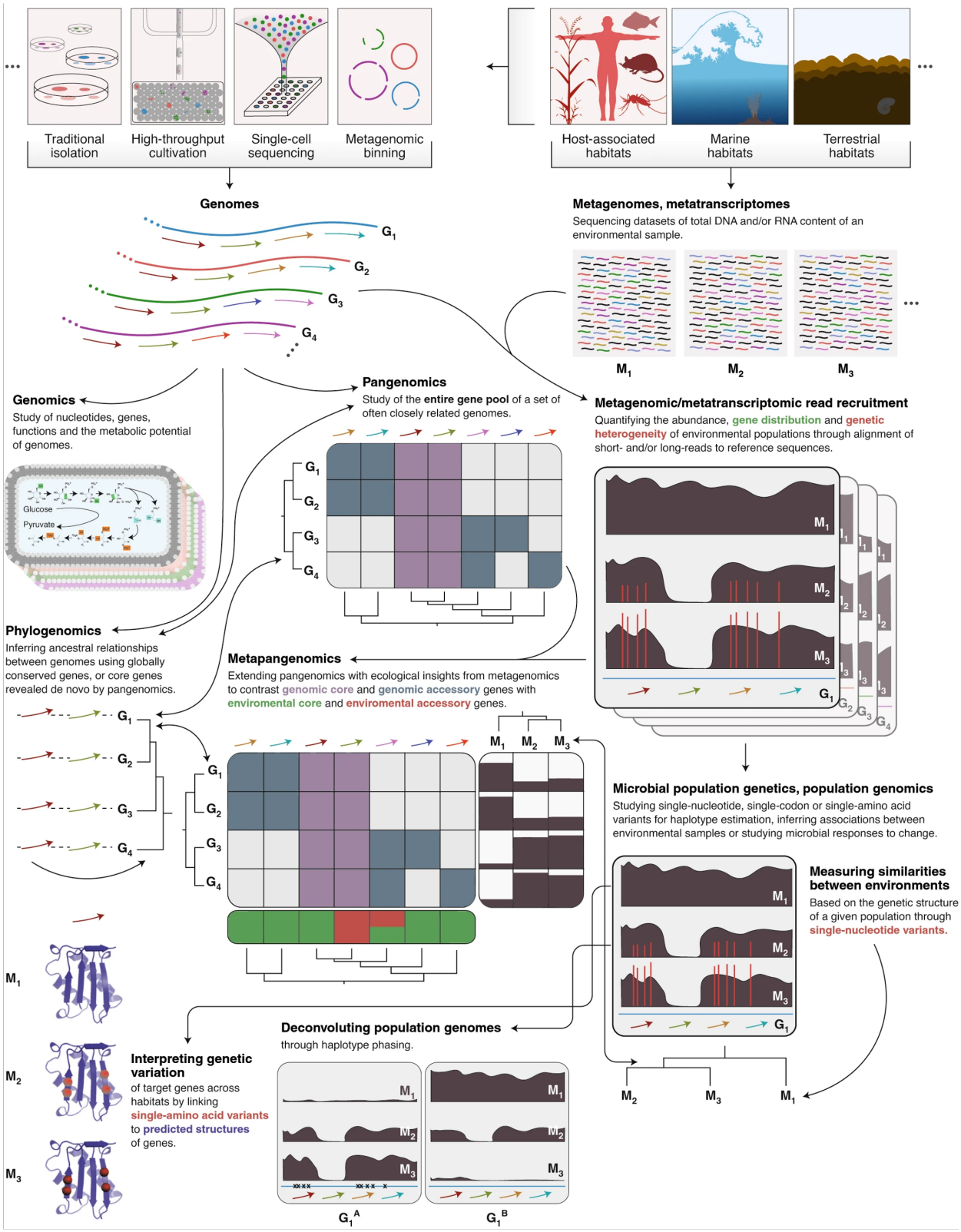
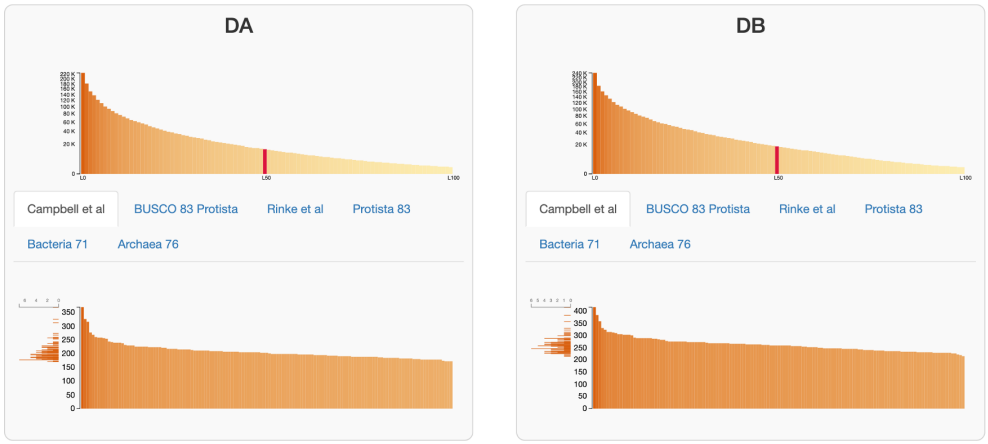


Figure 5.1: Schematic representing the modular anvi'o ecosystem and its myriad applications. [Eren et al., 2021]



Round big numbers Colorify output

Contigs Stats	DA	DB
Total Length	669,760,170	767,166,843
Num Contigs	118,736	118,080
Num Contigs > 100 kb	337	500
Num Contigs > 50 kb	1,546	2,212
Num Contigs > 20 kb	6,146	8,100
Num Contigs > 10 kb	14,073	16,772
Num Contigs > 5 kb	28,025	30,294
Num Contigs > 2.5 kb	53,891	54,311
Longest Contig	494,716	466,674
Shortest Contig	1,000	1,000
Num Genes (prodigal)	698,827	786,390
L50	10,158	9,011
L75	33,581	28,898
L90	70,538	65,156
N50	13,354	18,243
N75	4,129	5,287
N90	1,859	2,032

Raw number of HMM Hits	DA	DB
Archaea_76	7,444	9,136
BUSCO_83_Protista	806	953
Bacteria_71	14,262	17,982
Campbell_et_al	28,866	36,438
Protista_83	806	953
Ribosomal_RNAs	111	152
Rinke_et_al	18,477	22,698

Approx. number of genomes [?]	DA	DB
bacteria (Campbell_et_al)	179	248
archaea (Archaea_76)	0	0
eukarya (Protista_83)	0	0
eukarya (BUSCO_83_Protista)	0	0
bacteria (Bacteria_71)	184	236
archaea (Rinke_et_al)	0	0

Figure 5.2: The anvi-display-contigs-stats page compares two different metagenomic assemblies (DA and DB).

automatic binning can lead to poor quality, and even contaminated, MAGs [Chen et al., 2020]. An alternative is to bin contigs manually, something that anvi'o makes possible by visualizing all contigs and their coverages across samples, in an interactive dendrogram display that allows the user to cluster these contigs with different metrics, and even click on individual contigs to more closely inspect their coverage graphs and gene content. Manual binning in anvi'o is more time consuming than automatic binning, but inspecting my data closely not only taught me more about the structure of genomes in general, but it also allowed me to make scientific discoveries that would not have been possible if I had used an automatic binning algorithm. When binning the *A. shahii* MAG featured in chapter 3 of this dissertation, I identified several contigs that had clustered on a branch next to the contigs I had determined to be *A. shahii*. They were not clustered as closely to the other *A. shahii* contigs because their differential coverage across samples was not consistent with the other contigs. An automatic binning algorithm would have likely excluded these contigs from the *A. shahii* bin, but I decided to evaluate them further. I inspected the gene annotations on these contigs and noticed that they contained many prophage genes. Additionally, I saw that only one half of one of the contigs had coverage in some samples. Finally, I performed a nucleotide blast of the contig sequences to the NCBI's genome database, and saw that sections of some of the contigs were identical to sequences in known *A. shahii* genomes. These clues led me to suspect that these contigs included a prophage or other mobile element that was present in some samples and absent in others, and I decided to keep them in my *A. shahii* genome bin. Much later, through contig extension, I was able to prove that the prophage on these contigs was indeed linked to the rest of the *A. shahii* genome, which allowed me to accept my prophage hypothesis and prove that a donor *A. shahii* subpopulation had transferred this prophage to a recipient's native *A. shahii* subpopulation, a conclusion that perfectly explains the inconsistent differential coverage of this region. Had I used a black-box binning approach

that did not allow me to interrogate inconsistencies in my data instead of anvi'o, I would never have made the discovery that even transient colonizers of the gut can change the gene content of resident populations through horizontal gene transfer.

The stories that make up my dissertation are not the only examples of exciting discoveries that come as a result of empowering microbiologists with flexible computational tools. I was fortunate to be a part of a study where microbiologists used anvi'o to bin and analyze bacterial parasite MAGs from *Culex pipiens*, or mosquito, ovaries. The successful reconstruction of several *Wolbachia* genomes led to the discovery of the first *Wolbachia* plasmid. This plasmid has significant implications for potential genetic engineering of *Wolbachia* to control mosquito populations and limit transmission of mosquito-borne disease by [Reveillaud et al., 2019]. There are myriad other fascinating microbiological discoveries made using anvi'o, including the determinants of niche partitioning of bacteria in the human oral cavity, and the evolutionary forces acting on one of the world's most abundant ocean bacteria, SAR11 [Delmont et al., 2019, Shaiber et al., 2020]. When microbiologists collaborate with software developers and programmers on equal footing, empowering one another to harness and implement new computational tools rather than merely passing data back and forth, new and innovative research directions become possible.

REFERENCES

Varun Aggarwala, Ilaria Mogno, Zhihua Li, Chao Yang, Graham J Britton, Alice Chen-Liaw, Josephine Mitcham, Gerold Bongers, Dirk Gevers, Jose C Clemente, Jean-Frederic Colombel, Ari Grinspan, and Jeremiah Faith. Quantification of discrete gut bacterial strains following fecal transplantation for recurrent clostridioides difficile infection demonstrates long-term stable engraftment in non-relapsing recipients. September 2020.

Niyaz Ahmed, Ulrich Dobrindt, Jörg Hacker, and Seyed E Hasnain. Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. *Nat. Rev. Microbiol.*, 6(5):387–394, May 2008.

Stephen J Allen, Kathie Wareham, Duolao Wang, Caroline Bradley, Hayley Hutchings, Wyn Harris, Anjan Dhar, Helga Brown, Alwyn Foden, Michael B Gravenor, and Dietrich Mack. Lactobacilli and bifidobacteria in the prevention of antibiotic-associated diarrhoea and clostridium difficile diarrhoea in older inpatients (PLACIDE): a randomised, double-blind, placebo-controlled, multicentre trial. *Lancet*, 382(9900):1249–1257, October 2013.

Cátia Almeida, Rita Oliveira, Raquel Soares, and Pedro Barata. Influence of gut microbiota dysbiosis on brain function: a systematic review. *Porto Biomed J*, 5(2), March 2020.

Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson, and Christopher Quince. Binning metagenomic contigs by coverage and composition. *Nat. Methods*, 11 (11):1144–1146, November 2014.

S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, October 1990.

Amnon Amir, Daniel McDonald, Jose A Navas-Molina, Evguenia Kopylova, James T Morton, Zhenjiang Zech Xu, Eric P Kightley, Luke R Thompson, Embriette R Hyde, Antonio Gonzalez, and Rob Knight. Deblur rapidly resolves Single-Nucleotide community sequence patterns. *mSystems*, 2(2), March 2017.

Takuya Aramaki, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252, April 2020.

Silvia Arboleya, Claire Watkins, Catherine Stanton, and R Paul Ross. Gut bifidobacteria populations in human health and aging. *Front. Microbiol.*, 7:1204, August 2016.

Marie-Claire Arrieta, Jens Walter, and B Brett Finlay. Human Microbiota-Associated mice: A model with challenges. *Cell Host Microbe*, 19(5):575–578, May 2016.

Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R Mende, Gabriel R Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, Marcelo Bertalan, Natalia Borruel, Francesc Casellas, Leyden Fernandez, Laurent Gautier, Torben Hansen, Masahira Hattori, Tetsuya Hayashi, Michiel Kleerebezem, Ken Kurokawa, Marion Leclerc, Florence Levenez, Chaysavanh Manichanh, H Bjørn Nielsen, Trine Nielsen, Nicolas Pons, Julie Poulain, Junjie Qin, Thomas Sicheritz-Ponten, Sebastian Tims, David Torrents, Edgardo Ugarte, Erwin G Zoetendal, Jun Wang, Francisco Guarner, Oluf Pedersen, Willem M de Vos, Søren Brunak, Joel Doré, MetaHIT Consortium, María Antolín, François Artiguenave, Hervé M Blottiere, Mathieu Almeida, Christian Brechot, Carlos Cara, Christian Chervaux, Antonella Cultrone, Christine Delorme, Gérard Denariaz, Rozenn Dervyn, Konrad U Foerstner, Carsten Friss, Maarten van de Guchte, Eric Guedon, Florence Haimet, Wolfgang Huber, Johan van Hylckama-Vlieg, Alexandre Jamet, Catherine Juste, Ghalia Kaci, Jan Knol, Omar Lakhdari, Severine Layec, Karine Le Roux, Emmanuelle Maguin, Alexandre Mérieux, Raquel Melo Minardi, Christine M'rini, Jean Muller, Raish Oozeer, Julian Parkhill, Pierre Renault, Maria Rescigno, Nicolas Sanchez, Shinichi Sunagawa, Antonio Torrejon, Keith Turner, Gaetana Vandemeulebrouck, Encarna Varela, Yohanan Winogradsky, Georg Zeller, Jean Weissenbach, S Dusko Ehrlich, and Peer Bork. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, May 2011.

Gino J E Baart, Bert Zomer, Alex de Haan, Leo A van der Pol, E Coen Beuvery, Johannes Tramper, and Dirk E Martens. Modeling neisseria meningitidis metabolism: from genome to metabolic fluxes. *Genome Biol.*, 8(7):R136, 2007.

Fredrik Bäckhed, Claire M Fraser, Yehuda Ringel, Mary Ellen Sanders, R Balfour Sartor, Philip M Sherman, James Versalovic, Vincent Young, and B Brett Finlay. Defining a healthy human gut microbiome: current concepts, future directions, and clinical applications. *Cell Host Microbe*, 12(5):611–622, November 2012.

Aïda Bafeta, Mitsuki Koh, Carolina Riveros, and Philippe Ravaud. Harms reporting in randomized controlled trials of interventions aimed at modifying microbiota. *Ann. Intern. Med.*, 170(2):143–144, January 2019.

Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, 37(Web Server issue):W202–8, July 2009.

Rodolphe Barrangou, Elizabeth P Briczinski, Lindsay L Traeger, Joseph R Loquasto, Melissa Richards, Philippe Horvath, Anne-Claire Coûté-Monvoisin, Gregory Leyer, Snjezana Rendulic, James L Steele, Jeffery R Broadbent, Taylor Oberg, Edward G Dudley, Stephan Schuster, Dennis A Romero, and Robert F Roberts. Comparison of the complete genome sequences of bifidobacterium animalis subsp. lactis DSM 10140 and bl-04. *J. Bacteriol.*, 191(13):4144–4151, July 2009.

Daniel C Baumgart and Simon R Carding. Inflammatory bowel disease: cause and immunobiology. *Lancet*, 369(9573):1627–1640, May 2007.

Simone Becattini, Ying Taur, and Eric G Pamer. Antibiotic-Induced changes in the intestinal microbiota and disease. *Trends Mol. Med.*, 22(6):458–478, June 2016.

Julia Benckert, Nina Schmolka, Cornelia Kreschel, Markus Josef Zoller, Andreas Sturm, Bertram Wiedenmann, and Hedda Wardemann. The majority of intestinal IgA+ and IgG+ plasmablasts in the human gut are antigen-specific. *J. Clin. Invest.*, 121(5):1946–1955, May 2011.

Sean Benler, Ana Georgina Cobián-Güemes, Katelyn McNair, Shr-Hau Hung, Kyle Levi, Rob Edwards, and Forest Rohwer. A diversity-generating retroelement encoded by a globally ubiquitous bacteroides phage. *Microbiome*, 6(1):191, October 2018.

Gabriele Berg, Daria Rybakova, Doreen Fischer, Tomislav Cernava, Marie-Christine Champomier Vergès, Trevor Charles, Xiaoyulong Chen, Luca Cocolin, Kellye Eversole, Gema Herrero Corral, Maria Kazou, Linda Kinkel, Lene Lange, Nelson Lima, Alexander Loy, James A Macklin, Emmanuelle Maguin, Tim Mauchline, Ryan McClure, Birgit Mitter, Matthew Ryan, Inga Sarand, Hauke Smidt, Bettina Schelkle, Hugo Roume, G Seghal Kiran, Joseph Selvin, Rafael Soares Correa de Souza, Leo van Overbeek, Brajesh K Singh, Michael Wagner, Aaron Walsh, Angela Sessitsch, and Michael Schloter. Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, 8(1):103, June 2020.

Denis Bertrand, Jim Shaw, Manesh Kalathiyappan, Amanda Hui Qi Ng, M Senthil Kumar, Chenhao Li, Mirta Dvornicic, Janja Paliska Soldo, Jia Yu Koh, Chengxuan Tong, Oon Tek Ng, Timothy Barkham, Barnaby Young, Kalisvar Marimuthu, Kern Rei Chng, Mile Sikic, and Niranjan Nagarajan. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.*, 37(8):937–944, August 2019.

Marc Gh Besselink, Hjalmar C van Santvoort, Erik Buskens, Marja A Boermeester, Harry van Goor, Harro M Timmerman, Vincent B Nieuwenhuijs, Thomas L Bollen, Bert van Ramshorst, Ben Jm Witteman, Camiel Rosman, Rutger J Ploeg, Menno A Brink, Alexander Fm Schaapherder, Cornelis Hc Dejong, Peter J Wahab, Cees Jhm van Laarhoven, Erwin van der Harst, Casper Hj van Eijck, Miguel A Cuesta, Louis Ma Akkermans, Hein G Gooszen, and Dutch Acute Pancreatitis Study Group. Probiotic prophylaxis in predicted severe acute pancreatitis: a randomised, double-blind, placebo-controlled trial. *Lancet*, 371(9613):651–659, February 2008.

Hans K Biesalski. Nutrition meets the microbiome: micronutrients and the microbiota. *Ann. N. Y. Acad. Sci.*, 1372(1):53–64, May 2016.

Ran Blekhman, Julia K Goodrich, Katherine Huang, Qi Sun, Robert Bukowski, Jordana T Bell, Timothy D Spector, Alon Keinan, Ruth E Ley, Dirk Gevers, and Andrew G

Clark. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.*, 16(1):191, September 2015.

Robert M Bowers, Nikos C Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T B K Reddy, Frederik Schulz, Jessica Jarett, Adam R Rivers, Emiley A Eloe-Fadrosh, Susannah G Tringe, Natalia N Ivanova, Alex Copeland, Alicia Clum, Eric D Becraft, Rex R Malmstrom, Bruce Birren, Mircea Podar, Peer Bork, George M Weinstock, George M Garrity, Jeremy A Dodsworth, Shibu Yooseph, Granger Sutton, Frank O Glöckner, Jack A Gilbert, William C Nelson, Steven J Hallam, Sean P Jungbluth, Thijs J G Ettema, Scott Tighe, Konstantinos T Konstantinidis, Wen-Tso Liu, Brett J Baker, Thomas Rattei, Jonathan A Eisen, Brian Hedlund, Katherine D McMahon, Noah Fierer, Rob Knight, Rob Finn, Guy Cochrane, Ilene Karsch-Mizrachi, Gene W Tyson, Christian Rinke, Genome Standards Consortium, Alla Lapidus, Folker Meyer, Pelin Yilmaz, Donovan H Parks, A M Eren, Lynn Schriml, Jillian F Banfield, Philip Hugenholtz, and Tanja Woyke. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.*, 35(8):725–731, August 2017.

Felix Broecker, Jochen Klumpp, Markus Schuppler, Giancarlo Russo, Luc Biedermann, Michael Hombach, Gerhard Rogler, and Karin Moelling. Long-term changes of bacterial and viral compositions in the intestine of a recovered clostridium difficile patient after fecal microbiota transplantation. *Cold Spring Harb Mol Case Stud*, 2(1):a000448, January 2016.

Harald Brüssow. Probiotics and prebiotics in clinical tests: an update. *F1000Res.*, 8, July 2019.

Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, 12(1):59–60, January 2015.

Jeffrey J Bunker and Albert Bendelac. IgA responses to microbiota. *Immunity*, 49(2): 211–224, August 2018.

Jeffrey J Bunker, Theodore M Flynn, Jason C Koval, Dustin G Shaw, Marlies Meisel, Benjamin D McDonald, Isabel E Ishizuka, Alexander L Dent, Patrick C Wilson, Bana Jabri, Dionysios A Antonopoulos, and Albert Bendelac. Innate and adaptive humoral responses coat distinct commensal bacteria with immunoglobulin a. *Immunity*, 43(3): 541–553, September 2015.

Jeffrey J Bunker, Steven A Erickson, Theodore M Flynn, Carole Henry, Jason C Koval, Marlies Meisel, Bana Jabri, Dionysios A Antonopoulos, Patrick C Wilson, and Albert Bendelac. Natural polyreactive IgA antibodies coat the intestinal microbiota. *Science*, 358(6361), October 2017.

Jeffrey J Bunker, Christoph Drees, Andrea R Watson, Catherine H Plunkett, Cathryn R Nagler, Olaf Schneewind, A Murat Eren, and Albert Bendelac. B cell superantigens in the human intestinal microbiota. *Sci. Transl. Med.*, 11(507), August 2019.

Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.*, 11 (12):2639–2643, December 2017.

G Cammarota, L Masucci, G Ianiro, S Bibbò, G Dinoi, G Costamagna, M Sanguinetti, and A Gasbarrini. Randomised clinical trial: faecal microbiota transplantation by colonoscopy vs. vancomycin for the treatment of recurrent clostridium difficile infection. *Aliment. Pharmacol. Ther.*, 41(9):835–843, May 2015.

James H Campbell, Patrick O'Donoghue, Alisha G Campbell, Patrick Schwientek, Alexander Sczyrba, Tanja Woyke, Dieter Söll, and Mircea Podar. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc. Natl. Acad. Sci. U. S. A.*, 110(14):5540–5545, April 2013.

Paul Carini, Laura Steindler, Sara Beszteri, and Stephen J Giovannoni. Nutrient requirements for growth of the extreme oligotroph 'candidatus pelagibacter ubique' HTCC1062 on a defined medium. *ISME J.*, 7(3):592–602, March 2013.

Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*, November 2019.

Lin-Xing Chen, Karthik Anantharaman, Alon Shaiber, A Murat Eren, and Jillian F Banfield. Accurate and complete genomes from metagenomes. *Genome Res.*, 30(3):315–333, March 2020.

Hsiang-Lin Chien, Hui-Yu Huang, and Cheng-Chun Chou. Transformation of isoflavone phytoestrogens during the fermentation of soymilk with lactic acid bacteria and bifidobacteria. *Food Microbiol.*, 23(8):772–778, December 2006.

Janet Chow, Haiqing Tang, and Sarkis K Mazmanian. Pathobionts of the gastrointestinal microbiota and inflammatory disease. *Curr. Opin. Immunol.*, 23(4):473–480, August 2011.

Hachung Chung, Sünje J Pamp, Jonathan A Hill, Neeraj K Surana, Sanna M Edelman, Erin B Troy, Nicola C Reading, Eduardo J Villablanca, Sen Wang, Jorge R Mora, Yoshinori Umesaki, Diane Mathis, Christophe Benoist, David A Relman, and Dennis L Kasper. Gut immune maturation depends on colonization with a host-specific microbiota. *Cell*, 149(7):1578–1593, June 2012.

Marcus J Claesson, Siobhán Cusack, Orla O'Sullivan, Rachel Greene-Diniz, Heleen de Weerd, Edel Flannery, Julian R Marchesi, Daniel Falush, Timothy Dinan, Gerald Fitzgerald, Catherine Stanton, Douwe van Sinderen, Michael O'Connor, Norma Harnedy, Kieran O'Connor, Colm Henry, Denis O'Mahony, Anthony P Fitzgerald, Fergus Shanahan, Cillian Twomey, Colin Hill, R Paul Ross, and Paul W O'Toole. Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl. Acad. Sci. U. S. A.*, 108 Suppl 1:4586–4591, March 2011.

Adam G Clooney, Julia Eckenberger, Emilio Laserna-Mendieta, Kathryn A Sexton, Matthew T Bernstein, Kathy Vagianos, Michael Sargent, Feargal J Ryan, Carthage Moran, Donal Sheehan, Roy D Sleator, Laura E Targownik, Charles N Bernstein, Fergus Shanahan, and Marcus J Claesson. Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. *Gut*, 70(3):499–510, March 2021.

M Carmen Collado, Muriel Derrien, Erika Isolauri, Willem M de Vos, and Seppo Salminen. Intestinal integrity and akkermansia muciniphila, a mucin-degrading member of the intestinal microbiota present in infants, adults, and the elderly. *Appl. Environ. Microbiol.*, 73(23):7767–7770, December 2007.

Elizabeth K Costello, Keaton Stagaman, Les Dethlefsen, Brendan J M Bohannan, and David A Relman. The application of ecological theory toward an understanding of the human microbiome. *Science*, 336(6086):1255–1262, June 2012.

Michael J Coyne, Maria Chatzidaki-Livanis, Lawrence C Paoletti, and Laurie E Comstock. Role of glycan synthesis in colonization of the mammalian gut by the bacterial symbiont bacteroides fragilis. *Proc. Natl. Acad. Sci. U. S. A.*, 105(35):13099–13104, September 2008.

Carlos A Cuello-Garcia, Jan L Brożek, Alessandro Fiocchi, Ruby Pawankar, Juan José Yepes-Nuñez, Luigi Terracciano, Shreyas Gandhi, Arnav Agarwal, Yuan Zhang, and Holger J Schünemann. Probiotics for the prevention of allergy: A systematic review and meta-analysis of randomized controlled trials. *J. Allergy Clin. Immunol.*, 136(4): 952–961, October 2015.

C Cunningham-Rundles. Physiology of IgA and IgA deficiency. *J. Clin. Immunol.*, 21(5): 303–309, September 2001.

Simon D Goldenberg and Blair Merrick. The role of faecal microbiota transplantation: looking beyond clostridioides difficile infection. *Ther Adv Infect Dis*, 8: 2049936120981526, January 2021.

M R D'Aimmo, P Mattarelli, B Biavati, N G Carlsson, and T Andlid. The potential of bifidobacteria as a source of natural folate. *J. Appl. Microbiol.*, 112(5):975–984, May 2012.

Lawrence A David, Arne C Materna, Jonathan Friedman, Maria I Campos-Baptista, Matthew C Blackburn, Allison Perrotta, Susan E Erdman, and Eric J Alm. Host lifestyle affects human microbiota on daily timescales. *Genome Biol.*, 15(7):R89, 2014a.

Lawrence A David, Corinne F Maurice, Rachel N Carmody, David B Gootenberg, Julie E Button, Benjamin E Wolfe, Alisha V Ling, A Sloan Devlin, Yug Varma, Michael A Fischbach, Sudha B Biddinger, Rachel J Dutton, and Peter J Turnbaugh. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563, January 2014b.

Vicky De Preter, Veerle Bulteel, Peter Suenaert, Karen Paula Geboes, Gert De Hertogh, Anja Luypaerts, Karel Geboes, Kristin Verbeke, and Paul Rutgeerts. Pouchitis, similar to active ulcerative colitis, is associated with impaired butyrate oxidation by intestinal mucosa. *Inflamm. Bowel Dis.*, 15(3):335–340, March 2009.

Willem M de Vos and Elisabeth A J de Vos. Role of the intestinal microbiome in health and disease: from correlation to causation. *Nutr. Rev.*, 70 Suppl 1:S45–56, August 2012.

Zachariah DeFilipp, Patricia P Bloom, Mariam Torres Soto, Michael K Mansour, Mohamad R A Sater, Miriam H Huntley, Sarah Turbett, Raymond T Chung, Yi-Bin Chen, and Elizabeth L Hohmann. Drug-Resistant e. coli bacteremia transmitted by fecal microbiota transplant. *N. Engl. J. Med.*, 381(21):2043–2050, November 2019.

Tom O Delmont and A Murat Eren. Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ*, 4:e1839, March 2016.

Tom O Delmont, Christopher Quince, Alon Shaiber, Özcan C Esen, Sonny Tm Lee, Michael S Rappé, Sandra L McLellan, Sebastian Lücker, and A Murat Eren. Nitrogen-fixing populations of planctomycetes and proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol*, 3(7):804–813, July 2018.

Tom O Delmont, Evan Kiefl, Ozsel Kilinc, Ozcan C Esen, Ismail Uysal, Michael S Rappé, Steven Giovannoni, and A Murat Eren. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *Elife*, 8, September 2019.

Vincent J Denef. Peering into the genetic makeup of natural microbial populations using metagenomics. In Martin F Polz and Om P Rajora, editors, *Population Genomics: Microorganisms*, pages 49–75. Springer International Publishing, Cham, 2019.

Les Dethlefsen and David 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(Supplement 1):4554–4561, March 2011.

Maria G Dominguez-Bello, Elizabeth K Costello, Monica Contreras, Magda Magris, Glida Hidalgo, Noah Fierer, and Rob 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):11971–11975, June 2010.

Glen D'Souza, Shraddha Shitut, Daniel Preussger, Ghada Yousif, Silvio Waschina, and Christian Kost. Ecology and evolution of metabolic cross-feeding interactions in bacteria. *Nat. Prod. Rep.*, 35(5):455–488, May 2018.

Erik R Dubberke, Christine H Lee, Robert Orenstein, Sahil Khanna, Gail Hecht, and Dale N Gerding. Results from a randomized, Placebo-Controlled clinical trial of a RBX2660-A Microbiota-Based drug for the prevention of recurrent clostridium difficile infection. *Clin. Infect. Dis.*, 67(8):1198–1204, September 2018.

Grégory Dubourg, Jean-Christophe Lagier, Fabrice Armougom, Catherine Robert, Gilles Audoly, Laurent Papazian, and Didier Raoult. High-level colonisation of the human gut by verrucomicrobia following broad-spectrum antibiotic treatment. *Int. J. Antimicrob. Agents*, 41(2):149–155, February 2013.

Juliana Durack and Susan V Lynch. The gut microbiome: Relationships with disease and opportunities for therapy. *J. Exp. Med.*, 216(1):20–40, January 2019.

Sean R Eddy. Accelerated profile HMM searches. *PLoS Comput. Biol.*, 7(10):e1002195, October 2011.

Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, March 2004.

B Eiseman, W Silen, G S Bascom, and A J Kauvar. Fecal enema as an adjunct in the treatment of pseudomembranous enterocolitis. *Surgery*, 44(5):854–859, November 1958.

Michael Eisenstein. The hunt for a healthy microbiome. *Nature*, 577(7792):S6–S8, January 2020.

Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, Erik L L Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C E Tosatto, and Robert D Finn. The pfam protein families database in 2019. *Nucleic Acids Res.*, 47(D1):D427–D432, January 2019.

Gary W Elmer, Christina M Surawicz, and Lynne V McFarland. Biotherapeutic agents: A neglected modality for the treatment and prevention of selected intestinal and vaginal infections. *JAMA*, 275(11):870–876, March 1996.

A J Enright, S Van Dongen, and C A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30(7):1575–1584, April 2002.

A Murat Eren. Microbial 'omics: An introduction. `https://merenlab.org/momics/`. Accessed: 2021-6-15.

A Murat Eren, Loïs Maignien, Woo Jun Sul, Leslie G Murphy, Sharon L Grim, Hilary G Morrison, and Mitchell L Sogin. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.*, 4(12):1111–1119, 2013a.

A Murat Eren, Joseph H Vineis, Hilary G Morrison, and Mitchell L Sogin. A filtering method to generate high quality short reads using illumina paired-end technology. *PLoS One*, 8(6):e66643, June 2013b.

A Murat Eren, Hilary G Morrison, Pamela J Lescault, Julie Reveillaud, Joseph H Vineis, and Mitchell L Sogin. Minimum entropy decomposition: Unsupervised oligotyping for

sensitive partitioning of high-throughput marker gene sequences. *ISME J.*, 9(4):968–979, October 2014.

A Murat Eren, Özcan C Esen, Christopher Quince, Joseph H Vineis, Hilary G Morrison, Mitchell L Sogin, and Tom O Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, October 2015.

A Murat Eren, Evan Kiefl, Alon Shaiber, Iva Veseli, Samuel E Miller, Matthew S Schechter, Isaac Fink, Jessica N Pan, Mahmoud Yousef, Emily C Fogarty, Florian Trigodet, Andrea R Watson, Özcan C Esen, Ryan M Moore, Quentin Clayssen, Michael D Lee, Veronika Kivenson, Elaina D Graham, Bryan D Merrill, Antti Karkman, Daniel Blankenberg, John M Eppley, Andreas Sjödin, Jarrod J Scott, Xabier Vázquez-Campos, Luke J McKay, Elizabeth A McDaniel, Sarah L R Stevens, Rika E Anderson, Jessika Fuessel, Antonio Fernandez-Guerra, Lois Maignien, Tom O Delmont, and Amy D Willis. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol*, 6(1): 3–6, January 2021.

Yong Fan and Oluf Pedersen. Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.*, 19(1):55–71, January 2021.

Lihui Feng, Arjun S Raman, Matthew C Hibberd, Jiye Cheng, Nicholas W Griffin, Yangqing Peng, Semen A Leyn, Dmitry A Rodionov, Andrei L Osterman, and Jeffrey I Gordon. Identifying determinants of bacterial fitness in a model of human gut microbial succession. *Proc. Natl. Acad. Sci. U. S. A.*, 117(5):2622–2633, February 2020.

Robert D Finn, Jody Clements, and Sean R Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, 39(Web Server issue):W29–37, July 2011.

Mariel M Finucane, Thomas J Sharpton, Timothy J Laurent, and Katherine S Pollard. A taxonomic signature of obesity in the microbiome? getting to the guts of the matter. *PLoS One*, 9(1):e84689, January 2014.

Michael A Fischbach. Microbiome: Focus on causation and mechanism. *Cell*, 174(4): 785–790, August 2018.

Kevin R Foster, Jonas Schluter, Katharine Z Coyte, and Seth Rakoff-Nahoum. The evolution of the host microbiome as an ecosystem on a leash. *Nature*, 548(7665):43–51, August 2017.

Kei E Fujimura and Susan V Lynch. Microbiota in allergy and asthma and the emerging relationship with the gut microbiome. *Cell Host Microbe*, 17(5):592–602, May 2015.

Tadashi Fukami. Historical contingency in community assembly: Integrating niches, species pools, and priority effects. *Annu. Rev. Ecol. Evol. Syst.*, 46(1):1–23, December 2015.

Tadashi Fukami, Hubertus J E Beaumont, Xue-Xian Zhang, and Paul B Rainey. Immigration history controls diversification in experimental adaptive radiation. *Nature*, 446 (7134):436–439, March 2007.

Michael Y Galperin, Yuri I Wolf, Kira S Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V Koonin. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.*, 49(D1):D274–D281, January 2021.

Benedikt Geier, Emilia M Sogin, Dolma Michellod, Moritz Janda, Mario Kompauer, Bernhard Spengler, Nicole Dubilier, and Manuel Liebeke. Spatial metabolomics of in situ host-microbe interactions at the micrometre scale. *Nat Microbiol*, 5(3):498–510, March 2020.

G R Gibson and M B Roberfroid. Dietary modulation of the human colonic microbiota: introducing the concept of prebiotics. *J. Nutr.*, 125(6):1401–1412, June 1995.

Ana M P Gomes and F Xavier Malcata. Bifidobacterium spp. and lactobacillus acidophilus: biological, biochemical, technological and therapeutical properties relevant for use as probiotics. *Trends Food Sci. Technol.*, 10(4):139–157, April 1999.

Julia K Goodrich, Jillian L Waters, Angela C Poole, Jessica L Sutter, Omry Koren, Ran Blekhman, Michelle Beaumont, William Van Treuren, Rob Knight, Jordana T Bell, Timothy D Spector, Andrew G Clark, and Ruth E Ley. Human genetics shape the gut microbiome. *Cell*, 159(4):789–799, November 2014.

Martin J Grehan, Thomas Julius Borody, Sharyn M Leis, Jordana Campbell, Hazel Mitchell, and Antony Wettstein. Durable alteration of the colonic microbiota by the administration of donor fecal flora. *J. Clin. Gastroenterol.*, 44(8):551–561, September 2010.

Alice Y Guh, Yi Mu, Lisa G Winston, Helen Johnston, Danyel Olson, Monica M Farley, Lucy E Wilson, Stacy M Holzbauer, Erin C Phipps, Ghinwa K Dumyati, Zintars G Beldavs, Marion A Kainer, Maria Karlsson, Dale N Gerding, L Clifford McDonald, and Emerging Infections Program Clostridioides difficile Infection Working Group. Trends in U.S. burden of clostridioides difficile infection and outcomes. *N. Engl. J. Med.*, 382(14): 1320–1330, April 2020.

Huatao Guo, Diego Arambula, Partho Ghosh, and Jeff F Miller. Diversity-generating retroelements in phage and bacterial genomes. *Microbiol Spectr*, 2(6), December 2014.

Klaus M Herrmann and Lisa M Weaver. THE SHIKIMATE PATHWAY. *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, 50:473–503, June 1999.

Thomas C A Hitch, Afrizal Afrizal, Thomas Riedel, Antonios Kioukis, Dirk Haller, Ilias Lagkouvardos, Jörg Overmann, and Thomas Clavel. Recent advances in culture-based gut microbiome research. *Int. J. Med. Microbiol.*, 311(3):151485, April 2021.

Katharina Höfer and Andres Jäschke. Epitranscriptomics: RNA modifications in bacteria and archaea. *Microbiol Spectr*, 6(3), May 2018.

Elaine Holmes, James Kinross, Glenn R Gibson, Remy Burcelin, Wei Jia, Sven Pettersson, and Jeremy K Nicholson. Therapeutic modulation of microbiota-host metabolic interactions. *Sci. Transl. Med.*, 4(137):137rv6, June 2012.

Cristina Howard-Varona, Morgan M Lindback, G Eric Bastien, Natalie Solonenko, Ahmed A Zayed, Hobin Jang, Bill Andreopoulos, Heather M Brewer, Tijana Glavina Del Rio, Joshua N Adkins, Subhadeep Paul, Matthew B Sullivan, and Melissa B Duhaime. Phage-specific metabolic reprogramming of virocells. *ISME J.*, 14(4):881–895, April 2020.

Human Microbiome Jumpstart Reference Strains Consortium, Karen E Nelson, George M Weinstock, Sarah K Highlander, Kim C Worley, Heather Huot Creasy, Jennifer Russo Wortman, Douglas B Rusch, Makedonka Mitreva, Erica Sodergren, Asif T Chinwalla, Michael Feldgarden, Dirk Gevers, Brian J Haas, Ramana Madupu, Doyle V Ward, Bruce W Birren, Richard A Gibbs, Barbara Methe, Joseph F Petrosino, Robert L Strausberg, Granger G Sutton, Owen R White, Richard K Wilson, Scott Durkin, Michelle Gwinn Giglio, Sharvari Gujja, Clint Howarth, Chinnappa D Kodira, Nikos Kyrpides, Teena Mehta, Donna M Muzny, Matthew Pearson, Kymberlie Pepin, Amrita Pati, Xiang Qin, Chandri Yandava, Qiandong Zeng, Lan Zhang, Aaron M Berlin, Lei Chen, Theresa A Hepburn, Justin Johnson, Jamison McCorrison, Jason Miller, Pat Minx, Chad Nusbaum, Carsten Russ, Sean M Sykes, Chad M Tomlinson, Sarah Young, Wesley C Warren, Jonathan Badger, Jonathan Crabtree, Victor M Markowitz, Joshua Orvis, Andrew Cree, Steve Ferriera, Lucinda L Fulton, Robert S Fulton, Marcus Gillis, Lisa D Hemphill, Vandita Joshi, Christie Kovar, Manolito Torralba, Kris A Wetterstrand, Amr Abouellleil, Aye M Wollam, Christian J Buhay, Yan Ding, Shannon Dugan, Michael G FitzGerald, Mike Holder, Jessica Hostetler, Sandra W Clifton, Emma Allen-Vercoe, Ashlee M Earl, Candace N Farmer, Konstantinos Liolios, Michael G Surette, Qiang Xu, Craig Pohl, Katarzyna Wilczek-Boney, and Dianhui Zhu. A catalog of reference genomes from the human microbiome. *Science*, 328(5981):994–999, May 2010.

Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.

Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, March 2010.

Sandrine Isaac, Jose U Scher, Ana Djukovic, Nuria Jiménez, Dan R Littman, Steven B Abramson, Eric G Pamer, and Carles Ubeda. Short- and long-term effects of oral vancomycin on the human intestinal microbiota. *J. Antimicrob. Chemother.*, 72(1):128–136, January 2017.

Chirag Jain, Luis M Rodriguez-R, Adam M Phillippy, Konstantinos T Konstantinidis, and Srinivas Aluru. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, 9(1):5114, November 2018.

Jonna Jalanka, Eero Mattila, Hanne Jouhten, Jorn Hartman, Willem M de Vos, Perttu Arkkila, and Reetta Satokari. Long-term effects on luminal and mucosal microbiota and commonly acquired taxa in faecal microbiota transplantation for recurrent clostridium difficile infection. *BMC Med.*, 14(1):155, October 2016.

Elke Jaspers and Jörg Overmann. Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Appl. Environ. Microbiol.*, 70(8):4831–4839, August 2004.

Aashish R Jha, Emily R Davenport, Yoshina Gautam, Dinesh Bhandari, Sarmila Tandukar, Katharine M Ng, Gabriela K Fragiadakis, Susan Holmes, Guru Prasad Gautam, Jeff Leach, Jeevan Bahadur Sherchand, Carlos D Bustamante, and Justin L Sonnenburg. Gut microbiome transition across a lifestyle gradient in himalaya. *PLoS Biol.*, 16(11): e2005396, November 2018.

Xiaofang Jiang, A Brantley Hall, Timothy D Arthur, Damian R Plichta, Christian T Covington, Mathilde Poyet, Jessica Crothers, Peter L Moses, Andrew C Tolonen, Hera Vlamakis, Eric J Alm, and Ramnik J Xavier. Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. *Science*, 363(6423): 181–187, January 2019.

Marie Joossens, Geert Huys, Margo Cnockaert, Vicky De Preter, Kristin Verbeke, Paul Rutgeerts, Peter Vandamme, and Severine Vermeire. Dysbiosis of the faecal microbiota in patients with crohn's disease and their unaffected relatives. *Gut*, 60(5):631–637, May 2011.

Hanne Jouhten, Aki Ronkainen, Juhani Aakko, Seppo Salminen, Eero Mattila, Perttu Arkkila, and Reetta Satokari. Cultivation and genomics prove Long-Term colonization of donor's bifidobacteria in recurrent clostridioides difficile patients treated with fecal microbiota transplantation. *Front. Microbiol.*, 11:1663, July 2020.

Mikkel Jungersen, Anette Wind, Eric Johansen, Jeffrey E Christensen, Birgitte Stuer-Lauridsen, and Dorte Eskesen. The science behind the probiotic strain bifidobacterium animalis subsp. lactis BB-12(®). *Microorganisms*, 2(2):92–110, March 2014.

M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, January 2000.

Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, 42(Database issue):D199–205, January 2014.

Minoru Kanehisa, Yoko Sato, and Kanae Morishima. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.*, 428(4):726–731, February 2016.

Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, 45(D1):D353–D361, January 2017.

Dina Kao, Brandi Roach, Marisela Silva, Paul Beck, Kevin Rioux, Gilaad G Kaplan, Hsiu-Ju Chang, Stephanie Coward, Karen J Goodman, Huiping Xu, Karen Madsen, Andrew Mason, Gane Ka-Shu Wong, Juan Jovel, Jordan Patterson, and Thomas Louie. Effect of oral capsule- vs Colonoscopy-Delivered fecal microbiota transplantation on recurrent clostridium difficile infection: A randomized clinical trial. *JAMA*, 318(20):1985–1993, November 2017.

Andrew L Kau, Joseph D Planer, Jie Liu, Sindhuja Rao, Tanya Yatsunenko, Indi Trehan, Mark J Manary, Ta-Chiang Liu, Thaddeus S Stappenbeck, Kenneth M Maleta, Per Ashorn, Kathryn G Dewey, Eric R Houpt, Chyi-Song Hsieh, and Jeffrey I Gordon. Functional characterization of IgA-targeted bacterial taxa from undernourished malawian children that produce diet-dependent enteropathy. *Sci. Transl. Med.*, 7(276):276ra24, February 2015.

Colleen R Kelly, Alexander Khoruts, Christopher Staley, Michael J Sadowsky, Mortadha Abd, Mustafa Alani, Brianna Bakow, Patrizia Curran, Joyce McKenney, Allison Tisch, Steven E Reinert, Jason T Machan, and Lawrence 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, November 2016.

Alexander Khoruts and Michael J Sadowsky. Understanding the mechanisms of faecal microbiota transplantation. *Nat. Rev. Gastroenterol. Hepatol.*, 13(9):508–516, September 2016.

Alexander Khoruts, Johan Dicksved, Janet K Jansson, and Michael J Sadowsky. Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent clostridium difficile-associated diarrhea. *J. Clin. Gastroenterol.*, 44(5):354–360, May 2010.

Sohn G Kim, Simone Becattini, Thomas U Moody, Pavel V Shliaha, Eric R Littmann, Ruth Seok, Mergim Gjonbalaj, Vincent Eaton, Emily Fontana, Luigi Amoretti, Roberta Wright, Silvia Caballero, Zhong-Min X Wang, Hea-Jin Jung, Sejal M Morjaria, Ingrid M Leiner, Weige Qin, Ruben J J F Ramos, Justin R Cross, Seiko Narushima, Kenya Honda, Jonathan U Peled, Ronald C Hendrickson, Ying Taur, Marcel R M van den Brink, and Eric G Pamer. Microbiota-derived lantibiotic restores resistance against vancomycin-resistant enterococcus. *Nature*, 572(7771):665–669, August 2019.

Claas Kirchhelle. Pharming animals: a global history of antibiotics in food production (1935–2017). *Palgrave Communications*, 4(1):96, August 2018.

Michiel Kleerebezem and Elaine E Vaughan. Probiotic and gut lactobacilli and bifidobacteria: molecular approaches to study diversity and activity. *Annu. Rev. Microbiol.*, 63: 269–290, 2009.

Rob Knight, Chris Callewaert, Clarisse Marotz, Embriette R Hyde, Justine W Debelius, Daniel McDonald, and Mitchell L Sogin. The microbiome and human biology. *Annu. Rev. Genomics Hum. Genet.*, 18(1):65–86, August 2017.

Natalie C Knox, Jessica D Forbes, Gary Van Domselaar, and Charles N Bernstein. The gut microbiome as a target for IBD treatment: Are we there yet? *Curr. Treat. Options Gastroenterol.*, 17(1):115–126, March 2019.

Jeremy E Koenig, Aymé Spor, Nicholas Scalfone, Ashwana D Fricker, Jesse Stombaugh, Rob Knight, Largus T Angenent, and Ruth E Ley. Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U. S. A.*, 108 Suppl 1: 4578–4585, March 2011.

Alexander Koeppel, Elizabeth B Perry, Johannes Sikorski, Danny Krizanc, Andrew Warner, David M Ward, Alejandro P Rooney, Evelyne Brambilla, Nora Connor, Rodney M Ratcliff, Eviatar Nevo, and Frederick M Cohan. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proc. Natl. Acad. Sci. U. S. A.*, 105(7):2504–2509, February 2008.

Raivo Kolde, Eric A Franzosa, Gholamali Rahnavard, Andrew Brantley Hall, Hera Vlamakis, Christine Stevens, Mark J Daly, Ramnik J Xavier, and Curtis Huttenhower. Host genetic variation and its microbiome interactions within the human microbiome project. *Genome Med.*, 10(1):6, January 2018.

Mikhail Kolmogorov, Mikhail Rayko, Jeffrey Yuan, Evgeny Polevikov, and Pavel Pevzner. metaflye: scalable long-read metagenome assembly using repeat graphs. *bioRxiv*, page 637637, 2019.

Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, August 2012.

Mark Kowarsky, Joan Camunas-Soler, Michael Kertesz, Iwijn De Vlaminck, Winston Koh, Wenying Pan, Lance Martin, Norma F Neff, Jennifer Okamoto, Ronald J Wong, Sandhya Kharbanda, Yasser El-Sayed, Yair Blumenfeld, David K Stevenson, Gary M Shaw, Nathan D Wolfe, and Stephen R Quake. Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proc. Natl. Acad. Sci. U. S. A.*, August 2017.

Ranjit Kumar, Nengjun Yi, Degui Zhi, Peter Eipers, Kelly T Goldsmith, Paula Dixon, David K Crossman, Michael R Crowley, Elliot J Lefkowitz, J Martin Rodriguez, and Casey D Morrow. Identification of donor microbe species that colonize and persist long term in the recipient after fecal transplant for recurrent clostridium difficile. *NPJ Biofilms Microbiomes*, 3:12, June 2017.

Alexander Kurilshikov, Cisca Wijmenga, Jingyuan Fu, and Alexandra Zhernakova. Host genetics and gut microbiome: Challenges and perspectives. *Trends Immunol.*, 38(9): 633–647, September 2017.

Ilias Lagkouvardos, Rüdiger Pukall, Birte Abt, Bärbel U Foesel, Jan P Meier-Kolthoff, Neeraj Kumar, Anne Bresciani, Inés Martínez, Sarah Just, Caroline Ziegler, Sandrine Brugiroux, Debora Garzetti, Mareike Wenning, Thi P N Bui, Jun Wang, Floor Hugenholtz, Caroline M Plugge, Daniel A Peterson, Mathias W Hornef, John F Baines, Hauke Smidt, Jens Walter, Karsten Kristiansen, Henrik B Nielsen, Dirk Haller, Jörg Overmann, Bärbel Stecher, and Thomas Clavel. The mouse intestinal bacterial collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. *Nat Microbiol*, 1(10):16131, August 2016.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359, March 2012.

Nadja Larsen, Finn K Vogensen, Frans W J van den Berg, Dennis Sandris Nielsen, Anne Sofie Andreasen, Bente K Pedersen, Waleed Abu Al-Soud, Søren J Sørensen, Lars H Hansen, and Mogens Jakobsen. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One*, 5(2):e9085, February 2010.

Ramanan Laxminarayan, Adriano Duse, Chand Wattal, Anita K M Zaidi, Heiman F L Wertheim, Nithima Sumpradit, Erika Vlieghe, Gabriel Levy Hara, Ian M Gould, Herman Goossens, Christina Greko, Anthony D So, Maryam Bigdeli, Göran Tomson, Will Woodhouse, Eva Ombaka, Arturo Quizhpe Peralta, Farah Naz Qamar, Fatima Mir, Sam Kariuki, Zulfiqar A Bhutta, Anthony Coates, Richard Bergstrom, Gerard D Wright, Eric D Brown, and Otto Cars. Antibiotic resistance-the need for global solutions. *Lancet Infect. Dis.*, 13(12):1057–1098, December 2013.

S Melanie Lee, Gregory P Donaldson, Zbigniew Mikulski, Silva Boyajian, Klaus Ley, and Sarkis K Mazmanian. Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature*, 501(7467):426–429, September 2013.

Sonny T M Lee, Stacy A Kahn, Tom O Delmont, Alon Shaiber, Özcan C Esen, Nathaniel A Hubert, Hilary G Morrison, Dionysios A Antonopoulos, David T Rubin, and A Murat Eren. Tracking microbial colonization in fecal microbiota transplantation experiments via genome-resolved metagenomics. *Microbiome*, 5(1):50, May 2017.

Fernanda C Lessa, Yi Mu, Wendy M Bamberg, Zintars G Beldavs, Ghinwa K Dumyati, John R Dunn, Monica M Farley, Stacy M Holzbauer, James I Meek, Erin C Phipps, Lucy E Wilson, Lisa G Winston, Jessica A Cohen, Brandi M Limbago, Scott K Fridkin, Dale N Gerding, and L Clifford McDonald. Burden of clostridium difficile infection in the united states. *N. Engl. J. Med.*, 372(9):825–834, February 2015.

Maayan Levy, Aleksandra A Kolodziejczyk, Christoph A Thaiss, and Eran Elinav. Dysbiosis and the immune system. *Nat. Rev. Immunol.*, 17(4):219–232, April 2017.

Ruth E Ley, Peter J Turnbaugh, Samuel Klein, and Jeffrey I Gordon. Microbial ecology: human gut microbes associated with obesity. *Nature*, 444(7122):1022–1023, December 2006.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16): 2078–2079, August 2009.

Simone S Li, Ana Zhu, Vladimir Benes, Paul I Costea, Rajna Hercog, Falk Hildebrand, Jaime Huerta-Cepas, Max Nieuwdorp, Jarkko Salojärvi, Anita Y Voigt, Georg Zeller, Shinichi Sunagawa, Willem M de Vos, and Peer Bork. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science*, 352(6285):586–589, April 2016.

Mary E Lidstrom and Michael C Konopka. The role of physiological heterogeneity in microbial population behavior. *Nat. Chem. Biol.*, 6(10):705–712, October 2010.

Chang Liu, Nan Zhou, Meng-Xuan Du, Yu-Tong Sun, Kai Wang, Yu-Jing Wang, Dan-Hua Li, Hai-Ying Yu, Yuqin Song, Bing-Bing Bai, Yuhua Xin, Linhuan Wu, Cheng-Ying Jiang, Jie Feng, Hua Xiang, Yuguang Zhou, Juncai Ma, Jun Wang, Hong-Wei Liu, and Shuang-Jiang Liu. The mouse gut microbial biobank expands the coverage of cultured bacteria. *Nat. Commun.*, 11(1):79, January 2020.

Minghsun Liu, Rajendar Deora, Sergei R Doulatov, Mari Gingery, Frederick A Eiserling, Andrew Preston, Duncan J Maskell, Robert W Simons, Peggy A Cotter, Julian Parkhill, and Jeff F Miller. Reverse transcriptase-mediated tropism switching in bordetella bacteriophage. *Science*, 295(5562):2091–2094, March 2002.

Jason Lloyd-Price, Galeb Abu-Ali, and Curtis Huttenhower. The healthy human microbiome. *Genome Med.*, 8(1):51, April 2016.

Jason Lloyd-Price, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, Nadim J Ajami, Kevin S Bonham, Colin J Brislawn, David Casero, Holly Courtney, Antonio Gonzalez, Thomas G Graeber, A Brantley Hall, Kathleen Lake, Carol J Landers, Himel Mallick, Damian R Plichta, Mahadev Prasad, Gholamali Rahnavard, Jenny Sauk, Dmitry Shungin, Yoshiki Vázquez-Baeza, Richard A White, 3rd, IBDMDB Investigators, Jonathan Braun, Lee A Denson, Janet K Jansson, Rob Knight, Subra Kugathasan, Dermot P B McGovern, Joseph F Petrosino, Thaddeus S Stappenbeck, Harland S Winter, Clary B Clish, Eric A Franzosa, Hera Vlamakis, Ramnik J Xavier, and Curtis Huttenhower. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, May 2019.

Kate E Lynch, Emily C Parke, and Maureen A O'Malley. How causal are microbiomes? a comparison with the helicobacter pylori explanation of ulcers. *Biol. Philos.*, 34(6):62, November 2019.

Susan V Lynch and Oluf Pedersen. The human intestinal microbiome in health and disease. *N. Engl. J. Med.*, 375(24):2369–2379, December 2016.

Andrew J Macpherson, Bahtiyar Yilmaz, Julien P Limenitakis, and Stephanie C Ganal-Vonarburg. IgA function in relation to the intestinal microbiota. *Annu. Rev. Immunol.*, 36:359–381, April 2018.

Lisa Maier, Mihaela Pruteanu, Michael Kuhn, Georg Zeller, Anja Telzerow, Exene Erin Anderson, Ana Rita Brochado, Keith Conrad Fernandez, Hitomi Dose, Hirotada Mori, Kiran Raosaheb Patil, Peer Bork, and Athanasios Typas. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature*, 555(7698):623–628, March 2018.

Loïs Maignien, Emelia A DeForce, Meghan E Chafee, A Murat Eren, and Sheri L Simmons. Ecological succession and stochastic variation in the assembly of arabidopsis thaliana phyllosphere communities. *MBio*, 5(1):e00682–13, January 2014.

A Marcobal, P C Kashyap, T A Nelson, P A Aronov, M S Donia, A Spormann, M A Fischbach, and J L Sonnenburg. A metabolomic view of how the human gut microbiota impacts the host metabolome using humanized and gnotobiotic mice. *ISME J.*, 7(10): 1933–1943, October 2013.

Ioana A Marin, Jennifer E Goertz, Tiantian Ren, Stephen S Rich, Suna Onengut-Gumuscu, Emily Farber, Martin Wu, Christopher C Overall, Jonathan Kipnis, and Alban Gaultier. Microbiota alteration is associated with the development of stress-induced despair behavior. *Sci. Rep.*, 7:43859, March 2017.

Eric C Martens, Mareike Neumann, and Mahesh S Desai. Interactions of commensal and pathogenic microorganisms with the intestinal mucosal barrier. *Nat. Rev. Microbiol.*, 16 (8):457–470, August 2018.

J H Martens, H Barg, M J Warren, and D Jahn. Microbial production of vitamin B12. *Appl. Microbiol. Biotechnol.*, 58(3):275–285, March 2002.

Inés Martínez, Jaehyoung Kim, Patrick R Duffy, Vicki L Schlegel, and Jens Walter. Resistant starches types 2 and 4 have differential effects on the composition of the fecal microbiota in human subjects. *PLoS One*, 5(11):e15046, November 2010.

Inés Martínez, Maria X Maldonado-Gomez, João Carlos Gomes-Neto, Hatem Kittana, Hua Ding, Robert Schmaltz, Payal Joglekar, Roberto Jiménez Cardona, Nathan L Marsteller, Steven W Kembel, Andrew K Benson, Daniel A Peterson, Amanda E Ramer-Tait, and Jens Walter. Experimental evaluation of the importance of colonization history in early-life gut microbiota assembly. *Elife*, 7, September 2018.

Adam C Martiny, Kathleen Treseder, and Gordon Pusch. Phylogenetic conservatism of functional traits in microorganisms. *ISME J.*, 7(4):830–838, April 2013.

Michael I McBurney, Cindy Davis, Claire M Fraser, Barbara O Schneeman, Curtis Huttenhower, Kristin Verbeke, Jens Walter, and Marie E Latulippe. Establishing what constitutes a healthy human gut microbiome: State of the science, regulatory considerations, and future directions. *J. Nutr.*, 149(11):1882–1895, November 2019.

Bob Medhekar and Jeff F Miller. Diversity-generating retroelements. *Curr. Opin. Microbiol.*, 10(4):388–395, August 2007.

J S Messer, E R Liechty, O A Vogel, and E B Chang. Evolutionary and ecological forces that shape the bacterial communities of the human gut. *Mucosal Immunol.*, 10(3):567–579, May 2017.

André E Minoche, Juliane C Dohm, and Heinz Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on illumina HiSeq and genome analyzer systems. *Genome Biol.*, 12(11):R112, November 2011.

Clara Moon, Megan T Baldridge, Meghan A Wallace, Carey-Ann D, Burnham, Herbert W Virgin, and Thaddeus S Stappenbeck. Vertically transmitted faecal IgA levels determine extra-chromosomal phenotypic variation. *Nature*, 521(7550):90–93, May 2015.

Eli L Moss, Shannon B Falconer, Ekaterina Tkachenko, Mingjie Wang, Hannah Systrom, Jasmin Mahabamunuge, David A Relman, Elizabeth L Hohmann, and Ami S Bhatt. Long-term taxonomic and functional divergence from donor bacterial strains following fecal microbiota transplantation in immunocompromised patients. *PLoS One*, 12(8): e0182585, August 2017.

Lam-Tung Nguyen, Heiko A Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, 32(1):268–274, January 2015.

Agnieszka Olek, Marek Woynarowski, Irini Lazou Ahrén, Jarosław Kierkuś, Piotr Socha, Niklas Larsson, and Gunilla Önning. Efficacy and safety of lactobacillus plantarum DSM 9843 (LP299V) in the prevention of Antibiotic-Associated gastrointestinal symptoms in Children-Randomized, Double-Blind, Placebo-Controlled study. *J. Pediatr.*, 186:82–86, July 2017.

S J Ott, M Musfeldt, D F Wenderoth, J Hampe, O Brant, U R Fölsch, K N Timmis, and S Schreiber. Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. *Gut*, 53(5):685–693, May 2004.

Stephan J Ott, Georg H Waetzig, Ateequr Rehman, Jacqueline Moltzau-Anderson, Richa Bharti, Juris A Grasis, Liam Cassidy, Andreas Tholey, Helmut Fickenscher, Dirk Seegert, Philip Rosenstiel, and Stefan Schreiber. Efficacy of sterile fecal filtrate transfer for treating patients with clostridium difficile infection. *Gastroenterology*, 152(4): 799–811.e7, March 2017.

Noah W Palm, Marcel R de Zoete, Thomas W Cullen, Natasha A Barry, Jonathan Stefanowski, Liming Hao, Patrick H Degnan, Jianzhong Hu, Inga Peter, Wei Zhang, Elizabeth Ruggiero, Judy H Cho, Andrew L Goodman, and Richard A Flavell. Immunoglobulin a coating identifies colitogenic bacteria in inflammatory bowel disease. *Cell*, 158(5): 1000–1010, August 2014.

Bianca J Parker, Pamela A Wearsch, Alida C M Veloo, and Alex Rodriguez-Palacios. The genus alistipes: Gut bacteria with emerging implications to inflammation, cancer, and mental health. *Front. Immunol.*, 11:906, June 2020.

Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, 36(10):996–1004, November 2018.

Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, Maria Carmen Collado, Benjamin L Rice, Casey DuLong, Xochitl C Morgan, Christopher D Golden, Christopher Quince, Curtis Huttenhower, and Nicola Segata. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 176(3):649–662.e20, January 2019.

Blair G Paul, David Burstein, Cindy J Castelle, Sumit Handa, Diego Arambula, Elizabeth Czornyj, Brian C Thomas, Partho Ghosh, Jeff F Miller, Jillian F Banfield, and David L Valentine. Retroelement-guided protein diversification abounds in vast lineages of bacteria and archaea. *Nat Microbiol*, 2:17045, April 2017.

Noel T Pauli, Hwan Keun Kim, Fabiana Falugi, Min Huang, John Dulac, Carole Henry Dunand, Nai-Ying Zheng, Kaval Kaur, Sarah F Andrews, Yunping Huang, Andrea DeDent, Karen M Frank, Angella Charnot-Katsikas, Olaf Schneewind, and Patrick C Wilson. Staphylococcus aureus infection induces protein a-mediated immune evasion in humans. *J. Exp. Med.*, 211(12):2331–2339, November 2014.

Yu Peng, Henry C M Leung, S M Yiu, and Francis Y L Chin. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428, June 2012.

Laurent Philippot, Siv G E Andersson, Tom J Battin, James I Prosser, Joshua P Schimel, William B Whitman, and Sara Hallin. The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev. Microbiol.*, 8(7):523–529, July 2010.

Damian R Plichta, Daniel B Graham, Sathish Subramanian, and Ramnik J Xavier. Therapeutic opportunities in inflammatory bowel disease: Mechanistic dissection of Host-Microbiome relationships. *Cell*, 178(5):1041–1056, August 2019.

Daniel Podlesny and W Florian Fricke. Microbial strain engraftment, persistence and replacement after fecal microbiota transplantation. *medRxiv*, page 2020.09.29.20203638, September 2020.

Anna Pompei, Lisa Cordisco, Alberto Amaretti, Simona Zanoni, Diego Matteuzzi, and Maddalena Rossi. Folate production by bifidobacteria as a potential probiotic property. *Appl. Environ. Microbiol.*, 73(1):179–185, January 2007.

Nathan T Porter, Pablo Canales, Daniel A Peterson, and Eric C Martens. A subset of polysaccharide capsules in the human symbiont bacteroides thetaiotaomicron promote increased competitive fitness in the mouse gut. *Cell Host Microbe*, 22(4):494–506.e8, October 2017.

Morgan N Price, Paramvir S Dehal, and Adam P Arkin. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, 5(3):e9490, March 2010.

Leighton Pritchard, Rachel H Glover, Sonia Humphris, John G Elphinstone, and Ian K Toth. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal. Methods*, 8(1):12–24, 2016.

Sambhawa Priya and Ran Blekhman. Population dynamics of the human gut microbiome: change is the only constant. *Genome Biol.*, 20(1):150, July 2019.

Rodrigo Quera, Ricardo Espinoza, Camila Estay, and Daniela Rivera. Bacteremia as an adverse event of fecal microbiota transplantation in a patient with crohn's disease and recurrent clostridium difficile infection. *J. Crohns. Colitis*, 8(3):252–253, March 2014.

Eamonn M M Quigley and Prianka Gajula. Recent advances in modulating the microbiome. *F1000Res.*, 9, January 2020.

Christopher Quince, Umer Zeeshan Ijaz, Nick Loman, A Murat Eren, Delphine Saulnier, Julie Russell, Sarah J Haig, Szymon T Calus, Joshua Quick, Andrew Barclay, Martin Bertz, Michael Blaut, Richard Hansen, Paraic McGrogan, Richard K Russell, Christine A Edwards, and Konstantinos Gerasimidis. Extensive modulation of the fecal metagenome in children with crohn's disease during exclusive enteral nutrition. *Am. J. Gastroenterol.*, 110(12):1718–29; quiz 1730, December 2015.

Christopher Quince, Tom O Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E Darling, Gavin Collins, and A Murat Eren. DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.*, 18(1):181, September 2017a.

Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.*, 35(9): 833–844, September 2017b.

E Rackaityte, J Halkias, E M Fukui, V F Mendoza, C Hayzelden, E D Crawford, K E Fujimura, T D Burt, and S V Lynch. Viable bacterial colonization is highly limited in the human intestine in utero. *Nat. Med.*, 26(4):599–607, April 2020.

Julie Reveillaud, Sarah R Bordenstein, Corinne Cruaud, Alon Shaiber, Özcan C Esen, Mylène Weill, Patrick Makoundou, Karen Lolans, Andrea R Watson, Ignace Rakotoarivony, Seth R Bordenstein, and A Murat Eren. The wolbachia mobilome in culex pipiens includes a putative plasmid. *Nat. Commun.*, 10(1):1051, March 2019.

Filipe J Ribeiro, Dariusz Przybylski, Shuangye Yin, Ted Sharpe, Sante Gnerre, Amr Abouelleil, Aaron M Berlin, Anna Montmayeur, Terrance P Shea, Bruce J Walker, Sarah K Young, Carsten Russ, Chad Nusbaum, Iain MacCallum, and David B Jaffe. Finished bacterial genomes from shotgun sequence data. *Genome Res.*, 22(11):2270–2277, November 2012.

Christian Rinke, Patrick Schwientek, Alexander Sczyrba, Natalia N Ivanova, Iain J Anderson, Jan-Fang Cheng, Aaron Darling, Stephanie Malfatti, Brandon K Swan, Esther A Gies, Jeremy A Dodsworth, Brian P Hedlund, George Tsiamis, Stefan M Sievert, Wen-Tso Liu, Jonathan A Eisen, Steven J Hallam, Nikos C Kyrpides, Ramunas Stepanauskas, Edward M Rubin, Philip Hugenholtz, and Tanja Woyke. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437, July 2013.

Daphna Rothschild, Omer Weissbrod, Elad Barkan, Alexander Kurilshikov, Tal Korem, David Zeevi, Paul I Costea, Anastasia Godneva, Iris N Kalka, Noam Bar, Smadar Shilo, Dar Lador, Arnau Vich Vila, Niv Zmora, Meirav Pevsner-Fischer, David Israeli, Noa Kosower, Gal Malka, Bat Chen Wolf, Tali Avnit-Sagi, Maya Lotan-Pompan, Adina Weinberger, Zamir Halpern, Shai Carmi, Jingyuan Fu, Cisca Wijmenga, Alexandra Zhernakova, Eran Elinav, and Eran Segal. Environment dominates over host genetics in shaping human gut microbiota. *Nature*, 555(7695):210–215, March 2018.

June L Round and Sarkis K Mazmanian. The gut microbiota shapes intestinal immune responses during health and disease. *Nat. Rev. Immunol.*, 9(5):313–323, May 2009.

Simon Roux, Blair G Paul, Sarah C Bagby, Michelle A Allen, Graeme Attwood, Ricardo Cavicchioli, Ludmila Chistoserdova, Steven J Hallam, Maria E Hernandez, Matthias Hess, Wen-Tso Liu, Michelle A O'Malley, Xuefeng Peng, Virginia I Rich, Scott Saleska, and Emiley A Eloe-Fadrosh. Ecology and molecular targets of hypermutation in the global microbiome. April 2020.

J Graham Ruby, Priya Bellare, and Joseph L Derisi. PRICE: software for the targeted assembly of components of (meta) genomic sequence data. *G3*, 3(5):865–880, May 2013.

Mark A Schell, Maria Karmirantzou, Berend Snel, David Vilanova, Bernard Berger, Gabriella Pessi, Marie-Camille Zwahlen, Frank Desiere, Peer Bork, Michele Delley, R David Pridmore, and Fabrizio Arigoni. The genome sequence of bifidobacterium longum reflects its adaptation to the human gastrointestinal tract. *Proc. Natl. Acad. Sci. U. S. A.*, 99(22):14422–14427, October 2002.

Melanie Schirmer, Ashley Garner, Hera Vlamakis, and Ramnik J Xavier. Microbial genes and pathways in inflammatory bowel disease. *Nat. Rev. Microbiol.*, 17(8):497–511, August 2019.

Thomas S B Schmidt, Jeroen Raes, and Peer Bork. The human gut microbiome: From association to modulation. *Cell*, 172(6):1198–1215, March 2018.

Frank Schreiber and Martin Ackermann. Environmental drivers of metabolic heterogeneity in clonal microbial populations. *Curr. Opin. Biotechnol.*, 62:202–211, April 2020.

Michael H Schwartz, Haipeng Wang, Jessica N Pan, Wesley C Clark, Steven Cui, Matthew J Eckwahl, David W Pan, Marc Parisien, Sarah M Owens, Brian L Cheng, Kristina Martinez, Jinbo Xu, Eugene B Chang, Tao Pan, and A Murat Eren. Microbiome characterization by high-throughput transfer RNA sequencing and modification analysis. *Nat. Commun.*, 9(1):5353, December 2018.

Ron Sender, Shai Fuchs, and Ron Milo. Revised estimates for the number of human and bacteria cells in the body, August 2016.

Dea Shahinas, Michael Silverman, Taylor Sittler, Charles Chiu, Peter Kim, Emma Allen-Vercoe, Scott Weese, Andrew Wong, Donald E Low, and Dylan R Pillai. Toward an understanding of changes in diversity associated with fecal microbiome transplantation based on 16S rRNA gene deep sequencing. *MBio*, 3(5):e00338–12, October 2012.

Alon Shaiber, Amy D Willis, Tom O Delmont, Simon Roux, Lin-Xing Chen, Abigail C Schmid, Mahmoud Yousef, Andrea R Watson, Karen Lolans, Özcan C Esen, Sonny T M Lee, Nora Downey, Hilary G Morrison, Floyd E Dewhirst, Jessica L Mark Welch, and A Murat Eren. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol.*, 21(1):292, December 2020.

Itai Sharon, Michael J Morowitz, Brian C Thomas, Elizabeth K Costello, David A Relman, and Jillian F Banfield. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.*, 23 (1):111–120, January 2013.

Ravi U Sheth, Mingqiang Li, Weiqian Jiang, Peter A Sims, Kam W Leong, and Harris H Wang. Spatial metagenomic characterization of microbial biogeography in the gut. *Nat. Biotechnol.*, 37(8):877–883, August 2019.

Christopher S Smillie, Jenny Sauk, Dirk Gevers, Jonathan Friedman, Jaeyun Sung, Ilan Youngster, Elizabeth L Hohmann, Christopher Staley, Alexander Khoruts, Michael J Sadowsky, Jessica R Allegretti, Mark B Smith, Ramnik J Xavier, and Eric J Alm. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host Microbe*, 23(2):229–240.e5, February 2018.

Harry Sokol and Philippe Seksik. The intestinal microbiota in inflammatory bowel diseases: time to connect with the host. *Curr. Opin. Gastroenterol.*, 26(4):327–331, July 2010.

Paola R Solari, Patrick G Fairchild, Luis Junco Noa, and Mark R Wallace. Tempered enthusiasm for fecal transplant. *Clin. Infect. Dis.*, 59(2):319, July 2014.

Christopher Staley, Thomas Kaiser, Lalit K Beura, Matthew J Hamilton, Alexa R Weingarden, Aleh Bobr, Johnthomas Kang, David Masopust, Michael J Sadowsky, and Alexander Khoruts. Stable engraftment of human microbiota into mice with a single oral gavage following antibiotic conditioning. *Microbiome*, 5(1):87, August 2017.

G Paolo Strozzi and Luca Mogna. Quantification of folic acid in human feces after administration of bifidobacterium probiotic strains. *J. Clin. Gastroenterol.*, 42 Suppl 3 Pt 2: S179–84, September 2008.

Jotham Suez, Niv Zmora, Gili Zilberman-Schapira, Uria Mor, Mally Dori-Bachash, Stavros Bashiardes, Maya Zur, Dana Regev-Lehavi, Rotem Ben-Zeev Brik, Sara Federici, Max Horn, Yotam Cohen, Andreas E Moor, David Zeevi, Tal Korem, Eran Kotler, Alon Harmelin, Shalev Itzkovitz, Nitsan Maharshak, Oren Shibolet, Meirav Pevsner-Fischer, Hagit Shapiro, Itai Sharon, Zamir Halpern, Eran Segal, and Eran Elinav. Post-Antibiotic gut mucosal microbiome reconstitution is impaired by probiotics and improved by autologous FMT. *Cell*, 174(6):1406–1423.e16, September 2018.

Hirosuke Sugahara, Toshitaka Odamaki, Nanami Hashikura, Fumiaki Abe, and Jin-Zhong Xiao. Differences in folate production by bifidobacteria of different origins. *Biosci Microbiota Food Health*, 34(4):87–93, August 2015.

Alexander Swidsinski, Jutta Weber, Vera Loening-Baucke, Laura P Hale, and Herbert Lochs. Spatial organization and composition of the mucosal flora in patients with inflammatory bowel disease. *J. Clin. Microbiol.*, 43(7):3380–3389, July 2005.

Roman L Tatusov, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Boris Kiryutin, Eugene V Koonin, Dmitri M Krylov, Raja Mazumder, Sergei L Mekhedov, Anastasia N Nikolskaya, B Sridhar Rao, Sergei Smirnov, Alexander V Sverdlov, Sona Vasudevan, Yuri I Wolf, Jodie J Yin, and Darren A Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, September 2003.

Dominika Trzilova and Rita Tamayo. Site-Specific recombination - how simple DNA inversions produce complex phenotypic heterogeneity in bacterial populations. *Trends Genet.*, 37(1):59–72, January 2021.

Peter J Turnbaugh, Christopher Quince, Jeremiah J Faith, Alice C McHardy, Tanya Yatsunenko, Faheem Niazi, Jason Affourtit, Michael Egholm, Bernard Henrissat, Rob Knight, and Jeffrey I Gordon. Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc. Natl. Acad. Sci. U. S. A.*, 107(16):7503–7508, April 2010.

Daniel R Utter, Gary G Borisy, A Murat Eren, Colleen M Cavanaugh, and Jessica L Mark Welch. Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity. *Genome Biol.*, 21(1):293, December 2020.

Kees C H van der Ark, Ruben G A van Heck, Vitor A P Martins Dos Santos, Clara Belzer, and Willem M de Vos. More than just a gut feeling: constraint-based genome-scale metabolic models for predicting functions of human intestinal microbes. *Microbiome*, 5 (1):78, July 2017.

Els van Nood, Anne Vrieze, Max Nieuwdorp, Susana Fuentes, Erwin G Zoetendal, Willem M de Vos, Caroline E Visser, Ed J Kuijper, Joep F W M Bartelsman, Jan G P Tijssen, Peter Speelman, Marcel G W Dijkgraaf, and Josbert J Keller. Duodenal infusion of donor feces for recurrent clostridium difficile. *N. Engl. J. Med.*, 368(5):407–415, January 2013.

Chiara Vanni, Matthew S Schechter, Silvia G Acinas, Albert Barberán, Pier Luigi Buttigieg, Emilio O Casamayor, Tom O Delmont, Carlos M Duarte, A Murat Eren, Robert D Finn, Renzo Kottmann, Alex Mitchell, Pablo Sanchez, Kimmo Siren, Martin Steinegger, Frank Oliver Glöckner, and Antonio Fernandez-Guerra. Unifying the global coding sequence space enables the study of genes with unknown function across biomes. November 2020.

Mark Vellend. Conceptual synthesis in community ecology. *Q. Rev. Biol.*, 85(2):183–206, June 2010.

Marco Ventura, Sarah O'Flaherty, Marcus J Claesson, Francesca Turroni, Todd R Klaenhammer, Douwe van Sinderen, and Paul W O'Toole. Genome-scale analyses of health-promoting bacteria: probiogenomics. *Nat. Rev. Microbiol.*, 7(1):61–71, January 2009.

Angélica Thomaz Vieira, Claudio Fukumori, and Caroline Marcantonio Ferreira. New insights into therapeutic strategies for gut microbiota modulation in inflammatory diseases. *Clin Transl Immunology*, 5(6):e87, June 2016.

Monica Viladomiu, Charles Kivolowitz, Ahmed Abdulhamid, Belgin Dogan, Daniel Victorio, Jim G Castellanos, Viola Woo, Fei Teng, Nhan L Tran, Andrew Sczesnak, Christina Chai, Myunghoo Kim, Gretchen E Diehl, Nadim J Ajami, Joseph F Petrosino, Xi K Zhou, Sergio Schwartzman, Lisa A Mandl, Meira Abramowitz, Vinita Jacob, Brian Bosworth, Adam Steinlauf, Ellen J Scherl, Hsin-Jung Joyce Wu, Kenneth W Simpson, and Randy S Longman. IgA-coated e. coli enriched in crohn's disease spondyloarthritis promote TH17-dependent inflammation. *Sci. Transl. Med.*, 9(376), February 2017.

Joseph H Vineis, Daina L Ringus, Hilary G Morrison, Tom O Delmont, Sushila Dalal, Laura H Raffals, Dionysios A Antonopoulos, David T Rubin, A Murat Eren, Eugene B Chang, and Mitchell L Sogin. Patient-Specific *Bacteroides* Genome Variants in Pouchitis. *MBio*, 7(6):e01713–16, /mbio/7/6/e01713–16.atom, December 2016.

Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, and Ashlee M Earl. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9(11):e112963, November 2014.

Jens Walter, Anissa M Armet, B Brett Finlay, and Fergus Shanahan. Establishing or exaggerating causality for the gut microbiome: Lessons from human Microbiota-Associated rodents. *Cell*, 180(2):221–232, January 2020.

Jiunn-Wei Wang, Chao-Hung Kuo, Fu-Chen Kuo, Yao-Kuang Wang, Wen-Hung Hsu, Fang-Jung Yu, Huang-Ming Hu, Ping-I Hsu, Jaw-Yuan Wang, and Deng-Chyang Wu. Fecal microbiota transplantation: Review and update. *J. Formos. Med. Assoc.*, 118 Suppl 1:S23–S31, March 2019.

L Wang, C T Christophersen, M J Sorich, and others. Low relative abundances of the mucolytic bacterium akkermansia muciniphila and bifidobacterium spp. in feces of children with autism. *Applied and*, 2011.

Andrea R Watson, Jessika Fuessel, Iva Veseli, Johanna Zaal DeLongchamp, Marisela Silva, Florian Trigodet, Karen Lolans, Alon Shaiber, Emily Fogarty, Christopher Quince, Michael K Yu, Arda Soylev, Hilary G Morrison, Sonny T M Lee, David T Rubin, Bana Jabri, Thomas Louie, and A Murat Eren. Adaptive ecological processes and metabolic independence drive microbial colonization and resilience in the human gut, March 2021.

William J Watterson, Melikhan Tanyeri, Andrea R Watson, Candace M Cham, Yue Shan, Eugene B Chang, A Murat Eren, and Savaş Tay. Droplet-based high-throughput cultivation for accurate screening of antibiotic resistant gut microbes. *Elife*, 9:e56998, June 2020.

Aaron G Wexler and Andrew L Goodman. An insider's perspective: Bacteroides as a window into the microbiome, 2017.

S Whelan and N Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, 18(5): 691–699, May 2001.

Ryan R Wick, Louise M Judd, Claire L Gorrie, and Kathryn E Holt. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.*, 13(6):e1005595, June 2017.

Derrick E Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome Biol.*, 20(1):257, November 2019.

Guojun Wu, Naisi Zhao, Chenhong Zhang, Yan Y Lam, and Liping Zhao. Guild-based analysis for understanding gut microbiome in human health and diseases. *Genome Med.*, 13(1):22, February 2021.

Meng-Que Xu, Hai-Long Cao, Wei-Qiang Wang, Shan Wang, Xiao-Cang Cao, Fang Yan, and Bang-Mao Wang. Fecal microbiota transplantation broadening its application beyond intestinal disorders. *World J. Gastroenterol.*, 21(1):102–111, January 2015.

Fazhe Yan, Xuelin Yu, Zhongqu Duan, Jinyuan Lu, Ben Jia, Yuyang Qiao, Chen Sun, and Chaochun Wei. Discovery and characterization of the evolution, variation and functions of diversity-generating retroelements using thousands of genomes and metagenomes. *BMC Genomics*, 20(1):595, July 2019.

Koji Yasuda, Keunyoung Oh, Boyu Ren, Timothy L Tickle, Eric A Franzosa, Lynn M Wachtman, Andrew D Miller, Susan V Westmoreland, Keith G Mansfield, Eric J Vallender, Gregory M Miller, James K Rowlett, Dirk Gevers, Curtis Huttenhower, and Xochitl C Morgan. Biogeography of the intestinal mucosal and lumenal microbiome in the rhesus macaque. *Cell Host Microbe*, 17(3):385–391, March 2015.

Tanya Yatsunenko, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N Baldassano, Andrey P Anokhin, Andrew C Heath, Barbara Warner, Jens Reeder, Justin Kuczynski, J Gregory Caporaso, Catherine A Lozupone, Christian Lauber, Jose Carlos Clemente, Dan Knights, Rob Knight, and Jeffrey I Gordon. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227, May 2012.

David Zeevi, Tal Korem, Anastasia Godneva, Noam Bar, Alexander Kurilshikov, Maya Lotan-Pompan, Adina Weinberger, Jingyuan Fu, Cisca Wijmenga, Alexandra Zhernakova, and Eran Segal. Structural variation in the gut microbiome associates with host health. *Nature*, 568(7750):43–48, April 2019.

Li Zhang, Martin Iain Bahl, Henrik Munch Roager, Cilius Esmann Fonvig, Lars I Hellgren, Henrik Lauritz Frandsen, Oluf Pedersen, Jens-Christian Holm, Torben Hansen, and Tine Rask Licht. Environmental spread of microbes impacts the development of metabolic phenotypes in mice transplanted with microbial communities from humans. *ISME J.*, 11(3):676–690, March 2017.

Shijie Zhao, Tami D Lieberman, Mathilde Poyet, Kathryn M Kauffman, Sean M Gibbons, Mathieu Groussin, Ramnik J Xavier, and Eric J Alm. Adaptive evolution within gut microbiomes of healthy people. *Cell Host Microbe*, 25(5):656–667.e8, May 2019.